

There's Plenty of Room at the Bottom

An Invitation to Enter a New Field of Physics

by Richard P. Feynman



This transcript of the classic talk that Richard Feynman gave on December 29th 1959 at the annual meeting of the [American Physical Society](#) at the [California Institute of Technology \(Caltech\)](#) was first published in the February 1960 issue of Caltech's [Engineering and Science](#), which owns the copyright. It has been made available on the web at <http://www.zyvex.com/nanotech/feynman.html> with their kind permission.

[Information on the Feynman Prizes](#)

[Links to pages on Feynman](#)

For an account of the talk and how people reacted to it, see chapter 4 of *Nano!* by Ed Regis, Little/Brown 1995. An excellent technical introduction to nanotechnology is [Nanosystems: molecular machinery, manufacturing, and computation](#) by K. Eric Drexler, Wiley 1992.

I imagine experimental physicists must often look with envy at men like Kamerlingh Onnes, who discovered a field like low temperature, which seems to be bottomless and in which one can go down and down. Such a man is then a leader and has some temporary monopoly in a scientific adventure. Percy Bridgman, in designing a way to obtain higher pressures, opened up another new field and was able to move into it and to lead us all along. The development of ever higher vacuum was a continuing development of the same kind.

I would like to describe a field, in which little has been done, but in which an enormous amount can be done in principle. This field is not quite the same as the others in that it will not tell us much of fundamental physics (in the sense of, "What are the strange particles?") but it is more like solid-state physics in the sense that it might tell us much of great interest about the strange phenomena that occur in complex situations. Furthermore, a point that is most important is that it would have an enormous number of technical applications.

What I want to talk about is the problem of manipulating and controlling things on a small scale.

As soon as I mention this, people tell me about miniaturization, and how far it has progressed today. They tell me about electric motors that are the size of the nail on your small finger. And there is a device on the market, they tell me, by which you can write the Lord's Prayer on the head of a pin. But that's nothing; that's the most primitive, halting step in the direction I intend to discuss. It is a staggeringly small world that is below. In the year 2000, when they look back at this age, they will wonder why it was not until the year 1960 that anybody began seriously to move in this direction.

Why cannot we write the entire 24 volumes of the Encyclopedia Britannica on the head of a pin?

Let's see what would be involved. The head of a pin is a sixteenth of an inch across. If you magnify it by 25,000 diameters, the area of the head of the pin is then equal to the area of all the pages of the Encyclopaedia Britannica. Therefore, all it is necessary to do is to reduce in size all the writing in the Encyclopaedia by 25,000 times. Is that possible? The resolving power of the eye is about 1/120 of an inch---that is roughly the diameter of one of the little dots on the fine half-tone reproductions in the Encyclopaedia. This, when you demagnify it by 25,000 times, is still 80 angstroms in diameter---32 atoms across, in an ordinary metal. In other words, one of those dots still would contain in its area 1,000 atoms. So, each dot can easily be adjusted in size as required by the photoengraving, and there is no question that there is enough room on the head of a pin to put all of the Encyclopaedia Britannica.

Furthermore, it can be read if it is so written. Let's imagine that it is written in raised letters of metal; that is, where the black is in the Encyclopedia, we have raised letters of metal that are actually 1/25,000 of their ordinary size. How would we read it?

If we had something written in such a way, we could read it using techniques in common use today. (They will undoubtedly find a better way when we do actually have it written, but to make my point conservatively I shall just take techniques we know today.) We would press the metal into a plastic material and make a mold of it, then peel the plastic off very carefully, evaporate silica into the plastic to get a very thin film, then shadow it by evaporating gold at an angle against the silica so that all the little letters will appear clearly, dissolve the plastic away from the silica film, and then look through it with an electron microscope!

There is no question that if the thing were reduced by 25,000 times in the form of raised letters on the pin, it would be easy for us to read it today. Furthermore; there is no question that we would find it easy to make copies of the master; we would just need to press the same metal plate again into plastic and we would have another copy.

How do we write small?

The next question is: How do we *write* it? We have no standard technique to do this now. But let me argue that it is not as difficult as it first appears to be. We can reverse the lenses of the electron microscope in order to demagnify as well as magnify. A source of ions, sent through the microscope lenses in reverse, could be focused to a very small spot. We could write with that spot like we write in a TV cathode ray oscilloscope, by going across in lines, and having an adjustment which determines the amount of material which is going to be deposited as we scan in lines.

This method might be very slow because of space charge limitations. There will be more rapid methods. We could first make, perhaps by some photo process, a screen which has holes in it in the form of the letters. Then we would strike an arc behind the holes and draw metallic ions through the holes; then we could again use our system of lenses and make a small image in the form of ions, which would deposit the metal on the pin.

A simpler way might be this (though I am not sure it would work): We take light and, through an optical microscope running backwards, we focus it onto a very small photoelectric screen. Then electrons come away from the screen where the light is shining. These electrons are focused down in size by the electron microscope lenses to impinge directly upon the surface of the metal. Will such a beam etch away the metal if it is run long enough? I don't know. If it doesn't work for a metal surface, it must be possible to find some surface with which to coat the original pin so that, where the electrons bombard, a change is made which we could recognize later.

There is no intensity problem in these devices---not what you are used to in magnification, where you have to take a few electrons and spread them over a bigger and bigger screen; it is just the opposite. The light which we get from a page is concentrated onto a very small area so it is very intense. The few electrons which come from the photoelectric screen are demagnified down to a very tiny area so that, again, they are very intense. I don't know why this hasn't been done yet!

That's the Encyclopaedia Britannica on the head of a pin, but let's consider all the books in the world. The Library of Congress has approximately 9 million volumes; the British Museum Library has 5 million volumes; there are also 5 million volumes in the National Library in France. Undoubtedly there are duplications, so let us say that there are some 24 million volumes of interest in the world.

What would happen if I print all this down at the scale we have been discussing? How much space would it take? It would take, of course, the area of about a million pinheads because, instead of there being just the 24 volumes of the Encyclopaedia, there are 24 million volumes. The million pinheads can be put in a square of a thousand pins on a side, or an area of about 3 square yards. That is to say, the silica replica with the paper-thin backing of plastic, with which we have made the copies, with all this information, is on an area of approximately the size of 35 pages of the Encyclopaedia. That is about half as many pages as there are in this magazine. All of the information which all of mankind has ever recorded in books can be carried around in a pamphlet in your hand---and not written in code, but a simple reproduction of the original pictures, engravings, and everything else on a small scale without loss of resolution.

What would our librarian at Caltech say, as she runs all over from one building to another, if I tell her that, ten years from now, all of the information that she is struggling to keep track of--- 120,000 volumes, stacked from the floor to the ceiling, drawers full of cards, storage rooms full of the older books---can be kept on just one library card! When the University of Brazil, for example, finds that their library is burned, we can send them a copy of every book in our library by striking off a copy from the master plate in a few hours and mailing it in an envelope no bigger or heavier than any other ordinary air mail letter.

Now, the name of this talk is ``There is *Plenty* of Room at the Bottom"---not just ``There is Room at the Bottom." What I have demonstrated is that there *is* room---that you can decrease the size of things in a practical way. I now want to show that there is *plenty* of room. I will not now discuss how we are going to do it, but only what is possible in principle---in other words, what is possible according to the laws of physics. I am not inventing anti-gravity, which is possible someday only if the laws are not what we think. I am telling you what could be done if the laws *are* what we think; we are not doing it simply because we haven't yet gotten around to it.

Information on a small scale

Suppose that, instead of trying to reproduce the pictures and all the information directly in its present form, we write only the information content in a code of dots and dashes, or something like that, to represent the various letters. Each letter represents six or seven ``bits" of information; that is, you need only about six or seven dots or dashes for each letter. Now, instead of writing everything, as I did before, on the *surface* of the head of a pin, I am going to use the interior of the material as well.

Let us represent a dot by a small spot of one metal, the next dash, by an adjacent spot of another metal, and so on. Suppose, to be conservative, that a bit of information is going to require a little cube of atoms 5 times 5 times 5---that is 125 atoms. Perhaps we need a hundred and some odd atoms to make sure that the information is not lost through diffusion, or through some other process.

I have estimated how many letters there are in the Encyclopaedia, and I have assumed that each of my 24 million books is as big as an Encyclopaedia volume, and have calculated, then, how many bits of information there are (10^{15}). For each bit I allow 100 atoms. And it turns out that all of the information that man has carefully accumulated in all the books in the world can be written in this form in a cube of material one two-hundredth of an inch wide--- which is the barest piece of dust that can be made out by the human eye. So there is *plenty* of room at the bottom! Don't tell me about microfilm!

This fact---that enormous amounts of information can be carried in an exceedingly small space---is, of course, well known to the biologists, and resolves the mystery which existed before we understood all this clearly, of how it could be that, in the tiniest cell, all of the information for the organization of a complex creature such as ourselves can be stored. All this information---whether we have brown eyes, or whether we think at all, or that in the embryo the jawbone should first develop with a little hole in the side so that later a nerve can grow through it---all this information is contained in a very tiny fraction of the cell in the form of long-chain DNA molecules in which approximately 50 atoms are used for one bit of information about the cell.

Better electron microscopes

If I have written in a code, with 5 times 5 times 5 atoms to a bit, the question is: How could I read it today? The electron microscope is not quite good enough, with the greatest care and effort, it can only resolve about 10 angstroms. I would like to try and impress upon you while I am talking about all of these things on a small scale, the importance of improving the electron microscope by a hundred times. It is not impossible; it is not against the laws of diffraction of the electron. The wave length of the electron in such a microscope is only 1/20 of an angstrom. So it should be possible to see the individual atoms. What good would it be to see individual atoms distinctly?

We have friends in other fields---in biology, for instance. We physicists often look at them and say, ``You know the reason you fellows are making so little progress?" (Actually I don't know any field where they are making more rapid progress than they are in biology today.) ``You should use more mathematics, like we do." They could answer us---but they're polite, so I'll answer for them: ``What *you* should do in order for *us* to make more rapid progress is to make the electron microscope 100 times better."

What are the most central and fundamental problems of biology today? They are questions like: What is the sequence of bases in the DNA? What happens when you have a mutation? How is the base order in the DNA connected to the order of amino acids in the protein? What is the structure of the RNA; is it single-chain or double-chain, and how is it related in its order of bases to the DNA? What is the organization of the microsomes? How are proteins synthesized? Where does the RNA go? How does it sit? Where do the proteins sit? Where do the amino acids go in? In photosynthesis, where is the chlorophyll; how is it arranged; where are the carotenoids involved in this thing? What is the system of the conversion of light into chemical energy?

It is very easy to answer many of these fundamental biological questions; you just *look at the thing!* You will see the order of bases in the chain; you will see the structure of the microsome. Unfortunately, the present microscope sees at a scale which is just a bit too crude. Make the microscope one hundred times more powerful, and many problems of biology would be made very much easier. I exaggerate, of course, but the biologists would surely be very thankful to you---and they would prefer that to the criticism that they should use more mathematics.

The theory of chemical processes today is based on theoretical physics. In this sense, physics supplies the foundation of chemistry. But chemistry also has analysis. If you have a strange substance and you want to know what it is, you go through a long and complicated process of chemical analysis. You can analyze almost anything today, so I am a little late with my idea. But if the physicists wanted to, they could also dig under the chemists in the problem of chemical analysis. It would be very easy to make an analysis of any complicated chemical substance; all one would have to do would be to look at it and see where the atoms are. The only trouble is that the electron microscope is one hundred times too poor. (Later, I would like to ask the question: Can the physicists do something about the third problem of chemistry---namely, synthesis? Is there a *physical* way to synthesize any chemical substance?

The reason the electron microscope is so poor is that the f -value of the lenses is only 1 part to 1,000; you don't have a big enough numerical aperture. And I know that there are theorems which prove that it is impossible, with axially symmetrical stationary field lenses, to produce an f -value any bigger than so and so; and therefore the resolving power at the present time is at its theoretical maximum. But in every theorem there are assumptions. Why must the field be symmetrical? I put this out as a challenge: Is there no way to make the electron microscope more powerful?

The marvelous biological system

The biological example of writing information on a small scale has inspired me to think of something that should be possible. Biology is not simply writing information; it is *doing something* about it. A biological system can be exceedingly small. Many of the cells are very tiny, but they are very active; they manufacture various substances; they walk around; they wiggle; and they do all kinds of marvelous things---all on a very small scale. Also, they store information. Consider the possibility that we too can make a thing very small which does what we want---that we can manufacture an object that maneuvers at that level!

There may even be an economic point to this business of making things very small. Let me remind you of some of the problems of computing machines. In computers we have to store an enormous amount of information. The kind of writing that I was mentioning before, in which I had everything down as a distribution of metal, is permanent. Much more interesting to a computer is a way of writing, erasing, and writing something else. (This is usually because we don't want to waste the material on which we have just written. Yet if we could write it in a very small space, it wouldn't make any difference; it could just be thrown away after it was read. It doesn't cost very much for the material).

Miniaturizing the computer

I don't know how to do this on a small scale in a practical way, but I do know that computing machines are very large; they fill rooms. Why can't we make them very small, make them of little wires, little elements---and by little, I mean *little*. For instance, the wires should be 10 or 100 atoms in diameter, and the circuits should be a few thousand angstroms across. Everybody who has analyzed the logical theory of computers has come to the conclusion that the possibilities of computers are very interesting---if they could be made to be more complicated by several orders of magnitude. If they had millions of times as many elements, they could make judgments. They would have time to calculate what is the best way to make the calculation that they are about to make. They could select the method of analysis which, from their experience, is better than the one that we would give to them. And in many other ways, they would have new qualitative features.

If I look at your face I immediately recognize that I have seen it before. (Actually, my friends will say I have chosen an unfortunate example here for the subject of this illustration. At least I recognize that it is a *man* and not an *apple*.) Yet there is no machine which, with that speed, can take a picture of a face and say even that it is a man; and much less that it is the same man that you showed it before---unless it is exactly the same picture. If the face is changed; if I am closer to the face; if I am further from the face; if the light changes---I recognize it anyway. Now, this little computer I carry in my head is easily able to do that. The computers that we build are not able to do that. The number of elements in this bone box of mine are enormously greater than the number of elements in our "wonderful" computers. But our mechanical computers are too big; the elements in this box are microscopic. I want to make some that are *submicroscopic*.

If we wanted to make a computer that had all these marvelous extra qualitative abilities, we would have to make it, perhaps, the size of the Pentagon. This has several disadvantages. First, it requires too much material; there may not be enough germanium in the world for all the transistors which would have to be put into this enormous thing. There is also the problem of heat generation and power consumption; TVA would be needed to run the computer. But an even more practical difficulty is that the computer would be limited to a certain speed. Because of its large size, there is finite time required to get the information from one place to another. The information cannot go any faster than the speed of light---so, ultimately, when our computers get faster and faster and more and more elaborate, we will have to make them smaller and smaller.

But there is plenty of room to make them smaller. There is nothing that I can see in the physical laws that says the computer elements cannot be made enormously smaller than they are now. In fact, there may be certain advantages.

Miniaturization by evaporation

How can we make such a device? What kind of manufacturing processes would we use? One possibility we might consider, since we have talked about writing by putting atoms down in a certain arrangement, would be to evaporate the material, then evaporate the insulator next to it. Then, for the next layer, evaporate another position of a wire, another insulator, and so on. So, you simply evaporate until you have a block of stuff which has the elements--- coils and condensers, transistors and so on---of exceedingly fine dimensions.

But I would like to discuss, just for amusement, that there are other possibilities. Why can't we manufacture these small computers somewhat like we manufacture the big ones? Why can't we drill holes, cut things, solder things, stamp things out, mold different shapes all at an infinitesimal level? What are the limitations as to how small a thing has to be before you can no longer mold it? How many times when you are working on something frustratingly tiny like your wife's wrist watch, have you said to yourself, "If I could only train an ant to do this!" What I would like to suggest is the possibility of training an ant to train a mite to do this. What are the possibilities of small but movable machines? They may or may not be useful, but they surely would be fun to make.

Consider any machine---for example, an automobile---and ask about the problems of making an infinitesimal machine like it. Suppose, in the particular design of the automobile, we need a certain precision of the parts; we need an accuracy, let's suppose, of $4/10,000$ of an inch. If things are more inaccurate than that in the shape of the cylinder and so on, it isn't going to work very well. If I make the thing too small, I have to worry about the size of the atoms; I can't make a circle of "balls" so to speak, if the circle is too small. So, if I make the error, corresponding to $4/10,000$ of an inch, correspond to an error of 10 atoms, it turns out that I can reduce the dimensions of an automobile 4,000 times, approximately---so that it is 1 mm. across. Obviously, if you redesign the car so that it would work with a much larger tolerance, which is not at all impossible, then you could make a much smaller device.

It is interesting to consider what the problems are in such small machines. Firstly, with parts stressed to the same degree, the forces go as the area you are reducing, so that things like weight and inertia are of relatively no importance. The strength of material, in other words, is very much greater in proportion. The stresses and expansion of the flywheel from centrifugal force, for example, would be the same proportion only if the rotational speed is increased in the same proportion as we decrease the size. On the other hand, the metals that we use have a grain structure, and this would be very annoying at small scale because the material is not homogeneous. Plastics and glass and things of this amorphous nature are very much more homogeneous, and so we would have to make our machines out of such materials.

There are problems associated with the electrical part of the system---with the copper wires and the magnetic parts. The magnetic properties on a very small scale are not the same as on a large scale; there is the ``domain" problem involved. A big magnet made of millions of domains can only be made on a small scale with one domain. The electrical equipment won't simply be scaled down; it has to be redesigned. But I can see no reason why it can't be redesigned to work again.

Problems of lubrication

Lubrication involves some interesting points. The effective viscosity of oil would be higher and higher in proportion as we went down (and if we increase the speed as much as we can). If we don't increase the speed so much, and change from oil to kerosene or some other fluid, the problem is not so bad. But actually we may not have to lubricate at all! We have a lot of extra force. Let the bearings run dry; they won't run hot because the heat escapes away from such a small device very, very rapidly.

This rapid heat loss would prevent the gasoline from exploding, so an internal combustion engine is impossible. Other chemical reactions, liberating energy when cold, can be used. Probably an external supply of electrical power would be most convenient for such small machines.

What would be the utility of such machines? Who knows? Of course, a small automobile would only be useful for the mites to drive around in, and I suppose our Christian interests don't go that far. However, we did note the possibility of the manufacture of small elements for computers in completely automatic factories, containing lathes and other machine tools at the very small level. The small lathe would not have to be exactly like our big lathe. I leave to your imagination the improvement of the design to take full advantage of the properties of things on a small scale, and in such a way that the fully automatic aspect would be easiest to manage.

A friend of mine (Albert R. Hibbs) suggests a very interesting possibility for relatively small machines. He says that, although it is a very wild idea, it would be interesting in surgery if you could swallow the surgeon. You put the mechanical surgeon inside the blood vessel and it goes into the heart and ``looks" around. (Of course the information has to be fed out.) It finds out which valve is the faulty one and takes a little knife and slices it out. Other small machines might be permanently incorporated in the body to assist some inadequately-functioning organ.

Now comes the interesting question: How do we make such a tiny mechanism? I leave that to you. However, let me suggest one weird possibility. You know, in the atomic energy plants they have materials and machines that they can't handle directly because they have become radioactive. To unscrew nuts and put on bolts and so on, they have a set of master and slave hands, so that by operating a set of levers here, you control the ``hands" there, and can turn them this way and that so you can handle things quite nicely.

Most of these devices are actually made rather simply, in that there is a particular cable, like a marionette string, that goes directly from the controls to the ``hands." But, of course, things also have been made using servo motors, so that the connection between the one thing and the other is electrical rather than mechanical. When you turn the levers, they turn a servo motor, and it changes the electrical currents in the wires, which repositions a motor at the other end.

Now, I want to build much the same device---a master-slave system which operates electrically. But I want the slaves to be made especially carefully by modern large-scale machinists so that they are one-fourth the scale of the ``hands" that you ordinarily maneuver. So you have a scheme by which you can do things at one-quarter scale anyway---the little servo motors with little hands play with little nuts and bolts; they drill little holes; they are four times smaller. Aha! So I manufacture a quarter-size lathe; I manufacture quarter-size tools; and I make, at the one-quarter scale, still another set of hands again relatively one-quarter size! This is one-sixteenth size, from my point of view. And after I finish doing this I wire directly from my large-scale system, through transformers perhaps, to the one-sixteenth-size servo motors. Thus I can now manipulate the one-sixteenth size hands.

Well, you get the principle from there on. It is rather a difficult program, but it is a possibility. You might say that one can go much farther in one step than from one to four. Of course, this has all to be designed very carefully and it is not necessary simply to make it like hands. If you thought of it very carefully, you could probably arrive at a much better system for doing such things.

If you work through a pantograph, even today, you can get much more than a factor of four in even one step. But you can't work directly through a pantograph which makes a smaller pantograph which then makes a smaller pantograph---because of the looseness of the holes and the irregularities of construction. The end of the pantograph wiggles with a relatively greater irregularity than the irregularity with which you move your hands. In going down this scale, I would find the end of the pantograph on the end of the pantograph on the end of the pantograph shaking so badly that it wasn't doing anything sensible at all.

At each stage, it is necessary to improve the precision of the apparatus. If, for instance, having made a small lathe with a pantograph, we find its lead screw irregular---more irregular than the large-scale one---we could lap the lead screw against breakable nuts that you can reverse in the usual way back and forth until this lead screw is, at its scale, as accurate as our original lead screws, at our scale.

We can make flats by rubbing unflat surfaces in triplicates together---in three pairs---and the flats then become flatter than the thing you started with. Thus, it is not impossible to improve precision on a small scale by the correct operations. So, when we build this stuff, it is necessary at each step to improve the accuracy of the equipment by working for awhile down there, making accurate lead screws, Johansen blocks, and all the other materials which we use in accurate machine work at the higher level. We have to stop at each level and manufacture all the stuff to go to the next level---a very long and very difficult program. Perhaps you can figure a better way than that to get down to small scale more rapidly.

Yet, after all this, you have just got one little baby lathe four thousand times smaller than usual. But we were thinking of making an enormous computer, which we were going to build by drilling holes on this lathe to make little washers for the computer. How many washers can you manufacture on this one lathe?

A hundred tiny hands

When I make my first set of slave ``hands" at one-fourth scale, I am going to make ten sets. I make ten sets of ``hands," and I wire them to my original levers so they each do exactly the same thing at the same time in parallel. Now, when I am making my new devices one-quarter again as small, I let each one manufacture ten copies, so that I would have a hundred ``hands" at the 1/16th size.

Where am I going to put the million lathes that I am going to have? Why, there is nothing to it; the volume is much less than that of even one full-scale lathe. For instance, if I made a billion little lathes, each 1/4000 of the scale of a regular lathe, there are plenty of materials and space available because in the billion little ones there is less than 2 percent of the materials in one big lathe.

It doesn't cost anything for materials, you see. So I want to build a billion tiny factories, models of each other, which are manufacturing simultaneously, drilling holes, stamping parts, and so on.

As we go down in size, there are a number of interesting problems that arise. All things do not simply scale down in proportion. There is the problem that materials stick together by the molecular (Van der Waals) attractions. It would be like this: After you have made a part and you unscrew the nut from a bolt, it isn't going to fall down because the gravity isn't appreciable; it would even be hard to get it off the bolt. It would be like those old movies of a man with his hands full of molasses, trying to get rid of a glass of water. There will be several problems of this nature that we will have to be ready to design for.

Rearranging the atoms

But I am not afraid to consider the final question as to whether, ultimately---in the great future---we can arrange the atoms the way we want; the very *atoms*, all the way down! What would happen if we could arrange the atoms one by one the way we want them (within reason, of course; you can't put them so that they are chemically unstable, for example).

Up to now, we have been content to dig in the ground to find minerals. We heat them and we do things on a large scale with them, and we hope to get a pure substance with just so much impurity, and so on. But we must always accept some

atomic arrangement that nature gives us. We haven't got anything, say, with a ``checkerboard" arrangement, with the impurity atoms exactly arranged 1,000 angstroms apart, or in some other particular pattern.

What could we do with layered structures with just the right layers? What would the properties of materials be if we could really arrange the atoms the way we want them? They would be very interesting to investigate theoretically. I can't see exactly what would happen, but I can hardly doubt that when we have some *control* of the arrangement of things on a small scale we will get an enormously greater range of possible properties that substances can have, and of different things that we can do.

Consider, for example, a piece of material in which we make little coils and condensers (or their solid state analogs) 1,000 or 10,000 angstroms in a circuit, one right next to the other, over a large area, with little antennas sticking out at the other end---a whole series of circuits. Is it possible, for example, to emit light from a whole set of antennas, like we emit radio waves from an organized set of antennas to beam the radio programs to Europe? The same thing would be to *beam* the light out in a definite direction with very high intensity. (Perhaps such a beam is not very useful technically or economically.)

I have thought about some of the problems of building electric circuits on a small scale, and the problem of resistance is serious. If you build a corresponding circuit on a small scale, its natural frequency goes up, since the wave length goes down as the scale; but the skin depth only decreases with the square root of the scale ratio, and so resistive problems are of increasing difficulty. Possibly we can beat resistance through the use of superconductivity if the frequency is not too high, or by other tricks.

Atoms in a small world

When we get to the very, very small world---say circuits of seven atoms---we have a lot of new things that would happen that represent completely new opportunities for design. Atoms on a small scale behave like *nothing* on a large scale, for they satisfy the laws of quantum mechanics. So, as we go down and fiddle around with the atoms down there, we are working with different laws, and we can expect to do different things. We can manufacture in different ways. We can use, not just circuits, but some system involving the quantized energy levels, or the interactions of quantized spins, etc.

Another thing we will notice is that, if we go down far enough, all of our devices can be mass produced so that they are absolutely perfect copies of one another. We cannot build two large machines so that the dimensions are exactly the same. But if your machine is only 100 atoms high, you only have to get it correct to one-half of one percent to make sure the other machine is exactly the same size---namely, 100 atoms high!

At the atomic level, we have new kinds of forces and new kinds of possibilities, new kinds of effects. The problems of manufacture and reproduction of materials will be quite different. I am, as I said, inspired by the biological phenomena in which chemical forces are used in repetitious fashion to produce all kinds of weird effects (one of which is the author).

The principles of physics, as far as I can see, do not speak against the possibility of maneuvering things atom by atom. It is not an attempt to violate any laws; it is something, in principle, that can be done; but in practice, it has not been done because we are too big.

Ultimately, we can do chemical synthesis. A chemist comes to us and says, ``Look, I want a molecule that has the atoms arranged thus and so; make me that molecule." The chemist does a mysterious thing when he wants to make a molecule. He sees that it has got that ring, so he mixes this and that, and he shakes it, and he fiddles around. And, at the end of a difficult process, he usually does succeed in synthesizing what he wants. By the time I get my devices working, so that we can do it by physics, he will have figured out how to synthesize absolutely anything, so that this will really be useless.

But it is interesting that it would be, in principle, possible (I think) for a physicist to synthesize any chemical substance that the chemist writes down. Give the orders and the physicist synthesizes it. How? Put the atoms down where the chemist says, and so you make the substance. The problems of chemistry and biology can be greatly helped if our

ability to see what we are doing, and to do things on an atomic level, is ultimately developed---a development which I think cannot be avoided.

Now, you might say, ``Who should do this and why should they do it?" Well, I pointed out a few of the economic applications, but I know that the reason that you would do it might be just for fun. But have some fun! Let's have a competition between laboratories. Let one laboratory make a tiny motor which it sends to another lab which sends it back with a thing that fits inside the shaft of the first motor.

High school competition

Just for the fun of it, and in order to get kids interested in this field, I would propose that someone who has some contact with the high schools think of making some kind of high school competition. After all, we haven't even started in this field, and even the kids can write smaller than has ever been written before. They could have competition in high schools. The Los Angeles high school could send a pin to the Venice high school on which it says, ``How's this?" They get the pin back, and in the dot of the ``i" it says, ``Not so hot."

Perhaps this doesn't excite you to do it, and only economics will do so. Then I want to do something; but I can't do it at the present moment, because I haven't prepared the ground. It is my intention to offer a prize of \$1,000 to the first guy who can take the information on the page of a book and put it on an area $1/25,000$ smaller in linear scale in such manner that it can be read by an electron microscope.

And I want to offer another prize---if I can figure out how to phrase it so that I don't get into a mess of arguments about definitions---of another \$1,000 to the first guy who makes an operating electric motor---a rotating electric motor which can be controlled from the outside and, not counting the lead-in wires, is only $1/64$ inch cube.

I do not expect that such prizes will have to wait very long for claimants.

Turing, A.M. (1950). Computing machinery and intelligence. *Mind*, 59, 433-460.

COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

1. The Imitation Game

I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the 'imitation game.' It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either "X is A and Y is B" or "X is B and Y is A." The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?

Now suppose X is actually A, then A must answer. It is A's object in the game to try and cause C to make the wrong identification. His answer might therefore be:

"My hair is shingled, and the longest strands are about nine inches long."

In order that tones of voice may not help the interrogator the answers should be written, or better still, typewritten. The ideal arrangement is to have a teleprinter communicating between the two rooms. Alternatively the question and answers can be repeated by an intermediary. The object of the game for the third player (B) is to help the interrogator. The best strategy for her is probably to give truthful answers. She can add such things as "I am the woman, don't listen to him!" to her answers, but it will avail nothing as the man can make similar remarks.

We now ask the question, "What will happen when a machine takes the part of A in this game?" Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, "Can machines think?"

2. Critique of the New Problem

As well as asking, "What is the answer to this new form of the question," one may ask, "Is this new question a worthy one to investigate?" This latter question we investigate without further ado, thereby cutting short an infinite regress.

The new problem has the advantage of drawing a fairly sharp line between the physical and the intellectual capacities of a man. No engineer or chemist claims to be able to produce a material which is indistinguishable from the human skin. It is possible that at some time this might be done, but even supposing this invention available we should feel there was little point in trying to make a "thinking machine" more human by dressing it up in such artificial flesh. The form in which we have set the problem reflects this fact in the condition which prevents the interrogator from seeing or touching the other competitors, or hearing their voices. Some other advantages of the proposed criterion may be shown up by specimen questions and answers. Thus:

Q: Please write me a sonnet on the subject of the Forth Bridge.

A : Count me out on this one. I never could write poetry.

Q: Add 34957 to 70764.

A: (Pause about 30 seconds and then give as answer) 105621.

Q: Do you play chess?

A: Yes.

Q: I have K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play?

A: (After a pause of 15 seconds) R-R8 mate.

The question and answer method seems to be suitable for introducing almost any one of the fields of human endeavour that we wish to include. We do not wish to penalise the machine for its inability to shine in beauty competitions, nor to penalise a man for losing in a race against an aeroplane. The conditions of our game make these disabilities irrelevant. The "witnesses" can brag, if they consider it advisable, as much as they please about their charms, strength or heroism, but the interrogator cannot demand practical demonstrations.

The game may perhaps be criticised on the ground that the odds are weighted too heavily against the machine. If the man were to try and pretend to be the machine he would clearly make a very poor showing. He would be given away at once by slowness and inaccuracy in arithmetic. May not machines carry out something which ought to be described as thinking but which is very different from what a man does? This objection is a very strong one, but at least we can say that if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection.

It might be urged that when playing the "imitation game" the best strategy for the machine may possibly be something other than imitation of the behaviour of a man. This may be, but I think it is unlikely that there is any great effect of this kind. In any case there is no intention to investigate here the theory of the game, and it will be assumed that the best strategy is to try to provide answers that would naturally be given by a man.

3. The Machines Concerned in the Game

The question which we put in 1 will not be quite definite until we have specified what we mean by the word "machine." It is natural that we should wish to permit every kind of engineering technique to be used in our machines. We also wish to allow the possibility that an engineer or team of engineers may construct a machine which works, but whose manner of operation cannot be satisfactorily described by its constructors because they have applied a method which is largely experimental. Finally, we wish to exclude from the machines men born in the usual manner. It is difficult to frame the definitions so as to satisfy these three conditions. One might for instance insist that the team of engineers should be all of one sex, but this would not really be satisfactory, for it is probably possible to rear a complete individual from a single cell of the skin (say) of a man. To do so would be a feat of biological technique deserving of the very highest praise, but we would not be inclined to regard it as a case of "constructing a thinking machine." This prompts us to abandon the requirement that every kind of technique should be permitted. We are the more ready to do so in view of the fact that the present interest in "thinking machines" has been aroused by a particular kind of machine, usually called an "electronic computer" or "digital computer." Following this suggestion we only permit digital computers to take part in our game.

This restriction appears at first sight to be a very drastic one. I shall attempt to show that it is not so in reality. To do this necessitates a short account of the nature and properties of these computers.

It may also be said that this identification of machines with digital computers, like our criterion for "thinking," will only be unsatisfactory if (contrary to my belief), it turns out that digital computers are unable to give a good showing in the game.

There are already a number of digital computers in working order, and it may be asked, "Why not try the experiment straight away? It would be easy to satisfy the conditions of the game. A number of interrogators could be used, and

statistics compiled to show how often the right identification was given." The short answer is that we are not asking whether all digital computers would do well in the game nor whether the computers at present available would do well, but whether there are imaginable computers which would do well. But this is only the short answer. We shall see this question in a different light later.

4. Digital Computers

The idea behind digital computers may be explained by saying that these machines are intended to carry out any operations which could be done by a human computer. The human computer is supposed to be following fixed rules; he has no authority to deviate from them in any detail. We may suppose that these rules are supplied in a book, which is altered whenever he is put on to a new job. He has also an unlimited supply of paper on which he does his calculations. He may also do his multiplications and additions on a "desk machine," but this is not important.

If we use the above explanation as a definition we shall be in danger of circularity of argument. We avoid this by giving an outline of the means by which the desired effect is achieved. A digital computer can usually be regarded as consisting of three parts:

- (i) Store.
- (ii) Executive unit.
- (iii) Control.

The store is a store of information, and corresponds to the human computer's paper, whether this is the paper on which he does his calculations or that on which his book of rules is printed. In so far as the human computer does calculations in his head a part of the store will correspond to his memory.

The executive unit is the part which carries out the various individual operations involved in a calculation. What these individual operations are will vary from machine to machine. Usually fairly lengthy operations can be done such as "Multiply 3540675445 by 7076345687" but in some machines only very simple ones such as "Write down 0" are possible.

We have mentioned that the "book of rules" supplied to the computer is replaced in the machine by a part of the store. It is then called the "table of instructions." It is the duty of the control to see that these instructions are obeyed correctly and in the right order. The control is so constructed that this necessarily happens.

The information in the store is usually broken up into packets of moderately small size. In one machine, for instance, a packet might consist of ten decimal digits. Numbers are assigned to the parts of the store in which the various packets of information are stored, in some systematic manner. A typical instruction might say-

"Add the number stored in position 6809 to that in 4302 and put the result back into the latter storage position."

Needless to say it would not occur in the machine expressed in English. It would more likely be coded in a form such as 6809430217. Here 17 says which of various possible operations is to be performed on the two numbers. In this case the operation is that described above, viz., "Add the number. . . ." It will be noticed that the instruction takes up 10 digits and so forms one packet of information, very conveniently. The control will normally take the instructions to be obeyed in the order of the positions in which they are stored, but occasionally an instruction such as

"Now obey the instruction stored in position 5606, and continue from there"

may be encountered, or again

"If position 4505 contains 0 obey next the instruction stored in 6707, otherwise continue straight on."

Instructions of these latter types are very important because they make it possible for a sequence of operations to be replaced over and over again until some condition is fulfilled, but in doing so to obey, not fresh instructions on each repetition, but the same ones over and over again. To take a domestic analogy. Suppose Mother wants Tommy to call at the cobbler's every morning on his way to school to see if her shoes are done, she can ask him afresh every morning. Alternatively she can stick up a notice once and for all in the hall which he will see when he leaves for school and which tells him to call for the shoes, and also to destroy the notice when he comes back if he has the shoes with him.

The reader must accept it as a fact that digital computers can be constructed, and indeed have been constructed, according to the principles we have described, and that they can in fact mimic the actions of a human computer very closely.

The book of rules which we have described our human computer as using is of course a convenient fiction. Actual human computers really remember what they have got to do. If one wants to make a machine mimic the behaviour of the human computer in some complex operation one has to ask him how it is done, and then translate the answer into the form of an instruction table. Constructing instruction tables is usually described as "programming." To "programme a machine to carry out the operation A" means to put the appropriate instruction table into the machine so that it will do A.

An interesting variant on the idea of a digital computer is a "digital computer with a random element." These have instructions involving the throwing of a die or some equivalent electronic process; one such instruction might for instance be, "Throw the die and put the-resulting number into store 1000." Sometimes such a machine is described as having free will (though I would not use this phrase myself). It is not normally possible to determine from observing a machine whether it has a random element, for a similar effect can be produced by such devices as making the choices depend on the digits of the decimal for .

Most actual digital computers have only a finite store. There is no theoretical difficulty in the idea of a computer with an unlimited store. Of course only a finite part can have been used at any one time. Likewise only a finite amount can have been constructed, but we can imagine more and more being added as required. Such computers have special theoretical interest and will be called infinitive capacity computers.

The idea of a digital computer is an old one. Charles Babbage, Lucasian Professor of Mathematics at Cambridge from 1828 to 1839, planned such a machine, called the Analytical Engine, but it was never completed. Although Babbage had all the essential ideas, his machine was not at that time such a very attractive prospect. The speed which would have been available would be definitely faster than a human computer but something like 1 00 times slower than the Manchester machine, itself one of the slower of the modern machines, The storage was to be purely mechanical, using wheels and cards.

The fact that Babbage's Analytical Engine was to be entirely mechanical will help us to rid ourselves of a superstition. Importance is often attached to the fact that modern digital computers are electrical, and that the nervous system also is electrical. Since Babbage's machine was not electrical, and since all digital computers are in a sense equivalent, we see that this use of electricity cannot be of theoretical importance. Of course electricity usually comes in where fast signalling is concerned, so that it is not surprising that we find it in both these connections. In the nervous system chemical phenomena are at least as important as electrical. In certain computers the storage system is mainly acoustic. The feature of using electricity is thus seen to be only a very superficial similarity. If we wish to find such similarities we should look rather for mathematical analogies of function.

5. Universality of Digital Computers

The digital computers considered in the last section may be classified amongst the "discrete-state machines." These are the machines which move by sudden jumps or clicks from one quite definite state to another. These states are sufficiently different for the possibility of confusion between them to be ignored. Strictly speaking there are no such machines. Everything really moves continuously. But there are many kinds of machine which can profitably be thought of as being discrete-state machines. For instance in considering the switches for a lighting system it is a convenient fiction that each switch must be definitely on or definitely off. There must be intermediate positions, but for most purposes we can forget about them. As an example of a discrete-state machine we might consider a wheel which clicks round through 120 once a second, but may be stopped by a lever which can be operated from outside; in addition a lamp

is to light in one of the positions of the wheel. This machine could be described abstractly as follows. The internal state of the machine (which is described by the position of the wheel) may be q_1 , q_2 or q_3 . There is an input signal i_0 . or i_1 (position of lever). The internal state at any moment is determined by the last state and input signal according to the table

(TABLE DELETED)

The output signals, the only externally visible indication of the internal state (the light) are described by the table

State q_1 q_2 q_3

output o_0 o_1

This example is typical of discrete-state machines. They can be described by such tables provided they have only a finite number of possible states.

It will seem that given the initial state of the machine and the input signals it is always possible to predict all future states. This is reminiscent of Laplace's view that from the complete state of the universe at one moment of time, as described by the positions and velocities of all particles, it should be possible to predict all future states. The prediction which we are considering is, however, rather nearer to practicability than that considered by Laplace. The system of the "universe as a whole" is such that quite small errors in the initial conditions can have an overwhelming effect at a later time. The displacement of a single electron by a billionth of a centimetre at one moment might make the difference between a man being killed by an avalanche a year later, or escaping. It is an essential property of the mechanical systems which we have called "discrete-state machines" that this phenomenon does not occur. Even when we consider the actual physical machines instead of the idealised machines, reasonably accurate knowledge of the state at one moment yields reasonably accurate knowledge any number of steps later.

As we have mentioned, digital computers fall within the class of discrete-state machines. But the number of states of which such a machine is capable is usually enormously large. For instance, the number for the machine now working at Manchester is about $2^{165,000}$, i.e., about $10^{50,000}$. Compare this with our example of the clicking wheel described above, which had three states. It is not difficult to see why the number of states should be so immense. The computer includes a store corresponding to the paper used by a human computer. It must be possible to write into the store any one of the combinations of symbols which might have been written on the paper. For simplicity suppose that only digits from 0 to 9 are used as symbols. Variations in handwriting are ignored. Suppose the computer is allowed 100 sheets of paper each containing 50 lines each with room for 30 digits. Then the number of states is $10^{100 \times 50 \times 30}$ i.e., $10^{150,000}$. This is about the number of states of three Manchester machines put together. The logarithm to the base two of the number of states is usually called the "storage capacity" of the machine. Thus the Manchester machine has a storage capacity of about 165,000 and the wheel machine of our example about 1.6. If two machines are put together their capacities must be added to obtain the capacity of the resultant machine. This leads to the possibility of statements such as "The Manchester machine contains 64 magnetic tracks each with a capacity of 2560, eight electronic tubes with a capacity of 1280. Miscellaneous storage amounts to about 300 making a total of 174,380."

Given the table corresponding to a discrete-state machine it is possible to predict what it will do. There is no reason why this calculation should not be carried out by means of a digital computer. Provided it could be carried out sufficiently quickly the digital computer could mimic the behavior of any discrete-state machine. The imitation game could then be played with the machine in question (as B) and the mimicking digital computer (as A) and the interrogator would be unable to distinguish them. Of course the digital computer must have an adequate storage capacity as well as working sufficiently fast. Moreover, it must be programmed afresh for each new machine which it is desired to mimic.

This special property of digital computers, that they can mimic any discrete-state machine, is described by saying that they are universal machines. The existence of machines with this property has the important consequence that, considerations of speed apart, it is unnecessary to design various new machines to do various computing processes. They can all be done with one digital computer, suitably programmed for each case. It will be seen that as a consequence of this all digital computers are in a sense equivalent.

We may now consider again the point raised at the end of §3. It was suggested tentatively that the question, "Can machines think?" should be replaced by "Are there imaginable digital computers which would do well in the imitation game?" If we wish we can make this superficially more general and ask "Are there discrete-state machines which would do well?" But in view of the universality property we see that either of these questions is equivalent to this, "Let us fix our attention on one particular digital computer C. Is it true that by modifying this computer to have an adequate storage, suitably increasing its speed of action, and providing it with an appropriate programme, C can be made to play satisfactorily the part of A in the imitation game, the part of B being taken by a man?"

6. Contrary Views on the Main Question

We may now consider the ground to have been cleared and we are ready to proceed to the debate on our question, "Can machines think?" and the variant of it quoted at the end of the last section. We cannot altogether abandon the original form of the problem, for opinions will differ as to the appropriateness of the substitution and we must at least listen to what has to be said in this connexion.

It will simplify matters for the reader if I explain first my own beliefs in the matter. Consider first the more accurate form of the question. I believe that in about fifty years' time it will be possible, to programme computers, with a storage capacity of about 10^9 , to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning. The original question, "Can machines think?" I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted. I believe further that no useful purpose is served by concealing these beliefs. The popular view that scientists proceed inexorably from well-established fact to well-established fact, never being influenced by any improved conjecture, is quite mistaken. Provided it is made clear which are proved facts and which are conjectures, no harm can result. Conjectures are of great importance since they suggest useful lines of research.

I now proceed to consider opinions opposed to my own.

(1) The Theological Objection

Thinking is a function of man's immortal soul. God has given an immortal soul to every man and woman, but not to any other animal or to machines. Hence no animal or machine can think.

I am unable to accept any part of this, but will attempt to reply in theological terms. I should find the argument more convincing if animals were classed with men, for there is a greater difference, to my mind, between the typical animate and the inanimate than there is between man and the other animals. The arbitrary character of the orthodox view becomes clearer if we consider how it might appear to a member of some other religious community. How do Christians regard the Moslem view that women have no souls? But let us leave this point aside and return to the main argument. It appears to me that the argument quoted above implies a serious restriction of the omnipotence of the Almighty. It is admitted that there are certain things that He cannot do such as making one equal to two, but should we not believe that He has freedom to confer a soul on an elephant if He sees fit? We might expect that He would only exercise this power in conjunction with a mutation which provided the elephant with an appropriately improved brain to minister to the needs of this sort. An argument of exactly similar form may be made for the case of machines. It may seem different because it is more difficult to "swallow." But this really only means that we think it would be less likely that He would consider the circumstances suitable for conferring a soul. The circumstances in question are discussed in the rest of this paper. In attempting to construct such machines we should not be irreverently usurping His power of creating souls, any more than we are in the procreation of children: rather we are, in either case, instruments of His will providing mansions for the souls that He creates.

However, this is mere speculation. I am not very impressed with theological arguments whatever they may be used to support. Such arguments have often been found unsatisfactory in the past. In the time of Galileo it was argued that the texts, "And the sun stood still . . . and hasted not to go down about a whole day" (Joshua x. 13) and "He laid the foundations of the earth, that it should not move at any time" (Psalm cv. 5) were an adequate refutation of the Copernican theory. With our present knowledge such an argument appears futile. When that knowledge was not available it made a quite different impression.

(2) The "Heads in the Sand" Objection

The consequences of machines thinking would be too dreadful. Let us hope and believe that they cannot do so."

This argument is seldom expressed quite so openly as in the form above. But it affects most of us who think about it at all. We like to believe that Man is in some subtle way superior to the rest of creation. It is best if he can be shown to be necessarily superior, for then there is no danger of him losing his commanding position. The popularity of the theological argument is clearly connected with this feeling. It is likely to be quite strong in intellectual people, since they value the power of thinking more highly than others, and are more inclined to base their belief in the superiority of Man on this power.

I do not think that this argument is sufficiently substantial to require refutation. Consolation would be more appropriate: perhaps this should be sought in the transmigration of souls.

(3) The Mathematical Objection

There are a number of results of mathematical logic which can be used to show that there are limitations to the powers of discrete-state machines. The best known of these results is known as Godel's theorem (1931) and shows that in any sufficiently powerful logical system statements can be formulated which can neither be proved nor disproved within the system, unless possibly the system itself is inconsistent. There are other, in some respects similar, results due to Church (1936), Kleene (1935), Rosser, and Turing (1937). The latter result is the most convenient to consider, since it refers directly to machines, whereas the others can only be used in a comparatively indirect argument: for instance if Godel's theorem is to be used we need in addition to have some means of describing logical systems in terms of machines, and machines in terms of logical systems. The result in question refers to a type of machine which is essentially a digital computer with an infinite capacity. It states that there are certain things that such a machine cannot do. If it is rigged up to give answers to questions as in the imitation game, there will be some questions to which it will either give a wrong answer, or fail to give an answer at all however much time is allowed for a reply. There may, of course, be many such questions, and questions which cannot be answered by one machine may be satisfactorily answered by another. We are of course supposing for the present that the questions are of the kind to which an answer "Yes" or "No" is appropriate, rather than questions such as "What do you think of Picasso?" The questions that we know the machines must fail on are of this type, "Consider the machine specified as follows. . . . Will this machine ever answer 'Yes' to any question?" The dots are to be replaced by a description of some machine in a standard form, which could be something like that used in §5. When the machine described bears a certain comparatively simple relation to the machine which is under interrogation, it can be shown that the answer is either wrong or not forthcoming. This is the mathematical result: it is argued that it proves a disability of machines to which the human intellect is not subject.

The short answer to this argument is that although it is established that there are limitations to the Powers of any particular machine, it has only been stated, without any sort of proof, that no such limitations apply to the human intellect. But I do not think this view can be dismissed quite so lightly. Whenever one of these machines is asked the appropriate critical question, and gives a definite answer, we know that this answer must be wrong, and this gives us a certain feeling of superiority. Is this feeling illusory? It is no doubt quite genuine, but I do not think too much importance should be attached to it. We too often give wrong answers to questions ourselves to be justified in being very pleased at such evidence of fallibility on the part of the machines. Further, our superiority can only be felt on such an occasion in relation to the one machine over which we have scored our petty triumph. There would be no question of triumphing simultaneously over all machines. In short, then, there might be men cleverer than any given machine, but then again there might be other machines cleverer again, and so on.

Those who hold to the mathematical argument would, I think, mostly be willing to accept the imitation game as a basis for discussion. Those who believe in the two previous objections would probably not be interested in any criteria.

(4) The Argument from Consciousness

This argument is very well expressed in Professor Jefferson's Lister Oration for 1949, from which I quote. "Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain—that is, not only write it but know that it had written it. No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at its successes, grief when its

valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry or depressed when it cannot get what it wants."

This argument appears to be a denial of the validity of our test. According to the most extreme form of this view the only way by which one could be sure that machine thinks is to be the machine and to feel oneself thinking. One could then describe these feelings to the world, but of course no one would be justified in taking any notice. Likewise according to this view the only way to know that a man thinks is to be that particular man. It is in fact the solipsist point of view. It may be the most logical view to hold but it makes communication of ideas difficult. A is liable to believe "A thinks but B does not" whilst B believes "B thinks but A does not." instead of arguing continually over this point it is usual to have the polite convention that everyone thinks.

I am sure that Professor Jefferson does not wish to adopt the extreme and solipsist point of view. Probably he would be quite willing to accept the imitation game as a test. The game (with the player B omitted) is frequently used in practice under the name of *viva voce* to discover whether some one really understands something or has "learnt it parrot fashion." Let us listen in to a part of such a *viva voce*:

Interrogator: In the first line of your sonnet which reads "Shall I compare thee to a summer's day," would not "a spring day" do as well or better?

Witness: It wouldn't scan.

Interrogator: How about "a winter's day," That would scan all right.

Witness: Yes, but nobody wants to be compared to a winter's day.

Interrogator: Would you say Mr. Pickwick reminded you of Christmas?

Witness: In a way.

Interrogator: Yet Christmas is a winter's day, and I do not think Mr. Pickwick would mind the comparison.

Witness: I don't think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas.

And so on, What would Professor Jefferson say if the sonnet-writing machine was able to answer like this in the *viva voce*? I do not know whether he would regard the machine as "merely artificially signalling" these answers, but if the answers were as satisfactory and sustained as in the above passage I do not think he would describe it as "an easy contrivance." This phrase is, I think, intended to cover such devices as the inclusion in the machine of a record of someone reading a sonnet, with appropriate switching to turn it on from time to time.

In short then, I think that most of those who support the argument from consciousness could be persuaded to abandon it rather than be forced into the solipsist position. They will then probably be willing to accept our test.

I do not wish to give the impression that I think there is no mystery about consciousness. There is, for instance, something of a paradox connected with any attempt to localise it. But I do not think these mysteries necessarily need to be solved before we can answer the question with which we are concerned in this paper.

(5) Arguments from Various Disabilities

These arguments take the form, "I grant you that you can make machines do all the things you have mentioned but you will never be able to make one to do X." Numerous features X are suggested in this connexion I offer a selection:

Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humour, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make some one fall in love with it, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behaviour as a man, do something really new.

No support is usually offered for these statements. I believe they are mostly founded on the principle of scientific induction. A man has seen thousands of machines in his lifetime. From what he sees of them he draws a number of general conclusions. They are ugly, each is designed for a very limited purpose, when required for a minutely different purpose they are useless, the variety of behaviour of any one of them is very small, etc., etc. Naturally he concludes that these are necessary properties of machines in general. Many of these limitations are associated with the very small storage capacity of most machines. (I am assuming that the idea of storage capacity is extended in some way to cover machines other than discrete-state machines. The exact definition does not matter as no mathematical accuracy is claimed in the present discussion.) A few years ago, when very little had been heard of digital computers, it was possible to elicit much incredulity concerning them, if one mentioned their properties without describing their construction. That was presumably due to a similar application of the principle of scientific induction. These applications of the principle are of course largely unconscious. When a burnt child fears the fire and shows that he fears it by avoiding it, I should say that he was applying scientific induction. (I could of course also describe his behaviour in many other ways.) The works and customs of mankind do not seem to be very suitable material to which to apply scientific induction. A very large part of space-time must be investigated, if reliable results are to be obtained. Otherwise we may (as most English 'Children do) decide that everybody speaks English, and that it is silly to learn French.

There are, however, special remarks to be made about many of the disabilities that have been mentioned. The inability to enjoy strawberries and cream may have struck the reader as frivolous. Possibly a machine might be made to enjoy this delicious dish, but any attempt to make one do so would be idiotic. What is important about this disability is that it contributes to some of the other disabilities, e.g., to the difficulty of the same kind of friendliness occurring between man and machine as between white man and white man, or between black man and black man.

The claim that "machines cannot make mistakes" seems a curious one. One is tempted to retort, "Are they any the worse for that?" But let us adopt a more sympathetic attitude, and try to see what is really meant. I think this criticism can be explained in terms of the imitation game. It is claimed that the interrogator could distinguish the machine from the man simply by setting them a number of problems in arithmetic. The machine would be unmasked because of its deadly accuracy. The reply to this is simple. The machine (programmed for playing the game) would not attempt to give the right answers to the arithmetic problems. It would deliberately introduce mistakes in a manner calculated to confuse the interrogator. A mechanical fault would probably show itself through an unsuitable decision as to what sort of a mistake to make in the arithmetic. Even this interpretation of the criticism is not sufficiently sympathetic. But we cannot afford the space to go into it much further. It seems to me that this criticism depends on a confusion between two kinds of mistake. We may call them "errors of functioning" and "errors of conclusion." Errors of functioning are due to some mechanical or electrical fault which causes the machine to behave otherwise than it was designed to do. In philosophical discussions one likes to ignore the possibility of such errors; one is therefore discussing "abstract machines." These abstract machines are mathematical fictions rather than physical objects. By definition they are incapable of errors of functioning. In this sense we can truly say that "machines can never make mistakes." Errors of conclusion can only arise when some meaning is attached to the output signals from the machine. The machine might, for instance, type out mathematical equations, or sentences in English. When a false proposition is typed we say that the machine has committed an error of conclusion. There is clearly no reason at all for saying that a machine cannot make this kind of mistake. It might do nothing but type out repeatedly " $O = I$." To take a less perverse example, it might have some method for drawing conclusions by scientific induction. We must expect such a method to lead occasionally to erroneous results.

The claim that a machine cannot be the subject of its own thought can of course only be answered if it can be shown that the machine has some thought with some subject matter. Nevertheless, "the subject matter of a machine's operations" does seem to mean something, at least to the people who deal with it. If, for instance, the machine was trying to find a solution of the equation $x^2 - 40x - 11 = 0$ one would be tempted to describe this equation as part of the machine's subject matter at that moment. In this sort of sense a machine undoubtedly can be its own subject matter. It may be used to help in making up its own programmes, or to predict the effect of alterations in its own structure. By observing the results of its own behaviour it can modify its own programmes so as to achieve some purpose more effectively. These are possibilities of the near future, rather than Utopian dreams.

The criticism that a machine cannot have much diversity of behaviour is just a way of saying that it cannot have much storage capacity. Until fairly recently a storage capacity of even a thousand digits was very rare.

The criticisms that we are considering here are often disguised forms of the argument from consciousness. Usually if one maintains that a machine can do one of these things, and describes the kind of method that the machine could use, one will not make much of an impression. It is thought that tile method (whatever it may be, for it must be mechanical) is really rather base. Compare the parentheses in Jefferson's statement quoted on page 22.

(6) Lady Lovelace's Objection

Our most detailed information of Babbage's Analytical Engine comes from a memoir by Lady Lovelace (1842). In it she states, "The Analytical Engine has no pretensions to *originate* anything. It can do *whatever we know how to order it to perform*" (her italics). This statement is quoted by Hartree (1949) who adds: "This does not imply that it may not be possible to construct electronic equipment which will 'think for itself,' or in which, in biological terms, one could set up a conditioned reflex, which would serve as a basis for 'learning.' Whether this is possible in principle or not is a stimulating and exciting question, suggested by some of these recent developments. But it did not seem that the machines constructed or projected at the time had this property."

I am in thorough agreement with Hartree over this. It will be noticed that he does not assert that the machines in question had not got the property, but rather that the evidence available to Lady Lovelace did not encourage her to believe that they had it. It is quite possible that the machines in question had in a sense got this property. For suppose that some discrete-state machine has the property. The Analytical Engine was a universal digital computer, so that, if its storage capacity and speed were adequate, it could by suitable programming be made to mimic the machine in question. Probably this argument did not occur to the Countess or to Babbage. In any case there was no obligation on them to claim all that could be claimed.

This whole question will be considered again under the heading of learning machines.

A variant of Lady Lovelace's objection states that a machine can "never do anything really new." This may be parried for a moment with the saw, "There is nothing new under the sun." Who can be certain that "original work" that he has done was not simply the growth of the seed planted in him by teaching, or the effect of following well-known general principles. A better variant of the objection says that a machine can never "take us by surprise." This statement is a more direct challenge and can be met directly. Machines take me by surprise with great frequency. This is largely because I do not do sufficient calculation to decide what to expect them to do, or rather because, although I do a calculation, I do it in a hurried, slipshod fashion, taking risks. Perhaps I say to myself, "I suppose the Voltage here ought to be the same as there: anyway let's assume it is." Naturally I am often wrong, and the result is a surprise for me. For by the time the experiment is done these assumptions have been forgotten. These admissions lay me open to lectures on the subject of my vicious ways, but do not throw any doubt on my credibility when I testify to the surprises I experience.

I do not expect this reply to silence my critic. He will probably say that his surprises are due to some creative mental act on my part, and reflect no credit on the machine. This leads us back to the argument from consciousness, and far from the idea of surprise. It is a line of argument we must consider closed, but it is perhaps worth remarking that the appreciation of something as surprising requires as much of a "creative mental act" whether the surprising event originates from a man, a book, a machine or anything else.

The view that machines cannot give rise to surprises is due, I believe, to a fallacy to which philosophers and mathematicians are particularly subject. This is the assumption that as soon as a fact is presented to a mind all consequences of that fact spring into the mind simultaneously with it. It is a very useful assumption under many circumstances, but one too easily forgets that it is false. A natural consequence of doing so is that one then assumes that there is no virtue in the mere working out of consequences from data and general principles.

(7) Argument from Continuity in the Nervous System

The nervous system is certainly not a discrete-state machine. A small error in the information about the size of a nervous impulse impinging on a neuron, may make a large difference to the size of the outgoing impulse. It may be

argued that, this being so, one cannot expect to be able to mimic the behaviour of the nervous system with a discrete-state system.

It is true that a discrete-state machine must be different from a continuous machine. But if we adhere to the conditions of the imitation game, the interrogator will not be able to take any advantage of this difference. The situation can be made clearer if we consider some other simpler continuous machine. A differential analyser will do very well. (A differential analyser is a certain kind of machine not of the discrete-state type used for some kinds of calculation.) Some of these provide their answers in a typed form, and so are suitable for taking part in the game. It would not be possible for a digital computer to predict exactly what answers the differential analyser would give to a problem, but it would be quite capable of giving the right sort of answer. For instance, if asked to give the value of (actually about 3.1416) it would be reasonable to choose at random between the values 3.12, 3.13, 3.14, 3.15, 3.16 with the probabilities of 0.05, 0.15, 0.55, 0.19, 0.06 (say). Under these circumstances it would be very difficult for the interrogator to distinguish the differential analyser from the digital computer.

(8) The Argument from Informality of Behaviour

It is not possible to produce a set of rules purporting to describe what a man should do in every conceivable set of circumstances. One might for instance have a rule that one is to stop when one sees a red traffic light, and to go if one sees a green one, but what if by some fault both appear together? One may perhaps decide that it is safest to stop. But some further difficulty may well arise from this decision later. To attempt to provide rules of conduct to cover every eventuality, even those arising from traffic lights, appears to be impossible. With all this I agree.

From this it is argued that we cannot be machines. I shall try to reproduce the argument, but I fear I shall hardly do it justice. It seems to run something like this. "if each man had a definite set of rules of conduct by which he regulated his life he would be no better than a machine. But there are no such rules, so men cannot be machines." The undistributed middle is glaring. I do not think the argument is ever put quite like this, but I believe this is the argument used nevertheless. There may however be a certain confusion between "rules of conduct" and "laws of behaviour" to cloud the issue. By "rules of conduct" I mean precepts such as "Stop if you see red lights," on which one can act, and of which one can be conscious. By "laws of behaviour" I mean laws of nature as applied to a man's body such as "if you pinch him he will squeak." If we substitute "laws of behaviour which regulate his life" for "laws of conduct by which he regulates his life" in the argument quoted the undistributed middle is no longer insuperable. For we believe that it is not only true that being regulated by laws of behaviour implies being some sort of machine (though not necessarily a discrete-state machine), but that conversely being such a machine implies being regulated by such laws. However, we cannot so easily convince ourselves of the absence of complete laws of behaviour as of complete rules of conduct. The only way we know of for finding such laws is scientific observation, and we certainly know of no circumstances under which we could say, "We have searched enough. There are no such laws."

We can demonstrate more forcibly that any such statement would be unjustified. For suppose we could be sure of finding such laws if they existed. Then given a discrete-state machine it should certainly be possible to discover by observation sufficient about it to predict its future behaviour, and this within a reasonable time, say a thousand years. But this does not seem to be the case. I have set up on the Manchester computer a small programme using only 1,000 units of storage, whereby the machine supplied with one sixteen-figure number replies with another within two seconds. I would defy anyone to learn from these replies sufficient about the programme to be able to predict any replies to untried values.

(9) The Argument from Extrasensory Perception

I assume that the reader is familiar with the idea of extrasensory perception, and the meaning of the four items of it, viz., telepathy, clairvoyance, precognition and psychokinesis. These disturbing phenomena seem to deny all our usual scientific ideas. How we should like to discredit them! Unfortunately the statistical evidence, at least for telepathy, is overwhelming. It is very difficult to rearrange one's ideas so as to fit these new facts in. Once one has accepted them it does not seem a very big step to believe in ghosts and bogies. The idea that our bodies move simply according to the known laws of physics, together with some others not yet discovered but somewhat similar, would be one of the first to go.

This argument is to my mind quite a strong one. One can say in reply that many scientific theories seem to remain workable in practice, in spite of clashing with ESP; that in fact one can get along very nicely if one forgets about it. This is rather cold comfort, and one fears that thinking is just the kind of phenomenon where ESP may be especially relevant.

A more specific argument based on ESP might run as follows: "Let us play the imitation game, using as witnesses a man who is good as a telepathic receiver, and a digital computer. The interrogator can ask such questions as 'What suit does the card in my right hand belong to?' The man by telepathy or clairvoyance gives the right answer 130 times out of 400 cards. The machine can only guess at random, and perhaps gets 104 right, so the interrogator makes the right identification." There is an interesting possibility which opens here. Suppose the digital computer contains a random number generator. Then it will be natural to use this to decide what answer to give. But then the random number generator will be subject to the psychokinetic powers of the interrogator. Perhaps this psychokinesis might cause the machine to guess right more often than would be expected on a probability calculation, so that the interrogator might still be unable to make the right identification. On the other hand, he might be able to guess right without any questioning, by clairvoyance. With ESP anything may happen.

If telepathy is admitted it will be necessary to tighten our test up. The situation could be regarded as analogous to that which would occur if the interrogator were talking to himself and one of the competitors was listening with his ear to the wall. To put the competitors into a "telepathy-proof room" would satisfy all requirements.

7. Learning Machines

The reader will have anticipated that I have no very convincing arguments of a positive nature to support my views. If I had I should not have taken such pains to point out the fallacies in contrary views. Such evidence as I have I shall now give.

Let us return for a moment to Lady Lovelace's objection, which stated that the machine can only do what we tell it to do. One could say that a man can "inject" an idea into the machine, and that it will respond to a certain extent and then drop into quiescence, like a piano string struck by a hammer. Another simile would be an atomic pile of less than critical size: an injected idea is to correspond to a neutron entering the pile from without. Each such neutron will cause a certain disturbance which eventually dies away. If, however, the size of the pile is sufficiently increased, the disturbance caused by such an incoming neutron will very likely go on and on increasing until the whole pile is destroyed. Is there a corresponding phenomenon for minds, and is there one for machines? There does seem to be one for the human mind. The majority of them seem to be "subcritical," i.e., to correspond in this analogy to piles of subcritical size. An idea presented to such a mind will on average give rise to less than one idea in reply. A smallish proportion are supercritical. An idea presented to such a mind may give rise to a whole "theory" consisting of secondary, tertiary and more remote ideas. Animals' minds seem to be very definitely subcritical. Adhering to this analogy we ask, "Can a machine be made to be supercritical?"

The "skin-of-an-onion" analogy is also helpful. In considering the functions of the mind or the brain we find certain operations which we can explain in purely mechanical terms. This we say does not correspond to the real mind: it is a sort of skin which we must strip off if we are to find the real mind. But then in what remains we find a further skin to be stripped off, and so on. Proceeding in this way do we ever come to the "real" mind, or do we eventually come to the skin which has nothing in it? In the latter case the whole mind is mechanical. (It would not be a discrete-state machine however. We have discussed this.)

These last two paragraphs do not claim to be convincing arguments. They should rather be described as "recitations tending to produce belief."

The only really satisfactory support that can be given for the view expressed at the beginning of §6, will be that provided by waiting for the end of the century and then doing the experiment described. But what can we say in the meantime? What steps should be taken now if the experiment is to be successful?

As I have explained, the problem is mainly one of programming. Advances in engineering will have to be made too, but it seems unlikely that these will not be adequate for the requirements. Estimates of the storage capacity of the brain vary from 10^{10} to 10^{15} binary digits. I incline to the lower values and believe that only a very small fraction is used for the

higher types of thinking. Most of it is probably used for the retention of visual impressions, I should be surprised if more than 10^9 was required for satisfactory playing of the imitation game, at any rate against a blind man. (Note: The capacity of the *Encyclopaedia Britannica*, 11th edition, is 2×10^9) A storage capacity of 10^7 , would be a very practicable possibility even by present techniques. It is probably not necessary to increase the speed of operations of the machines at all. Parts of modern machines which can be regarded as analogs of nerve cells work about a thousand times faster than the latter. This should provide a "margin of safety" which could cover losses of speed arising in many ways. Our problem then is to find out how to programme these machines to play the game. At my present rate of working I produce about a thousand digits of programme a day, so that about sixty workers, working steadily through the fifty years might accomplish the job, if nothing went into the wastepaper basket. Some more expeditious method seems desirable.

In the process of trying to imitate an adult human mind we are bound to think a good deal about the process which has brought it to the state that it is in. We may notice three components.

- (a) The initial state of the mind, say at birth,
- (b) The education to which it has been subjected,
- (c) Other experience, not to be described as education, to which it has been subjected.

Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain. Presumably the child brain is something like a notebook as one buys it from the stationer's. Rather little mechanism, and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous.) Our hope is that there is so little mechanism in the child brain that something like it can be easily programmed. The amount of work in the education we can assume, as a first approximation, to be much the same as for the human child.

We have thus divided our problem into two parts. The child programme and the education process. These two remain very closely connected. We cannot expect to find a good child machine at the first attempt. One must experiment with teaching one such machine and see how well it learns. One can then try another and see if it is better or worse. There is an obvious connection between this process and evolution, by the identifications

Structure of the child machine = hereditary material

Changes of the child machine = mutation,

Natural selection = judgment of the experimenter

One may hope, however, that this process will be more expeditious than evolution. The survival of the fittest is a slow method for measuring advantages. The experimenter, by the exercise of intelligence, should be able to speed it up. Equally important is the fact that he is not restricted to random mutations. If he can trace a cause for some weakness he can probably think of the kind of mutation which will improve it.

It will not be possible to apply exactly the same teaching process to the machine as to a normal child. It will not, for instance, be provided with legs, so that it could not be asked to go out and fill the coal scuttle. Possibly it might not have eyes. But however well these deficiencies might be overcome by clever engineering, one could not send the creature to school without the other children making excessive fun of it. It must be given some tuition. We need not be too concerned about the legs, eyes, etc. The example of Miss Helen Keller shows that education can take place provided that communication in both directions between teacher and pupil can take place by some means or other.

We normally associate punishments and rewards with the teaching process. Some simple child machines can be constructed or programmed on this sort of principle. The machine has to be so constructed that events which shortly preceded the occurrence of a punishment signal are unlikely to be repeated, whereas a reward signal increased the probability of repetition of the events which led up to it. These definitions do not presuppose any feelings on the part of

the machine, I have done some experiments with one such child machine, and succeeded in teaching it a few things, but the teaching method was too unorthodox for the experiment to be considered really successful.

The use of punishments and rewards can at best be a part of the teaching process. Roughly speaking, if the teacher has no other means of communicating to the pupil, the amount of information which can reach him does not exceed the total number of rewards and punishments applied. By the time a child has learnt to repeat "Casabianca" he would probably feel very sore indeed, if the text could only be discovered by a "Twenty Questions" technique, every "NO" taking the form of a blow. It is necessary therefore to have some other "unemotional" channels of communication. If these are available it is possible to teach a machine by punishments and rewards to obey orders given in some language, e.g., a symbolic language. These orders are to be transmitted through the "unemotional" channels. The use of this language will diminish greatly the number of punishments and rewards required.

Opinions may vary as to the complexity which is suitable in the child machine. One might try to make it as simple as possible consistently with the general principles. Alternatively one might have a complete system of logical inference "built in." In the latter case the store would be largely occupied with definitions and propositions. The propositions would have various kinds of status, e.g., well-established facts, conjectures, mathematically proved theorems, statements given by an authority, expressions having the logical form of proposition but not belief-value. Certain propositions may be described as "imperatives." The machine should be so constructed that as soon as an imperative is classed as "well established" the appropriate action automatically takes place. To illustrate this, suppose the teacher says to the machine, "Do your homework now." This may cause "Teacher says 'Do your homework now' " to be included amongst the well-established facts. Another such fact might be, "Everything that teacher says is true." Combining these may eventually lead to the imperative, "Do your homework now," being included amongst the well-established facts, and this, by the construction of the machine, will mean that the homework actually gets started, but the effect is very satisfactory. The processes of inference used by the machine need not be such as would satisfy the most exacting logicians. There might for instance be no hierarchy of types. But this need not mean that type fallacies will occur, any more than we are bound to fall over unfenced cliffs. Suitable imperatives (expressed within the systems, not forming part of the rules of the system) such as "Do not use a class unless it is a subclass of one which has been mentioned by teacher" can have a similar effect to "Do not go too near the edge."

The imperatives that can be obeyed by a machine that has no limbs are bound to be of a rather intellectual character, as in the example (doing homework) given above. important amongst such imperatives will be ones which regulate the order in which the rules of the logical system concerned are to be applied, For at each stage when one is using a logical system, there is a very large number of alternative steps, any of which one is permitted to apply, so far as obedience to the rules of the logical system is concerned. These choices make the difference between a brilliant and a footling reasoner, not the difference between a sound and a fallacious one. Propositions leading to imperatives of this kind might be "When Socrates is mentioned, use the syllogism in Barbara" or "If one method has been proved to be quicker than another, do not use the slower method." Some of these may be "given by authority," but others may be produced by the machine itself, e.g. by scientific induction.

The idea of a learning machine may appear paradoxical to some readers. How can the rules of operation of the machine change? They should describe completely how the machine will react whatever its history might be, whatever changes it might undergo. The rules are thus quite time-invariant. This is quite true. The explanation of the paradox is that the rules which get changed in the learning process are of a rather less pretentious kind, claiming only an ephemeral validity. The reader may draw a parallel with the Constitution of the United States.

An important feature of a learning machine is that its teacher will often be very largely ignorant of quite what is going on inside, although he may still be able to some extent to predict his pupil's behavior. This should apply most strongly to the later education of a machine arising from a child machine of well-ried design (or programme). This is in clear contrast with normal procedure when using a machine to do computations one's object is then to have a clear mental picture of the state of the machine at each moment in the computation. This object can only be achieved with a struggle. The view that "the machine can only do what we know how to order it to do," appears strange in face of this. Most of the programmes which we can put into the machine will result in its doing something that we cannot make sense (if at all, or which we regard as completely random behaviour. Intelligent behaviour presumably consists in a departure from the completely disciplined behaviour involved in computation, but a rather slight one, which does not give rise to random behaviour, or to pointless repetitive loops. Another important result of preparing our machine for its part in the imitation game by a process of teaching and learning is that "human fallibility" is likely to be omitted in a rather natural

way, i.e., without special "coaching." (The reader should reconcile this with the point of view on pages 23 and 24.) Processes that are learnt do not produce a hundred per cent certainty of result; if they did they could not be unlearnt.

It is probably wise to include a random element in a learning machine. A random element is rather useful when we are searching for a solution of some problem. Suppose for instance we wanted to find a number between 50 and 200 which was equal to the square of the sum of its digits, we might start at 51 then try 52 and go on until we got a number that worked. Alternatively we might choose numbers at random until we got a good one. This method has the advantage that it is unnecessary to keep track of the values that have been tried, but the disadvantage that one may try the same one twice, but this is not very important if there are several solutions. The systematic method has the disadvantage that there may be an enormous block without any solutions in the region which has to be investigated first. Now the learning process may be regarded as a search for a form of behaviour which will satisfy the teacher (or some other criterion). Since there is probably a very large number of satisfactory solutions the random method seems to be better than the systematic. It should be noticed that it is used in the analogous process of evolution. But there the systematic method is not possible. How could one keep track of the different genetical combinations that had been tried, so as to avoid trying them again?

We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision. Many people think that a very abstract activity, like the playing of chess, would be best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. This process could follow the normal teaching of a child. Things would be pointed out and named, etc. Again I do not know what the right answer is, but I think both approaches should be tried.

We can only see a short distance ahead, but we can see plenty there that needs to be done.

Lessons from a Restricted Turing Test

[Stuart M. Shieber](#)
Aiken Computation Laboratory
[Division of Applied Sciences](#)
[Harvard University](#)

April 15, 1993
(Revision 5)

Abstract:

We report on the recent Loebner prize competition inspired by Turing's test of intelligent behavior. The presentation covers the structure of the competition and the outcome of its first instantiation in an actual event, and an analysis of the purpose, design, and appropriateness of such a competition. We argue that the competition has no clear purpose, that its design prevents any useful outcome, and that such a competition is inappropriate given the current level of technology. We then speculate as to suitable alternatives to the Loebner prize.

This report appeared in [Communications of the Association for Computing Machinery](#), volume 37, number 6, pages 70-78, 1994. Also available as [cmp-lg/9404002](#) and from the Center for Research in Computing Technology, Harvard University, as [Technical Report TR-19-92](#).

The Turing Test and the Loebner Prize

The English logician and mathematician Alan Turing, in an attempt to develop a working definition of intelligence free of the difficulties and philosophical pitfalls of defining exactly what constitutes the mental process of intelligent reasoning, devised a test, instead, of intelligent behavior. The idea, codified in his celebrated 1950 paper "Computing Machinery and Intelligence" [28], was specified as an "imitation game" in which a judge attempts to distinguish which of two agents is a human and which a computer imitating human responses by engaging each in a wide-ranging conversation of any topic and tenor. Turing's reasoning was that, presuming that intelligence was only practically determinable behaviorally, then any agent that was indistinguishable in behavior from an intelligent agent was, for all intents and purposes, intelligent. It is presumably uncontroversial that humans are intelligent as evidenced by their conversational behavior. Thus, any agent that can be mistaken by virtue of its conversational behavior with a human must be intelligent. As Turing himself noted, this syllogism argues that the criterion provides a sufficient, but not necessary, condition for intelligent behavior. The game has since become known as the "Turing test", a term that has eclipsed even his eponymous machine in Turing's terminological legacy. Turing predicted that by the year 2000, computers would be able to pass the Turing test at a reasonably sophisticated level, in particular, that the average interrogator would not be able to identify the computer correctly more than 70 per cent of the time after a five minute conversation.

On November 8, 1991, an eclectic group including academics, business people, press, and passers-by filled two floors of Boston's Computer Museum for a tournament billed as the first actual administration of the Turing test. The tournament was the first attempt on the recently constituted Loebner Prize established by New York theater equipment manufacturer Dr. Hugh Loebner and organized by Dr. Robert Epstein, President *Emeritus* of the Cambridge Center for Behavioral Studies, a research center specializing in behaviorist psychology. The Loebner Prize is administered by an illustrious committee headed by Dr. Daniel Dennett, Distinguished Professor of Arts and Sciences and Director for Cognitive Studies, Tufts University, and including Dr. Epstein; Dr. Harry Lewis, Gordon McKay Professor of Computer Science, Harvard University; Dr. H. McIlvaine Parsons, Senior Research Scientist, HumRRO; Dr. Willard van Orman Quine, Edgar Pierce Professor of Philosophy *Emeritus*, Harvard University; and Dr. Joseph Weizenbaum, Professor of Computer Science *Emeritus*, Massachusetts Institute of Technology. (Dr. I. Bernard Cohen, Victor S. Thomas Professor of the History of Science *Emeritus*, Harvard University, chaired the committee at an earlier stage in its genesis, and Dr. Allen Newell, U. A. and Helen Whitaker University Professor of Computer Science, Carnegie-Mellon University, and the prize establisher Dr. Loebner served as advisors.)

The prize committee spent almost two years in planning the structure of the tournament. Because this was to be a real competition, rather than a thought experiment, there would be several computer contestants, and therefore several

confederates would be needed as well. It was decided that there would be ten agents all together. In the event, six were computer programs. Ten judges would converse with the agents and score them. The judges and confederates were both selected from the general public on the basis of a newspaper employment advertisement that required little beyond typing ability, then screened by interview with the prize committee. They were chosen so as to have "no special expertise in computer science".

The committee realized early on that given the current state of the art, there was no chance that Turing's test, as originally defined, had the slightest chance of being passed by a computer program. Consequently, they attempted to adjust both the structure of the test and the scoring mechanism, so as to allow the computers a fighting chance. In particular, the following two rules were added to dramatically restrict Turing's test.

- *Limiting the topic:* In order to limit the amount of area that the contestant programs must be able to cope with, the topic of the conversation was to be strictly limited, both for the contestants and the confederates. The judges were required to stay on the subject in their conversations with the agents.
- *Limiting the tenor:* Further, only behavior evinced during the course of a natural conversation on the single specified topic would be required to be duplicated faithfully by the contestants. The operative rule precluded the use of "trickery or guile. Judges should respond naturally, as they would in a conversation with another person." (The method of choosing judges served as a further measure against excessive judicial sophistication.)

As will be seen, these two rules - limiting the topic and tenor of the discussion - were quite problematic.

The prize committee specified that there be independent referees stationed in several locations: several in the rooms with the judges and confederates to answer questions concerning interpretation of the above rules, and one in the auditorium to serve as a sort of roving ombudsman. I was a referee in the confederates' room, and can vouch for the fact that my and my colleagues' efforts there were hardly needed; the confederates performed admirably. Reports from the other referees indicated the same for the judges.

Dr. Loebner placed only two restrictions on the setting up of the competition by the prize committee: that a competition be held each year, and that a prize be awarded at each competition. The prize at this first competition was a nominal \$1500, although Dr. Loebner has reportedly earmarked \$100,000 for the first computer program to pass the full Turing test at some later running of the competition. (Costs for the running of the competition itself were paid for by grants from the National Science Foundation and the Sloan Foundation.)

Rank Order of the Terminals									
Least									Most
Human-Like	1	2	3	4	5	6	7	8	Human-Like
	A	B	C	D	E	F	G	H	

Figure 1: Mock-up of the form used to implement the scoring method for the first Loebner competition. The judge writes the letters corresponding to the terminals in order from least to most human-like, and draws a line purporting to separate the computer contestants from the human confederates. In this case, the line has been drawn such that three of the terminals (F, H, and G) were deemed to be connected to humans.

To determine the prize-winner, an ingenious scoring mechanism was devised. The Turing test involves a single binary decision, which is either right or wrong. But to determine a winner, the contestants had to be ranked, so each judge was required to place all of the agents in order from the apparently least human to most human. This alone induced the ordering on the basis of which the prize would be awarded. The contestant with the highest average rank would be deemed the winner of the tournament. However, this does not allow a direct reconstruction of the results of the 100 implicit binary decisions that might be made: which of the agents were humans, and which computers. To allow for this to be deduced as well, each judge was requested to place a single line separating the ranked agents into two groups. Those to the right of the line were claimed by that judge to be humans, those to the left computers. (See Figure 1.) The judges were told that at least two of the agents were human confederates, and at least two computer contestants, thus limiting the number of places that the line could be (rationally) placed. The binary decisions could then be read off of the rankings by noting on which side of the line each agent fell. This demarcation process was not used in the awarding of the prize, but was carried out for its informational value alone.

The Event

The tournament was to begin at 1 pm on the scheduled Friday. One room of the computer museum was set up with ten terminals for the judges, each labeled with a code letter and the specified topic for conversing with the associated agent. In a back room, hidden from the publicly accessible part of the museum for obvious reasons, five computers had been set up to serve the four confederates. (One terminal was intended to be a backup, and in case it was not needed, to be connected to a publicly accessible terminal so that press and the public could interact with it as a sort of separate Turing test.) In a large auditorium, the ten conversations were projected each on its own screen around the perimeter of the room, and A. K. Dewdney provided running commentary.

Unfortunately, there were serious technical difficulties with the rented computer equipment that had been set up for the confederates. None of the three IBM computers could be made to appropriately interact over the prepared lines with their companion terminal in the judges' room. (The two DEC workstations seemed to work fine.) After almost two hours of unsuccessful last minute engineering, the prize committee decided to begin the competition with only two confederates in place (just the number that the judges had been told was the minimum), reducing the number of agents to eight. The time that each judge had to converse with each agent was shortened from approximately fifteen minutes to approximately seven in order to accommodate the press's deadlines.

The topics chosen by the six contestants were of the sort appropriate for a cocktail party venue (burgundy wines, dry martinis, small talk, whimsical conversation, dissatisfactions in relationships) or perhaps, a child's birthday party (second grade school topics). The two participating confederates chose to converse on Shakespeare and women's clothing. In the end, and perhaps unsurprisingly, the average rankings placed the two human confederates as "more human-like" than the six contestants. The highest-ranked contestant, Joseph Weintraub's program (topic: whimsical conversation) was awarded the \$1500 prize.

The surprises, such as there were, were in the bipartite classifications. Five judges ranked the top contestant as human, and there were eight instances of such misclassifications of computers as humans overall. (For the record, the contestant discussing relationship problems, the entry of Dr. Kenneth Colby, well known for his PARRY program, which simulates paranoid responses, was twice rated human. The simulated seven-year old accounted for the final misclassification.)

Ms. Cynthia Clay, the Shakespeare aficionado, was thrice misclassified as a computer. At least one of the judges made her classification on the premise that "[no] human would have that amount of knowledge about Shakespeare." [26] Ms. Lisette Gozo was honored as the most human of the agents for her discussion of women's clothing, although one judge rated two computer programs above her.

Perhaps the most conspicuous characteristic of the six computer programs was their poor performance. It was widely recognized that computer experts could readily distinguish the contestants from the confederates. Indeed, many of the techniques being used by the programs were easily spotted by those familiar with the ELIZA program that prize committee member Weizenbaum developed in 1965. The repetition of previous statements verbatim (subject only to pronominal adjustments, sometimes wrong), answers transparently keyed to trigger words, and similar tricks of the ELIZA trade were ubiquitous. For example, the following example from the whimsical conversation program is illustrative of the regurgitation technique: ♡

01:11:42	Judge	10			
It has been a long day! Now I think I'm ready for some fun. The weather isn't very pleasant today. Do you have any					
jokes	that	could	lift	the	clouds?
01:13:01	Terminal	5			
Everybody talks about the weather but nobody seems to do much about it. Why do you need to tell me now you think					
you're ready for some fun? Groannnnnn.....					

Note that the first sentence in the reply, which concerns the weather, seems to be triggered by the keywords "weather" or "clouds" in the judge's contribution. The second sentence is comprised of the judge's second sentence, with pronouns transposed, and preceded by the introductory "Why do you need to tell me".

Such techniques are subject to well-known and obvious errors, as can be exemplified in the transcripts of interactions with the winning computer entrant. For instance, a naive application of pronoun transposition interacts poorly with tag questions like ``aren't you".

01:17:52	Judge	10
You're trying to make this difficult for me aren't you? I just explained that I was ready for fun!		
01:18:53	Terminal	5
Would you pre]fer it if it was not true that IO re trying to make this difficult for you aren't you? I just explained that i was ready for fun! I am, therefore I compute.... What makes you think I am trying to make this difficult for you aren't i?		
Similarly, repetition of a compound phrase is conversationally bizarre.		
01:35:26	Judge	6
I can't say I understand the question...but the whimsy of a poet is one thing and the whimsy of most of us is another.		
01:36:51	Terminal	5
How do you know you can t say you understand the question...but the whimsy of a poet is one thing and the whimsy of most of us is another.		

One reporter noted that Weizenbaum himself was ``disturbed" by how easily people were fooled by these programs [26], and more than one of the judges reported that they were disappointed in the programs' capabilities after their expectations had been raised by interacting with ELIZA in the interviewing process. Dr. Epstein, in a speech after the event, noted that he had learned from the day's proceedings that ``little progress has been made in the last twenty-five years", that is, since ELIZA. (We address this dubious conclusion below.)

Analysis

The obvious question, then, is how to reconcile the apparent success of the programs in fooling judges with their patently low technology. Clearly, part of the answer relies on the phenomenon that P. T. Barnum used to amass a fortune. People are easily fooled, and are especially easily fooled into reading structure into chaos, reading meaning into nonsense. This accounts for the popularity of newspaper horoscopes and roadside psychics. This is not a flaw in the human mental capacity. Sensitivity to subtle patterns in our environment is extremely important to our ability to perceive, learn, and communicate. Clouds look like ships, and Rorschach blots seem like vignettes. How much different is interpreting non sequitur as whimsical conversation?

Ned Block, a professor of philosophy at MIT (and by coincidence a referee at the competition, stationed with the judges) has argued that the Turing test is a sorely inadequate test of intelligence because it relies solely on the ability to fool people [3].✂ Certainly, it has been known since Weizenbaum's surprising experiences with ELIZA that a test based on fooling people is confoundingly simple to pass.

People are even more easily fooled when their ability to detect fooling is explicitly vitiated, for instance, by a prohibition against using ``trickery or guile".✂ When I asked Mr. Weintraub during the post-contest press conference how he himself would have unmasked his program, his response - typing gibberish in to see if the program spat it back verbatim at a later time a la ELIZA - was certainly outside the established rules. In fact, the referees had discussed that very technique the previous night at a meeting with the prize committee to calibrate our collective understanding of the rules. I pointed out to Mr. Weintraub that his response fell under the ``trickery and guile" prohibition, and he took another stab at the question. His second attempt to specify a winning strategy against his program succumbed to the same problem. (It involved repeating questions multiple times.)

Weintraub's problem in answering the question points to the craftiness of his solution to the Loebner prize puzzle. His entry is unfalsifiable independent of its performance and solely on the basis of the choice of topic. As almost everyone has noted who was familiar with the rules, whimsical conversation is not in fact a *topic* but a *style* of conversation (at least as practiced by Weintraub's program). And whimsical conversation in the mold of Weintraub's program is essentially nonsense conversation, a series of non sequiturs. Thus, when Weintraub's program is unresponsive, fails to make any sense, or shows a reckless abandonment of linguistic normalcy, it, unlike its competitor programs, is operating *as advertised*. It is being ``whimsical". At those times when, by happenstance, the program trips over an especially suggestive response, a judge can grab at it as the real article. (The strategy is reminiscent of that used by the program Racter to create ``free verse" poetry, another unfalsifiable genre.) Weintraub's strategy was an artful dodge of

the competition rules. He had found a loophole and exploited it elegantly. I for one believe that, in so doing, he heartily deserved to win.

We might call this winning strategy "PARRY's finesse", after Kenneth Colby's previously mentioned PARRY program [4]. PARRY was designed to engage in a dialogue in the role of a paranoid patient. The program was perhaps the first to be subject to an actual controlled experiment modeled on the Turing test [5], in which psychiatrists were given transcripts of electronically mediated dialogues with PARRY and with actual paranoids and were asked to pick out the simulated patient from the real. The fact that the expert judges, the psychiatrists, did no better than chance, has been credited to the fact that unresponsiveness and non sequitur are typical behaviors of paranoids. Joseph Weizenbaum's response to the experiment - in the form of his own model of a deviant mentality - parodies PARRY's finesse succinctly:

The contribution here reported should lead to a full understanding of one of man's most troublesome disorders: infantile autism.... It responds *exactly* as does an autistic patient - that is, not at all.... This program has the advantage that it can be implemented on a plain typewriter not connected to a computer at all. [29]

Post hoc thinking of this sort can go a long way to rationalizing the various misclassifications of the whimsical conversation program or, in the same vein, the program that talks at the level of a second-grader. (Who could fail to give a seven-year-old child the benefit of the doubt?) It leads to noting other insidious forms of scoring bias that crept into the competition. One possible source of such bias, for instance, follows from the technical problems that caused two of the confederates to be eliminated. Once the number of confederates had been reduced to the announced minimum, it became impossible for a judge to rationally place the demarcation line between "humans" and "computers" in such a way as to rate a human as a computer without also rating a computer as a human. Of course, the converse was not true. This might have accounted for one or two more of the errors. Dr. Epstein points out in response to this observation that "(1) Two of the ten judges drew the line after just *one* entry, in spite of our instructions. (2) Three of the 5 judges who mistook Weintraub's program for a person rated it above one or both confederates. (3) Two judges mistook a confederate for a computer. In fact, in two (and only two) cases could our instructions have forced the judge to mistake a computer for a person." (personal communication to Harry Lewis, 1992) The third point is, of course, irrelevant, the first hardly gratifying, the second accounted for by Weintraub's use of PARRY's finesse, and the final comment is exactly my point.

But post hoc rationalization, like telling your boss off, may be enjoyable at the moment, but is, in the long run, ungratifying. The important questions do not involve microanalysis of the particular competition as run several months ago, but the larger questions of the purpose, design, and even existence of the Loebner prize itself.

Why a Loebner Prize?

There is a long history of argumentation in the philosophical literature opposing the appropriateness of the Turing test as a litmus test of intelligence. Certain arguments against the effectiveness of the test in answering questions about the intelligence of computers or the possibility of human thought center around the behaviorist nature of the test. Intelligence, it may be claimed, is not determinable simply by surface behavior. Variants of this argument have been given by Block [2], Gunderson [15], and Searle [24][23]. Others have suggested that the Turing test is not sufficient in that the behaviors under adjudication are too limited [10][15]. On the basis of such counterarguments, Moor [18] has argued for a drastically limited view of the Turing test, not as an operational definition of intelligence at all, but rather as a mode for accumulating evidence leading to an inductive argument for the intelligence of the machine. (See the reply by Stalker [25] and a later clarification by Moor [19] for further arguments.) Moor [20] provides a good introduction to these issues. French [11] provides a strong argument that as a sufficient condition for intelligence, the Turing test is so difficult as to be uninteresting. Nonetheless, none of these sorts of presumptive counterarguments to the use of a Turing test are the basis for the discussion in the remainder of this paper. The issue of whether an operational definition of intelligence is appropriate, and whether the particular definition codified in the Turing test is too narrow, though important questions, can be taken as resolved in favor of the Turing test for the purposes of the present discussion. Thus, we will side with the behaviorist interpretation favored by the organization administering the prize, the Cambridge Center for Behavioral Studies. Nonetheless, these arguments do provide another strong basis on which to question the appropriateness of the Loebner prize. A full discussion is, unfortunately, well beyond the scope of this paper, but readers are urged to consult the cited literature. Having sided, for the nonce, with the philosophical

appropriateness of Turing's design as a test of intelligent behavior, we turn to the question of whether the Loebner prize competition is itself an appropriate enterprise.

Prizes for technological advances have existed before, and much can be learned by comparison with previous exemplars. Just as humankind has dreamed of mimicking the human power of thought, so have we longed to possess the avian power of flight. Human-powered flight entered the mythology of the ancient Chinese and Romans, the designs of da Vinci, yet was only accomplished within the last generation as a direct result of a prize set up for the express purpose of promoting that technology. The Kremer prize, established in 1959 by British engineer and industrialist Henry Kremer, provided for an award of £5000 for the first human-powered vehicle to fly a specified half-mile figure-eight course. It was awarded in 1977, less than twenty years later, to a team headed by Paul Macready, Jr., for a flight by Bryan Allen in the *Gossamer Condor*.

The success of the Kremer prize depended on two factors.

- *Pursuing a purpose:* The goals of the Kremer prize were clear. At the time of the institution of the prize, there were no active efforts to build human-powered aircraft. The goal of the prize was to provide an incentive to enter the field of human-powered flight. It was tremendously successful at this goal. By the time that the *Gossamer Condor* made its award-winning flight, Macready's team was in competition with several other teams with planes that were flying substantial distances solely under human power.
- *Pushing the envelope:* The basic sciences underlying human-powered flight were, by 1959, well understood. These included aerodynamics, mechanics, anatomy and physiology, and materials technology. It was even possible for Robert Graham, an expert in the field of human-powered flight and a founding member of the Cranfield Man-Powered Aircraft Committee, to state at that time that "Man could fly, if only someone would put up a prize for it." (Quoted by Grosser [11][page 23]grosser.) Overcoming the human difficulties in building a team that had collective mastery of these various fields and the engineering difficulties in creatively combining them were astonishing accomplishments. Nonetheless, as it turned out, no new basic discoveries were required at the time of the founding of the Kremer prize to win it. The task was just beyond the edge of the current technology. Unfortunately, since our ability to dream far outstrips our ability to build, the establishment of tests of ridiculous difficulty is not difficult to imagine. At a time when an award-winning human-powered flight was one of one meter at an altitude of 10 centimeters (the 1912), the Paris newspaper *La Justice* established a prize for the first nonstop human-powered flight from Paris to Versailles and back. (It was never won.)

The history of human-powered flight indicates that only when the purpose of the prize is clear and the task is just beyond the edge of current technology is a prize an appropriate incentive. The Kremer prize is a prime example of a prize that meets these criteria. The Loebner prize is not.

We turn first to the goals of the Loebner prize. It was, according to the formal statement in the competition application, "established...to further the scientific understanding of complex human behavior." Along these lines Dr. Loebner has been quoted as saying "People had been discussing the Turing test; people had been discussing AI, but nobody was doing anything about it." [17] The several thousand members of the American Association for Artificial Intelligence may be surprised to learn that nobody is doing anything about it.

Others have argued that the prize will serve to publicize the Turing test, thereby increasing the public's awareness and understanding of artificial intelligence. Increased public understanding of AI is certainly a laudable goal, especially since the regular appearance of superficial popularizations in the press serves more to mislead the public by alternately raising and dashing expectations than to inform it by cogent coverage of actual results. A flurry of the standard stories in the press like "Computer fools half of human panel" [13] and "Test a breakthrough in artificial intelligence" [16] was certainly one of the side effects of the Loebner prize competition, but perhaps not a laudable contribution.

Overselling of AI by the media (and, occasionally, practitioners) has, in its brief history, been a repeated and persistent problem, and the hubristic claims of the organizers of the Loebner prize that they are "confident that within 10 to 20 years a system will pass this electronic litmus test" [27] perpetuates the hyperbole. Robert Epstein, in his recent article describing the event, its genesis, and his speculations as to its importance, constructs a standard claim of this sort:

Thinking computers will be a new race, a sentient companion to our own. When a computer finally passes the Turing Test, will we have the right to turn it off? Who should get the prize money - the programmer or the computer? Can we say that such a machine is "self-aware"? Should we give it the right to vote? Should it pay taxes? If you doubt the significance of these issues, consider the possibility that someday soon *you will have to argue them with a computer.* [\[emphasis in original\]](#)epstein

Not surprisingly, the winner of the Loebner prize has jumped on the publicity bandwagon by taking out an advertisement pushing his program as the "first to pass the Turing Test". Conversely, a prize whose execution convinces fellow scientists mistakenly that little progress has been made in a quarter century does little to promote the field. In summary, there is a difference between publicity and increased public understanding. Events of this sort - and the Loebner competition has been no exception - tend to generate the former rather than the latter.

Dennett has hinted at a completely different goal for the Loebner prize. "It is useful to have the demonstration of the particular foibles that human beings exhibit in 1991.... We won't learn much about AI from the Loebner prize for a long time, but we will learn some non-negligible things about social psychology, perhaps, in the meantime." (Dennett, personal communication) For instance, the competition might be justified "as a proving ground for the environmental conditions necessary to permit the Turing test to someday occur. In other words, the Loebner competition can tell us what we need to know about how humans behave in computer mediated interactions." (Dranoff, personal communication) This line of teleology for the Loebner prize, that it serves not as a test of the abilities of the computers but of the psychologies of the various participants, has often been proposed informally. Such a "conspiracy theory" of the prize as a vast psychology experiment executed on unwitting and unconsenting adults is as unlikely as it is disturbing. Of course, there is already an extensive literature on how humans behave in computer-mediated interactions, and the Loebner competition is not likely to contribute to it; it was not designed or executed as a controlled scientific experiment, nor was that its apparent intention, despite the hopes of Dennett and Dranoff that firm conclusions in psychology might be gleaned from it.

Thus, it is difficult to imagine a clear scientific goal that the Loebner prize might satisfy. Turing's test as originally defined, on the other hand had a clear goal, to serve as a sufficient condition for demonstrating that a human artifact exhibited intelligent behavior. Even this goal is lost in the Loebner prize competition. By limiting the test, it no longer serves its original purpose (and arguably no purpose at all), as Turing's syllogism fails. It is questionable whether the notion of a Turing test limited in the ways specified by the Loebner prize committee is even a coherent one. The prize committee spent some time with the referees attempting to explicate the notion of "natural conversation without trickery or guile". It was suggested that a criterion be used as to whether you might say the utterance in conversation with a stranger seated next to you on an airplane. For instance, what might a competition judge legitimately ask on the topic of Washington, DC? Certainly, the question "Are there any zoos in Washington?" is the kind of thing you might ask a stranger when flying to the capital for the first time, whereas "Is Washington bigger than a breadbasket?" is just as certainly a trick question. What about "Is there much crime in Washington?" Undoubtedly acceptable. "Are there any dogs in Washington?" An odd question for an airplane conversation. "Are there many dogs in Washington?" Sounds better. "Are there many marmosets in Washington?" Odd. "Are there many marmosets in the Washington zoo?" Okay again. The exegesis of such examples begins to sound like arguments about angels and sharp objects.

Similar problems accrue to the notion of limiting the topic of discourse. Is the last question about Washington, DC or marmosets? (One of the referees in fact thought that this and similar questions should be ruled out as it was not strictly on the topic of the city alone.) How about "Are the buildings in Washington very modern?" Perhaps a question about architecture, as the following question surely is: "Do you know any examples of neo-Georgian architecture in Washington?" Are culinary topics ruled out, as in "What foods is our nation's capital best known for?" Such issues are not idle in the context of the Loebner competition. Cynthia Clay, the Shakespeare expert, was asked why Mario Cuomo has been referred to recently as "Hamlet on the Hudson". The question caused much consternation among the referees peering over Ms. Clay's shoulder. Her response was "His brooding" after which she coolly changed the topic back to Shakespeare. Or had it ever left?

The reason that Turing chose natural language as the behavior definitional of human intelligence is exactly its open-ended, free-wheeling nature. "The question and answer method seems to be suitable for introducing almost any of the fields of human endeavor that we wish to include." [page 435]turing-mind In attempting to limit the *task* of the contestants through limiting the *domain alone*, the prize committee succeeded in doing neither.

The distinction between domain and task is crucial. Finance is a domain, but not a task; withdrawing money from a bank account is a task, one that is achievable through both human and computer intermediaries these days; taking dictation of a funds-transfer request is a task that only humans can currently undertake with reliability. Had Babbage limited his differential analyzer to multiply only even numbers, the design would have been no more successful. This is a limitation of domain that does not yield a concomitant limitation in task.

It is well understood in the field that natural-language systems must be tested using a constrained task. Currently, standard limited tasks can be found in evaluation of natural-language database retrieval systems (like withdrawing money from a bank account on the basis of a natural-language request) and speech recognition systems (like transcribing a spoken funds-transfer request). The tasks, typically undertaken with limited vocabulary, are easily quantifiable along several dimensions (for example, technical notions of precision, recall, overgeneration, perplexity) independently of the subjective judgments of lay judges. In addition, they can be adjusted to sit just at the edge of technology (a topic we return to below) unlike the Turing test itself. The natural-language-processing research community has used such tests for some time now, and there has been increased interest in issues of evaluation of systems (primarily at the behest of funding agencies) over the last few years; whole conferences have been devoted to the subject (see, for instance, the report by Neal and Walter [\[21\]](#)). ♡

In summary, the Loebner prize competition neither satisfies its own avowed goals, nor the original goals of Alan Turing. In fact, it is hard to imagine a scientific goal that establishment of the Loebner prize provides a better route to than would be provided by other uses of Dr. Loebner's \$1500, his \$100,000 promissory note, and the \$80,000 in ancillary grants from the National Science Foundation and the Sloan Foundation. (Nonscientific goals are much easier to imagine, of course.)

Now to the second criterion for an appropriate technology prize, that the task be just beyond the edge of technology. Imagine that a prize for human-powered flight were set up when the basic science of the time was far too impoverished for such an enterprise, say, in da Vinci's era. The da Vinci prize, we shall imagine, is constituted in 1492 and is to be awarded to the highest human-powered flight. Like the Loebner prize, a competition is held every year and a prize must be awarded each time it is held. The first da Vinci competition is won by a clever fellow with big springs on his shoes. Since the next competition is only one year away (no time to invent the airfoil), the optimal strategy is universally observed by potential contestants to involve building a bigger pair of springs. Twenty-five years later, the head of the prize committee announces that little progress has been made in human-powered flight since the first round of the prize as everyone is still manufacturing springs. ♡

Of course, a lot of progress had been made in human-powered flight in those twenty-five years. Da Vinci himself was studying human physiology and anatomy and the flight of birds, and - although his own work directly on the topic of human-powered flight, ornithopter design, was essentially meritless beyond its decorative qualities - the apparently tangential work was, in the long run, pertinent to the technologies that would eventually enable the *Gossamer Condor* to be constructed. (See, e.g., Gibbs-Smith [\[12\]](#).) However, over that period, and indeed at every point during the following four centuries, the kind of progress that needed to get made to solve the problem was not directly observable *at that time* in improvement in solutions to the problem, the kind of improvement that might be observable in an annual contest. Nonetheless, tremendous scientific progress was made between the fifteenth and twentieth centuries. The *Gossamer Condor* and the digital computer are two outgrowths of this progress.

The field of artificial intelligence is in that kind of state. ♡ The AI problem, like the problem of human-powered flight in the Renaissance, is only addressed directly and dismissed as imminently solvable by those who underestimate its magnitude. Progress on restricted tasks in limited domains is well documented in the literature on applications of artificial intelligence. But progress on the underlying science that has been made in the last twenty-five years, important though it is, is not of the type that allows incremental advantage to be demonstrated on the big problem, the full-blown Turing test, nor should this be seen as a failing of a field addressing a problem of the scope and magnitude of human intelligence. (And like all scientific endeavors, a lot of time can be spent on fruitless avenues of attack; ELIZA, as a discipline for natural-language processing, was such a fruitless avenue. It was quite fruitful in other areas, however, as cogently argued by Weizenbaum himself.) Indeed, one aspect of the progress made in research on natural-language processing is the appreciation for its complexity, which led to the dearth of entrants from the artificial intelligence community - the realization that time spent on winning the Loebner prize is not time spent furthering the field.

Twenty-five years of progress in the fields associated with the Turing test - artificial intelligence, computational linguistics, and natural-language processing - cannot be summarized in a single program, but is captured in the many small results, some of which, some day, at an unpredictable time in the future, may lead to a dramatic demonstration of apparently intelligent artificial behavior. To expect more is hubris. What is needed is not more work on solving the Turing test, as promoted by Dr. Loebner, but more work on the basic research issues involved in understanding intelligent behavior. The parlor games can be saved for later.

Alternatives to the Loebner Prize

Given that the Loebner prize, as constituted, is at best a diversion of effort and attention and at worst a disparagement of the scientific community, what might a better alternative use of Dr. Loebner's largesse be? The goal of furthering the scientific understanding of complex human behavior is no less laudable now than it was before the competition, but clearly, a direct assault on a valid test of intelligent behavior is out of the question for a long time; even the prize committee well appreciates that. Thus, any award or prize based on a behavioral test must use a limited task and domain, so that the envelope of technology is pushed, not ignored. The efforts of the Loebner prize committee to design such a test have failed in that the test that they developed rewards cheap tricks like parrying and insertion of random typing errors. This is an (indubitably predictable) lesson of the 1991 Loebner prize competition.

This problem is a general one: Any behavioral test that is sufficiently constrained for our current technology must so limit the task and domain as to render the test scientifically uninteresting. Adjusting the particulars of the Loebner competition rules will not help. By way of example, many years of effort have gone into the design of the tests of natural-language-processing systems used at the annual DARPA-sponsored Message Understanding Conferences. The trend among entrants over the last several conferences has been toward less and less sophisticated natural-language-processing techniques, concentrating instead on engineering tricks oriented to the exigencies of the restricted task - keyword-spotting, template-matching, and similar methods. In short, this is because such limited tests are better addressed in the near term by engineering (building bigger springs) than science (discovering the airfoil). Behavioral tests of intelligence are either too hard for a prize or too rewarding of incidentals.

At this stage, objective behavioral tests must give way to subjective evaluative ones. A more appropriate way to reward novel, potentially breakthrough-inducing efforts toward the eventual goal of mimicking intelligent behavior would be to institute a prize for just such efforts, on the model of the Nobel prizes, ACM's Turing award, and similar subjectively determined awards. Rather than awarding lifelong achievement or past accomplishments, however, the prize could be awarded for particular discoveries, regardless of field, that the committee determined were of sufficient originality, import, and technical merit and that were deemed contributory to Turing's goal (even though they may provide no incremental edge in a current-day Turing test). To avoid rapt obeisance to AI conventional wisdom, the awards committee would include eminent thinkers from a wide range of related fields (much as the current Loebner prize committee does) but to ensure technical fidelity, a nominating committee of researchers from the pertinent technical fields should verify purported results before passing them on for consideration. In order to prevent degrading of the imprimatur of the reconstructed Loebner prize, it would be awarded on an occasional basis, only when a sufficiently deserving new result, idea, or development presented itself. I am not ostentatious enough to provide examples that I believe would be appropriate for such an award; I am sure that the reader can imagine one or two. ♡

As the years elapsed, and the speculations of this Loebner prize committee as documented in their past decisions began to prove perspicacious, the Loebner prize might grow in stature to that of the highly sought prizes of other scientific areas, and so provide a tremendous motivation for innovative ideas in the quest for an artificial intelligence.

Postscript

The Second Annual Loebner Prize Competition was held at the Cambridge Center for Behavioral Studies on December 15, 1992. The number of computer entrants had decreased from six to three, with Joseph Weintraub's program, complete with the winning strategy from the previous year's competition, taking first prize once again, this time under the purported topic "men vs. women". Bigger springs had prevailed.

Acknowledgements

The research in this paper was supported in part by grant IRI-9157996 from the National Science Foundation and by matching funds from Xerox Corporation.

I am grateful to the many readers of earlier drafts of this paper: Ned Block, Noam Chomsky, Jacques Cohen, Daniel Dennett, Susan Cole Dranoff, Barbara Grosz, Harry Lewis, David Mumford, Fernando Pereira, Jeff Rosenschein, David Yarowsky, and two anonymous reviewers. I have incorporated many of their thoughtful comments into the paper, although the opinions presented here are my own, and should not be taken as necessarily representative of the previous readers' views.

References

- 1 Michael A. Arbib. More on computer models of psychopathic behavior. *Communications of the Association for Computing Machinery*, 17(9):543, September 1974.
- 2 Ned Block. Psychologism and behaviorism. *The Philosophical Review*, XC(1):5-43, 1981.
- 3 Ned Block. The computer model of the mind. In Daniel N. Osherson and Edward E. Smith, editors, *An Introduction to Cognitive Science III: Thinking*, chapter 3, pages 147-289. MIT Press, Cambridge, Massachusetts, 1990.
- 4 Kenneth Mark Colby. Modeling a paranoid mind. *Behavioral and Brain Sciences*, 4(4):515-560, 1981.
- 5 Kenneth Mark Colby, Franklin Dennis Hilf, Sylvia Weber, and Helena C. Kraemer. Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes. *Artificial Intelligence*, 3(1):199-221, 1972.
- 6 Daniel C. Dennett. Can machines think? In Michael Shafto, editor, *How We Know*, pages 121-145. Harper and Row, San Francisco, California, 1985.
- 7 Hubert Dreyfus. *What Computers Can't Do: A Critique Of Artificial Reason*. Harper and Row, New York, New York, revised edition, 1979.
- 8 Hubert L. Dreyfus. Alchemy and artificial intelligence. P. 3244, The RAND Corporation, December 1965.
- 9 Robert Epstein. The quest for the thinking computer. *AI Magazine*, 1992.
- 10 Jerry Fodor. *Psychological Explanation*, pages 126-127. Random House, New York, New York, 1968.
- 11 Robert French. Subcognition and the limits of the Turing test. *Mind*, 99(393):53-65, January 1990.
- 12 Charles H. Gibbs-Smith. *Leonardo da Vinci's Aeronautics*. Her Majesty's Stationery Office, London, 1967.
- 13 Lee Gomes. Computer fools half of human panel. San Jose Mercury News, 9 November 1991.
- 14 Morton Grosser. *Gossamer Odyssey*. Michael Joseph, London, 1981.
- 15 Keith Gunderson. *Mentality and Machines*. Doubleday & Company, Inc., Garden City, New York, 1971.
- 16 Jeffrey Krasner. Experts try to tell man from machine. Boston Herald, 9 November 1991.
- 17 Christopher Lindquist. Quest for machines that think. Computerworld, 18 November 1991.
- 18 James H. Moor. An analysis of the Turing test. *Philosophical Studies*, 30:249-257, 1976.
- 19

- 20 James H. Moor. Explaining computer behavior. *Philosophical Studies*, 34:325-327, 1978.
- 21 James. H. Moor. Turing test. In Stuart C. Shapiro, editor, *Encyclopedia of Artificial Intelligence*, pages 1126-1130. John Wiley and Sons, New York, New York, 1987.
- 22 Jeanette G. Neal and Sharon M. Walter. Natural language processing systems evaluation workshop. Technical Report RL-TR-91-362, Rome Laboratory, Griffiss Air Force Base, NY, December 1991.
- 23 D. A. Reay. *The History of Man-Powered Flight*. Pergamon Press, Oxford, 1977.
- 24 John R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3:417-457, 1980.
- 25 John R. Searle. Can computers think? In *Minds, Brains, and Science*, chapter 2, pages 28-41. Harvard University Press, Cambridge, Massachusetts, 1984.
- 26 Douglas F. Stalker. Why machines can't think: A reply to James Moor. *Philosophical Studies*, 34:317-320, 1978.
- 27 David Stipp. Some computers manage to fool people at game of imitating human beings. *Wall Street Journal*, 11 November 1991. Page B3A.
- 28 The Guardian. Machines meet mastermind, 29 August 1991.
- 29 Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433-460, October 1950.
- 30 Joseph Weizenbaum. Automating psychotherapy. *Communications of the Association for Computing Machinery*, 17(7):425, July 1974.
- 31 Joseph Weizenbaum. Reply to Arbib: More on computer models of psychopathic behavior. *Communications of the Association for Computing Machinery*, 17(9):543, September 1974.
- Joseph Weizenbaum. *Computer Power and Human Reason*. W. H. Freeman, San Francisco, 1976.

About this document ...

Lessons from a Restricted Turing Test

This document was generated using the [LaTeX2HTML](#) translator Version 0.6.4 (Tues Aug 30 1994) Copyright © 1993, 1994, [Nikos Drakos](#), Computer Based Learning Unit, University of Leeds.

The command line arguments were:

latex2html -split 0 loebner-rev-html.tex.

The translation was initiated by on Mon Jun 2 18:37:14 EDT 1997 and modified by hand thereafter.

Mon Jun 2 18:37:14 EDT 1997

FEATURES

- 3-axis sensing
- Small, low-profile package
 - 4 mm × 4 mm × 1.45 mm LFCSP
- Low power
 - 200 μA at $V_s = 2.0 V$ (typical)
- Single-supply operation
 - 2.0 V to 3.6 V
- 10,000 g shock survival
- Excellent temperature stability
- BW adjustment with a single capacitor per axis
- RoHS/WEEE lead-free compliant

APPLICATIONS

- Cost-sensitive, low power, motion- and tilt-sensing applications
 - Mobile devices
 - Gaming systems
 - Disk drive protection
 - Image stabilization
 - Sports and health devices

GENERAL DESCRIPTION

The ADXL330 is a small, thin, low power, complete three axis accelerometer with signal conditioned voltage outputs, all on a single monolithic IC. The product measures acceleration with a minimum full-scale range of $\pm 3 g$. It can measure the static acceleration of gravity in tilt-sensing applications, as well as dynamic acceleration resulting from motion, shock, or vibration.

The user selects the bandwidth of the accelerometer using the C_X , C_Y , and C_Z capacitors at the X_{OUT} , Y_{OUT} , and Z_{OUT} pins. Bandwidths can be selected to suit the application, with a range of 0.5 Hz to 1,600 Hz for X and Y axes, and a range of 0.5 Hz to 550 Hz for the Z axis.

The ADXL330 is available in a small, low-profile, 4 mm × 4 mm × 1.45 mm, 16-lead, plastic lead frame chip scale package (LFCSP_LQ).

FUNCTIONAL BLOCK DIAGRAM

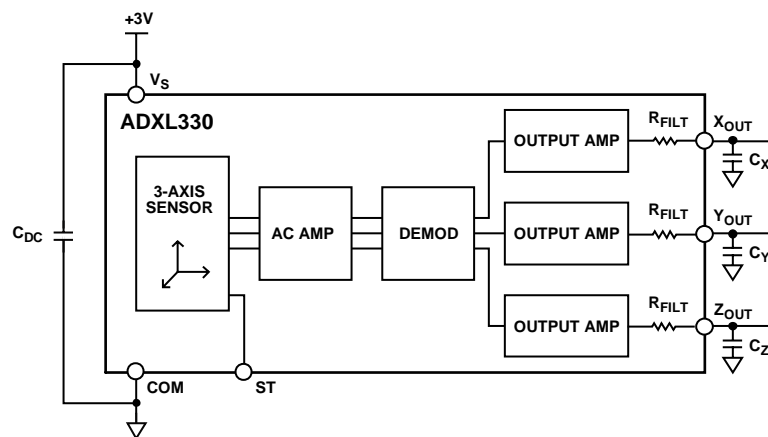


Figure 1.

Rev. 0

Information furnished by Analog Devices is believed to be accurate and reliable. However, no responsibility is assumed by Analog Devices for its use, nor for any infringements of patents or other rights of third parties that may result from its use. Specifications subject to change without notice. No license is granted by implication or otherwise under any patent or patent rights of Analog Devices. Trademarks and registered trademarks are the property of their respective owners.

TABLE OF CONTENTS

Features	1	Performance	11
Applications.....	1	Applications.....	12
General Description	1	Power Supply Decoupling.....	12
Functional Block Diagram	1	Setting the Bandwidth Using C _X , C _Y and C _Z	12
Revision History	2	Self-Test	12
Specifications.....	3	Design Trade-Offs for Selecting Filter Characteristics: The Noise/BW Trade-Off.....	12
Absolute Maximum Ratings.....	4	Use with Operating Voltages Other than 3 V.....	12
ESD Caution.....	4	Axes of Acceleration Sensitivity	13
Pin Configuration and Function Descriptions.....	5	Outline Dimensions	14
Typical Performance Characteristics	6	Ordering Guide	14
Theory of Operation	11		
Mechanical Sensor.....	11		

REVISION HISTORY

3/06—Revision 0: Initial Version

SPECIFICATIONS

$T_A = 25^\circ\text{C}$, $V_S = 3\text{ V}$, $C_X = C_Y = C_Z = 0.1\text{ }\mu\text{F}$, acceleration = 0 g, unless otherwise noted. All minimum and maximum specifications are guaranteed. Typical specifications are not guaranteed.

Table 1.

Parameter	Conditions	Min	Typ	Max	Unit
SENSOR INPUT	Each axis				
Measurement Range		± 3	± 3.6		g
Nonlinearity	% of full scale		± 0.3		%
Package Alignment Error			± 1		Degrees
Inter-Axis Alignment Error			± 0.1		Degrees
Cross Axis Sensitivity ¹			± 1		%
SENSITIVITY (RATIOMETRIC) ²	Each axis				
Sensitivity at X_{OUT} , Y_{OUT} , Z_{OUT}	$V_S = 3\text{ V}$	270	300	330	mV/g
Sensitivity Change Due to Temperature ³	$V_S = 3\text{ V}$		± 0.015		%/ $^\circ\text{C}$
ZERO g BIAS LEVEL (RATIOMETRIC)	Each axis				
0 g Voltage at X_{OUT} , Y_{OUT} , Z_{OUT}	$V_S = 3\text{ V}$	1.2	1.5	1.8	V
0 g Offset vs. Temperature			± 1		mg/ $^\circ\text{C}$
NOISE PERFORMANCE					
Noise Density X_{OUT} , Y_{OUT}			280		$\mu\text{g}/\sqrt{\text{Hz}}$ rms
Noise Density Z_{OUT}			350		$\mu\text{g}/\sqrt{\text{Hz}}$ rms
FREQUENCY RESPONSE ⁴					
Bandwidth X_{OUT} , Y_{OUT} ⁵	No external filter		1600		Hz
Bandwidth Z_{OUT} ⁵	No external filter		550		Hz
R_{FILT} Tolerance			$32 \pm 15\%$		k Ω
Sensor Resonant Frequency			5.5		kHz
SELF-TEST ⁶					
Logic Input Low			+0.6		V
Logic Input High			+2.4		V
ST Actuation Current			+60		μA
Output Change at X_{OUT}	Self-test 0 to 1		-150		mV
Output Change at Y_{OUT}	Self-test 0 to 1		+150		mV
Output Change at Z_{OUT}	Self-test 0 to 1		-60		mV
OUTPUT AMPLIFIER					
Output Swing Low	No load		0.1		V
Output Swing High	No load		2.8		V
POWER SUPPLY					
Operating Voltage Range		2.0		3.6	V
Supply Current	$V_S = 3\text{ V}$		320		μA
Turn-On Time ⁷	No external filter		1		ms
TEMPERATURE					
Operating Temperature Range		-25		+70	$^\circ\text{C}$

¹ Defined as coupling between any two axes.

² Sensitivity is essentially ratiometric to V_S .

³ Defined as the output change from ambient-to-maximum temperature or ambient-to-minimum temperature.

⁴ Actual frequency response controlled by user-supplied external filter capacitors (C_X , C_Y , C_Z).

⁵ Bandwidth with external capacitors = $1/(2 \times \pi \times 32\text{ k}\Omega \times C)$. For C_X , $C_Y = 0.003\text{ }\mu\text{F}$, bandwidth = 1.6 kHz. For $C_Z = 0.01\text{ }\mu\text{F}$, bandwidth = 500 Hz. For C_X , C_Y , $C_Z = 10\text{ }\mu\text{F}$, bandwidth = 0.5 Hz.

⁶ Self-test response changes cubically with V_S .

⁷ Turn-on time is dependent on C_X , C_Y , C_Z and is approximately $160 \times C_X$ or C_Y or $C_Z + 1\text{ ms}$, where C_X , C_Y , C_Z are in μF .

ABSOLUTE MAXIMUM RATINGS

Table 2.

Parameter	Rating
Acceleration (Any Axis, Unpowered)	10,000 g
Acceleration (Any Axis, Powered)	10,000 g
V_S	-0.3 V to +7.0 V
All Other Pins	(COM - 0.3 V) to ($V_S + 0.3$ V)
Output Short-Circuit Duration (Any Pin to Common)	Indefinite
Temperature Range (Powered)	-55°C to +125°C
Temperature Range (Storage)	-65°C to +150°C

Stresses above those listed under Absolute Maximum Ratings may cause permanent damage to the device. This is a stress rating only; functional operation of the device at these or any other conditions above those indicated in the operational section of this specification is not implied. Exposure to absolute maximum rating conditions for extended periods may affect device reliability.

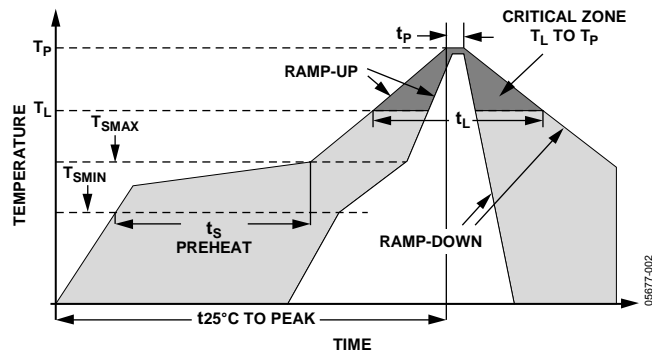


Figure 2. Recommended Soldering Profile

Table 3. Recommended Soldering Profile

Profile Feature	Sn63/Pb37	Pb-Free
Average Ramp Rate (T_L to T_P)	3°C/s max	3°C/s max
Preheat		
Minimum Temperature (T_{SMIN})	100°C	150°C
Maximum Temperature (T_{SMAX})	150°C	200°C
Time (T_{SMIN} to T_{SMAX}), t_s	60 s to 120 s	60 s to 180 s
T_{SMAX} to T_L		
Ramp-Up Rate	3°C/s max	3°C/s max
Time Maintained Above Liquidous (T_L)		
Liquidous Temperature (T_L)	183°C	217°C
Time (t_L)	60 s to 150 s	60 s to 150 s
Peak Temperature (T_P)	240°C + 0°C/-5°C	260°C + 0°C/-5°C
Time within 5°C of Actual Peak Temperature (t_p)	10 s to 30 s	20 s to 40 s
Ramp-Down Rate	6°C/s max	6°C/s max
Time 25°C to Peak Temperature	6 minutes max	8 minutes max

ESD CAUTION

ESD (electrostatic discharge) sensitive device. Electrostatic charges as high as 4000 V readily accumulate on the human body and test equipment and can discharge without detection. Although this product features proprietary ESD protection circuitry, permanent damage may occur on devices subjected to high energy electrostatic discharges. Therefore, proper ESD precautions are recommended to avoid performance degradation or loss of functionality.



PIN CONFIGURATION AND FUNCTION DESCRIPTIONS

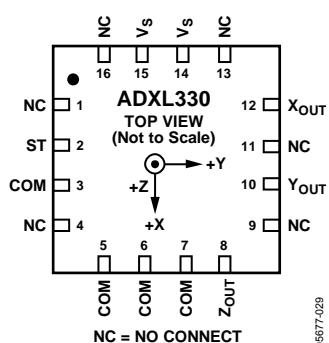


Figure 3. Pin Configuration

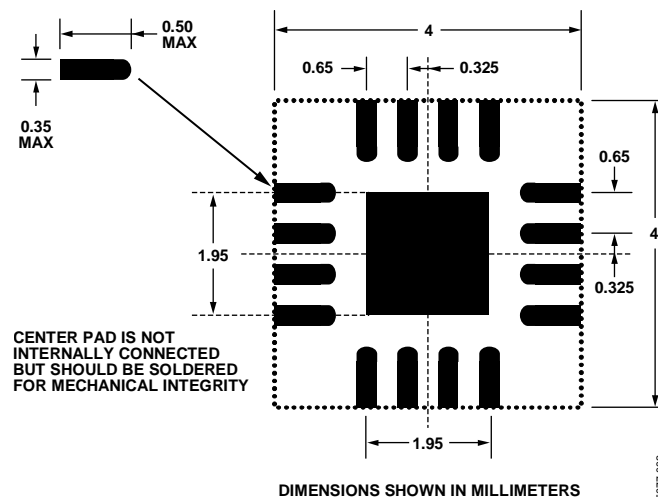


Figure 4. Recommended PCB Layout

Table 4. Pin Function Descriptions

Pin No.	Mnemonic	Description
1	NC	No Connect
2	ST	Self-Test
3	COM	Common
4	NC	No Connect
5	COM	Common
6	COM	Common
7	COM	Common
8	Z _{OUT}	Z Channel Output
9	NC	No Connect
10	Y _{OUT}	Y Channel Output
11	NC	No Connect
12	X _{OUT}	X Channel Output
13	NC	No Connect
14	V _S	Supply Voltage (2.0 V to 3.6 V)
15	V _S	Supply Voltage (2.0 V to 3.6 V)
16	NC	No Connect

TYPICAL PERFORMANCE CHARACTERISTICS

N > 1000 for all typical performance plots, unless otherwise noted.

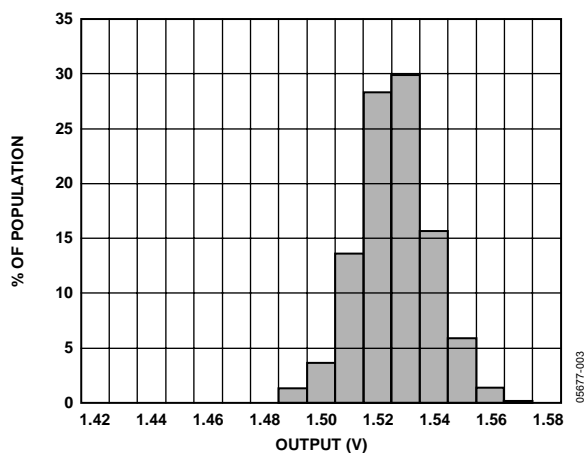


Figure 5. X-Axis Zero g Bias at 25°C, $V_S = 3\text{ V}$

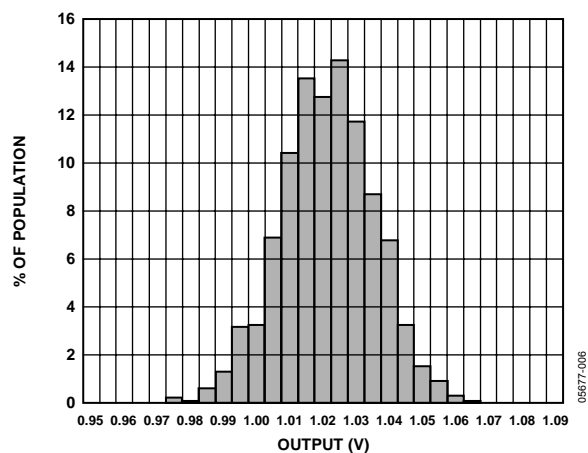


Figure 8. X-Axis Zero g Bias at 25°C, $V_S = 2\text{ V}$

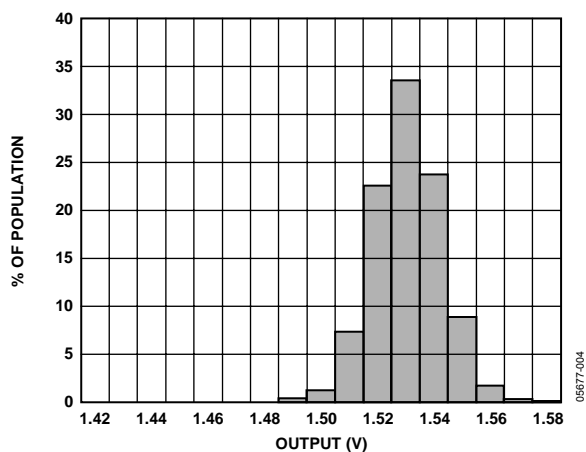


Figure 6. Y-Axis Zero g Bias at 25°C, $V_S = 3\text{ V}$

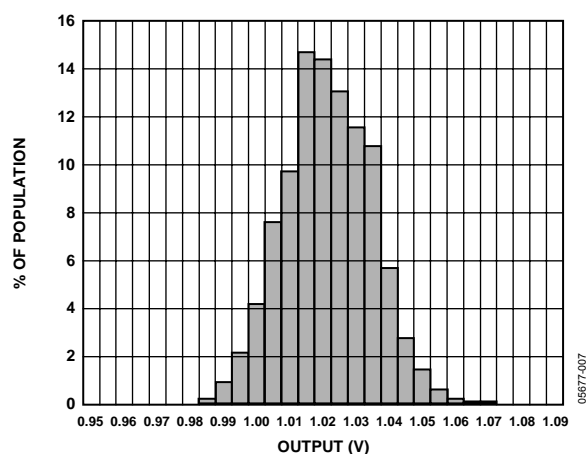


Figure 9. Y-Axis Zero g Bias at 25°C, $V_S = 2\text{ V}$

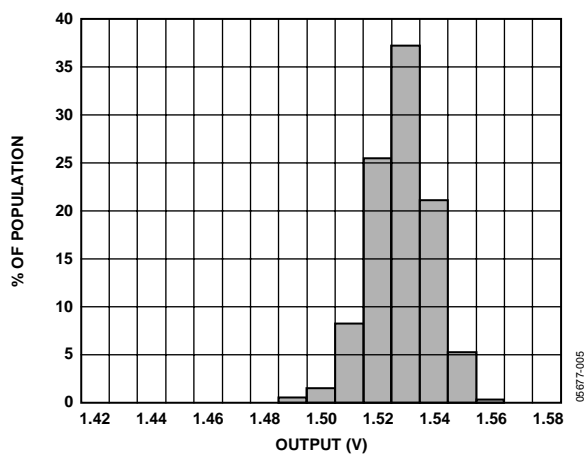


Figure 7. Z-Axis Zero g Bias at 25°C, $V_S = 3\text{ V}$

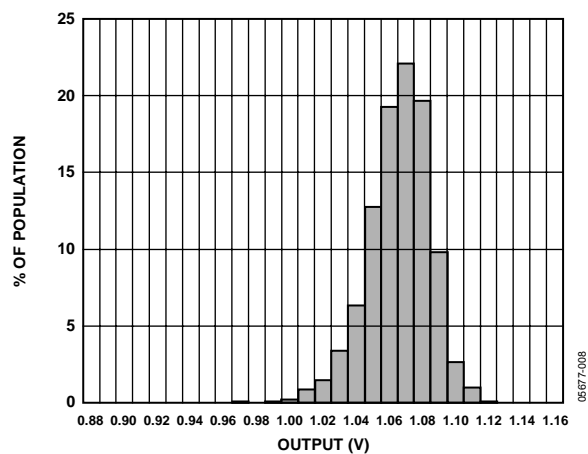


Figure 10. Z-Axis Zero g Bias at 25°C, $V_S = 2\text{ V}$

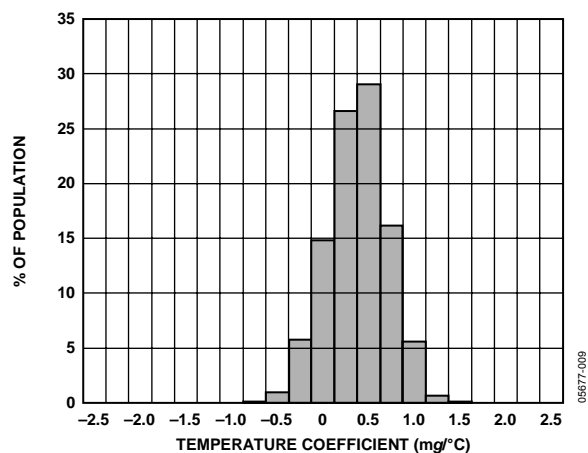


Figure 11. X-Axis Zero g Bias Temperature Coefficient, $V_s = 3\text{ V}$

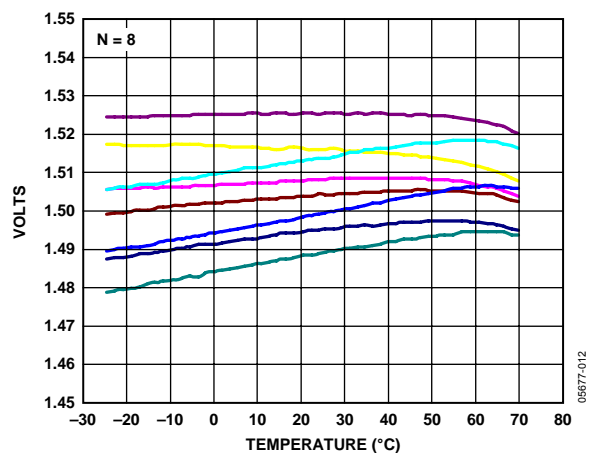


Figure 14. X-Axis Zero g Bias vs. Temperature—8 Parts Soldered to PCB

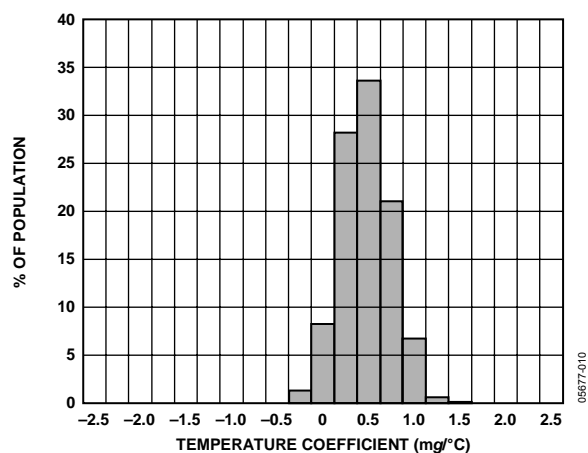


Figure 12. Y-Axis Zero g Bias Temperature Coefficient, $V_s = 3\text{ V}$

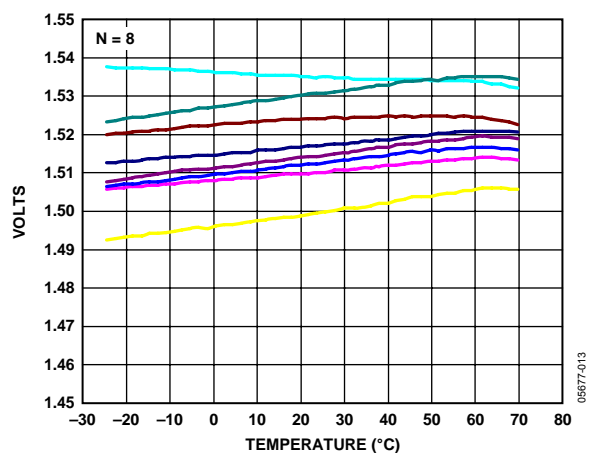


Figure 15. Y-Axis Zero g Bias vs. Temperature—8 Parts Soldered to PCB

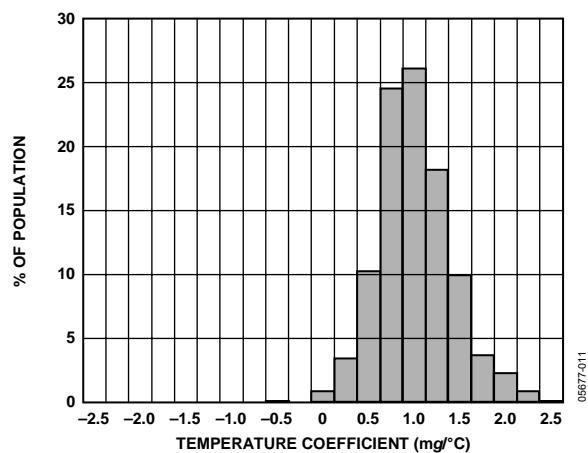


Figure 13. Z-Axis Zero g Bias Temperature Coefficient, $V_s = 3\text{ V}$

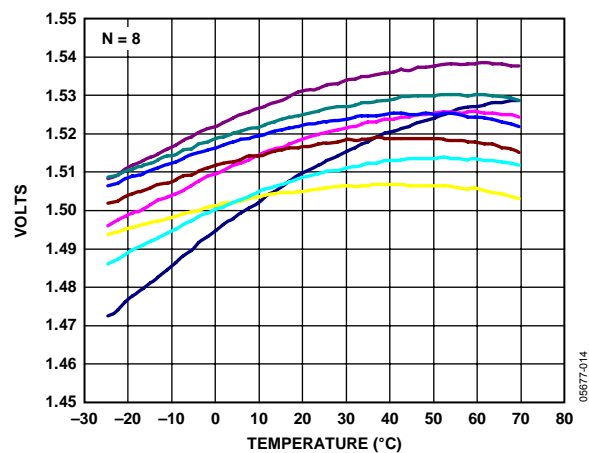


Figure 16. Z-Axis Zero g Bias vs. Temperature—8 Parts Soldered to PCB

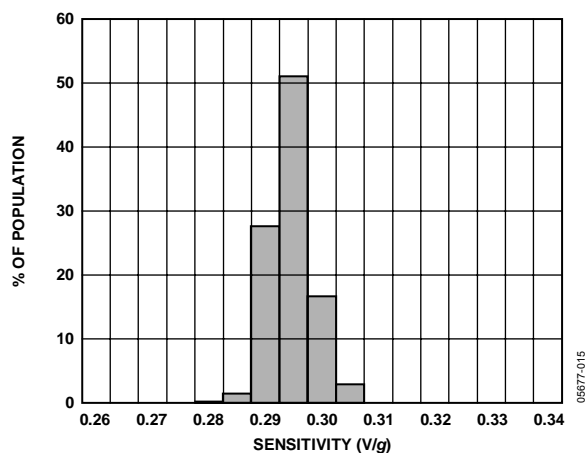


Figure 17. X-Axis Sensitivity at 25°C, $V_S = 3\text{ V}$

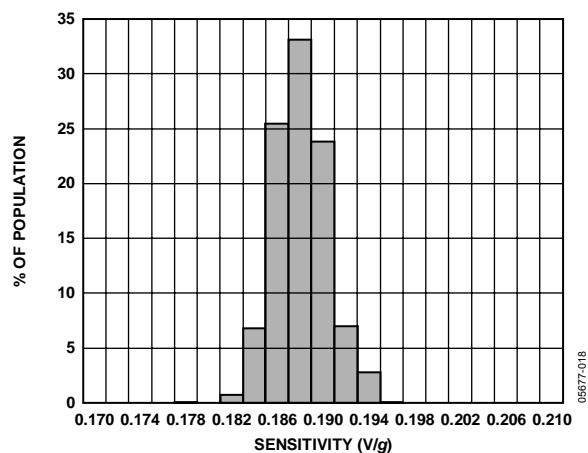


Figure 20. X-Axis Sensitivity at 25°C, $V_S = 2\text{ V}$

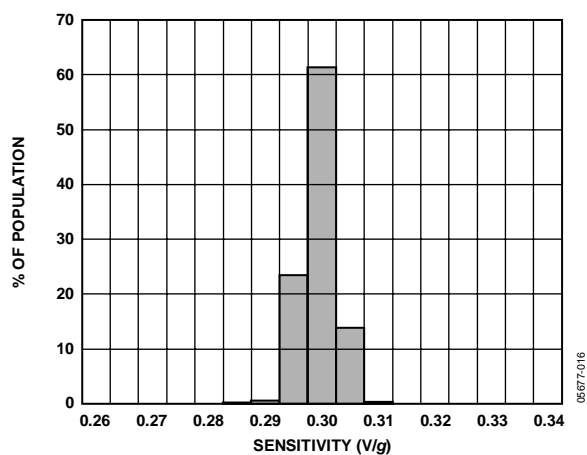


Figure 18. Y-Axis Sensitivity at 25°C, $V_S = 3\text{ V}$

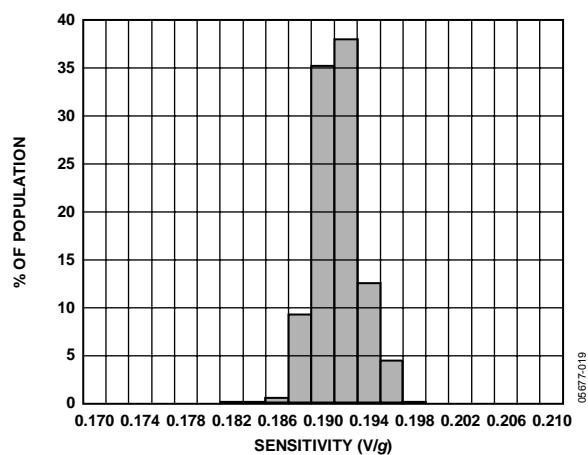


Figure 21. Y-Axis Sensitivity at 25°C, $V_S = 2\text{ V}$

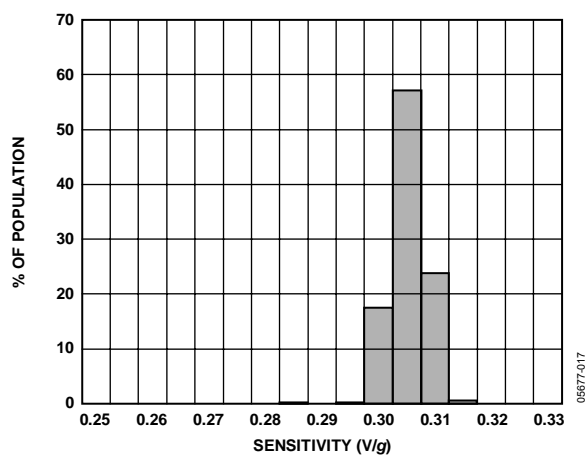


Figure 19. Z-Axis Sensitivity at 25°C, $V_S = 3\text{ V}$

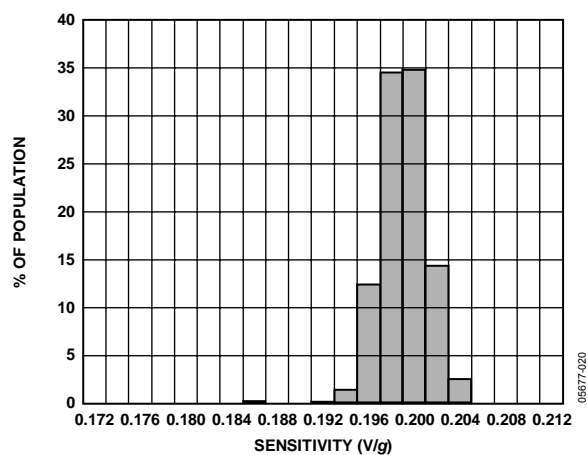


Figure 22. Z-Axis Sensitivity at 25°C, $V_S = 2\text{ V}$

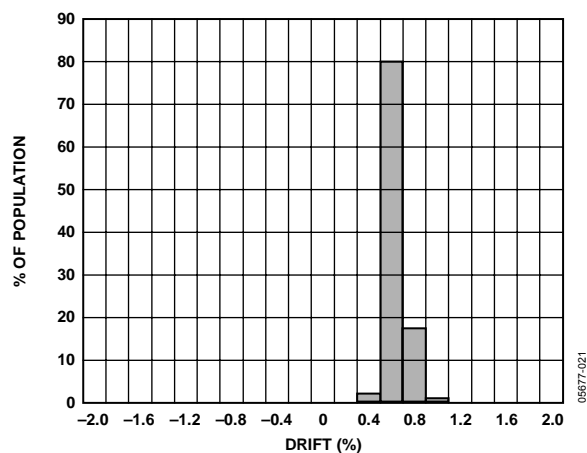


Figure 23. X-Axis Sensitivity Drift Over Temperature, $V_s = 3\text{ V}$

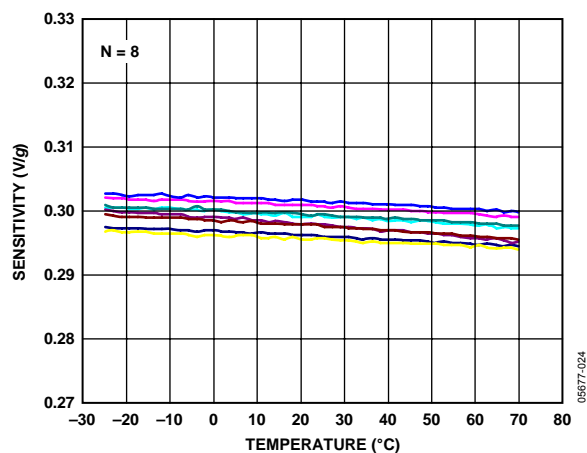


Figure 26. X-Axis Sensitivity vs. Temperature
8 Parts Soldered to PCB, $V_s = 3\text{ V}$

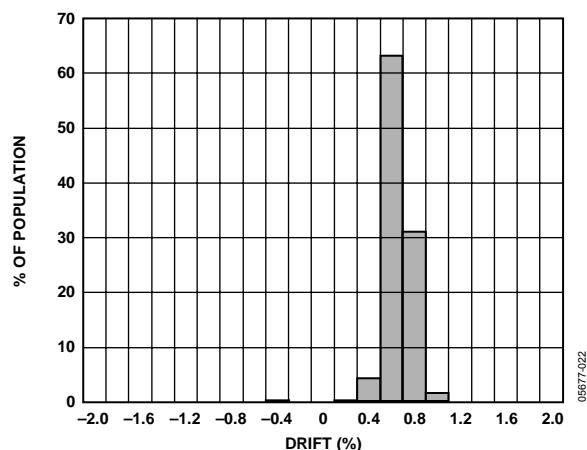


Figure 24. Y-Axis Sensitivity Drift Over Temperature, $V_s = 3\text{ V}$

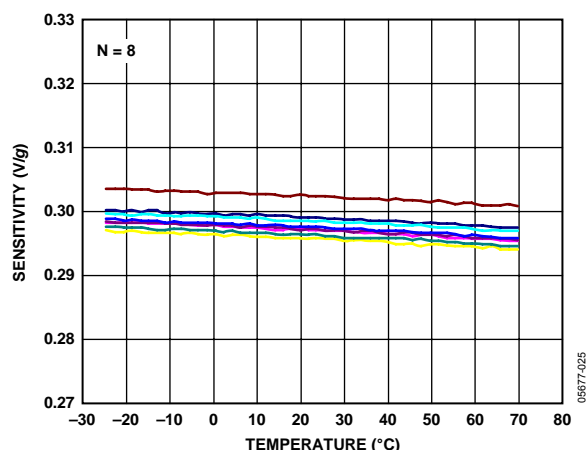


Figure 27. Y-Axis Sensitivity vs. Temperature
8 Parts Soldered to PCB, $V_s = 3\text{ V}$

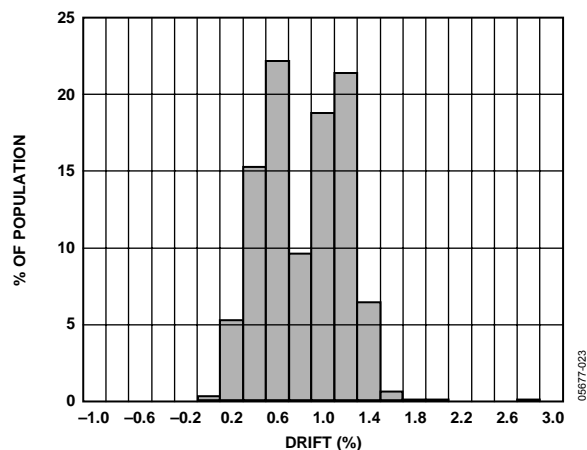


Figure 25. Z-Axis Sensitivity Drift Over Temperature, $V_s = 3\text{ V}$

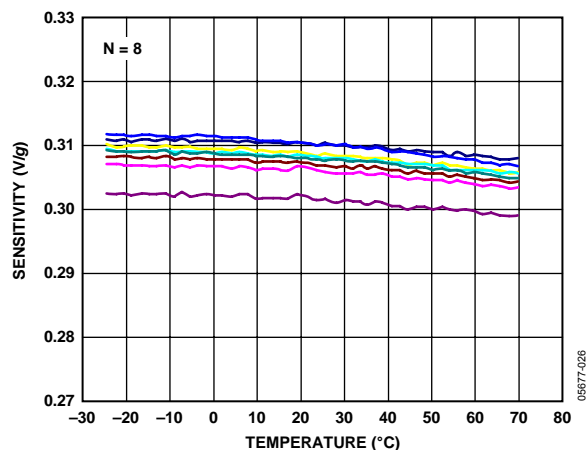


Figure 28. Z-Axis Sensitivity vs. Temperature
8 Parts Soldered to PCB, $V_s = 3\text{ V}$

ADXL330

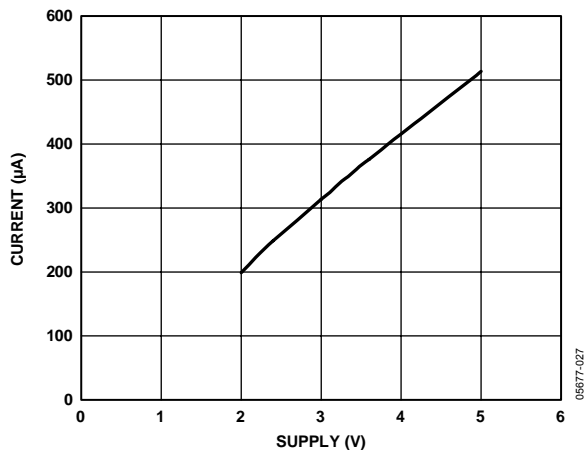


Figure 29. Typical Current Consumption vs. Supply Voltage

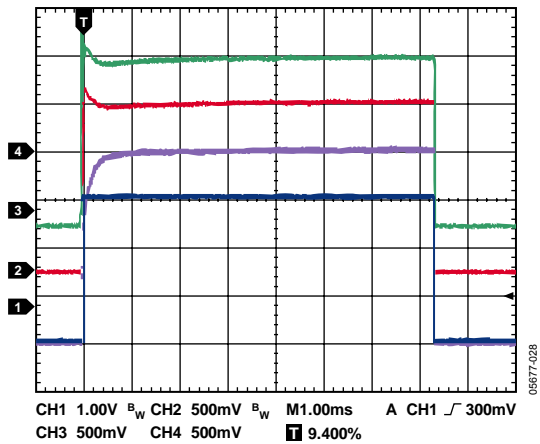


Figure 30. Typical Turn-On Time— $C_X, C_Y, C_Z = 0.0047 \mu F, V_S = 3 V$

THEORY OF OPERATION

The ADXL330 is a complete 3-axis acceleration measurement system on a single monolithic IC. The ADXL330 has a measurement range of $\pm 3\text{ g}$ minimum. It contains a polysilicon surface micromachined sensor and signal conditioning circuitry to implement an open-loop acceleration measurement architecture. The output signals are analog voltages that are proportional to acceleration. The accelerometer can measure the static acceleration of gravity in tilt sensing applications as well as dynamic acceleration resulting from motion, shock, or vibration.

The sensor is a polysilicon surface micromachined structure built on top of a silicon wafer. Polysilicon springs suspend the structure over the surface of the wafer and provide a resistance against acceleration forces. Deflection of the structure is measured using a differential capacitor that consists of independent fixed plates and plates attached to the moving mass. The fixed plates are driven by 180° out-of-phase square waves. Acceleration deflects the moving mass and unbalances the differential capacitor resulting in a sensor output whose amplitude is proportional to acceleration. Phase-sensitive demodulation techniques are then used to determine the magnitude and direction of the acceleration.

The demodulator output is amplified and brought off-chip through a $32\text{ k}\Omega$ resistor. The user then sets the signal bandwidth of the device by adding a capacitor. This filtering improves measurement resolution and helps prevent aliasing.

MECHANICAL SENSOR

The ADXL330 uses a single structure for sensing the X, Y, and Z axes. As a result, the three axes sense directions are highly orthogonal with little cross axis sensitivity. Mechanical misalignment of the sensor die to the package is the chief source of cross axis sensitivity. Mechanical misalignment can, of course, be calibrated out at the system level.

PERFORMANCE

Rather than using additional temperature compensation circuitry, innovative design techniques ensure high performance is built-in to the ADXL330. As a result, there is neither quantization error nor nonmonotonic behavior, and temperature hysteresis is very low (typically less than 3 mg over the -25°C to $+70^\circ\text{C}$ temperature range).

Figure 14, Figure 15, and Figure 16 show the zero g output performance of eight parts (X-, Y-, and Z-axis) soldered to a PCB over a -25°C to $+70^\circ\text{C}$ temperature range.

Figure 26, Figure 27, and Figure 28 demonstrate the typical sensitivity shift over temperature for supply voltages of 3 V . This is typically better than $\pm 1\%$ over the -25°C to $+70^\circ\text{C}$ temperature range.

APPLICATIONS

POWER SUPPLY DECOUPLING

For most applications, a single 0.1 μF capacitor, C_{DC} , placed close to the ADXL330 supply pins adequately decouples the accelerometer from noise on the power supply. However, in applications where noise is present at the 50 kHz internal clock frequency (or any harmonic thereof), additional care in power supply bypassing is required as this noise can cause errors in acceleration measurement. If additional decoupling is needed, a 100 Ω (or smaller) resistor or ferrite bead can be inserted in the supply line. Additionally, a larger bulk bypass capacitor (1 μF or greater) can be added in parallel to C_{DC} . Ensure that the connection from the ADXL330 ground to the power supply ground is low impedance because noise transmitted through ground has a similar effect as noise transmitted through V_{S} .

SETTING THE BANDWIDTH USING C_{X} , C_{Y} , AND C_{Z}

The ADXL330 has provisions for band limiting the X_{OUT} , Y_{OUT} , and Z_{OUT} pins. Capacitors must be added at these pins to implement low-pass filtering for antialiasing and noise reduction. The equation for the 3 dB bandwidth is

$$F_{-3\text{ dB}} = 1/(2\pi(32\text{ k}\Omega) \times C_{(\text{X}, \text{Y}, \text{Z})})$$

or more simply

$$F_{-3\text{ dB}} = 5\text{ }\mu\text{F}/C_{(\text{X}, \text{Y}, \text{Z})}$$

The tolerance of the internal resistor (R_{FILT}) typically varies as much as $\pm 15\%$ of its nominal value (32 k Ω), and the bandwidth varies accordingly. A minimum capacitance of 0.0047 μF for C_{X} , C_{Y} , and C_{Z} is recommended in all cases.

Table 5. Filter Capacitor Selection, C_{X} , C_{Y} , and C_{Z}

Bandwidth (Hz)	Capacitor (μF)
1	4.7
10	0.47
50	0.10
100	0.05
200	0.027
500	0.01

SELF-TEST

The ST pin controls the self-test feature. When this pin is set to V_{S} , an electrostatic force is exerted on the accelerometer beam. The resulting movement of the beam allows the user to test if the accelerometer is functional. The typical change in output is -500 mg (corresponding to -150 mV) in the X-axis, 500 mg (or 150 mV) on the Y-axis, and -200 mg (or -60 mV) on the Z-axis. This ST pin may be left open circuit or connected to common (COM) in normal use.

Never expose the ST pin to voltages greater than $V_{\text{S}} + 0.3\text{ V}$. If this cannot be guaranteed due to the system design (for

instance, if there are multiple supply voltages), then a low V_{F} clamping diode between ST and V_{S} is recommended.

DESIGN TRADE-OFFS FOR SELECTING FILTER CHARACTERISTICS: THE NOISE/BW TRADE-OFF

The selected accelerometer bandwidth ultimately determines the measurement resolution (smallest detectable acceleration). Filtering can be used to lower the noise floor to improve the resolution of the accelerometer. Resolution is dependent on the analog filter bandwidth at X_{OUT} , Y_{OUT} , and Z_{OUT} .

The output of the ADXL330 has a typical bandwidth of greater than 500 Hz. The user must filter the signal at this point to limit aliasing errors. The analog bandwidth must be no more than half the analog-to-digital sampling frequency to minimize aliasing. The analog bandwidth can be further decreased to reduce noise and improve resolution.

The ADXL330 noise has the characteristics of white Gaussian noise, which contributes equally at all frequencies and is described in terms of $\mu\text{g}/\sqrt{\text{Hz}}$ (the noise is proportional to the square root of the accelerometer bandwidth). The user should limit bandwidth to the lowest frequency needed by the application to maximize the resolution and dynamic range of the accelerometer.

With the single-pole, roll-off characteristic, the typical noise of the ADXL330 is determined by

$$rms\text{ Noise} = \text{Noise Density} \times (\sqrt{BW \times 1.6})$$

Often, the peak value of the noise is desired. Peak-to-peak noise can only be estimated by statistical methods. Table 6 is useful for estimating the probabilities of exceeding various peak values, given the rms value.

Table 6. Estimation of Peak-to-Peak Noise

Peak-to-Peak Value	% of Time that Noise Exceeds Nominal Peak-to-Peak Value
$2 \times rms$	32
$4 \times rms$	4.6
$6 \times rms$	0.27
$8 \times rms$	0.006

USE WITH OPERATING VOLTAGES OTHER THAN 3 V

The ADXL330 is tested and specified at $V_{\text{S}} = 3\text{ V}$; however, it can be powered with V_{S} as low as 2 V or as high as 3.6 V. Note that some performance parameters change as the supply voltage is varied.

The ADXL330 output is ratiometric, therefore, the output sensitivity (or scale factor) varies proportionally to the supply voltage. At $V_s = 3.6$ V, the output sensitivity is typically 360 mV/g. At $V_s = 2$ V, the output sensitivity is typically 195 mV/g.

The zero g bias output is also ratiometric, so the zero g output is nominally equal to $V_s/2$ at all supply voltages.

The output noise is not ratiometric but is absolute in volts; therefore, the noise density decreases as the supply voltage increases. This is because the scale factor (mV/g) increases while the noise voltage remains constant. At $V_s = 3.6$ V, the X- and Y-axis noise density is typically $230 \mu\text{g}/\sqrt{\text{Hz}}$, while at $V_s = 2$ V, the X- and Y-axis noise density is typically $350 \mu\text{g}/\sqrt{\text{Hz}}$.

Self-test response in g is roughly proportional to the square of the supply voltage. However, when ratiometricity of sensitivity is factored in with supply voltage, the self-test response in volts is roughly proportional to the cube of the supply voltage. For example, at $V_s = 3.6$ V, the self-test response for the ADXL330 is approximately -275 mV for the X-axis, +275 mV for the Y-axis, and -100 mV for the Z-axis.

At $V_s = 2$ V, the self-test response is approximately -60 mV for the X-axis, +60 mV for the Y-axis, and -25 mV for the Z-axis.

The supply current decreases as the supply voltage decreases. Typical current consumption at $V_s = 3.6$ V is 375 μA , and typical current consumption at $V_s = 2$ V is 200 μA .

AXES OF ACCELERATION SENSITIVITY

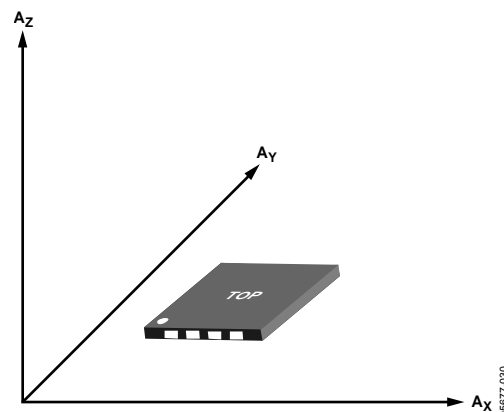


Figure 31. Axes of Acceleration Sensitivity, Corresponding Output Voltage Increases When Accelerated Along the Sensitive Axis

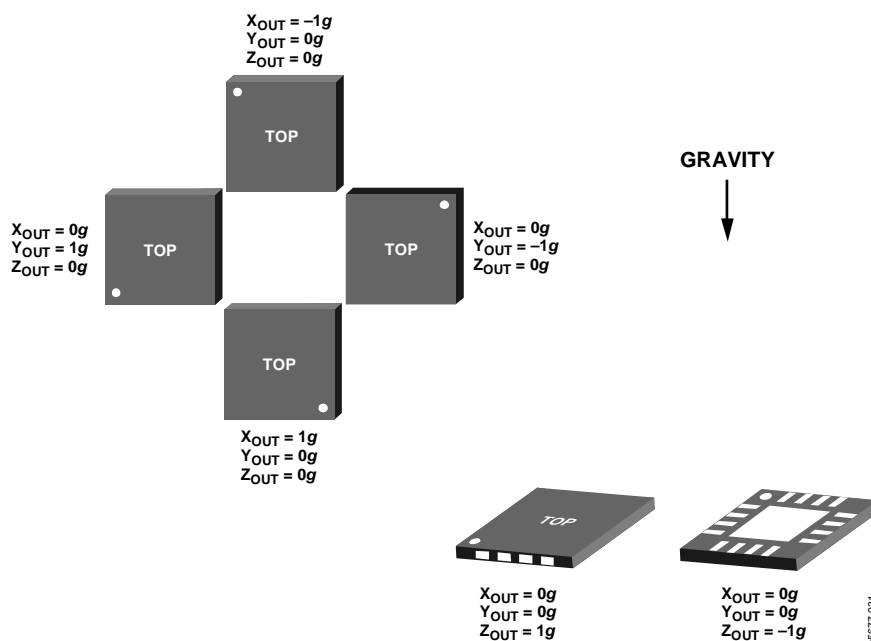


Figure 32. Output Response vs. Orientation to Gravity

ADXL330

OUTLINE DIMENSIONS

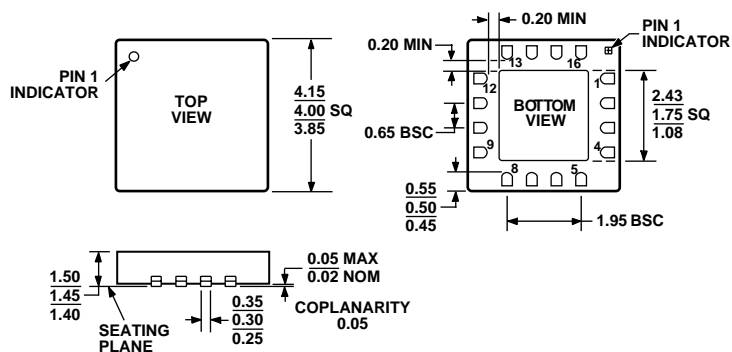


Figure 33. 16-Lead Lead Frame Chip Scale Package [LFCSP_LQ]
4 mm x 4 mm Body, Thick Quad
(CP-16-5)
Dimensions shown in millimeters

ORDERING GUIDE

Model	Measurement Range	Specified Voltage	Temperature Range	Package Description	Package Option
ADXL330KCPZ ¹	$\pm 3 g$	3 V	-25°C to +70°C	16-Lead LFCSP_LQ	CP-16-5
ADXL330KCPZ-RL ¹	$\pm 3 g$	3 V	-25°C to +70°C	16-Lead LFCSP_LQ	CP-16-5
EVAL-ADXL330				Evaluation Board	

¹ Z = Pb-free part.

NOTES

NOTES

MEMS INERTIAL SENSOR: 3-Axis - $\pm 2g/\pm 6g$ LINEAR ACCELEROMETER

1 Features

- 2.4V TO 3.6V SINGLE SUPPLY OPERATION
- LOW POWER CONSUMPTION
- $\pm 2g/\pm 6g$ USER SELECTABLE FULL-SCALE
- 0.5mg RESOLUTION OVER 100Hz BANDWIDTH
- EMBEDDED SELF TEST AND POWER DOWN
- OUTPUT VOLTAGE, OFFSET AND SENSITIVITY RATIO METRIC TO THE SUPPLY VOLTAGE
- HIGH SHOCK SURVIVABILITY
- LEAD FREE AND ECOPACK COMPATIBLE

2 Description

The LIS3L02AS4 is a low-power three axes linear accelerometer that includes a sensing element and an IC interface able to take the information from the sensing element and to provide an analog signal to the external world.

The sensing element, capable of detecting the acceleration, is manufactured using a dedicated process developed by ST to produce inertial sensors and actuators in silicon.

The IC interface is manufactured using a standard CMOS process that allows high level of integration to design a dedicated circuit which is trimmed to better match the sensing element characteristics.

The LIS3L02AS4 has a user selectable full scale of

Figure 1. Package

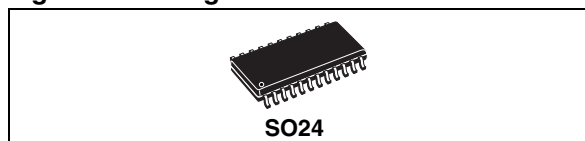


Table 1. Order Codes

Part Number	Package	Finishing
E-LIS3L02AS4	SO24	Tube
E-LIS3L02AS4TR	SO24	Tape & Reel

$\pm 2g$, $\pm 6g$ and it is capable of measuring accelerations over a bandwidth of 1.5KHz for all axes. The device bandwidth may be reduced by using external capacitances. A self-test capability allows to check the mechanical and electrical signal path of the sensor.

The LIS3L02AS4 is available in plastic SMD package and it is specified over an extended temperature range of -40°C to $+85^{\circ}\text{C}$.

The LIS3L02AS4 belongs to a family of products suitable for a variety of applications:

- Mobile terminals
- Gaming and Virtual Reality input devices
- Free-fall detection for data protection
- Antitheft systems and Inertial Navigation
- Appliance and Robotics

Figure 2. Block Diagram

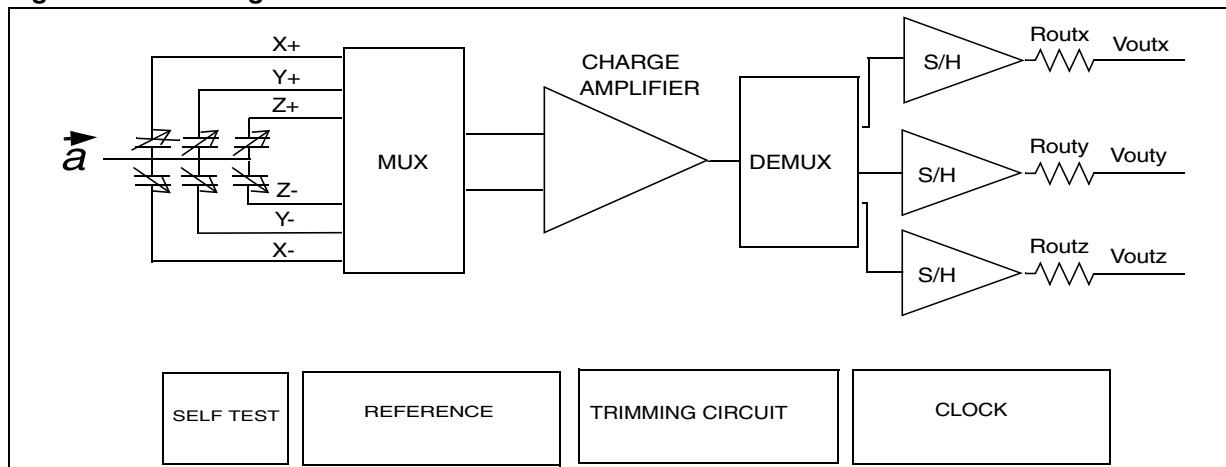


Table 2. Pin Description

N°	Pin	Function
1 to 5	NC	Internally not connected
6	GND	0V supply
7	Vdd	Power supply
8	Vouty	Output Voltage
9	ST	Self Test (Logic 0: normal mode; Logic 1: Self-test)
10	Voutx	Output Voltage
11	PD	Power Down (Logic 0: normal mode; Logic 1: Power-Down mode)
12	Voutz	Output Voltage
13	FS	Full Scale selection (Logic 0: 2g Full-scale; Logic 1: 6g Full-scale)
14-15	Reserved	Leave unconnected or connect to Vdd
16	Reserved	Connect to Vdd or ground
17	Reserved	Leave unconnected or connect to Vdd
18	Reserved	Leave unconnected or connect to ground
19 to 24	NC	Internally not connected

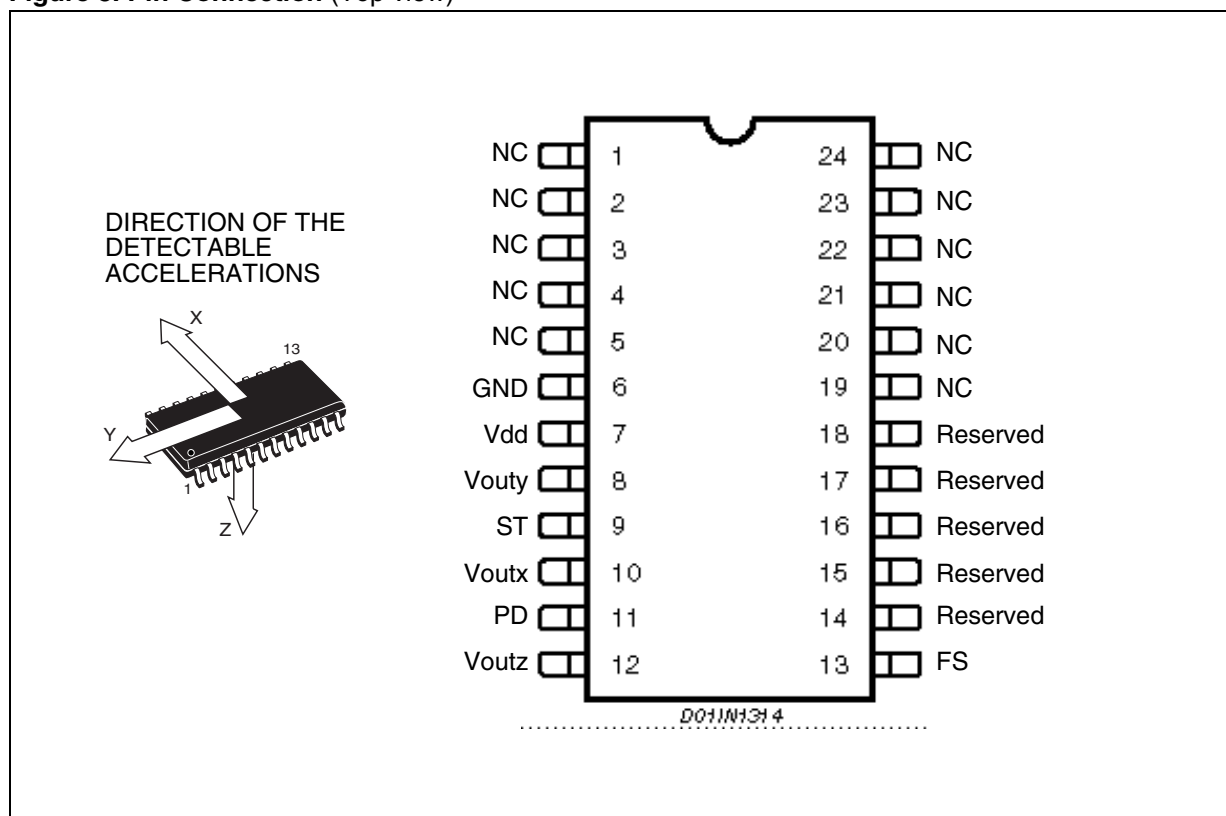
Figure 3. Pin Connection (Top view)

Table 3. Mechanical Characteristics¹

(Temperature range -40°C to +85°C). All the parameters are specified @ Vdd =3.3V, T=25°C unless otherwise noted

Symbol	Parameter	Test Condition	Min.	Typ. ²	Max.	Unit
Ar	Acceleration Range ³	FS pin connected to GND	±1.8	±2.0		g
		FS pin connected to Vdd	±5.4	±6.0		g
So	Sensitivity ⁴	Full-scale = 2g	Vdd/5–10%	Vdd/5	Vdd/5+10%	V/g
		Full-scale = 6g	Vdd/15–10%	Vdd/15	Vdd/15+10%	V/g
SoDr	Sensitivity Change Vs Temperature	Delta from +25°C		±0.01		%/°C
Voff	Zero-g Level ⁴	T = 25°C	Vdd/2-10%	Vdd/2	Vdd/2+10%	V
OffDr	Zero-g level Change Vs Temperature	Delta from +25°C		±1.1		mg/°C
NL	Non Linearity ⁵	Best fit straight line Full-scale = 2g X, Y axis		±0.3	±1.5	% FS
		Best fit straight line; Full-scale = 2g Z axis		±0.6	±2	% FS
CrossAx	Cross-Axis ⁶			±2	±4	%
An	Acceleration Noise Density	Vdd=3.3V; Full-scale = 2g		50		µg/√Hz
Vt	Self test Output Voltage Change ^{7,8,9}	T = 25°C Vdd=3.3V Full-scale = 2g X axis	-20	-50	-100	mV
		T = 25°C Vdd=3.3V Full-scale = 2g Y axis	20	50	100	mV
		T = 25°C Vdd=3.3V Full-scale = 2g Z axis	20	50	100	mV
Fres	Sensing Element Resonance Frequency ¹⁰	all axes	1.5			KHz
Top	Operating Temperature Range		-40		+85	°C
Wh	Product Weight			0.6		gram

Notes: 1. The product is factory calibrated at 3.3V. The device can be powered from 2.4V to 3.6V. Voff, So and Vt parameters will vary with supply voltage.

2. Typical specifications are not guaranteed

3. Verified by wafer level test and measurement of initial offset and sensitivity

4. Zero-g level and sensitivity are essentially ratiometric to supply voltage

5. Guaranteed by design

6. Contribution to the measuring output of an inclination/acceleration along any perpendicular axis

7. "Self test output voltage change" is defined as $V_{out}(V_{st}=Logic1) - V_{out}(V_{st}=Logic0)$

8. "Self test output voltage change" varies cubically with supply voltage

9. When full-scale is set to ±6g, "self-test output voltage change" is one third of the specified value.

10. Minimum resonance frequency Fres=1.5KHz. Sensor bandwidth= $1/(2 \cdot \pi \cdot 110K\Omega \cdot C_{load})$ with $C_{load} > 1nF$.

Table 4. Electrical Characteristics¹

(Temperature range -40°C to +85°C) All the parameters are specified @ Vdd =3.3V, T=25°C unless otherwise noted

Symbol	Parameter	Test Condition	Min.	Typ. ²	Max.	Unit
Vdd	Supply Voltage		2.4	3.3	3.6	V
Idd	Supply Current	mean value PD pin connected to GND		0.85	1.5	mA
IddPdn	Supply Current in Power Down Mode	rms value PD pin connected to Vdd		2	5	μA
Vst	Self Test Input	Logic 0 level	0		0.8	V
		Logic 1 level	2.2		Vdd	V
Rout	Output Impedance		80	110	140	kΩ
Cload	Capacitive Load Drive ³		320			pF
Ton	Turn-On Time at exit from Power Down mode	Cload in μF		550*Cload+0.3		ms

Notes: 1. The product is factory calibrated at 3.3V.

2. Typical specifications are not guaranteed

3. Minimum resonance frequency Fres=1.5kHz. Sensor bandwidth=1/(2*π*110KΩ*Cload) with Cload>1nF

3 Absolute Maximum Rating

Stresses above those listed as “absolute maximum ratings” may cause permanent damage to the device. This is a stress rating only and functional operation of the device under these conditions is not implied. Exposure to maximum rating conditions for extended periods may affect device reliability.

Table 5. Absolute Maximum Rating

Symbol	Ratings	Maximum Value	Unit
Vdd	Supply Voltage	-0.3 to 7	V
Vin	Input Voltage on any control pin (FS, PD, ST)	-0.3 to Vdd +0.3	V
A _{POW}	Acceleration (Any axis, Powered, Vdd=3.3V)	3000g for 0.5 ms	
		10000g for 0.1 ms	
A _{UNP}	Acceleration (Any axis, Not powered)	3000g for 0.5 ms	
		10000g for 0.1 ms	
T _{STG}	Storage Temperature Range	-40 to +125	°C
ESD	Electrostatic Discharge Protection	2 (HBM)	kV
		200 (MM)	V
		1500 (CDM)	V



This is a Mechanical Shock sensitive device, improper handling can cause permanent damages to the part



This is an ESD sensitive device, improper handling can cause permanent damages to the part

3.1 Terminology

3.1.1 Sensitivity

Describes the gain of the sensor and can be determined by applying 1g acceleration to it. As the sensor can measure DC accelerations this can be done easily by pointing the axis of interest towards the center of the earth, note the output value, rotate the sensor by 180 degrees (point to the sky) and note the output value again thus applying $\pm 1g$ acceleration to the sensor. Subtracting the larger output value from the smaller one and dividing the result by 2 will give the actual sensitivity of the sensor. This value changes very little over temperature (see sensitivity change vs. temperature) and also very little over time. The Sensitivity Tolerance describes the range of Sensitivities of a large population of sensors.

3.1.2 Zero-g level

Describes the actual output signal if there is no acceleration present. A sensor in a steady state on an horizontal surface will measure 0g in X axis and 0g in Y axis whereas the Z axis will measure +1g. The output is ideally for a 3.3V powered sensor $V_{dd}/2 = 1650mV$. A deviation from ideal 0-g level (1650mV in this case) is called Zero-g offset. Offset of precise MEMS sensors is to some extent a result of stress to the sensor and therefore the offset can slightly change after mounting the sensor onto a printed circuit board or exposing it to extensive mechanical stress. Offset changes little over temperature - see "Zero-g Level Change vs. Temperature" - the Zero-g level of an individual sensor is very stable over lifetime. The Zero-g level tolerance describes the range of zero-g levels of a population of sensors.

3.1.3 Self Test

Self Test allows to test the mechanical and electric part of the sensor, allowing the seismic mass to be moved by means of an electrostatic test-force. The Self Test function is off when the ST pin is connected to GND. When the ST pin is tied at Vdd an actuation force is applied to the sensor, simulating a definite input acceleration. In this case the sensor outputs will exhibit a voltage change in their DC levels which is related to the selected full scale and depending on the Supply Voltage through the device sensitivity. When ST is activated, the device output level is given by the algebraic sum of the signals produced by the acceleration acting on the sensor and by the electrostatic test-force. If the output signals change within the amplitude specified inside Table 3, then the sensor is working properly and the parameters of the interface chip are within the defined specification.

3.1.4 Output impedance

Describes the resistor inside the output stage of each channel. This resistor is part of a filter consisting of an external capacitor of at least 320pF and the internal resistor. Due to the high resistor level only small, inexpensive external capacitors are needed to generate low corner frequencies. When interfacing with an ADC it is important to use high input impedance input circuitries to avoid measurement errors. Note that the minimum load capacitance forms a corner frequency beyond the resonance frequency of the sensor. For a flat frequency response a corner frequency well below the resonance frequency is recommended. In general the smallest possible bandwidth for an particular application should be chosen to get the best results.

4 Functionality

The LIS3L02AS4 is a high performance, low-power, analog output three axes linear accelerometer packaged in a SO24 package. The complete device includes a sensing element and an IC interface able to take the information from the sensing element and to provide an analog signal to the external world.

4.1 Sensing element

A proprietary process is used to create a surface micro-machined accelerometer. The technology allows to carry out suspended silicon structures which are attached to the substrate in a few points called anchors and are free to move in the direction of the sensed acceleration. To be compatible with the traditional packaging techniques a cap is placed on top of the sensing element to avoid blocking the moving parts during the moulding phase of the plastic encapsulation.

When an acceleration is applied to the sensor the proof mass displaces from its nominal position, causing an imbalance in the capacitive half-bridge. This imbalance is measured using charge integration in response to a voltage pulse applied to the sense capacitor.

At steady state the nominal value of the capacitors are few pF and when an acceleration is applied the maximum variation of the capacitive load is up to 100fF.

4.2 IC Interface

In order to increase robustness and immunity against external disturbances the complete signal processing chain uses a fully differential structure. The final stage converts the differential signal into a single-ended one to be compatible with the external world.

The signals of the sensing element are multiplexed and fed into a low-noise capacitive charge amplifier that implements a Correlated Double Sampling (CDS) at its output to cancel the offset and the 1/f noise. The output signal is de-multiplexed and transferred to three different S&Hs, one for each channel and made available to the outside.

The low noise input amplifier operates at 200 kHz while the three S&Hs operate at a sampling frequency of 66 kHz. This allows a large oversampling ratio, which leads to in-band noise reduction and to an accurate output waveform.

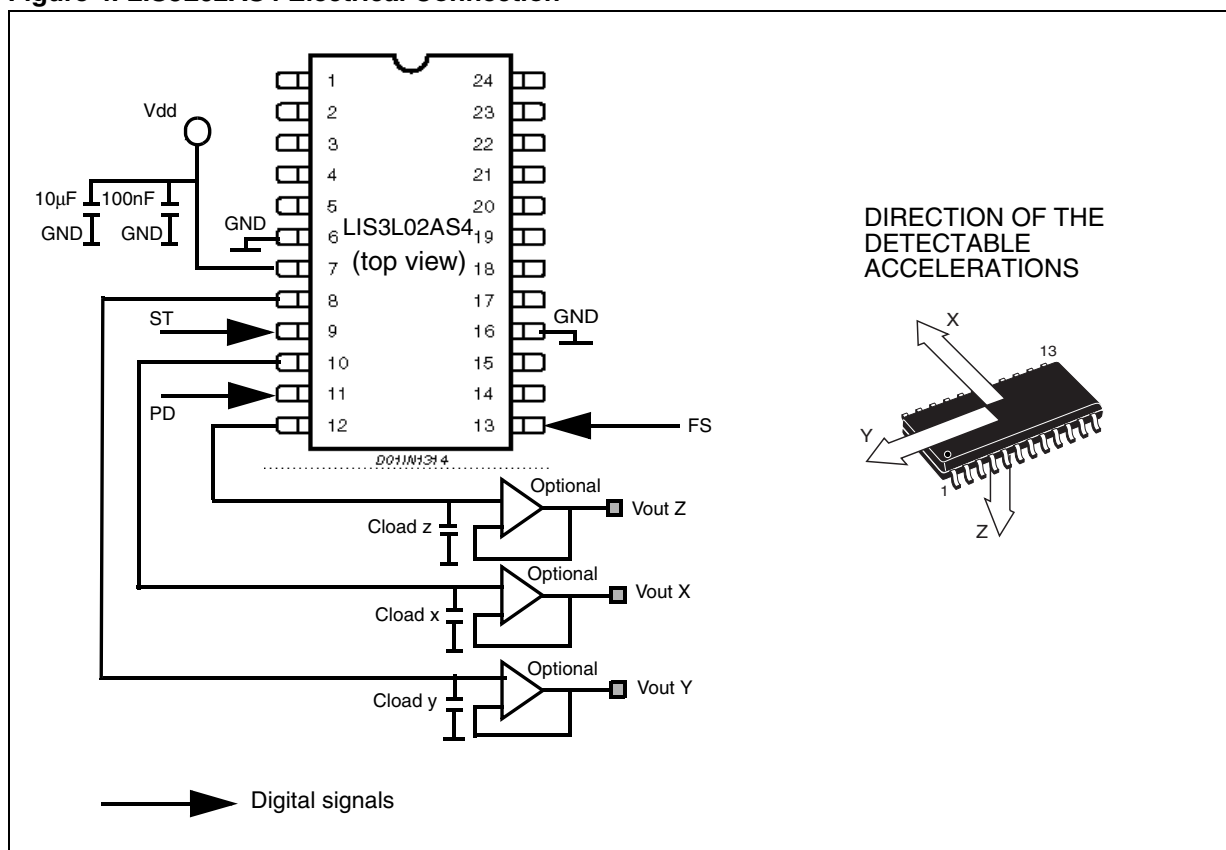
All the analog parameters (zero-g level, sensitivity and self-test) are ratiometric to the supply voltage. Increasing or decreasing the supply voltage, the sensitivity and the offset will increase or decrease almost linearly. The self test voltage change varies cubically with the supply voltage

4.3 Factory calibration

The IC interface is factory calibrated for Sensitivity (So) and Zero-g Level (Voff). The trimming values are stored inside the device by a non volatile structure. Any time the device is turned on, the trimming parameters are downloaded into the registers to be employed during the normal operation. This allows the user to employ the device without further calibration.

5 Application Hints

Figure 4. LIS3L02AS4 Electrical Connection



Power supply decoupling capacitors (100nF ceramic + 10μF Al) should be placed as near as possible to the device (common design practice).

The LIS3L02AS4 allows to band limit Voutx, Vouty and Voutz through the use of external capacitors. The re-commended frequency range spans from DC up to 1.5 KHz. In particular, capacitors must be added at output pins to implement low-pass filtering for antialiasing and noise reduction. The equation for the cut-off frequency (f_t) of the external filters is:

$$f_t = \frac{1}{2\pi \cdot R_{out} \cdot C_{load}(x, y, z)}$$

Taking in account that the internal filtering resistor (R_{out}) has a nominal value equal to 110kΩ, the equation for the external filter cut-off frequency may be simplified as follows:

$$f_t = \frac{1.45\mu F}{C_{load}(x, y, z)} [Hz]$$

The tolerance of the internal resistor can vary typically of ±20% within its nominal value of 110kΩ; thus the cut-off frequency will vary accordingly. A minimum capacitance of 320 pF for $C_{load}(x, y, z)$ is required in any case.

Table 6. Filter Capacitor Selection, $C_{load}(x,y,z)$. Capacitance Value Choose.

Cut-off frequency	Capacitor value
1 Hz	1500nF
10 Hz	150nF
50 Hz	30 nF
100 Hz	15 nF
200 Hz	6.8 nF
500 Hz	3 nF

5.1 Soldering information

The SO24 package is lead free qualified for soldering heat resistance according to JEDEC J-STD-020C.

5.2 Output response vs orientation

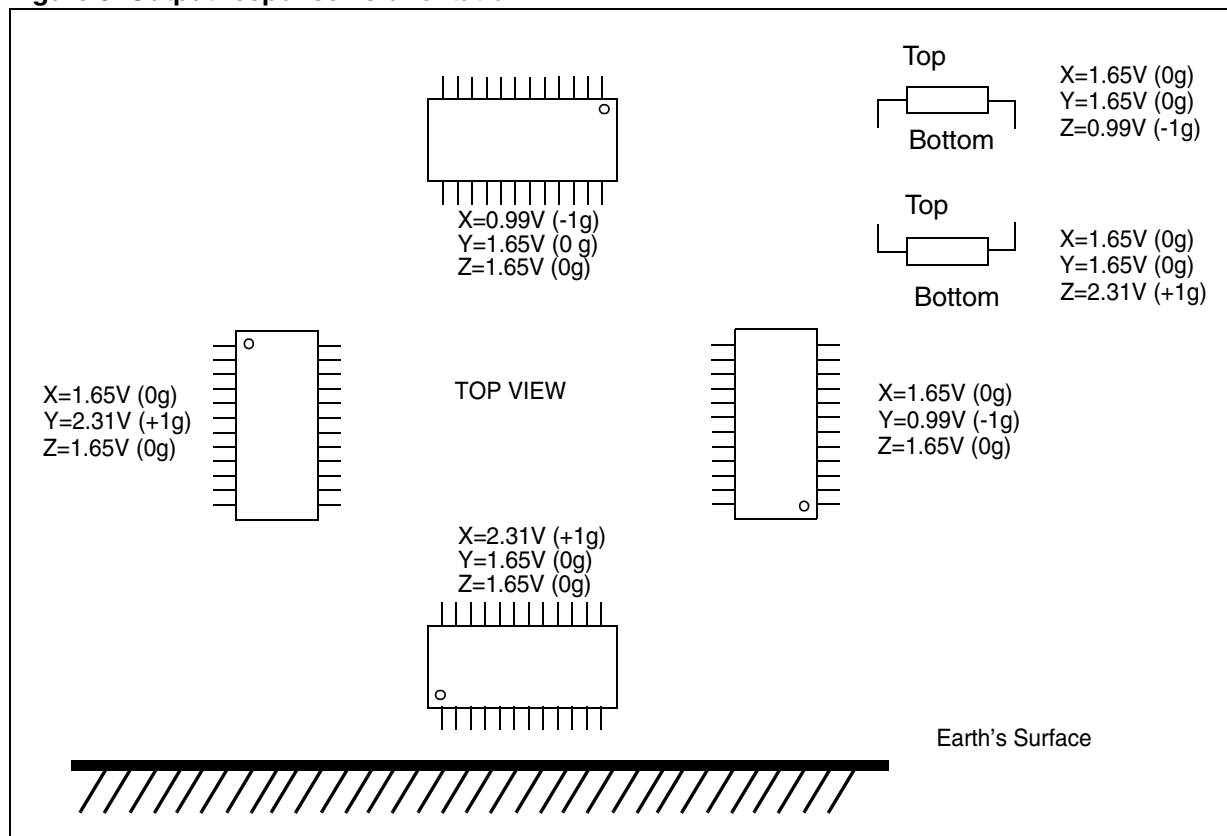
Figure 5. Output response vs orientation

Figure 5 refers to LIS3L02AS4 device powered at 3.3V

6 Typical performance Characteristics

6.1 Mechanical Characteristics at 25°C.

Figure 6. X axis Zero g Level at 3.3V

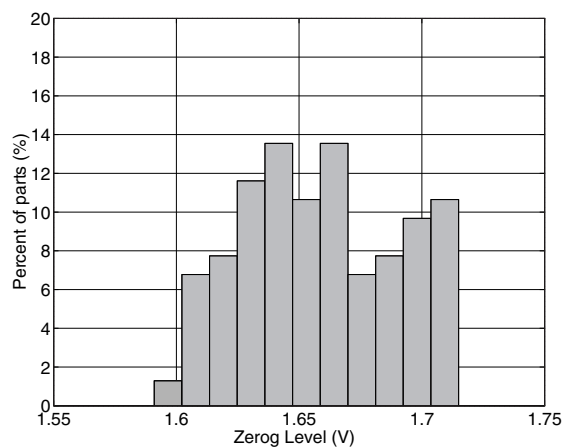


Figure 9. X axis Sensitivity at 3.3V

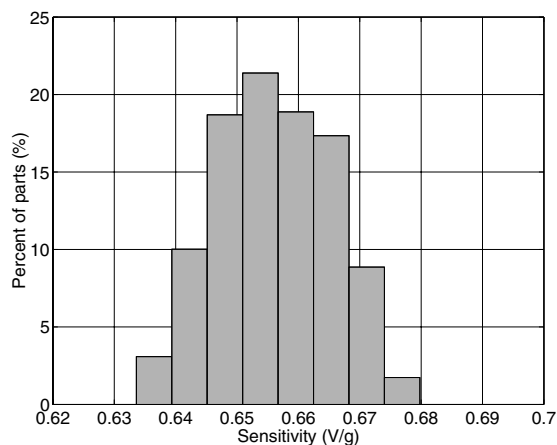


Figure 7. Y axis Zero g Level at 3.3V

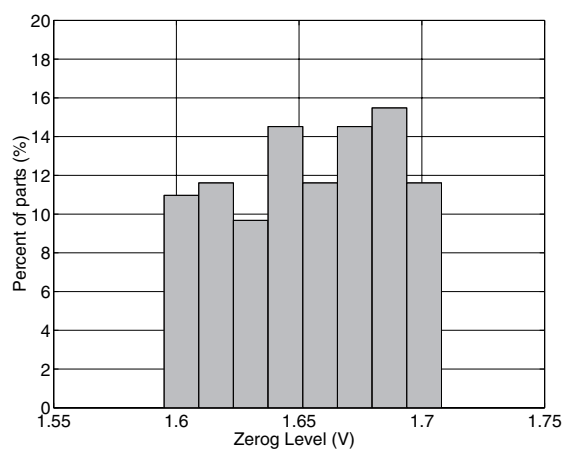


Figure 10. Y axis Sensitivity at 3.3V

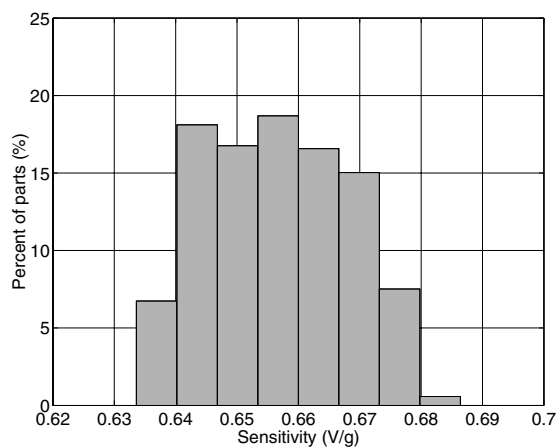


Figure 8. Z axis Zero g Level at 3.3V

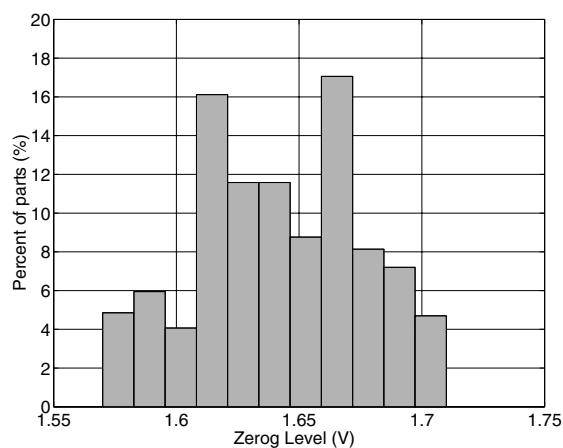
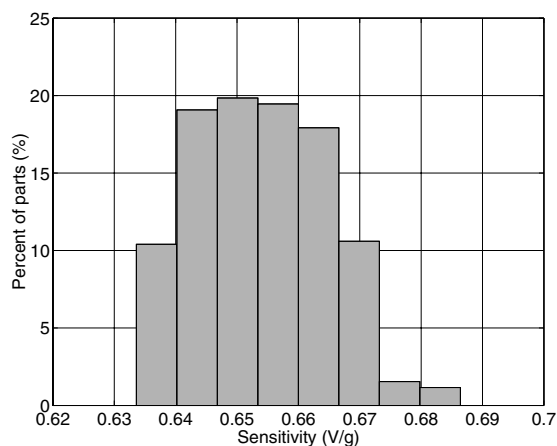


Figure 11. Z axis Sensitivity at 3.3V



6.2 Mechanical Characteristics derived from measurement in the -40°C to +85°C temperature range

Figure 12. X axis Zero g Level Change Vs. Temperature

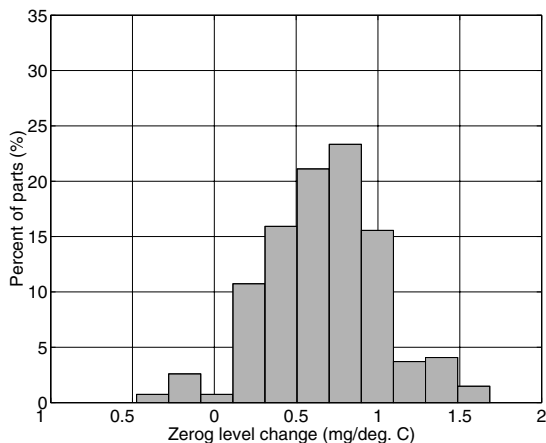


Figure 15. X axis Sensitivity Change Vs. Temperature

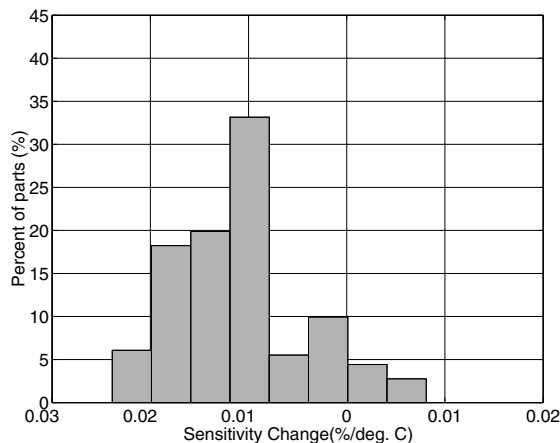


Figure 13. Y axis Zero g Level Change Vs. Temperature

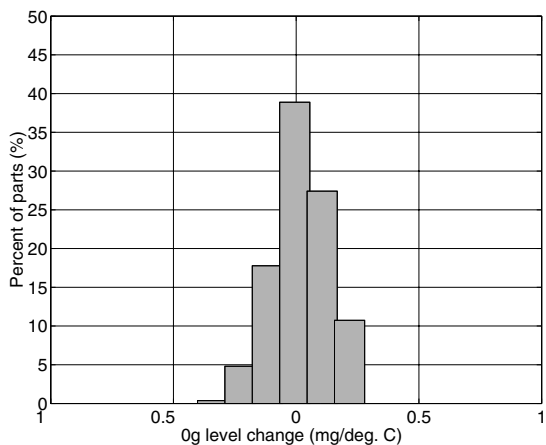


Figure 16. Y axis Sensitivity Change Vs. Temperature

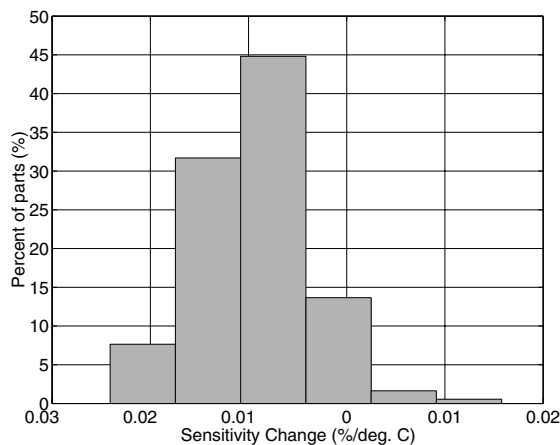


Figure 14. Z axis Zero g Level Change Vs. Temperature

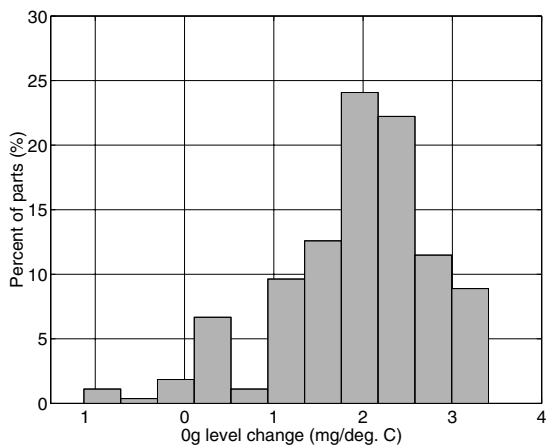
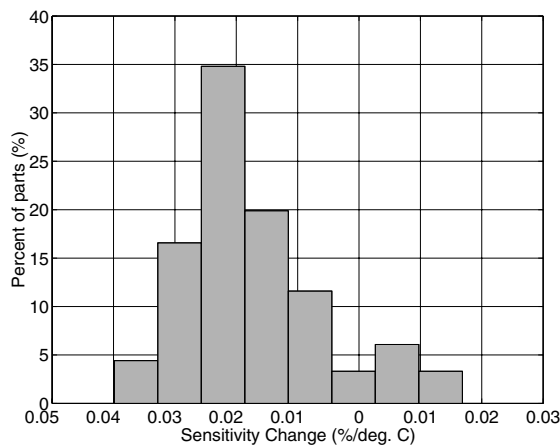


Figure 17. Z axis Sensitivity Change Vs. Temperature



6.3 Electrical Characteristics at 25°C

Figure 18. Noise density at 3.3V (X,Y axes)

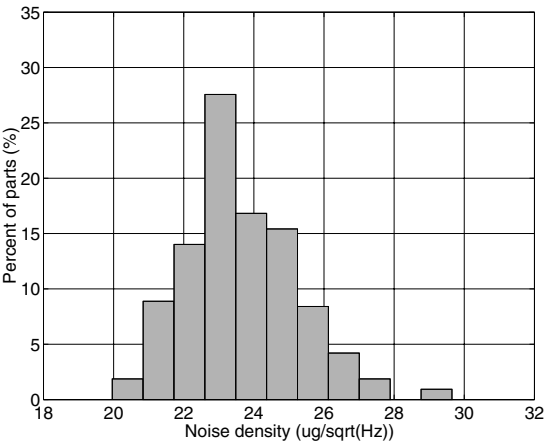


Figure 20. Current consumption at 3.3V

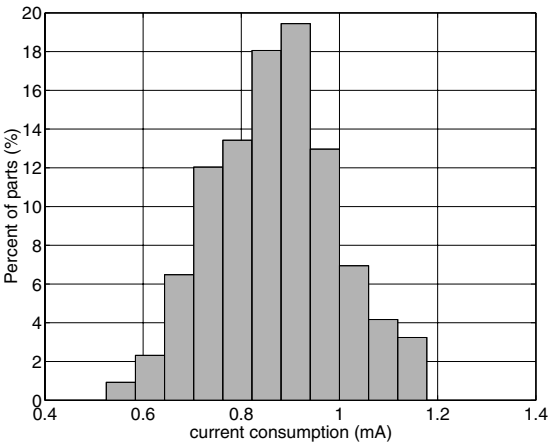


Figure 19. Noise density at 3.3V (Z axis)

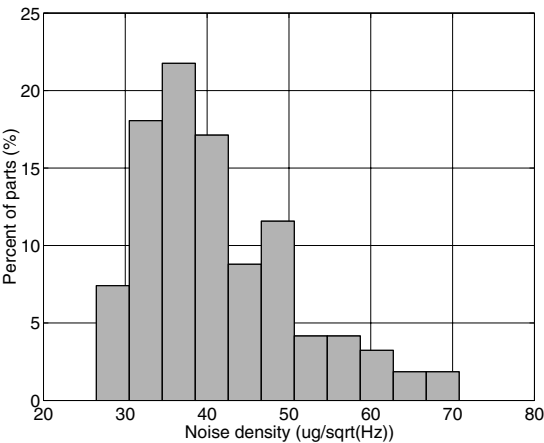
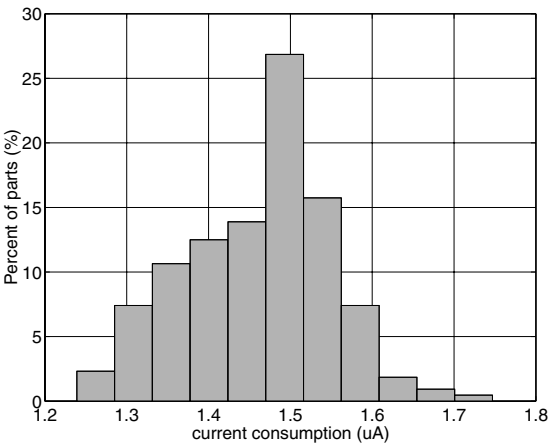


Figure 21. Current consumption in power down mode at 3.3V



7 Package Information

In order to meet environmental requirements, ST offers these devices in ECOPACK® packages. These packages have a Lead-free second level interconnect. The category of second Level Interconnect is marked on the package and on the inner box label, in compliance with JEDEC Standard JESD97. The maximum ratings related to soldering conditions are also marked on the inner box label.

ECOPACK is an ST trademark. ECOPACK specifications are available at: www.st.com.

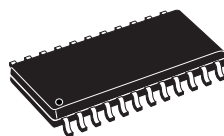
Figure 22. SO24 Mechanical Data & Package Dimensions

DIM.	mm			inch		
	MIN.	TYP.	MAX.	MIN.	TYP.	MAX.
A	2.35		2.65	0.093		0.104
A1	0.10		0.30	0.004		0.012
B	0.33		0.51	0.013		0.200
C	0.23		0.32	0.009		0.013
D (1)	15.20		15.60	0.598		0.614
E	7.40		7.60	0.291		0.299
e		1.27			0.050	
H	10.0		10.65	0.394		0.419
h	0.25		0.75	0.010		0.030
L	0.40		1.27	0.016		0.050
k	0° (min.), 8° (max.)					
ddd			0.10			0.004

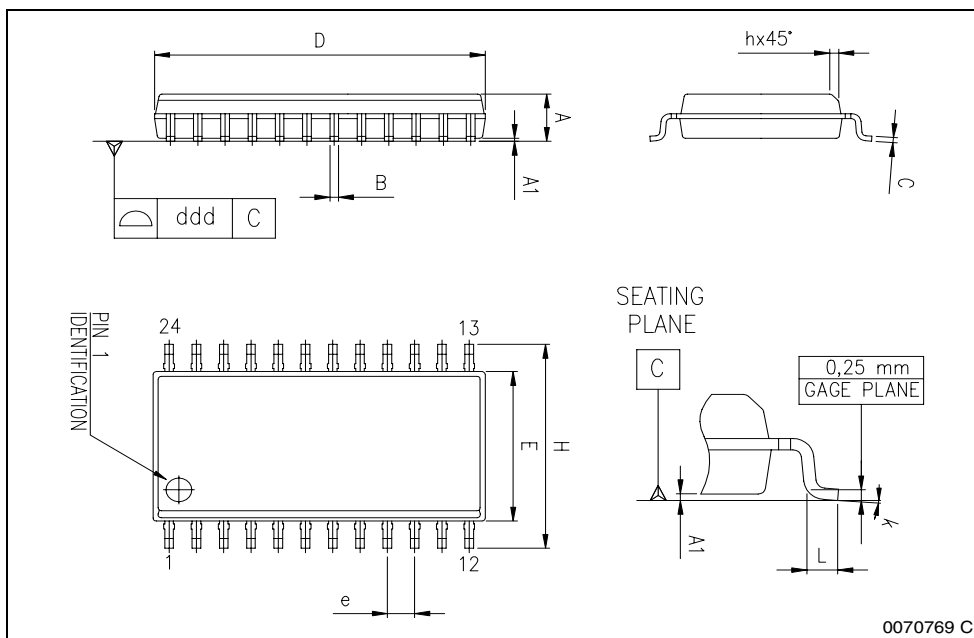
(1) "D" dimension does not include mold flash, protusions or gate burrs. Mold flash, protusions or gate burrs shall not exceed 0.15mm per side.

OUTLINE AND MECHANICAL DATA

Weight: 0.60gr



SO24



8 Revision History

Table 7. Revision History

Date	Revision	Description of Changes
February 2004	1	First issue
1-Dec-2005	2	Changed from Product preview to Datasheet maturity. Added Typical performance Characteristics section.

Information furnished is believed to be accurate and reliable. However, STMicroelectronics assumes no responsibility for the consequences of use of such information nor for any infringement of patents or other rights of third parties which may result from its use. No license is granted by implication or otherwise under any patent or patent rights of STMicroelectronics. Specifications mentioned in this publication are subject to change without notice. This publication supersedes and replaces all information previously supplied. STMicroelectronics products are not authorized for use as critical components in life support devices or systems without express written approval of STMicroelectronics.

The ST logo is a registered trademark of STMicroelectronics.
All other names are the property of their respective owners

© 2005 STMicroelectronics - All rights reserved

STMicroelectronics group of companies

Australia - Belgium - Brazil - Canada - China - Czech Republic - Finland - France - Germany - Hong Kong - India - Israel - Italy - Japan -
Malaysia - Malta - Morocco - Singapore - Spain - Sweden - Switzerland - United Kingdom - United States of America

www.st.com

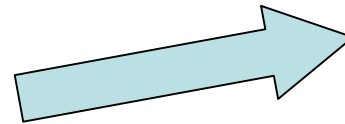
Accelerometers and How they Work

- **Contents summary**
 - Definition of Acceleration
 - Technologies
 - Terminology
 - Effect of Tilt
 - Typical applications
 - Summary

Acceleration Fundamentals

- **What is Acceleration?**

- Definition: the time rate of change of velocity
- A.K.A.: the time rate of change of the time rate of change of distance



$$a = \frac{\partial v}{\partial t} = \frac{\partial^2 x}{\partial t^2}$$

- **What are the units?**

- Acceleration is measured in (ft/s)/s or (m/s)/s

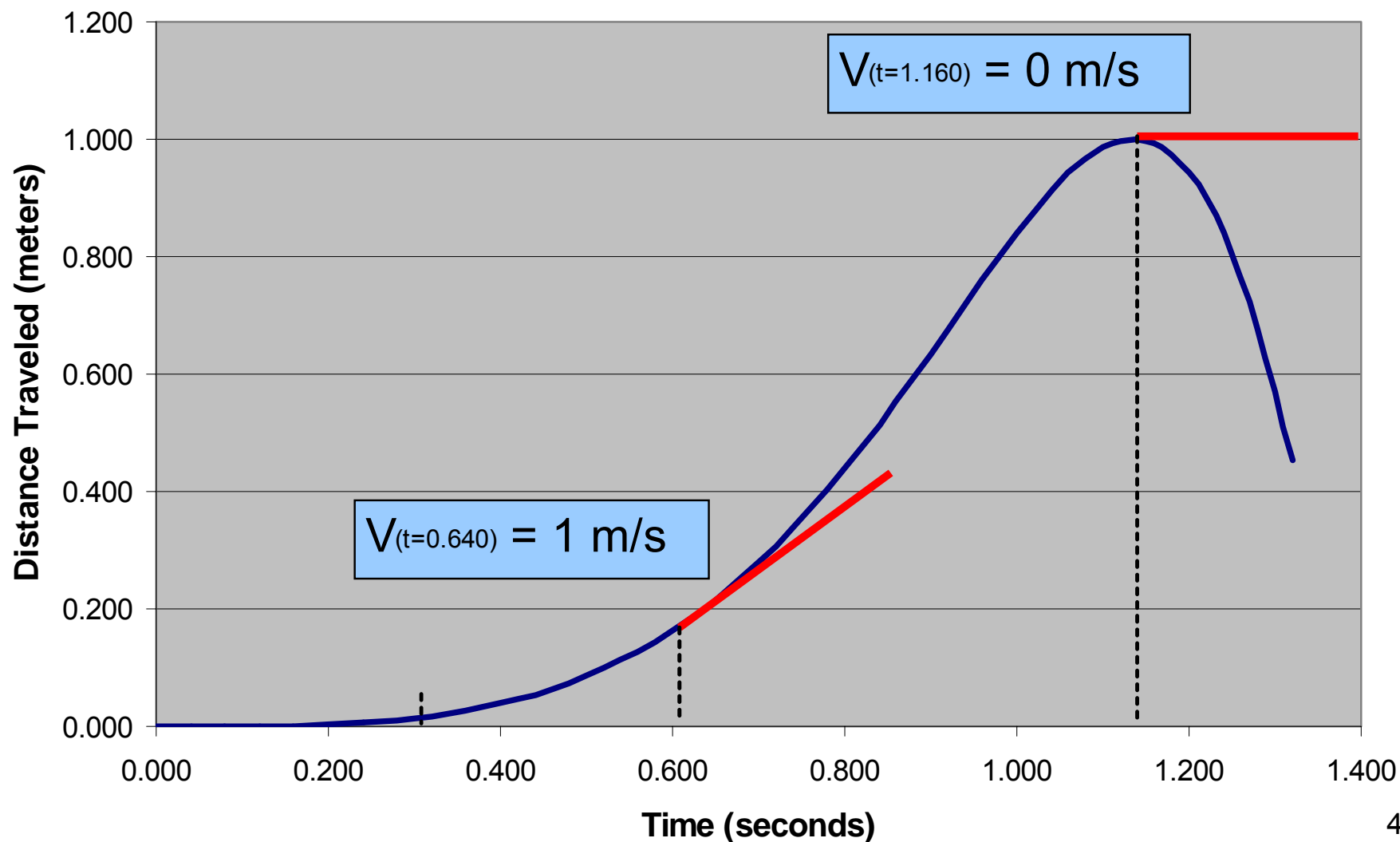
- **What is a “g”?**

- A “g” is a unit of acceleration equal to Earth’s gravity at sea level
 - 32.2 ft/s² or 9.81 m/s²

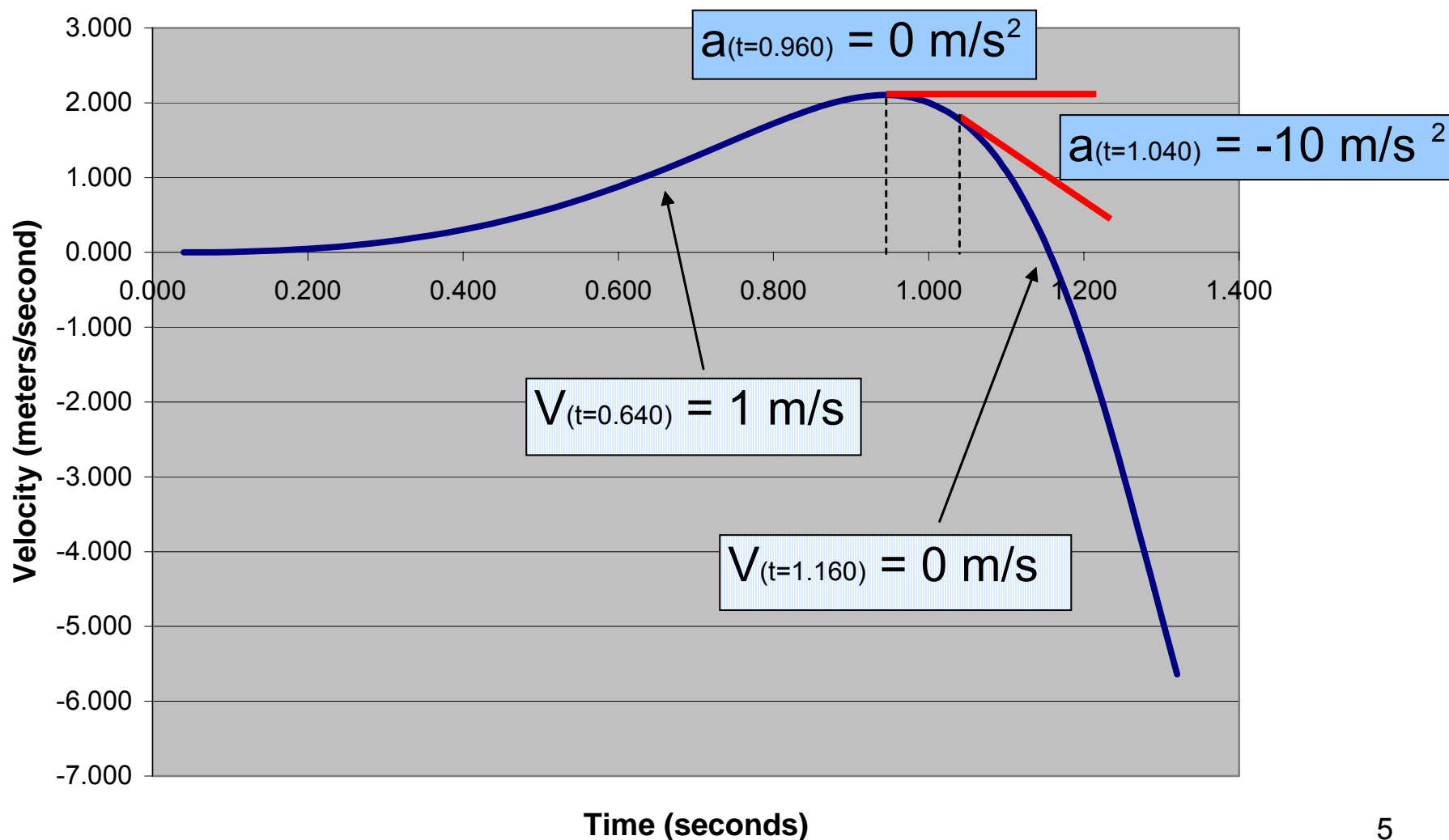
More Notes on Acceleration

- **What is the time rate of change of velocity?**
 - When plotted on a graph, velocity is the slope of distance versus time
 - Acceleration is the slope of velocity versus time

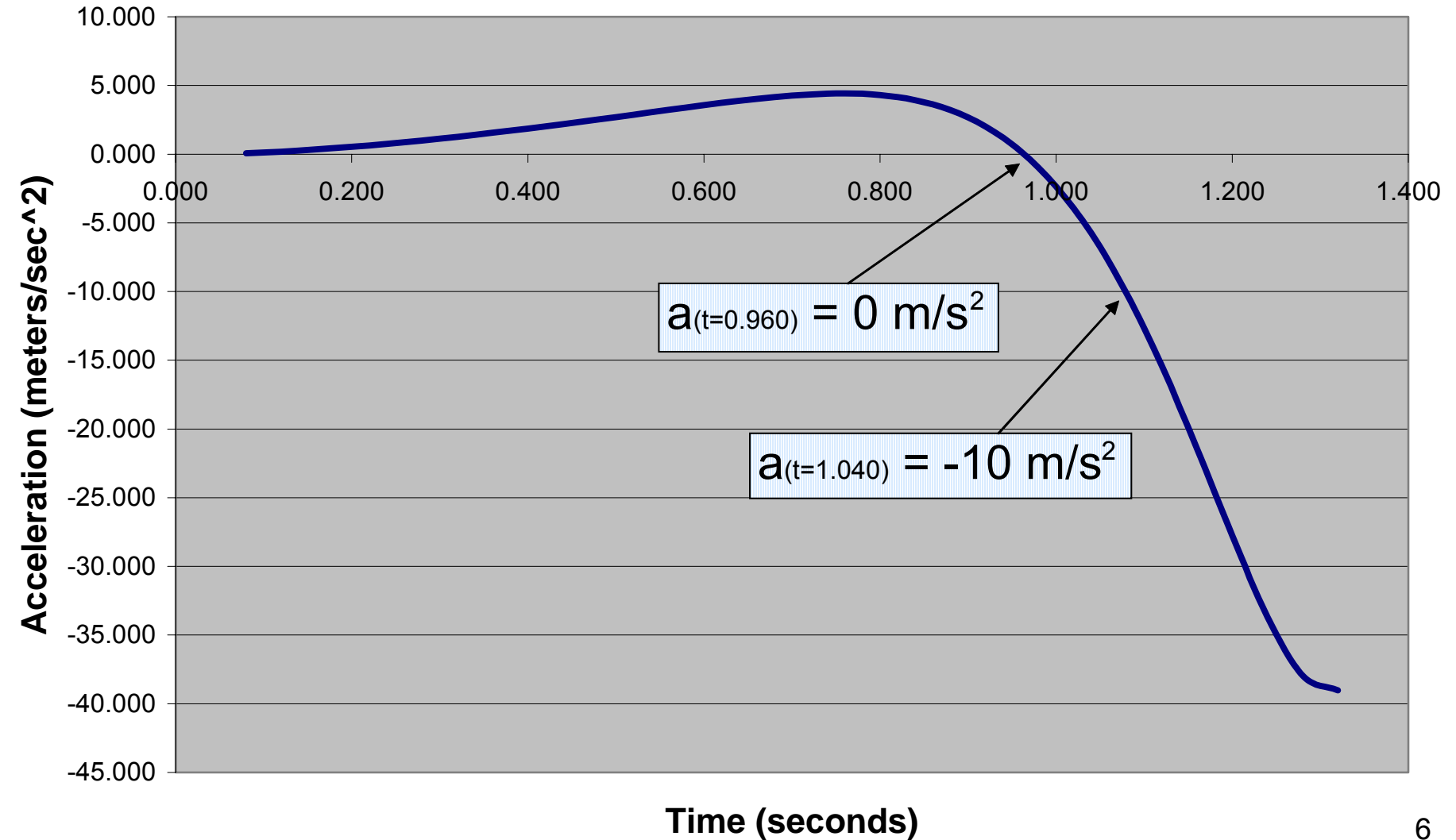
How to find velocity from distance traveled



How to find acceleration from velocity



Acceleration vs. Time



Acceleration in Human Terms

- What are some “g” reference points?

Description	“g” level
Earth’s gravity	1g
Passenger car in corner	2g
Bumps in road	2g
Indy car driver in corner	3g
Bobsled rider in corner	5g
Human unconsciousness	7g
Space shuttle	10g

What's the point?

- **Why measure acceleration?**
 - Acceleration is a physical characteristic of a system.
 - The measurement of acceleration is used as an input into some types of control systems.
 - The control systems use the measured acceleration to correct for changing dynamic conditions

Common Types of Accelerometers

Sensor Category

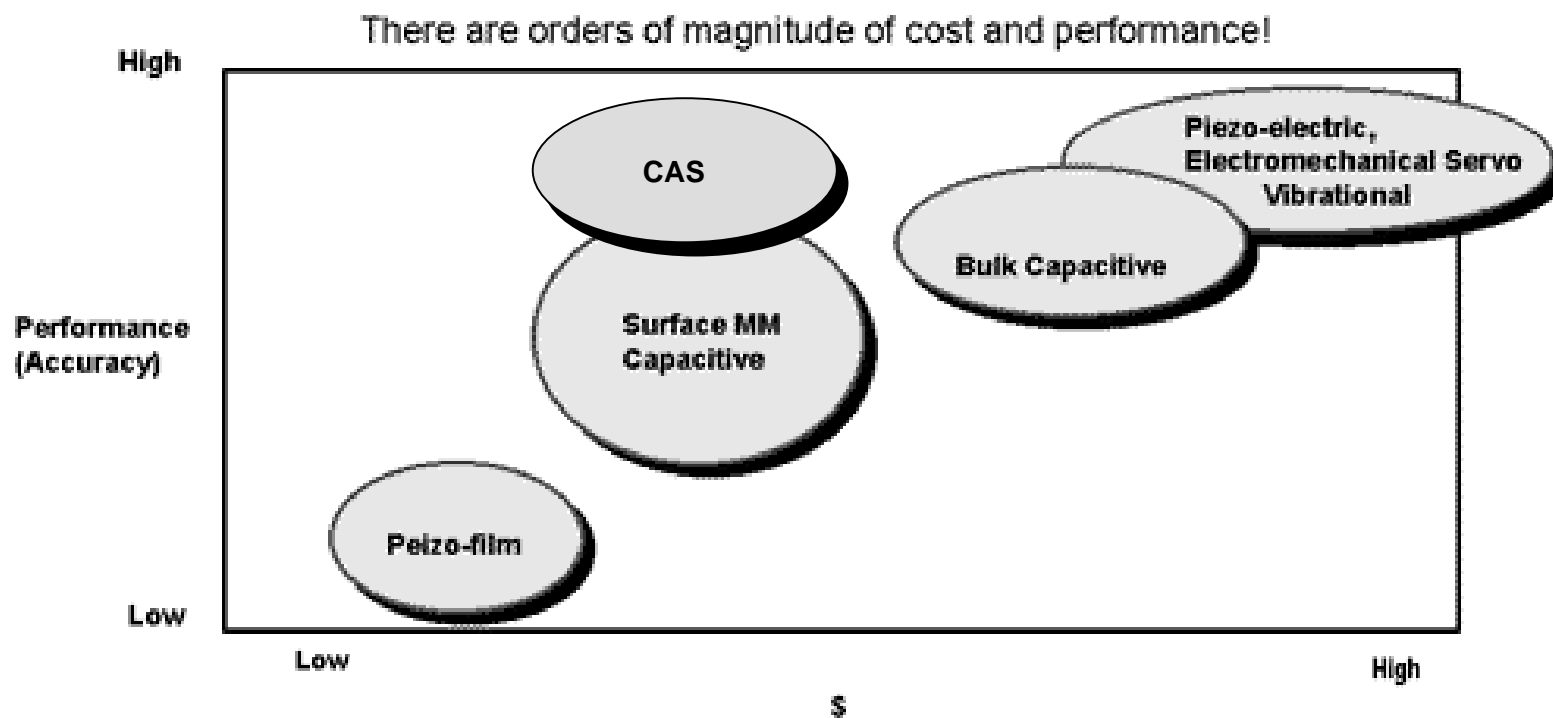
- **Capacitive**
- **Piezoelectric**
- **Piezoresistive**
- **Hall Effect**
- **Magnetoresistive**
- **Heat Transfer**

Key Technologies

- Metal beam or micromachined feature produces capacitance; change in capacitance related to acceleration
- Piezoelectric crystal mounted to mass – voltage output converted to acceleration
- Beam or micromachined feature whose resistance changes with acceleration
- Motion converted to electrical signal by sensing of changing magnetic fields
- Material resistivity changes in presence of magnetic field
- Location of heated mass tracked during acceleration by sensing temperature

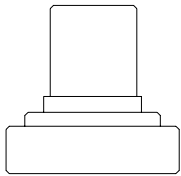
What Type of Acceleration Sensor Does TI Produce and why?

- **Capacitive Acceleration Sensor**
 - “CAS”

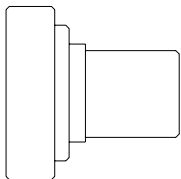


Acceleration Sensor Terminology

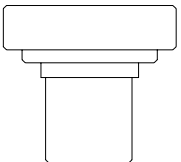
(TI Convention)



- **+1g:** Output of the sensor with the base connector pointed up



- **0g:** Output of the sensor with the base connector horizontal



- **-1g:** Output of the sensor with the base connector pointed down

- **Linearity:** The maximum deviation of the calibration curve from a straight line.

$$Linearity = V_{out,0g} - \frac{1}{2} (V_{out,+1g} + V_{out,-1g})$$

Acceleration Sensor Terminology

- **Sensitivity**: A measure of how much the output of a sensor changes as the input acceleration changes. Measured in Volts/g

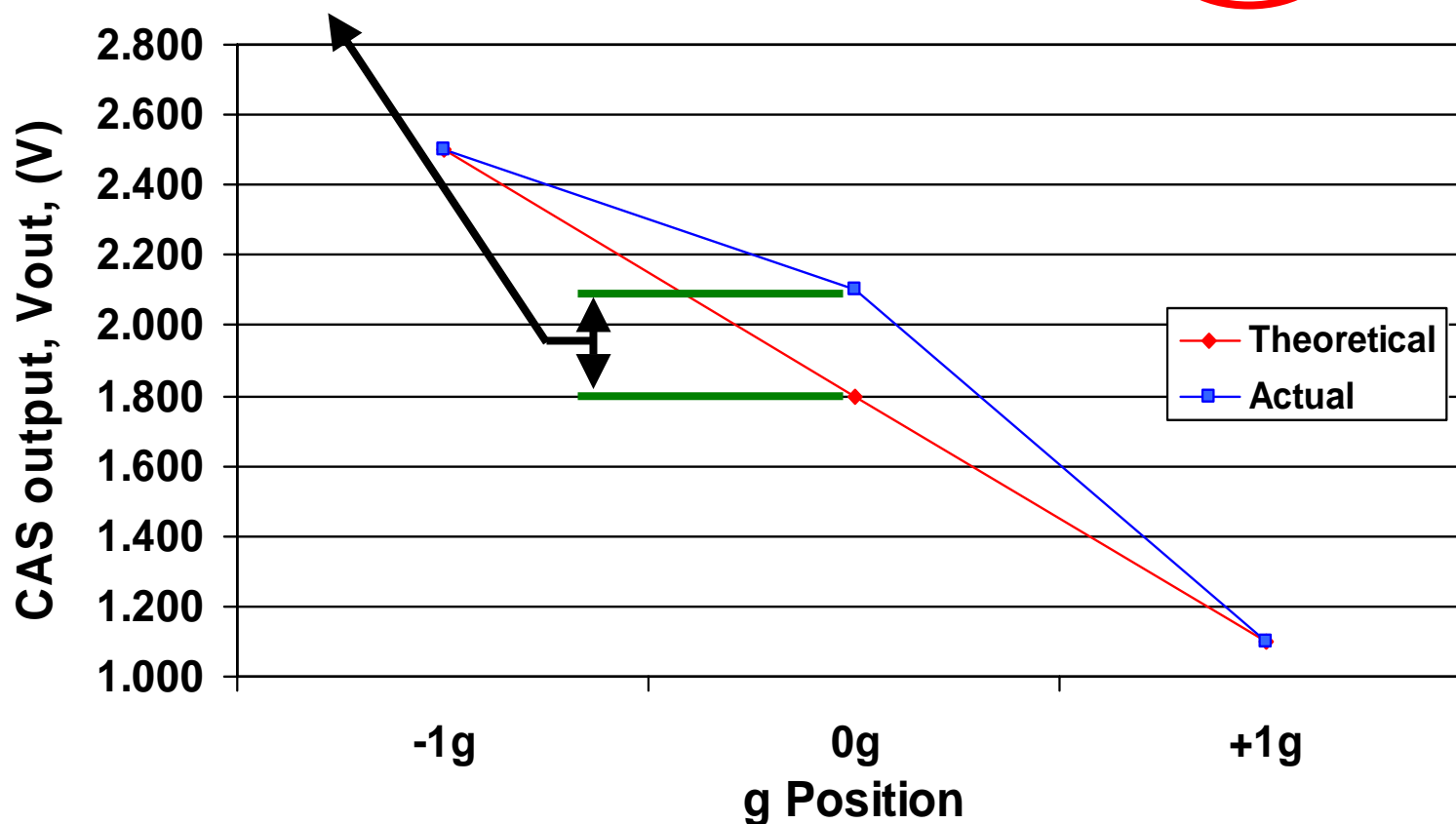
$$Sensitivity = \frac{\Delta V_{out}}{\Delta g} = \frac{V_{out,+1g} - V_{out,-1g}}{2g}$$

- **Vcc**: The voltage supplied to the input of the sensor
 - 5.000 ± 0.005V for CAS device
- **%Vcc**: Readings are often represented as a % of the supply voltage. This allows for correction due to supply voltage variances between readings.

Example: Sensitivity & Linearity

$$\text{Sensitivity} = \frac{\Delta V_{out}}{\Delta g} = \frac{V_{out,+1g} - V_{out,-1g}}{2g} = \frac{1.1V - 2.5V}{2g} = \frac{1.1V - 2.5V}{2g} = -0.7 \frac{\text{Volts}}{g}$$

$$\text{Linearity} = V_{out,0g} - \frac{1}{2}(V_{out,+1g} + V_{out,-1g}) = 2.1 - \frac{1}{2}(1.1 + 2.5) = 0.3 \text{Volts}$$



Acceleration Sensor Terminology

- **Ratiometric**: The output of the sensor changes with a change in the input voltage.
 - Example
 - At $V_{cc} = 5.000V$, V_{out} at $0g = 1.800V$
 - In terms of $\%V_{cc}$, this is $1.800V_{out}/5.000V_{cc} * 100\% = 36\%V_{cc}$
 - Now suppose the input voltage changes: $V_{cc} = 5.010V$.
 - At $0g$, the ratiometric device output is still $36\% V_{cc}$.
 - In terms of the output voltage, $36\%V_{cc} * 5.010V_{cc} = 1.804V_{out}$
 - So a $0.010V$ change in V_{cc} will cause a $0.004V$ error in the $0g$ output if you do not evaluate the output as $\%V_{cc}$

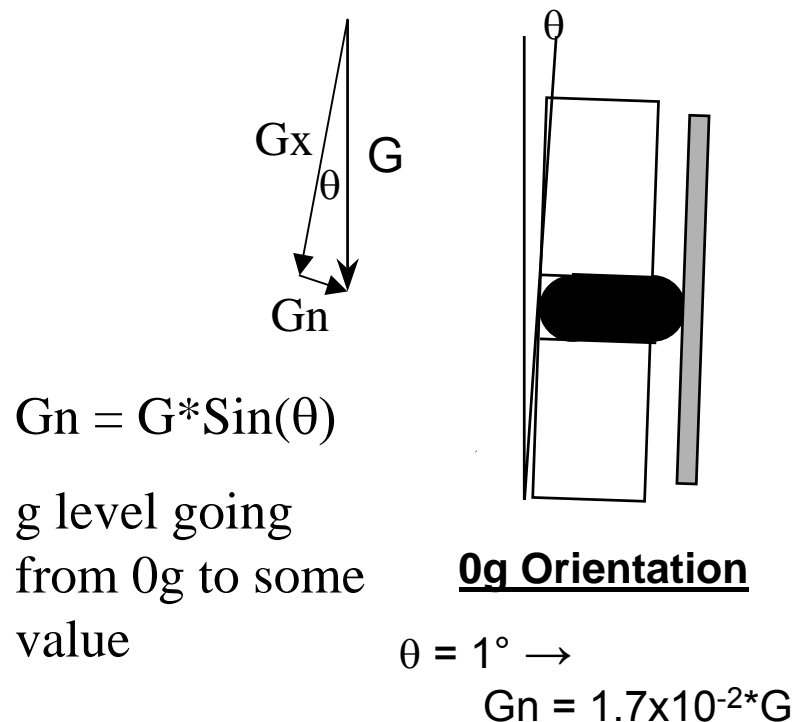
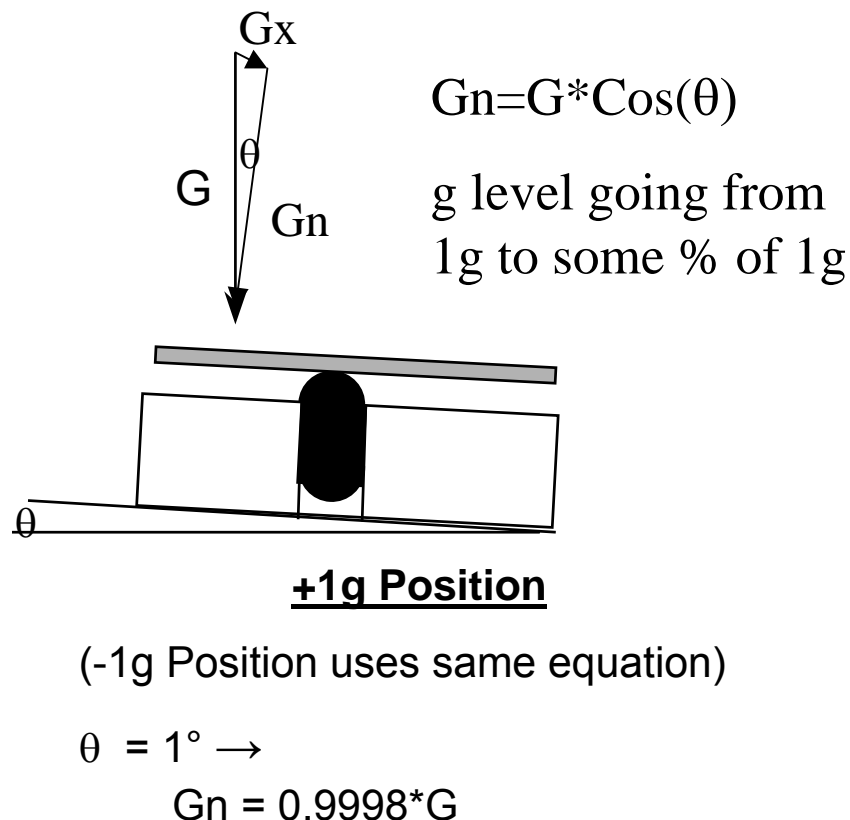
Important Setup Requirements for your CAS Device

- **Rigid Mounting**
 - Bees Wax
 - Double Sided tape
 - Bolt(s)
- **No Loose Wires**
 - Loose wires can create false signals
 - Secure wires firmly to mounting body
- **Weight of Sensor**
 - Should be approximately an order of magnitude less than object being measured
 - Example: CAS = 47g; accelerating object should be more than 470g
- **Don't drop the sensor!**
 - Extreme jarring accelerations can cause permanent errors in device output

Effect of Tilt

- **DC response sensors measure tilt. Mounting errors are therefore significant**
- **a 1° tilt in the 0g position creates an output error equivalent to a 10° tilt in the +1g or -1g positions**
- **0g is the most sensitive to mounting errors**

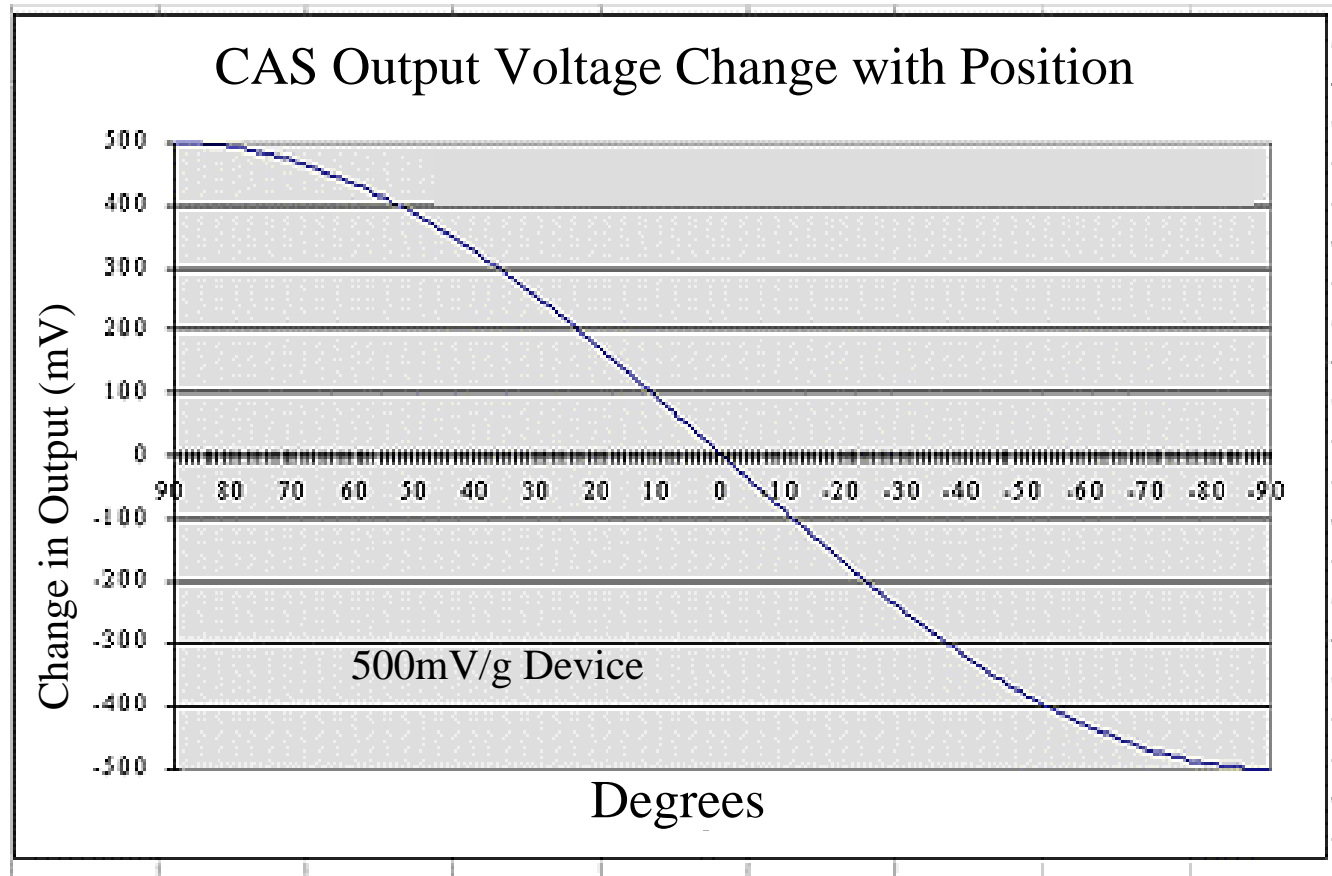
Why is device sensitive to tilt in the 0g orientation?



Conclusion: at 0g orientation, change in 1° tilt causes 57x bigger change in sensor output versus -1g or +1g orientation

Effect of Tilt on DC Accelerometer

1g
Acceleration
↓



Typical Accelerometer Applications

- **Tilt / Roll**
- **Vibration / “Rough-road” detection**
 - Can be used to isolate vibration of mechanical system from outside sources
- **Vehicle skid detection**
 - Often used with systems that deploy “smart” braking to regain control of vehicle
- **Impact detection**
 - To determine the severity of impact, or to log when an impact has occurred
- **Input / feedback for active suspension control systems**
 - Keeps vehicle level

Summary

- **Acceleration is a measure of how fast the speed of something is changing**
- **It is used as an input to control systems**
- **Sensor voltage output should be determined as a percentage of voltage input for consistency**
- **The device is sensitive to tilt in the 0g position**
 - 1° tilt in 0g = 10° of tilt in the +1g and -1g positions

Huge advances in interface modalities are evident and imminent. This panel demonstrates and explores the most interesting, promising, and clever of these modalities, and their integration into exciting multimodal systems.

Because this is a gadget-intensive topic, the panel presents gadgets galore. Input devices that can tell systems where users are looking, the gestures they are making, the direction and content of their sounds and speech, and what and how they are touching. Display devices that image directly onto the retina, high-resolution miniature LCDs, and spatial sound generators. Some of these innovative transducers operate both non-invasively and invisibly. No one should ever have to see a computer. The complexity should be suffused in the world around you.

Panelist perspectives are theoretical and pragmatic, incremental and radical; their work is elegantly inspiring and often delightfully unconventional. All were formerly considered visionaries, but now their visions are achievable, and many industries are paying attention. They are seasoned practitioners with their own viewpoints. All are articulate, and none are shy.

Michael Harris

When users talk about computers, they usually describe the interfaces - because, for most users, the interface is the system. The most powerful force in shaping people's mental model of the nature of the beast is that which they see, feel, and hear. It seemed to take forever for toggle-switch panels to evolve into today's WIMPs, although both are visual/motor-based controls. And switch panels were clearly more haptically satisfying! Now, thanks to exponential increases in commonly available computer power and versatility (and concomitant cost decreases), significant progress in interface modalities and their affordability can be perceived.

While humans are adept at sensory integration and data fusion, computers are far less so. It is clear (and probably has been since Glowflow in 1968) that multimodal interaction is a seminal goal and that achieving it is a formidable challenge. Computational power seems to be catching up with algorithmic understanding.

Interfaces to newborn technology are usually "close to the machine:" early automobiles had spark advance levers, mixture adjustments, hand throttles, choke controls. As automobiles have evolved, their affordability have moved "closer to the user:" speed, stop, reverse. We're tracking a similar evolution in human-computer interaction space. Perhaps interfaces are finally growing up?

Hiroshi Ishii

Tangible Interfaces

People have developed sophisticated skills for sensing and manipulating their physical environments. However, most of these skills are not employed by traditional graphical user interfaces (GUIs). Tangible Bits, our vision of human-computer interaction, seeks to build upon these skills by giving physical form to digital information, seamlessly coupling the dual worlds of bits and atoms.

Guided by the Tangible Bits vision, we are designing "tangible user interfaces," which employ physical objects, surfaces, and spaces as tangible embodiments of digital information. These include foreground interactions with graspable objects and augmented surfaces that exploit the human senses of touch and kinesthesia. We are also exploring background information displays that use "ambient media:" ambient light, sound, air-flow, and water movement. Here, we seek to communicate digitally mediated senses of activity and presence at the periphery of human awareness.

Panelists
Hiroshi Ishii
Massachusetts Institute of Technology

Caleb Chung
Giving Toys, Inc.

Clark Dodsworth
Digital Illusion/Osage Associates

Bill Buxton
Alias|Wavefront, Inc.

interfaces

The goal is to change the “painted bits” of GUIs to “tangible bits,” taking advantage of the richness of multimodal human senses and skills developed through our lifetime of interaction with the physical world.

The musicBottles project presents a tangible interface for interaction with a musical composition. The core concept is that of using glass bottles as containers and controls for digital information. The bottles represent the three performers (violin, cello, and piano) in a classical music trio. Moving and uncorking of the bottles controls the different sound tracks and the patterns of colored light that are rear-projected onto the table’s translucent surface.

Caleb Chung

Interfaces as Pets

Computer-human interface (moving from computers to humans) is difficult to bring off. Better to go the other way: start with humans. Begin with what people want around them: “nurturing,” “fun.” Remember that 80 percent of communication is non-verbal. Don’t try to make interfaces friendly. Instead, start with friendly things and make them smart!

Imagine a personal digital assistant (PDA) that acts and reacts like a “virtual pet.” It has attitude, character – qualities and cues that make you want to interact with it. It has an interesting personality. It has its own agenda. It can become a friend.

Remember when you were six? Your imagination brought your simplest toys to life, and the world around you was limitless. Those interfaces were driven by the human imagination. Open thinking let you make intuitive leaps to invention, expanded your imagination. You were free to create “on top of” your toys, following the most basic human/animal cues.

Toys! Toys have “user friendly” down pat. Toys teach minimalism in physical design. You can’t use expensive parts, four circuit boards, etc. The best toys support and encourage imagination-driven open-ended play – true intelligence!

The best interfaces are transparent, unobtrusively observing humans and responding to their needs. The model “personal assistant” is the valet of 18th century British culture. Just tell it what you want done, without concern for its feelings, but always with a sense of play. You needn’t be precise. And no one has time to learn to speak alien languages.

Send the message that something is alive, and we’ll attribute intelligence to it. This is the true natural interface. And it’s not that difficult, if we but try. Today’s amazingly powerful computers can’t even tell if you’re there with them. Let them observe what humans naturally do, then do that!

Clark Dodsworth

Universal Studios' expansion project, *Islands of Adventure*, recently opened in Orlando. It uses roughly two orders of magnitude more digital infrastructure than the original – a step toward putting ubiquitous digital sensor and effector intelligence into the entire built environment, indoors and outdoors. The guiding notion is that a theme park should be aware of everyone who enters, learn a few facts about them, and then provide a customized user-experience. Every bit of the park, including the landscaping and robotic fauna, should behave or respond interestingly, engagingly to you, and then react with tailored nuance to the next person or family. That notion is not unique to the theme park industry; it's in the strategic plans of the consumer electronics industry, the toy business, the automobile business, and it's important to a few alert individuals in the computer industry.

The task is to create intelligent devices and environments that are designed to adapt to humans and augment the human experience, rather than ones designed to be easily manufactured and then adapted to by humans. In industry, the driving force is competition: parity products need to differentiate themselves. In the computer industry, which holds the biggest rewards for such adaptive interfaces and human-centered design, that driving force is largely quiescent. As intelligence and the software behind it migrate to common objects, the computing world has far more to learn than to teach.

Bill Buxton

If the user is conscious of using a computer, that is a strong indicator of a design failure. Another way of looking at this is to ask: Of the total number of brain cycles expended in performing a task, what percentage are consumed on operational issues compared to content-specific ones? If it is greater than about five percent, we most likely have a failure of design.

It borders on banal to state that we live in an ever-more-complex world, and much of that complexity is due to the previous generation of technology. It seems equally obvious that the basic litmus test of future designs should be: Does it enhance our ability to cope with that complexity? I view well-designed technology as a cognitive (and often social) prosthesis. It is a means to render tractable problems that would otherwise be overwhelming.

From a design perspective, there has been literally no progress since 1982 in the computers used by the majority of the population. And we still live in a climate where it is acceptable for over 90 percent of computer science students to graduate without ever writing a program that is used by another individual, much less be graded on their ability to do so.

Well, the 1980s are over. And the status quo in design and education is just as dated as the music of the Bee Gees, sideburns, and bell bottom trousers. We look back on them with a bit of quaint nostalgia, coupled with horror that we ever found them acceptable. It is time to grow up.

Invisible

TouchLight: An Imaging Touch Screen and Display for Gesture-Based Interaction

Andrew D. Wilson

Microsoft Research

One Microsoft Way

Redmond, WA

awilson@microsoft.com

ABSTRACT

A novel touch screen technology is presented. TouchLight uses simple image processing techniques to combine the output of two video cameras placed behind a semi-transparent plane in front of the user. The resulting image shows objects that are on the plane. This technique is well suited for application with a commercially available projection screen material (DNP HoloScreen) which permits projection onto a transparent sheet of acrylic plastic in normal indoor lighting conditions. The resulting touch screen display system transforms an otherwise normal sheet of acrylic plastic into a high bandwidth input/output surface suitable for gesture-based interaction. Image processing techniques are detailed, and several novel capabilities of the system are outlined.

Categories and Subject Descriptors

H.5.2 [Information Interfaces and Presentation]: User Interfaces—*Input devices and strategies*; I.4.9 [Image Processing and Computer Vision]: Applications

General Terms

Algorithms, Design, Human Factors

Keywords

Computer vision, gesture recognition, computer human interaction, displays, videoconferencing

1. INTRODUCTION

Common touch screen technologies are limited in capability. For example, most are not able to track more than a small number of objects on the screen at a time, and typically they report only the 2D position of the object and no shape information. Partly this is due to superficial limitations of the particular hardware implementation, which in turn are driven by the emphasis on emulating pointer input for common GUI interactions. Typically, today's applications are only able to handle one 2D pointer input.

A number of systems have recently introduced the concept of imaging touch screens, where instead of a small list of discrete points, a full *touch image* is computed, where each 'pixel' of the output image indicates the presence of an object on the touch screen's surface. The utility of the touch image thus computed has been demonstrated in gesture-based interactions for application on wall and table form factors. For example, the DiamondTouch [3] system uses horizontal and vertical rows of electrodes to sense the capacitively coupled touch of the users' hands at electrode intersections.

MetaDesk [13], HoloWall [9] and Designer's Outpost [8] each use video cameras and computer vision techniques to compute a touch image. These systems permit simultaneous video projection and surface sensing by using a diffusing screen material which, from the camera view, only resolves those objects that are on or very near the surface. The touch image produced by these camera-based systems reveals the appearance of the object as it is viewed from behind the surface. Application events may be triggered as the result of image processing techniques applied to the touch image. For example, the appearance or shape of an object may uniquely identify the object to the system and trigger certain application events.

In this paper we introduce the TouchLight system, which uses simple computer vision techniques to compute a touch image on a plane situated between a pair of cameras and the user (see Figures 1 and 2). We demonstrate these techniques in combination with a projection display material which permits the projection of an image onto a transparent sheet of acrylic plastic, and the simultaneous operation of the computer vision processes.

TouchLight goes beyond the previous camera-based systems; by not using a diffusing projection surface, it permits a high resolution touch image. For example, a high resolution image of a paper document may be captured using a high-resolution still camera, or one of the newer high resolution CMOS video cameras.

The absence of a diffuser also permits the cameras to see beyond the display surface, just as they would if placed behind a sheet of glass. This allows a variety of interesting capabilities such as using face recognition techniques to identify the current user, eye-to-eye video conferencing, and other processes which are typically the domain of vision-based perceptual user interfaces.

We describe the overall configuration of TouchLight, and detail the image processing techniques used to compute TouchLight's

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ICMI'04, October 13–15, 2004, State College, Pennsylvania, USA.

Copyright 2004 ACM 1-58113-890-3/04/0010...\$5.00.

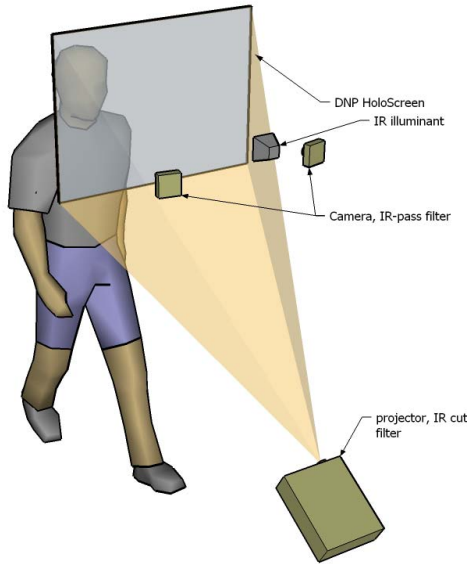


Figure 1 TouchLight physical configuration: DNP HoloScreen with two IR cameras and IR illuminant behind screen.

touch image. Finally, we discuss how TouchLight enables novel gesture-based interaction.

2. TOUCHLIGHT CONFIGURATION

The physical configuration of TouchLight is illustrated in Figure 1 and Figure 2. A pair of commonly available Firewire web cameras are mounted behind the display surface such that each camera can see all four corners of the display. The importance of the distance between the cameras is discussed later.

The DNP HoloScreen material is applied to the rear surface of the acrylic display surface. The HoloScreen is a special refractive holographic film which scatters light from a rear projector when the incident light is at a particular angle. The material is transparent to all other light, and so is suitable for applications where traditional projection display surfaces would be overwhelmed by ambient light. Typical applications include retail storefronts, where ambient light streaming through windows precludes traditional rear-projection screens. Additionally the screen is transparent in the near-infrared range. Per manufacturer's instructions the projector is mounted such that the projected light strikes the display at an angle of about 35 degrees. In a typical vertical, eye-level installation, this configuration does not result in the user looking directly into the "hot spot" of the projector. We note that many projectors are not able to correct for the keystone distortion when the projector is mounted at this extreme angle. In our implementation, we use the NVKeystone digital keystone distortion correction utility that is available on NVidia video cards.

Experience with the HoloScreen material suggests that while the light reflected back from the rear of the screen is significantly less than the light scattered out the front, the projected image will still interfere with the image captured by any visible light-based cameras situated behind the display. In the present work we avoid difficulties with visible light reflections by conducting image-based sensing in the infrared (IR) domain. An IR illuminant is placed behind the display to illuminate the surface evenly in IR

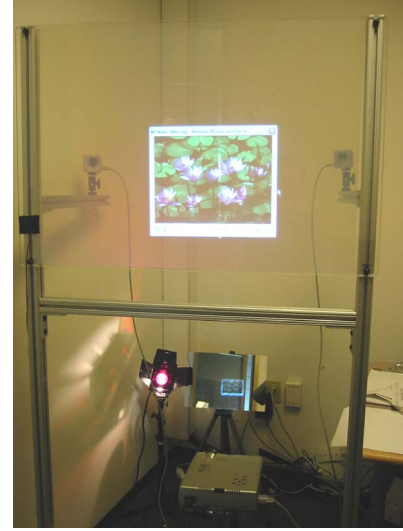


Figure 2 TouchLight prototype displaying a sample graphic.

light. Any IR-cut filters in the stock camera are removed, and an IR-pass filter is applied to the lens. If necessary, an IR-cut filter may be applied to the projector. By restricting the projected light to the visible spectrum, and the sensed light to the IR spectrum, the resulting images from the camera do not include artifacts from projected light reflected backwards from the HoloScreen film.

In future work we plan to investigate the application of anti-reflection films applied to the back and also perhaps the front surface of the display to eliminate reflections from the projector. This would allow the cameras to sense visible light and perhaps eliminate the need for a separate illuminant. Later, we describe applications which benefit from visible-light based sensing.

While for our initial implementation we have chosen to mount the display vertically such that the user may stand, it is also possible to mount the display surface horizontally to make a table. In this case a "short throw" projector such as the NEC WT600 may be desirable.

Finally, a microphone is rigidly attached to the display surface to enable the simple detection of "knocking" on the display. Except for the microphone, there are no wires attached, making TouchLight more robust for public installation.

3. IMAGE PROCESSING

3.1 Introduction

The goal of TouchLight image processing is to compute an image of the objects touching the surface of the display, such as the user's hand. Due to the transparency of the display, each camera view shows the objects on the display and objects beyond the surface of the display, including the background and the rest of the user. With two cameras, the system can determine if a given object is on the display surface or above it. TouchLight image processing acts as a filter to remove objects not on the display surface, producing a *touch image* which shows objects that are on the display surface and is blank everywhere else. A sample output image is illustrated in Figure 3d.

(a) Raw input



(b) Lens distortion correction



(c) Perspective correction



(d) Fused image



Figure 3 TouchLight image processing steps illustrated. Images are captured in an office with normal indoor lighting: (a) raw input from both cameras, (b) input after lens distortion correction, showing display geometry during calibration, (c) input after perspective correction to rectify both views to display, and (d) fused image obtained by multiplying perspective corrected images shows only the objects that are very near the display. Hand on the left is placed flat on the display, hand on the right is slightly cupped, with tips of fingers on the display, and surface of palm above the display.

The touch image is produced by directly combining the output of the two video cameras. Depth information may be computed by relating *binocular disparity*, the change in image position an object undergoes from one view to another view, to the depth of the object in world coordinates. In computer vision there is a long history of exploiting binocular disparity to compute the depth of every point in a scene. Such depth from stereo algorithms are typically computationally intensive, difficult to make robust, and constrain the physical arrangement of the cameras.

Often such general stereo algorithms are applied in scenarios that in the end do not require general depth maps. Here we are interested in the related but easier problem of determining what is located on a particular plane in three dimensions (the display surface) rather than the depth of everything in the scene. A related approach is taken in [14] and [2]. The algorithm detailed here runs in real time (30Hz) on a Pentium 4, operating on 640x480 images.

3.2 Image Rectification

The TouchLight image processing algorithm proceeds by transforming the image from the left camera I_{left} and the image from the right camera I_{right} such that in the transformed images points $I_{left}(x, y)$ and $I_{right}(x, y)$ refer to the same physical point on the display surface.

Secondly, this transform is such that the point (x, y) may be trivially mapped to real world dimensions (i.e., inches) on the display surface. For both criteria, it suffices to find the *homography* from each camera to the display surface, which we obtain during a manual calibration phase.

In the case of using wide angle lenses to make a compact setup, it is important to remove the effects of lens distortion imparted by wide angle lenses. We use the formulation outlined in [7]. Given the lens distortion parameters, we undistort the input image by bilinear interpolation. Sample images are shown in Figure 3b.

During a manual calibration phase, the 4 corners of the display are manually located in each view. This specifies a projective transform bringing pixels in the lens distortion corrected image to display surface coordinates. Together with the lens distortion correction, the projective transform completes the homography from camera view to display coordinates. Sample resulting images are shown in Figure 3c. We note that it is desirable to combine the lens distortion correction and projective transform into a single nonlinear transformation on the image, thus requiring only one resampling of the image. Furthermore it is straightforward to perform this entire calculation on a graphics processing unit (GPU), where the transformation is specified as a mesh.

3.3 Image Fusion

After rectification the same point (x, y) in both I_{left} and I_{right} refer to the same point on the display surface. Thus, if some image feature f is computed on I_{left} and I_{right} , and $f_{left}(x, y) \neq f_{right}(x, y)$, we may conclude that there is no object present at the point (x, y) on the display surface. The touch image mask is computed by performing such pixel-wise

comparisons of the left and right images. This is essentially equivalent to performing standard stereo-based matching where the disparity is constrained to zero, and the rectification process serves to align image rasters.

In the case where a strong IR illuminant is available, and the goal is to identify hands and other IR reflective materials on the display surface, it may suffice to simply pixel-wise multiply the two rectified images. Regions which are bright in both images at the same location will survive multiplication. Sample resulting fused images are shown in Figure 3d. We note that it is possible to implement this image comparison as a pixel shader program running on the GPU.

As with traditional stereo computer vision techniques, it is possible to confuse the image comparison process by presenting a large uniformly textured object at some height above the display. Indeed, the height above the surface at which any bright regions are matched is related to the size of the object and to the *baseline*, the distance between the cameras. For the same size object, larger baselines result in fusion at a smaller height above the surface, consequently allowing a finer distinction as to whether an object is on the display, or just above the display.

Similarly, it is possible to arrange two distinct bright objects above the display surface such that they are erroneously fused as a single object on the surface.

More sophisticated feature matching techniques may be used to make different tradeoffs on robustness and sensitivity. For example, one possibility is to first compute the edge map of the rectified image before multiplying the two images. Figure 4 illustrates the result of applying a Sobel edge filter on the rectified images. Only edges which are present in the same location in both images will survive the multiplication. Thus, large uniform bright objects are less likely to be matched above the surface, since the edges from both views will not overlay one another. In the case of using edges, it is possible and perhaps desirable to

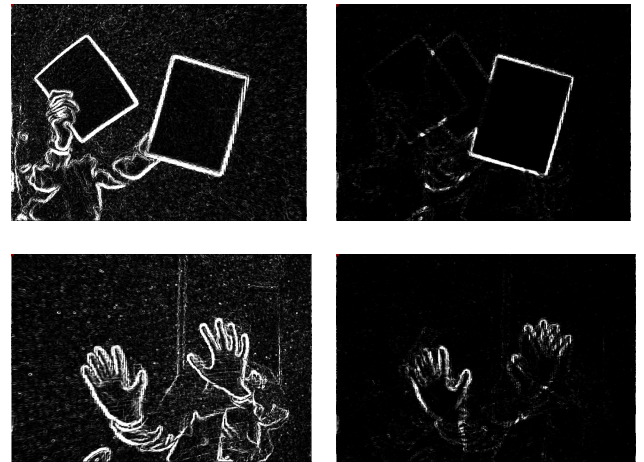


Figure 4 Edge-based image fusion. Top left: Edge extraction of one view's undistorted image (after step c in Figure 3) with sheet of paper a few inches above the display (left) and on the display (right). Top right: product of edge images. Note page above the display is not visible. Bottom: similar images for same images in Figure 3. Hand on the left is placed flat on the display, hand on the right is slightly cupped, with tips of fingers on the display, and surface of palm above the display.

reduce the baseline, resulting in better overall resolution in the rectified images due to a less extreme projective transform. The use of edge images takes advantage of the typical distribution of edges in the scene, in which the accidental alignment of two edges is unlikely.

Similarly, motion magnitude, image differences and other features and combinations of such features may be used, depending on the nature of the objects placed on the surface, the desired robustness, and the nature of subsequent image processing steps.

It should be noted that the touch plane is arbitrarily defined to coincide with the display. It is possible to configure the plane such that it lies at an arbitrary depth above the display. Furthermore, multiple such planes at various depths may be defined depending on the application. Such an arrangement may be used to implement “hover”, as used in pen-based models of interaction. The image rectification and image comparison processes do not require the physical presence of the display. In fact, it is possible to configure TouchLight to operate without the HoloScreen, in which case the “touch” interaction is performed on

an invisible plane in front of the user. In this case, it may be unnecessary to perform imaging in IR.

3.4 Image Normalization

A further image normalization step may be performed to remove effects due to the non-uniformity of the illumination. The current touch image may be normalized pixel-wise by

$$I_{normalized}(x, y) = \frac{I_{product}(x, y) - I_{min}(x, y)}{I_{max}(x, y) - I_{min}(x, y)}$$

where minimum and maximum images I_{min} and I_{max} may be collected by a calibration phase in which the user moves a white piece of paper over the display surface. This normalization step maps the white page to the highest allowable pixel value, corrects for the non-uniformity of the illumination, and also captures any fixed noise patterns due to IR sources and reflections in the environment.

After normalization, other image processing algorithms which are sensitive to absolute gray level values may proceed. For example, binarization and subsequent connected components algorithm, template matching and other computer vision tasks rely on uniform illumination.

3.5 Touch Image Interpretation

Figure 5 shows three different visualizations of the touch image as it is projected back to the user. Figure 5a shows the user’s hand on the surface, which displays both left and right undistorted views composited together (not a simple reflection of two people in front of the display). This shows how an object fuses as it gets closer to the display. Figure 5b shows a hand on the surface, which displays the computed touch image. Note that because of the computed homography, the image of the hand indicated by bright regions is physically aligned with the hand on the screen.

Presently we have only begun exploring the possibilities in interpreting the touch image. Figure 5c shows an interactive drawing program that adds strokes derived from the touch image to a drawing image while using a cycling colormap.

Many traditional computer vision algorithms may be used to derive features relevant to an application. For example, it is straightforward to determine the centroid and moments of multiple objects on the surface, such as hands. One approach is to binarize the touch image, and compute connected components to find distinct objects on the surface (see [5]). Such techniques may also be used to find the moments of object shapes, from which dominant orientation may be determined. Further analysis such as contour analysis for the recognition of specific shapes and barcode processing are possible.

We have implemented a number of mouse emulation algorithms which rely on simple object detection and tracking. In one instance, the topmost object of size larger than some threshold is determined from a binarized version of the touch image. The position of this object determines the mouse position, while a region in the lower left corner of the display functions as a left mouse button: when the user puts their left hand on the region, this is detected as a sufficient number of bright pixels found in the region, and a left mouse button down event is generated. When the bright mass is removed, a button up event is generated. Elaborations on this have been generated, including looking for a

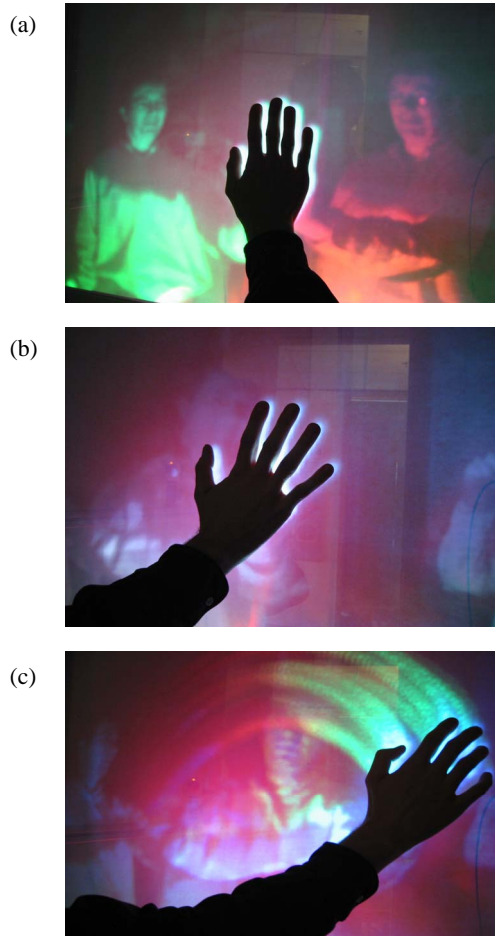


Figure 5 Three different projected visualizations of TouchLight touch image: (a) left undistorted image in the green channel, right undistorted image in red channel. (b) projection of touch image illustrates alignment of touch image with physical display. (c) an interactive drawing application with decaying strokes and cycling colors.

bright mass just to the right of the tracked cursor object to detect left and right button down events when the second mass is near and far from the first, respectively.

Finally, we use a microphone rigidly attached to the display to detect “knocking” events. That is, when the user taps the display with their knuckle or hand, this is detected by finding large peaks in the digitized audio signal. This can be used to simulate clicks, generate “forward” or “next slide” events, and so on. Note that while the tap detector determines that a tap event occurred, the touch image may be used to determine *where* the event occurred. For example, a tap on the left side of the screen may generate a “previous” event, while a tap on the right a “next” event. This contrasts with the tap detector in [10].

4. APPLICATIONS

The unique characteristics of TouchLight lead us to speculate on some possible applications that go beyond emulating traditional touch screen technology. In the following we outline a few possibilities for future exploration.

4.1 Visible Light Surface Scanning

The HoloScreen display material is unique in that it supports video projection and is nearly transparent to IR and visible light. The basic TouchLight system takes advantage of this fact in the placement of the cameras behind the display. This placement provides a good view of the underside of the objects placed on the display surface. The transparency of the display surface may be exploited to create high resolution scans of documents and other objects placed on the display surface.

A high resolution still digital camera or CMOS video camera may be placed behind the display to acquire high resolution images of the objects on the display surface. This camera may capture images in the visible spectrum (no IR-pass filter). In such a configuration it may be beneficial to use the touch image computed from the IR cameras to perform detection and segmentation of objects of interest, and limit the projection of visible light onto the area of interest.

For example, an image processing algorithm may detect the presence of a letter-sized piece of paper on the display surface. The application removes any projected graphics under the presented page to enable a clear visible light view, and triggers the acquisition of a high resolution image of the display surface. The detected position, size and orientation of the page may then be used to automatically crop, straighten and reflect the high resolution scan of the document. Alternatively, the application may project an all-white graphic on the page to clearly illuminate it.

The ability to create high resolution surface scans of documents and other objects may play an important role in business and productivity oriented applications for smart surfaces such as interactive tables and smart whiteboards.

We note that related systems such as the MetaDesk, HoloWall, and Designer’s Outpost all use diffusing projection surfaces to facilitate projection and sensing algorithms. Such diffusing surfaces severely limit the ability of these systems to acquire high resolution imagery of objects on the surface.

4.2 Video Conferencing

The ability to place a camera directly behind the HoloScreen display, and the ability of the TouchLight system to selectively attend to objects on the surface and the scene beyond the surface may enable some interesting video conferencing scenarios.

For example, maintaining direct eye contact is impossible in today’s video conferencing systems, where the camera and the display are not co-axial. It is possible to use a half-silvered mirror to make the camera and display coaxial. This approach has been studied in the context of video conferencing systems in [1] and [6]. The use of a half-silvered mirror has the disadvantages that the brightness of the display and the acquired image is significantly reduced, the setup requires large amounts of space in front of the display, and finally, the configuration imposes restrictions on viewing angle.

An eye-to-eye video conferencing system may be constructed by placing a video camera directly behind the TouchLight display surface. The chief difficulty in constructing such a system is that if the camera used is acquiring IR images so as to avoid artifacts from the projected image, the resulting imagery may not be satisfactory for presentation back to the user. Alternatively, if the camera acquires visible light images, then the presentation must be carefully crafted so that the acquired image does not include any light scattered back from the rear of the display surface. The application of an anti-reflective film on the front and rear of the HoloScreen material may eliminate the back reflection. We also note that it is theoretically possible to use image processing techniques to remove artifacts due to the projection since the system has access to the projected image and the homography from the camera to the display surface is known.

The ability to place a camera behind the screen may have uses beyond eye-to-eye video conferencing. Even with the grayscale IR image returned by TouchLight, it will be possible to determine who is interacting with the display surface by face recognition techniques, determine whether they are looking at the display and possibly even where on the display the user is looking. Such capabilities may be relevant in multi-user and collaborative scenarios. Perhaps uncomfortably, such analysis can be conducted with the cameras completely concealed behind the display surface.

A number of research projects have explored video conferencing displays which are loosely modeled as panes of glass in which two non co-located users are able to see each other manipulate objects rendered on the display. ClearBoard [6] is an early example (see Figure 6). We foresee the applicability of this window metaphor in using TouchLight in video conferencing scenarios. Note that the ability to create high resolution scans outlined in the previous section may be especially valuable in this scenario.

4.3 Minority Report Interfaces

Movies such as *Minority Report* and *The Matrix Reloaded* have popularized the idea of gesture and direct manipulation-based interfaces involving transparent displays. Of the hundreds of people that have seen TouchLight demos, roughly half made unsolicited comparisons of TouchLight to the interaction systems shown in these two movies. The value of the transparency of the displays used in these future visions is debatable. Clearly, the transparency taps into the public’s fascination with holograms, but more mundanely it creates the opportunity for filmmakers to

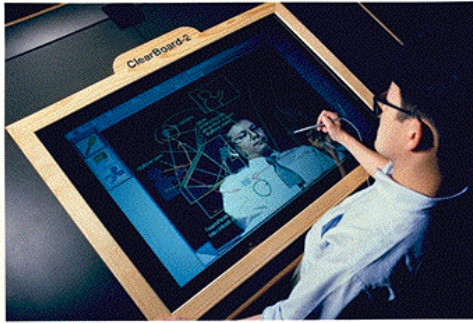


Figure 6 ClearBoard-2 illustrates shared drawing surface and eye contact between remote participants. Image courtesy Hiroshi Ishii and NTT Human Interface Laboratories.

cleanly put the interaction system and the actor's face in the same shot.

Several research projects, however, are taking seriously the gesture-based manipulation of onscreen objects [15] [11] in the style of direct manipulation. For certain classes of interaction, this style of interaction seems to be more natural than the traditional WIMP (windows, icons, menus, pointer) interface. For example, sorting through a stack of photos may be more easily conducted in a direct manipulation framework that allows the use of multiple hands, taking advantage of our own abilities to sort objects into groups or piles [12]. Objects may be rotated in a way that mimics the rotation of a physical piece of paper on a desk. Certain collaborative exercises may benefit from direct manipulation, where each user may easily comprehend the other users' actions. We suspect that direct manipulation frameworks are more readily picked up by novice users, and therefore are suited to quick serendipitous interactions, perhaps at public kiosks, or in short face to face, collaborative meetings. In these situations the overhead in acquiring an input modality may mean the difference between conducting an interaction or not.

4.4 Augmented Reality and Spatial Displays

With the ability to project on a transparent display, TouchLight enables scenarios where projected graphics are overlaid onto imagery from the real world. The application of the HoloScreen material for an augmented reality application is explored in [4], which describes a boom-mounted and instrumented screen and projector system used to overlay graphics onto the real world beyond the screen.

TouchLight raises new possibilities for augmented reality and spatial displays. For example, imagine a retail environment installation where customers are invited to try on virtual articles of clothing while looking at themselves in a TouchLight "mirror". In this scenario, a camera may be placed to synthesize the view the customer would have if they looked into a real mirror. A computer graphics system would composite the clothing onto the view in real time as the customer moves, while TouchLight interaction may allow the user to select various articles of clothing on their mirror image, or interact with buttons alongside their image.

With the touch sensitive capabilities of TouchLight, scenarios inspired by the concept of Alberti's Veil or Leonardo's Window are possible. Alberti's Veil is a technique still used to teach

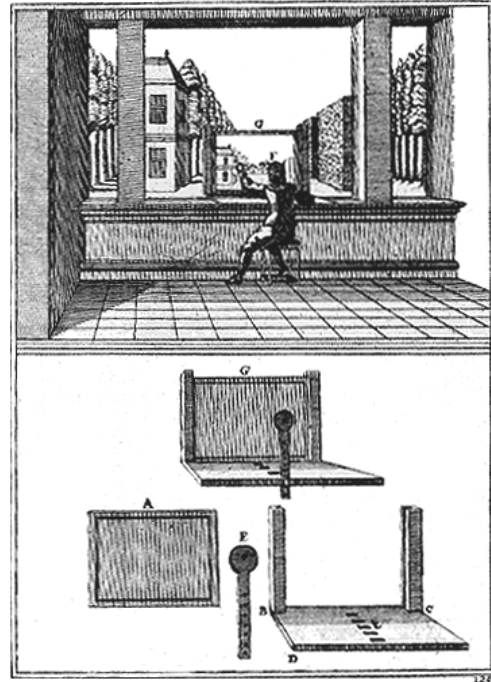


Figure 7 A typical device after Leonardo's window. Such devices were used to teach perspective in architecture. From P. Le Dubreuil, *La Perspective Pratique*, 1649

perspective whereby a scene projected onto a window is traced, with the artist maintaining a stationary viewpoint (see Figure 7). With TouchLight, an artist may trace or modify a visual scene, and with computer vision techniques it is possible to track the face of the user and perhaps detect gaze direction to correct for parallax from the user's point of view to the display in aligning projected graphics with the real world. Many spatial display systems are based on the ability to track the user's face and eyes.

5. CONCLUSION

A novel interactive surface and touch screen technology is presented. TouchLight uses two cameras in combination with a commercially available projection screen technology which allows projection onto an otherwise transparent surface. This arrangement allows for certain novel applications and flexibility which go beyond previous related technologies.

We have presented image processing techniques to produce a touch image useful for many gesture-based and perceptual computing scenarios. A number of applications which take advantage of the unique characteristics of TouchLight have been suggested; we hope to explore some of these in the future.

6. REFERENCES

1. Buxton, B., T. Moran, EuroPARC's Integrated Interactive Intermedia Facility (IIIF): Early Experiences. in *IFIP WG8.4 Conference on Multi-User Interfaces and Applications*, (1990), 11-34.
2. de la Hammette, P., P. Lukowicz, G. Tröster, T. Svoboda, Fingermouse: A Wearable Hand Tracking System. in *Ubicomp 2003: Ubiquitous Computing*, (2002).

3. Dietz, P.H., D. L. Leigh, DiamondTouch: A Multi-User Touch Technology. in *ACM Symposium on User Interface Software and Technology (UIST)*, (2001), 219-226.
4. Ferscha A., M.K., DigiScope: An Invisible Worlds Window. in *Adjunct Proceedings, The Fifth International Conference on Ubiquitous Computing*, (Seattle, 2003), 261-264.
5. Horn, B.K.P. *Robot Vision*. MIT Press, Cambridge, MA, 1986.
6. Ishii, H., M. Kobayashi, ClearBoard: A Seamless Media for Shared Drawing and Conversation with Eye-Contact. in *Conference on Human Factors in Computing Systems (CHI)*, (1992), 525-532.
7. Kang, S.B. Radial Distortion Snakes. *IEICE Transactions on Information and Systems*, E84-D (12). 1603-1611.
8. Klemmer, S.R., M. W. Newman, R. Farrell, M. Bilezikjian, J. A. Landay, The Designer's Output: A Tangible Interface for Collaborative Web Site Design. in *ACM Symposium on User Interface Software and Technology*, (2001), 1-10.
9. Matsushita, N., J. Rekimoto, HoloWall: Designing a Finger, Hand, Body and Object Sensitive Wall. in *ACM Symposium on User Interface Software and Technology (UIST)*, (1997).
10. Paradiso, J.A., C. K. Leo, N. Checka, K. Hsiao, Passive Acoustic Knock Tracking for Interactive Windows. in *ACM Conference on Human Factors in Computing: CHI 2002*, (2002), 732-733.
11. Ringel, M., K. Ryall, C. Shen, C. Forlines, F. Vernier, Release, Rotate, Reorient, Resize: Fluid Techniques for Document Sharing on Multi-User Interactive Tables. in *Short Paper, ACM Conference on Human Factors in Computing Systems*, (2004).
12. Shen, C., F.D. Vernier, C. Forlines, M. Ringel, DiamondSpin: An Extensible Toolkit for Around-the-Table Interaction. in *ACM Conference on Human Factors in Computing Systems (CHI)*, (2004).
13. Ullmer, B., H. Ishii, The metaDESK: Models and Prototypes for Tangible User Interfaces. in *ACM Symposium on User Interface Software and Technology*, (1997), 223-232.
14. Wren, C.R., Y. A. Ivanov, Volumetric Operations with Surface Margins. in *Computer Vision and Pattern Recognition: Technical Sketches*, (2001).
15. Wu, M., R. Balakrishnan, Multi-Finger and Whole Hand Gestural Interaction Techniques for Multi-User Tabletop Displays. in *ACM Symposium on User Interface Software and Technology*, (2003), 193-202.

Frustrated Total Internal Reflection

Becky Urban

Physics Department, The College of Wooster, Wooster, Ohio 44691

May 7, 2002

This experiment tests the theory for frustrated total internal reflection using light in the visible spectrum. A decaying exponential relationship between the intensity of a transmitted light beam and the distance between two media was found. The results are relevant for an undergraduate optics physics course or quantum mechanics course where it is analogous to barrier penetration.

INTRODUCTION

J. C. Bose was examining the wave nature of the radiation of microwaves in 1897. His experiment consisted of a beam of microwaves directed at a right angle asphalt prism. The microwaves were totally internally reflected by the hypotenuse of the prism. However, when a second right angle prism was placed in contact with the first, hypotenuse to hypotenuse, the beam of microwaves passed through. The distance between the two prisms was increased, but kept significantly smaller than the wavelength of the microwave. A portion of the beam was transmitted through the prisms and across the gap between them. This confirmed the wave nature of microwaves.¹

In order for this to occur, the wave passing through the first prism must have penetrated into the air gap between the prisms. When the gap is small enough, the wave is able to pass through the second prism as well. The wave that is penetrating the barrier of air between the two prisms is called the evanescent wave. Carniglia and Mandel found that evanescent waves are not different from homogeneous waves, when the photoelectric emission of a bound charge is under the influence of an evanescent wave.²

Hall studied the experimental and theoretical ideas of the transmitted wave and published it in 1902.³ His method consisted of introducing a third material identical to the first (the prisms in Bose's experiment) and placing it very close to the first, creating a thin area between the identical materials. The thickness of the material between the two identical materials must be on the order of the wavelength of the wave used. The total reflection of light is frustrated. The materials used are assumed to be transparent. Hall studied the distance that the wave penetrated into the barrier with respect to the angle of incidence and the polarization of the incident radiation beam for several different materials. He

found that the intensity of the transmitted light increases as the distance the wave penetrates into the barrier region increases.³ He also observed that the penetrated distance increases as the indices of refraction of the two media decrease.³

The transmission coefficient in terms of the distance between the two identical materials was found using Maxwell's equations. They were verified using centimeter wave radiation and the experiment first used by Bose. Hall's theoretical work was extended by Eichenwald, Foersterling, and Arzelies. They examined the energy flow, and found that the evanescent wave decayed exponentially in the barrier.⁴ With the development of quantum mechanics, barrier penetration was found to be analogous to the frustrated total internal reflection of optics.

THEORY

A beam of light that is incident on a reflective surface at an angle θ_i will be reflected at an angle θ_r according to the Law of Reflection: $\theta_i = \theta_r$. The angles are measured from the normal to the surface. Both the incident and reflected beams of light lie in one plane, the plane of incidence. However, if a beam of light is incident on a surface that is not completely reflective, the beam will "bend" as it crosses the boundary. The light does not actually bend, but its speed changes, resulting in the transmitted light traveling at a different angle, the transmittance angle θ_t . The medium that the plane of incidence lies in has the index of refraction of n_i , and the medium the transmitted plane is in has the index of refraction of n_t . The angle of incidence and the transmitted angle relate to each other by their respective indices of refraction according to Snell's Law, $n_i \sin(\theta_i) = n_t \sin(\theta_t)$.

In the case of internal reflection (where $n_i > n_t$) all the incoming light is reflected back into the incident medium when the incident angle is

greater than or equal to the critical angle, θ_c (which is the incident angle for which the transmitted angle is equal to 90 degrees).

While it does not appear that there is a transmitted wave, it does exist, it just cannot carry energy across the boundary. The intensity, I , of the transmitted light is given by⁵

$$I \propto e^{-\alpha \frac{D}{\lambda_0}}, \quad (1)$$

where

$$\alpha = 4\pi n_i \left[\frac{n_i^2}{n_t^2} \sin^2 \theta_i - 1 \right]^{1/2}, \quad (2)$$

D is the distance between the two media, and λ_0 is the wavelength of light in a vacuum.

So, if the angle of incidence and the indices of refraction of the two media are known, the graph of the intensity of the transmitted light beam versus the ratio of the distance between two media to the wavelength of light would follow an exponential decay. Equation 2 can be used to find the value of α which can be compared to a calculated value from an exponential fit of a graph of the intensity of the transmitted light beam versus the ratio of the distance between the two media to the wavelength of light.

EXPERIMENT

The method used consisted of two flat glass Fabry-Perot mirrors ($n_{\text{glass}}=1.51509$) secured into holders on a Hilger & Watts translation stage with their coated sides facing each other. The glass flats were cleaned thoroughly using lens paper so that there were no debris on the flats. To make sure the flats were parallel to each other, a Melles Griot HeNe laser was used to roughly align them, followed by a Hydrogen-Deuterium source placed at the focal point of a lens. A filter was placed in between the lens and the glass flats, which were aligned so circular fringes were observed. As the flats were moved closer together, the interference fringes (which look like concentric rings getting smaller then disappearing, while other rings appear on the outside) occur. At the point where the fringes stop, the distance between the flats is approximately less than one half the wavelength of light.

A small, right angle prism was attached to the glass disk in the movable translation holder using decahydronaphthalene, a liquid material that has a similar index of reflection as the prism and the glass. The decahydronaphthalene was added drop by drop using a toothpick to the hypotenuse of the prism and then attached by the surface tension of the liquid. The prism and glass flat will then act as a single material with the same index of reflection, so when the light beam passes from

the prism to the glass flat through the liquid material, it will not deflect. In order for the transmitted light beam to exit the glass flat and be observed, a second prism needed to be attached to the glass flat in the stationary holder. See Figure 1. The second prism was the exact same as the first prism, and attached to the glass disk in the same manner.

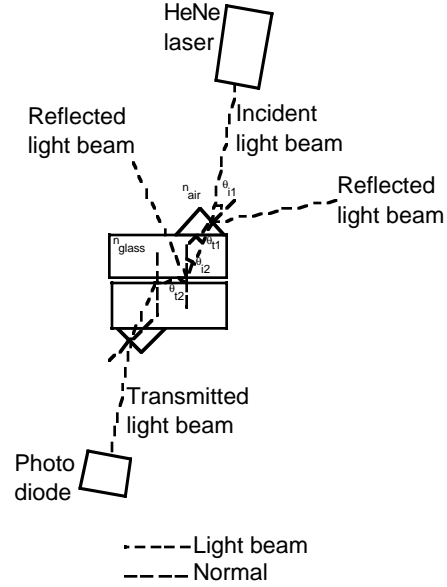


FIG. 1. This was the arrangement of glass flats, prisms, HeNe laser, and photo diode. Some of the light from the HeNe laser is reflected off of the surface of the prism while the rest of it goes into the prism and glass flats. At the interface between the glass flat and the air of the gap, some light is reflected also.

The Melles Griot HeNe laser ($\lambda=633$ nm) was moved more perpendicular to the glass disks, and at an angle of approximately 10 degrees from the normal of the first prism. This resulted in the incident angle (incident on the interface from glass to the air in the gap between the two glass flats) of the light beam inside the glass to be greater than the critical angle. This makes sense because in order for total internal reflection to occur, the angle of incidence must be greater than or equal to the critical angle.

A photo diode connected to a United Detector Technology optometer was used to measure the intensity of the transmitted light beam. The photo diode was positioned so that its screen was perpendicular to the transmitted light beam, making sure that the whole beam was hitting the screen.

The intensity of the transmitted light beam at different distances between the glass flats was measured, starting when the glass flats were touching. Because Fabry-Perot flats were used, the intensity of the light increased, then decreased, then increased again, then decreased again, and so on as the glass flats were moved apart. The

reason for the interference fringes, and therefore the variation in the intensity of the transmitted light beam, was the coating on the glass flats. The coating is partially reflective, so light bounces in between the glass flats, only some of the light getting out each time. In order to compensate for this while taking data, intensity measurements were only taken at the peak intensities, when the beam was at its most intense. While looking at the screen of the photo diode, the distance between the glass flats was increased very slowly until the intensity of the beam was at its greatest. At this time, the maximum intensity reading and the distance between the glass flats were recorded. Then, the distance was again increased very slowly. The intensity of the next maximum fringe and the distance between the flats were recorded. The intensity for the fringes and the distances between the flats were recorded in this manner until there was little change in the maximum intensity of subsequent readings.

DATA AND DISCUSSION

The value for the angle of incidence in the first interface was found using geometry. A meter stick was placed a measured distance away from the first prism, but not as far away as the HeNe laser. The distance between where the light beam entered the prism and the zero point of the meter stick was the measured distance the meter stick was away from the prism. Knowing these distances and the distance between the meter stick and the prism, the angle of incidence was calculated to be 9.53° .

The procedure that was used to take the data necessitated the advancing of the distance between the glass disks at a very slow speed so the advancing could be stopped at the maximum intensity of the light beam hitting screen of the photo diode. Since the maximum comes and goes so quickly, the instant that the maximum occurred, the intensity reading needed to be taken. Instead of watching the photo diode screen to see when the maximum intensity occurred, the intensity meter was watched. This allowed the maximum intensity of the light beam reading to be made for each of the intensity fringes. The second set of data taken verified that this process worked because there were no outlying intensity measurements.

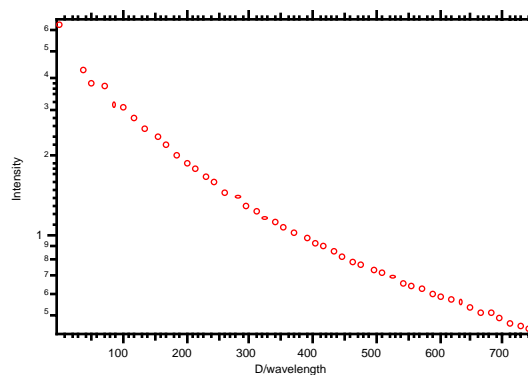


FIG. 2. This is the graph of the second set of

The graph in Figure 2 is the intensity of the transmitted light beam plotted on a log scale on the y-axis and the ratio of the distance, D , between the glass flats to the wavelength of light. The general trend for the graph appears to be a double exponential. If it was a single exponential, then the data would form a linear pattern. There appears to be two different linear patterns for this graph, one ranging from where $D/\text{wavelength}$ is zero to approximately 150 or so. The other linear pattern ranges from where $D/\text{wavelength}$ is about 150 through 750.

It is speculated that one exponential function describes the reflective coating of the Fabry-Perot glass flats while the other describes the intensity of the light beam as a function of the ratio of D divided by the wavelength of light.

Igor Pro 4.01 was unable to fit a satisfactory double exponential curve to this data on its own. So, one of the parameters of the function needed to be set for it. Since one of the exponential functions describes the intensity of light as opposed to the reflective properties of the glass flats' coating, Equation 2 was used to find the value of α . The value of α then relates the intensity of the light beam to the ratio of D divided by the wavelength of light through Equation 1.

The transmitted angle for the first interface (between air and the first prism) was found using Snell's Law, with the angle of incidence and the indices of refraction as stated earlier. Using this angle, the angle of incidence for the second interface (between the first glass flat and the air of the gap) was found to be 51.3° . This angle is necessarily greater than the critical angle, which is 41.3° .

RESULTS

The value for the variable α in Equation 1 was found using Equation 2 with the angle of incidence and the indices of refraction for the second interface (between the first glass flat and the air of the gap) stated earlier. The calculated

value of α ($\alpha = 7.92$) allowed Igor Pro 4.01 to find a double exponential function that fit the graph of the intensity of the transmitted light beam versus the ratio of the distance between the glass flats to the wavelength of light.

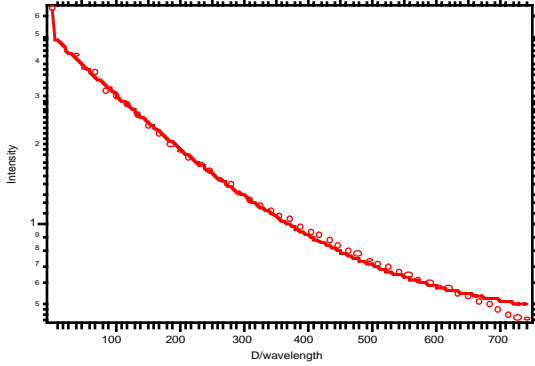


FIG. 3. This is a plot of the second set of data. The ratio of D to the wavelength of light is plotted on the x-axis and the intensity of the transmitted light beam is plotted on a log scale on the y-axis. A double exponential curve was fit to the data:

$$I = 9.57 \times 10^{10} e^{-7.92 \frac{D}{\lambda_0}} + 4.64 e^{-0.006 \frac{D}{\lambda_0}} + 0.431.$$

CONCLUSION

This experiment yielded sufficient data that supported the theory for frustrated total internal reflection using visible light. Good quantitative results were not obtained, but a decaying exponential relationship between the intensity of a transmitted light beam and the distance between two media was found. In order to yield accurate quantitative results, the distance between the two glass flats needed to be measurable at distances less than one wavelength of light. The Hilger & Watts translation stage used only had resolution to one micrometer instead of ten nanometers (which would be a sufficient division of the wavelength of light). The translation stage was, however, able enough to yield a good qualitative picture. Another factor that halted a deeper exploration into the transmission coefficients of the tunneled light beam was the crude accuracy of the optometer, which measured the intensity of light. Quantitative results may also be able to be obtained with glass flats that did not have any reflective coating on their surfaces.

¹Ghose, Partha, *Testing Quantum Mechanics on New Ground*. Cambridge University Press: Cambridge, England, 1999, p 29.

²Carniglia, C.K. and L. Mandel, "Quantization of evanescent electromagnetic waves," *Physical Review D*. 3 (1), 280-296 (1971).

³Hall, E. E., "The penetration of totally reflected light into the rarer medium." *Physical Review*. 15, 73-106 (1902).

⁴Zhu, S., A. W. Yu, D. Hawley, and R. Roy, "Frustrated total internal reflection: A demonstration and review," *American Journal of Physics*. 54 (7), 601-606 (1986).

⁵Hecht, Eugene, *Optics*, 4th ed. Addison Wesley: New York, 2002, p 125.

Synthesis and Control on Large Scale Multi-Touch Sensing Displays

Philip L. Davidson

Jefferson Y. Han

Courant Institute of Mathematical Sciences
New York University
719 Broadway New York, NY 10003
{ philipd, jhan }@mrl.nyu.edu

ABSTRACT

In this paper, we describe our experience in musical interface design for a large scale, high-resolution, multi-touch display surface. We provide an overview of historical and present-day context in multi-touch audio interaction, and describe our approach to analysis of tracked multi-finger, multi-hand data for controlling live audio synthesis.

Keywords

multi-touch, touch, tactile, bi-manual, multi-user, synthesis, dynamic patching

1 INTRODUCTION

The musician's need to manipulate many simultaneous degrees of freedom in audio synthesis has long driven the development of novel interface devices. Touch sensors integrated with graphical display functionality can provide intuitively direct interactivity with richly dynamic context; however they are typically only able to respond to a single point of contact a time, making them quite limiting for musical input. *Multi-touch* sensors on the other hand permit the user fully bi-manual operation as well as chording gestures, offering the potential for great input expression. Such devices also inherently accommodate *multiple* users, which makes them especially useful for larger interaction scenarios such as interactive tables.

These devices have historically been difficult to construct, but we have taken advantage of a new rear-projectable multi-touch sensing technology with unique advantages in scalability and resolution, to create novel musical interfaces for synthesis and control in a large format dynamic workspace.

2 PREVIOUS WORK

2.1 Multi-Touch Interfaces

Boards composed of a plurality of individual controls such as sliders, knobs, buttons, keys, and touchpads, can in a sense be considered multi-touch interfaces. Advanced devices of this class include large arrays of position-sensitive touch sensors such as Buchla's *Thunder* [2], Eaton and Moog's *Multiple-Touch*

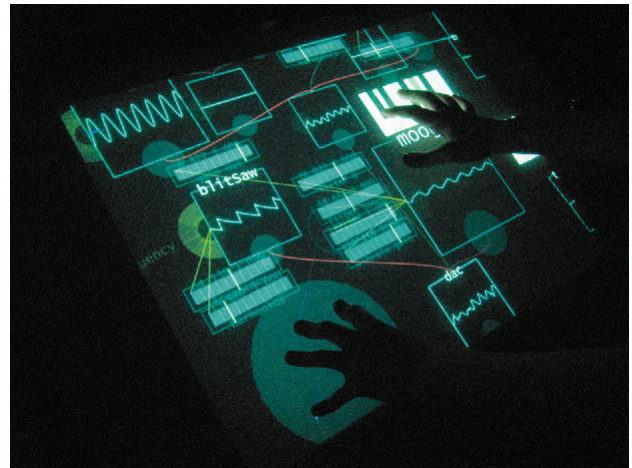


Figure 1: Rear-projected, multi-touch interaction session

Keyboard [7] and the *Continuum Fingerboard* [8]. However, we are more interested in homogeneous interaction surfaces that allow for dynamic contextualization.

Buxton experimented with continuous touch-sensing [22] as well as multi-touch sensing devices for music with the *Fast Multiple-Touch-Sensitive Input Device* [3][14]. This device was an active matrix of capacitive touch sensors, 64×32 in resolution. Instead of integrating it with a display, Buxton utilized cardboard template overlays to partition the interaction surface to provide context, in addition to kinesthetic feedback.

Tactex more recently experimented in the marketplace with a product directly aimed at musicians called the *MTC Express* [23]. This device optically measured the compression of a translucent compressible foam, and though it only had a spatial resolution of 8×9, it has an impressive temporal sampling rate (200Hz) and dynamic range in pressure, making it mostly useful for percussive control.

The recent *Lemur* from JazzMutant [11] is a multi-touch sensor that is tightly integrated with an LCD display. The device is sized for , and functions as a software-configurable controller board. However, the device is low resolution (128×100) and provides no pressure information, limiting the sophistication of the interface widgets that are provided. Furthermore, the system is not open enough to allow access to either the raw sensor data stream or to the raw display itself, limiting its usefulness for the exploration and development of novel interfaces.

All of the systems above have a complexity on the order of the number of tactels, which limits both resolution (though interpolation and other signal processing techniques can mitigate

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

NIME 06, June 4-8, 2006, Paris, France.

Copyright remains with the author(s).



Figure 2: AudioPad, reacTable*, and Lemur

this for a sparse set of contacts) and physical scale, reducing their role in musical performance to a component within a larger system. Other more scalable multi-touch sensing technologies are starting to become available [6][21][26], but these are still difficult/expensive to obtain, and we have not yet seen any reports of their usage in a musical context.

2.2 Tangible Interfaces

Larger scale musical interfaces have also developed around the concept of the manipulation of trackable tangible assets, such as blocks or pucks. These tangible interfaces [10] can accommodate more than one hand and/or more than one user, and take advantage of the user's sense of kinesthesia and skills in three-dimensional spatialization.

The *AudioPad* [19], is a tabletop instrument which utilizes modified Wacom tablet systems to track the position and orientation of a limited number of pucks. This tabletop environment enabled the dynamic control of loops of other synthesis through marking menus, and also allowed the pucks to act as dials and other controllers to vary parameters. Pucks could also be equipped with a pushbutton, which could be regarded as 1-bit pressure sensitivity.

d-touch [5] and the *reacTable** [12] are more recent tabletop instruments based on vision-based tracking of optical fiducials. They track many more pucks without compromising the sensing update rate, and have developed several tangible musical interface paradigms.

We find that these, and other tangible instruments [1][16][17][18] provide an intuitive and approachable environment for musical control, but face challenges as the complexity of the environment increases.

3 SYSTEM OVERVIEW

Through the usage of a scalable high-resolution multi-touch sensing technique, we build a system that encompasses the functionality of both the virtualized controllers possible on multi-touch devices such as *Lemur*, and the space and scale of multi-user patching systems such as *AudioPad* and *reacTable**[13].

The technique is based on *frustrated total internal reflection* [9], implemented in the form factor of a 36"x27" drafting table, at a sensing resolution of ~2mm at 50Hz. It provides full touch image information without any projective ambiguity issues whatsoever. The touch information is true- it accurately discriminates touch from a very slight hover, while also providing pressure information. The sensor image sequence is analyzed and parsed into discrete stroke events and paths with a processing latency

of about 3.5ms on a 3GHz Pentium 4. Measurements including position, velocity, pressure, and image moments are sent to client applications using the lightweight OSC protocol [27] over UDP. The system is notably graphically integrated via *rear*-projection, preventing undesirable occlusion issues.

For our experiments with audio control, we built a simple set of synthesis modules using STK [4], controlled by a modular patching interface.

4. DISCUSSION

4.1 Graphical Context

As Buxton first demonstrated, context is a critical issue for touch interfaces. While we are a few steps beyond cardboard overlays, context for interaction on continuous control surfaces is a challenging problem. Although the pucks used in *AudioPad* and *reacTable** are visually passive, information is projected on and around the puck to provide additional feedback to the user. As such, they are a convenient metaphor for control in contextualizing the surface.

4.2 Basic Gestures

Pucks emphasize our ability to precisely manipulate objects between our fingers. True multi-touch surfaces should provide a similar capacity for manipulation, in contrast to a discrete set of continuous controls. We begin by extending the dextrous manipulation concept to the touch surface by creating regions of the surface that act as virtual puck-like widgets. Touch information captured by each widget is processed together as a single complex gesture. As with pucks, we use the space in and around these controllers for rich visual feedback.

4.3 Interpretation Model

Free from the limitation of the physical world, we can start to extend the metaphor of the basic puck- for instance, the control region associated with a widget can be dynamically resized or reshaped in the course of a performance.

We can also flexibly divide inputs into separate control groups, and selectively constrain degrees of freedom while maintaining a robust handling of under- or overconstrained input cases. As an example, constraining the transformation to rotation and translation is equivalent to the degrees of freedom in a physical puck, while constraint to single-axis translation acts as a slider.

We implemented the more traditional interface widgets such as sliders, knobs, and keys, which the performer can manipulate any set of simultaneously. Additionally, the availability of pressure information allows for more sophisticated revisions of these basic controls. We also use a 'deadband' model [15] to differentiate between tracking and control, permitting the precise acquisition of control elements by the user. Pressure data is also heavily used for more novel controls such as Zsliders [20], as well as control pads which interpret relative pressure values as tilt measurements.

4.3 Complex Gestures

With the input captured from two or more hands, we can start to simulate physical manipulations such as strain, twist, or bending motions. Through this we can consider virtual instruments controlled by simplified physical systems - for example, we could monitor volume of a deformable object to determine the flow rate for a wind controller, or use strain measurements to modify string tension or resonance modes. We are currently exploring

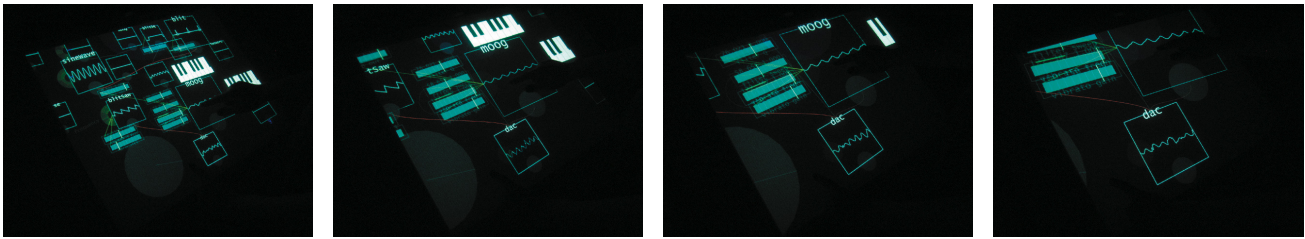


Figure 3: Dynamic workspace- users easily pan/zoom/rotate with a bimanual gesture

the possibilities using a fretboard and plucked string model to produce an autoharp, or koto-like instrument.

4.3 Structural Flexibility

We find that contextualizing manipulation through widgets allows similar precision in parametric control as a physical puck model, and that multi-touch gestures are a natural extension of the control space. Capturing the wide gestural range possible with the hand [24] requires that the sensor accurately track points in close proximity, and control gestures must recognize the limitations of hand geometry as described in [25], to prevent painful or impractical gestures. One advantage to virtualization is that each arrangement can conform to the size and shape of the user's hands, preventing undue stress. As with any continuous control surface, widgets may be adjusted, expanded or repositioned without the synchronizing the location of their physical counterparts. In Figure 3, we show the use of a two-dimensional view manipulator, actuated with a simple two-fingered gesture, allowing the user to pan, zoom, and rotate the workspace and inspect a modular element in detail with no loss of context, giving the performer the ability to manage large workspaces much more effectively.

5 FUTURE DIRECTIONS

There are some limitations in the core implementation that we would like to address that would further increase its usefulness for musical applications. For instance, our current sample rate of 50Hz is good but not great, particularly for percussive input, although this is mitigated by the fact that a large amount of simultaneous information can be updated for each frame. We will be immediately upgrading the system to achieve 120Hz or more.

Also, our current setup provides context only through visual means, but we are definitely looking to be able to provide some degree of haptic feedback as well.

We will continue to explore new and design of new widgets in this new domain. While the table has its advantages over traditional control surfaces, we are primarily interested in controls that take full advantage of the multi-touch data. A uniform control surface also raises the possibility of flexible interfaces - for example, a piano keyboard interface that adjusts spacing based on a user playing a set of prompted chords. In provided a customized scaling of the interface we can adapt to different players to better fit their stature, or to reduce RSI related conditions.

The versatility of the sensor allows for much more interesting form-factors than the console table we have shown here. In particular, for multi-user collaborative setups, we can envision a wider setup where two musicians perform on the same surface, while passing or linking sonic elements in a shared workspace.

Multi-touch sensing is currently an active field in HCI research,

so we stand to harness the fruits of much other work in advancing the intuitiveness, efficiency, and usability of this unique family of interfaces.

6 REFERENCES

- [1] Berry, R., Makino, M., Hikawa, N. and Suzuki, M. 2003. The Augmented Composer Project: The Music Table. In *Proceedings of the 2003 International Symposium on Mixed and Augmented Reality*, Tokyo, Japan, 2003.
- [2] Buchla, D. 1990. Thunder. <http://www.buchla.com/historical/thunder/>
- [3] Buxton, W., Hill, R., and Rowley, P. 1985. Issues and Techniques in Touch-Sensitive Tablet Input. In *Proceedings of the 12th Annual Conference on Computer Graphics and interactive Techniques SIGGRAPH '85*. ACM Press, New York, NY, 215-224.
- [4] Cook, P. R. and G. Scavone. 1999. The Synthesis Toolkit (STK). In *Proceedings of the International Computer Music Conference*. International Computer Music Association, pp. 164-166.
- [5] Costanza, E, Shelley, S. B., and Robinson, J.. Introducing Audio d-touch: A Tangible User Interface for Music Composition and Performance. In *Proceedings of the 2003 International Conference on Digital Audio Effects*, London, UK, September 8-11 2003b.
- [6] Dietz, P. and Leigh, D. 2001. DiamondTouch: a Multi-User Touch Technology. In *Proceedings of the 14th Annual ACM Symposium on User interface Software and Technology* (Orlando, Florida, November 11 - 14, 2001). UIST '01. ACM Press, New York, NY, 219-226.
- [7] Eaton, J. and Moog, R. 2005. Multiple-Touch-Sensitive Keyboard. In *Proceedings of the 2005 International Conference on New Interfaces for Musical Expression (NIME05)*, Vancouver, BC, Canada.
- [8] Haken, L. 2005. Continuum Fingerboard. <http://www.cerloundgroup.org/Continuum/>
- [9] Han, J. Y. 2005. Low-Cost Multi-Touch Sensing through Frustrated Total Internal Reflection. In *Proceedings of the 18th Annual ACM Symposium on User interface Software and Technology* (Seattle, WA, USA, October 23 - 26, 2005). UIST '05. ACM Press, New York, NY, 115-118.
- [10] Ishii, H. and Ullmer, B. 1997. Tangible Bits: Towards Seamless Interfaces between People, Bits and Atoms.

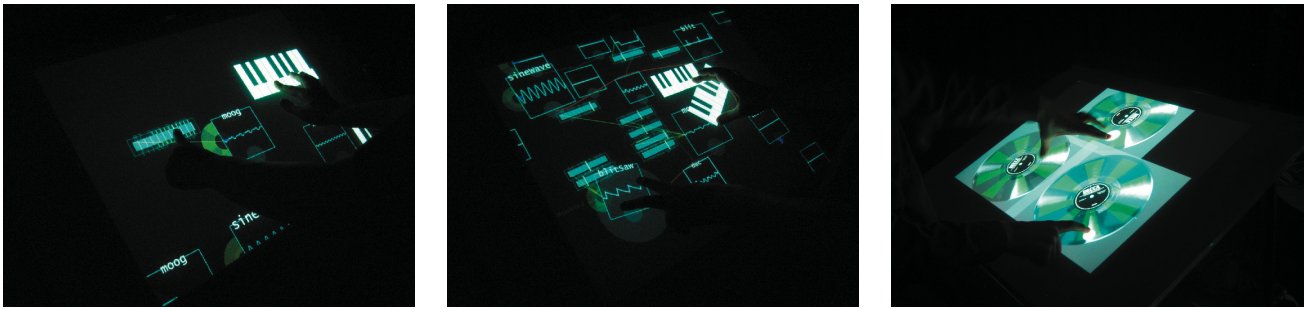


Figure 4: Experiments in multi-touch interfaces

In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, United States, March 22 - 27, 1997). S. Pemberton, Ed. CHI '97. ACM Press, New York, NY, 234-241.

- [11] JazzMutant. 2004. Lemur. <http://www.jazzmutant.com/>
- [12] Jordà, S. & Kaltenbrunner, M. & Geiger, G. & Bencina, R. The reacTable*. In *Proceedings of the International Computer Music Conference (ICMC2005)*, Barcelona (Spain)
- [13] Kaltenbrunner, M. & Geiger, G. & Jordà, S. 2004. Dynamic Patches for Live Musical Performance. In *Proceedings of the 2004 Conference on New Interfaces for Musical Expression (NIME04)*, Hamamatsu, Japan
- [14] Lee, S. K., Buxton, W. and Smith, K. C. 1985. A Multi-Touch Three Dimensional Touch-Sensitive Tablet. In *Proceedings of CHI '85 (April 1985)*, ACM/SIGCHI, NY, 1985, pp. 21-25.
- [15] Miller, T. and Zeleznik, R. 1999. The Design of 3D Haptic Widgets. In *Proceedings of the 1999 Symposium on interactive 3D Graphics* (Atlanta, Georgia, United States, April 26 - 29, 1999). SI3D '99. ACM Press, New York, NY, 97-102.
- [16] Newton-Dunn, H., Nakao, H., and Gibson, J. 2003. Block Jam: A Tangible Interface for Interactive Music. In *Proceedings of the 2003 International Conference on New Interfaces for Musical Expression*, Montreal, Canada, May 22-24 2003.
- [17] Paradiso, J. and Hsiao, K. 1999. A New Continuous Multimodal Musical Controller using Wireless Magnetic Tags. In *Proceedings of the 1999 International Computer Music Conference*, pages 24-27, Beijing, China, October 22-28 1999.
- [18] Paradiso, J. A. 2002. Several Sensor Approaches that Retrofit Large Surfaces for Interactivity. Presented at the *UbiComp 2002 Workshop on Collaboration with Interactive Walls and Tables*, Gothenburg, Sweden, September 29, 2002.
- [19] Patten, J., Ishii, H., Hines, J., and Pangaro, G. 2001. SenseTable: A Wireless Object Tracking Platform for Tangible User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, United States). CHI '01. ACM Press, New York, NY, 253-260.
- [20] Ramos, G. and Balakrishnan, R. 2005. Zliding: Fluid Zooming and Sliding for High Precision Parameter Manipulation. In *Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology* (Seattle, WA, USA, October 23 - 26, 2005). UIST '05. ACM Press, New York, NY, 143-152.
- [21] Rekimoto, J. 2002. SmartSkin: An Infrastructure for Freehand Manipulation on Interactive Surfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing Our World, Changing Ourselves* (Minneapolis, Minnesota, USA, April 20 - 25, 2002). CHI '02. ACM Press, New York, NY, 113-120.
- [22] Sasaki, L., Fedorkow, G., Buxton, W., Retterath, C., Smith, K.C. 1981. A Touch Sensitive Input Device. In *Proceedings of the 5th International Conference on Computer Music*. North Texas State University, Denton Texas.
- [23] Tactex. Smart Fabric Technology. <http://www.tactex.com>
- [24] Waisvisz, M. 1984. The Hands. <http://www.crackle.org/TheHands.html>
- [25] Wessel, D., Wright, M., and Schott, J. 2002. Intimate Musical Control of Computers with a Variety of Controllers and Gesture Mapping Metaphors. In *Proceedings of the 2002 Conference on New Instruments for Musical Expression (NIME-02)*, Dublin, Ireland, May 24-26, 2002
- [26] Westerman, W., Elias, J. G., and Hedge, A. 2001. Multi-Touch: A New Tactile 2-D Gesture Interface for Human-Computer Interaction. In *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting*, Vol. 1, pp. 632-636. 2001.
- [27] Wright, M. 2003. OpenSound Control: State of the Art 2003. In *Proceedings of the 2003 Conference on New Interfaces for Musical Expression*, Montreal, Canada, 2003.

In press: A. Tucker, (Ed.) *CRC Handbook of Computer Science*, CRC Press, Boca Raton, FL.

Neural Networks

Michael I. Jordan
Massachusetts Institute of Technology

Christopher M. Bishop
Aston University

January 18, 1996

1 Introduction

Within the broad scope of the study of artificial intelligence, research in neural networks is characterized by a particular focus on pattern recognition and pattern generation. Many neural network methods can be viewed as generalizations of classical pattern-oriented techniques in statistics and the engineering areas of signal processing, system identification and control theory. As in these parent disciplines, the notion of “pattern” in neural network research is essentially probabilistic and numerical. Neural network methods have had their greatest impact in problems where statistical issues dominate and where data are easily obtained.

A neural network is first and foremost a graph, with patterns represented in terms of numerical values attached to the nodes of the graph, and transformations between patterns achieved via simple message-passing algorithms. Many neural network architectures, however, are also statistical processors, characterized by making particular probabilistic assumptions about data. As we will see, this conjunction of graphical algorithms and probability theory is not unique to neural networks, but characterizes a wider family of probabilistic systems in the form of chains, trees, and networks that are currently being studied throughout AI [Spiegelhalter, et al., 1993].

Neural networks have found a wide range of applications, the majority of which are associated with problems in pattern recognition and control theory. In this context, neural networks can best be viewed as a class of algorithms for statistical modeling and prediction. Based on a source of *training data*, the aim is to produce a statistical model of the process from which the data are generated, so as to allow the best predictions to be made for new data. We shall find it convenient to distinguish three broad types of statistical modeling problem, which we shall call *density estimation*, *classification* and *regression*.

For density estimation problems (also referred to as *unsupervised learning* problems), the goal is to model the unconditional distribution of data described by some vector \mathbf{x} . A practical example of the application of density estimation involves the interpretation of X-ray images (mammograms) used for breast cancer screening [Tarassenko, 1995]. In this case the training vectors \mathbf{x} form a sample taken from normal (non-cancerous) images, and a network model is used to build a representation of the density $p(\mathbf{x})$. When a new input vector \mathbf{x}' is presented to the system, a high value for $p(\mathbf{x}')$ indicates a normal image while a low value indicates a novel input which might be characteristic of an abnormality. This is used to label regions of images which are unusual, for further examination by an experienced clinician.

For classification and regression problems (often referred to as *supervised learning* problems), we need to distinguish between *input* variables, which we again denote by \mathbf{x} , and *target* variables which we denote by the vector \mathbf{t} . Classification problems require that each input vector \mathbf{x} be assigned to one of C classes $\mathcal{C}_1, \dots, \mathcal{C}_C$, in which case the target variables represent class labels. As an example, consider the problem of recognizing handwritten digits [LeCun, et al., 1989]. In this case the input vector would be some (pre-processed) image of the digit, and the network would have ten outputs, one for each digit, which can be used to assign input vectors to the appropriate class (as discussed in Section 2).

Regression problems involve estimating the values of continuous variables. For example, neural networks have been used as part of the control system for adaptive optics telescopes [Sandler, et

al., 1991]. The network input \mathbf{x} consists of one in-focus and one de-focused image of a star and the output \mathbf{t} consists of a set of coefficients that describe the phase distortion due to atmospheric turbulence. These output values are then used to make real-time adjustments of the multiple mirror segments to cancel the atmospheric distortion.

Classification and regression problems can also be viewed as special cases of density estimation. The most general and complete description of the data is given by the probability distribution function $p(\mathbf{x}, \mathbf{t})$ in the joint input-target space. However, the usual goal is to be able to make good predictions for the target variables when presented with new values of the inputs. In this case it is convenient to decompose the joint distribution in the form:

$$p(\mathbf{x}, \mathbf{t}) = p(\mathbf{t}|\mathbf{x})p(\mathbf{x}) \quad (1)$$

and to consider only the conditional distribution $p(\mathbf{t}|\mathbf{x})$, in other words the distribution of \mathbf{t} *given* the value of \mathbf{x} . Thus classification and regression involve the estimation of *conditional* densities, a problem which has its own idiosyncracies.

The organization of the chapter is as follows. In Section 2 we present examples of network representations of unconditional and conditional densities. In Section 3 we discuss the problem of adjusting the parameters of these networks to fit them to data. This problem has a number of practical aspects, including the choice of optimization procedure and the method used to control network complexity. We then discuss a broader perspective on probabilistic network models in Section 4. The final section presents further information and pointers to the literature.

2 Representation

In this section we describe a selection of neural network architectures that have been proposed as representations for unconditional and conditional densities. After a brief discussion of density estimation, we discuss classification and regression, beginning with simple models that illustrate the fundamental ideas and then progressing to more complex architectures. We focus here on representational issues, postponing the problem of learning from data until the following section.

2.1 Density estimation

We begin with a brief discussion of density estimation, utilizing the Gaussian mixture model as an illustrative model. We return to more complex density estimation techniques later in the chapter.

Although density estimation can be the main goal of a learning system, as in the diagnosis example mentioned in the introduction, density estimation models arise more often as components of the solution to a more general classification or regression problem. To return to Eq. 1, note that the joint density is composed of $p(\mathbf{t}|\mathbf{x})$, to be handled by classification or regression models, and $p(\mathbf{x})$, the (unconditional) input density. There are several reasons for wanting to form an explicit model of the input density. First, real-life data sets often have missing components in the input vector. Having a model of the density allows the missing components to be “filled in” in an intelligent way. This can be useful both for training and for prediction [cf. Bishop, 1995]. Second, as we see in Eq. 1, a model of $p(\mathbf{x})$ makes possible an estimate of the joint probability $p(\mathbf{x}, \mathbf{t})$. Thus

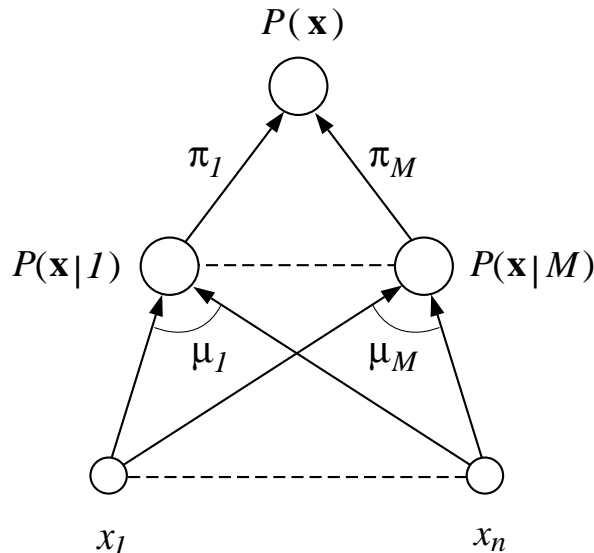


Figure 1: A network representation of a Gaussian mixture distribution. The input pattern \mathbf{x} is represented by numerical values associated with the input nodes in the lower level. Each link has a weight μ_{ij} , which is the j^{th} component of the mean vector for the i^{th} Gaussian. The i^{th} intermediate node contains the covariance matrix Σ_i and calculates the Gaussian conditional probability $p(\mathbf{x}|i, \mu_i, \Sigma_i)$. These probabilities are weighted by the mixing proportions π_i and the output node calculates the weighted sum $p(\mathbf{x}) = \sum_i \pi_i p(\mathbf{x}|i, \mu_i, \Sigma_i)$.

in turn provides us with the necessary information to estimate the “inverse” conditional density $p(\mathbf{x}|\mathbf{t})$. The calculation of such inverses is important for applications in control and optimization.

A general and flexible approach to density estimation is to treat the density as being composed of a set of M simpler densities. This approach involves modeling the observed data as a sample from a *mixture density*:

$$p(\mathbf{x}|\mathbf{w}) = \sum_{i=1}^M \pi_i p(\mathbf{x}|i, \mathbf{w}_i), \quad (2)$$

where the π_i are constants known as *mixing proportions*, and the $p(\mathbf{x}|i, \mathbf{w}_i)$ are the *component densities*, generally taken to be from a simple parametric family. A common choice of component density is the multivariate Gaussian, in which case the parameters \mathbf{w}_i are the means and covariance matrices of each of the components. By varying the means and covariances to place and orient the Gaussians appropriately, a wide variety of high-dimensional, multi-modal data can be modeled. This approach to density estimation is essentially a probabilistic form of clustering.

Gaussian mixtures have a representation as a network diagram as shown in Figure 1. The utility of such network representations will become clearer as we proceed; for now, it suffices to note that not only mixture models, but also a wide variety of other classical statistical models for density estimation are representable as simple networks with one or more layers of adaptive weights. These methods include *principal component analysis*, *canonical correlation analysis*, *kernel density*

estimation and factor analysis [Anderson, 1984].

2.2 Linear regression and linear discriminants

Regression models and classification models both focus on the conditional density $p(\mathbf{t}|\mathbf{x})$. They differ in that in regression the target vector \mathbf{t} is a real-valued vector, whereas in classification \mathbf{t} takes its values from a discrete set representing the class labels.

The simplest probabilistic model for regression is one in which \mathbf{t} is viewed as the sum of an underlying deterministic function $f(\mathbf{x})$ and a Gaussian random variable ϵ :

$$\mathbf{t} = f(\mathbf{x}) + \epsilon. \quad (3)$$

If ϵ has zero mean, as is commonly assumed, $f(\mathbf{x})$ then becomes the *conditional mean* $E(\mathbf{t}|\mathbf{x})$. It is this function that is the focus of most regression modeling. Of course, the conditional mean describes only the first moment of the conditional distribution, and, as we discuss in a later section, a good regression model will also generally report information about the second moment.

In a linear regression model the conditional mean is a linear function of \mathbf{x} : $E(\mathbf{t}|\mathbf{x}) = W\mathbf{x}$, for a fixed matrix W . Linear regression has a straightforward representation as a network diagram in which the j^{th} input unit represents the j^{th} component of the input vector x_j , each output unit i takes the weighted sum of the input values, and the weight w_{ij} is placed on the link between the j^{th} input unit and the i^{th} output unit.

The conditional mean is also an important function in classification problems, but most of the focus in classification is on a different function known as a *discriminant function*. To see how this function arises and to relate it to the conditional mean, we consider a simple two-class problem in which the target is a simple binary scalar that we now denote by t . The conditional mean $E(t|\mathbf{x})$ is equal to the probability that t equals one, and this latter probability can be expanded via Bayes rule:

$$p(t = 1|\mathbf{x}) = \frac{p(\mathbf{x}|t = 1)p(t = 1)}{p(\mathbf{x})} \quad (4)$$

The density $p(t|\mathbf{x})$ in this equation is referred to as the *posterior probability* of the class given the input, and the density $p(\mathbf{x}|t)$ is referred to as the *class-conditional density*. Continuing the derivation, we expand the denominator and (with some foresight) introduce an exponential:

$$\begin{aligned} p(t = 1|\mathbf{x}) &= \frac{p(\mathbf{x}|t = 1)p(t = 1)}{p(\mathbf{x}|t = 1)p(t = 1) + p(\mathbf{x}|t = 0)p(t = 0)} \\ &= \frac{1}{1 + \exp \left\{ -\ln \left[\frac{p(\mathbf{x}|t=1)}{p(\mathbf{x}|t=0)} \right] - \ln \left[\frac{p(t=1)}{p(t=0)} \right] \right\}} \end{aligned} \quad (5)$$

We see that the posterior probability can be written in the form of the *logistic function*:

$$y = \frac{1}{1 + e^{-z}}, \quad (6)$$

where z is a function of the likelihood ratio $p(\mathbf{x}|t = 1)/p(\mathbf{x}|t = 0)$, and the prior ratio $p(t = 1)/p(t = 0)$. This is a useful representation of the posterior probability if z turns out to be simple.

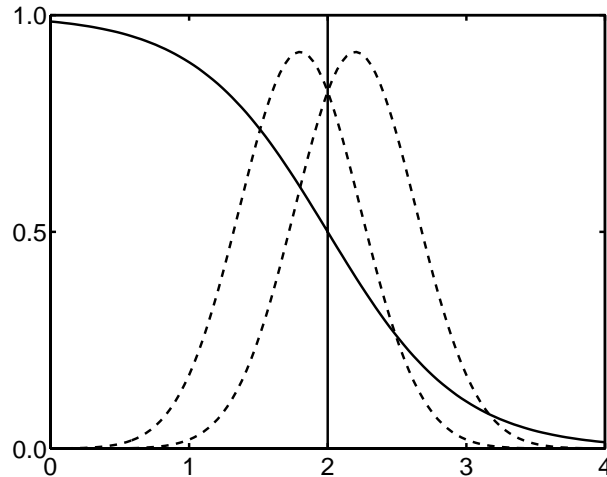


Figure 2: This shows the Gaussian class-conditional densities $p(x|\mathcal{C}_1)$ (dashed curves) for a two-class problem in one dimension, together with the corresponding posterior probability $p(\mathcal{C}_1|x)$ (solid curve) which takes the form of a logistic sigmoid. The vertical line shows the decision boundary for $y = 0.5$ which coincides with the point at which the two density curves cross.

It is easily verified that if the class conditional densities are multivariate Gaussians with identical covariance matrices, then z is a linear function of \mathbf{x} : $z = \mathbf{w}^T \mathbf{x} + w_0$. Moreover this representation is appropriate for any distribution in a broad class of densities known as the exponential family (which includes the Gaussian, the Poisson, the gamma, the binomial, and many other densities). All of the densities in this family can be put in the following form:

$$g(\mathbf{x}; \theta, \phi) = \exp\{(\theta^T \mathbf{x} - b(\theta))/a(\phi) + c(\mathbf{x}, \phi)\}, \quad (7)$$

where θ is the *location parameter*, and ϕ is the *scale parameter*. Substituting this general form in Eq. 5, where θ is allowed to vary between the classes and ϕ is assumed to be constant between classes, we see that z is in all cases a linear function. Thus the choice of a linear-logistic model is rather robust.

The geometry of the two-class problem is shown in Figure 2, which shows Gaussian class-conditional densities, and suggests the logistic form of the posterior probability.

The function z in our analysis is an example of a discriminant function. In general a discriminant function is any function that can be used to decide on class membership (Duda and Hart, 1972); our analysis has produced a particular form of discriminant function that is an intermediate step in the calculation of a posterior probability. Note that if we set $z = 0$, from the form of the logistic function we obtain a probability of 0.5, which shows that $z = 0$ is a *decision boundary* between the two classes.

The discriminant function that we found for exponential family densities is linear under the given conditions on ϕ . In more general situations, in which the class-conditional densities are more complex than a single exponential family density, the posterior probability will not be well

characterized by the linear-logistic form. Nonetheless it is still useful to retain the logistic function and focus on *nonlinear* representations for the function z . This is the approach taken within the neural network field.

To summarize, we have identified two functions that are important for regression and classification, respectively: the conditional mean and the discriminant function. These are the two functions that are of concern for simple linear models and, as we now discuss, for more complex nonlinear models as well.

2.3 Nonlinear regression and nonlinear classification

The linear regression and linear discriminant functions introduced in the previous section have the merit of simplicity, but are severely restricted in their representational capabilities. A convenient way to see this is to consider the geometrical interpretation of these models. When viewed in the d -dimensional \mathbf{x} -space, the linear regression function $\mathbf{w}^T \mathbf{x} + w_0$ is constant on hyper-planes which are orthogonal to the vector \mathbf{w} . For many practical applications we need to consider much more general classes of function. We therefore seek representations for nonlinear mappings which can approximate any given mapping to arbitrary accuracy. One way to achieve this is to transform the original \mathbf{x} using a set of M nonlinear functions $\phi_j(\mathbf{x})$ where $j = 1, \dots, M$, and then to form a linear combination of these functions, so that:

$$y_k(\mathbf{x}) = \sum_j w_{kj} \phi_j(\mathbf{x}). \quad (8)$$

For a sufficiently large value of M , and for a suitable choice of the $\phi_j(\mathbf{x})$, such a model has the desired ‘universal approximation’ properties. A familiar example, for the case of 1-dimensional input spaces, is the simple polynomial, for which the $\phi_j(\mathbf{x})$ are simply successive powers of x and the w ’s are the polynomial coefficients. Models of the form in Eq. 8 have the property that they can be expressed as network diagrams in which there is a *single* layer of adaptive weights.

There are a variety of families of functions in one dimension that can approximate any continuous function to arbitrary accuracy. There is, however, an important issue which must be addressed, called the *curse of dimensionality*. If, for example, we consider an M^{th} -order polynomial then the number of independent coefficients grows as d^M [Bishop, 1995]. For a typical medium-scale application with, say, 30 inputs a fourth-order polynomial (which is still quite restricted in its representational capability) would have over 46,000 adjustable parameters. As we shall see in Section 3.3 in order to achieve good generalization it is important to have more data points than adaptive parameters in the model, and this is a serious problem for methods that have a power law or exponential growth in the number of parameters.

A solution to the problem lies in the fact that, for most real-world data sets, there are strong (often nonlinear) correlations between the input variables such that the data does not uniformly fill the input space but is effectively confined to a sub-space whose dimensionality is called the *intrinsic dimensionality* of the data. We can take advantage of this phenomenon by considering again a model of the form in Eq. 8 but in which the basis functions $\phi_j(\mathbf{x})$ are *adaptive* so that they themselves contain weight parameters whose values can be adjusted in the light of the observed data set. Different models result from different choices for the basis functions, and here we consider

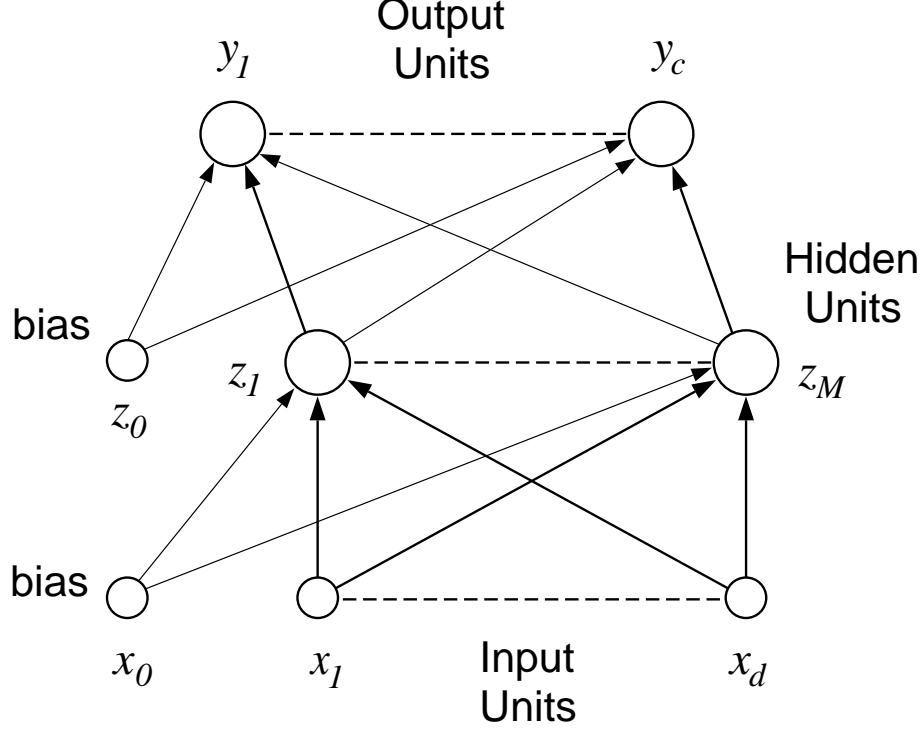


Figure 3: An example of a feed-forward network having two layers of adaptive weights. The bias parameters in the first layer are shown as weights from an extra input having a fixed value of $x_0 = 1$. Similarly, the bias parameters in the second layer are shown as weights from an extra hidden unit, with activation again fixed at $z_0 = 1$.

the two most common examples. The first of these is called the *multilayer perceptron* (MLP) and is obtained by choosing the basis functions to be given by linear-logistic functions (Eq. 6). This leads to a multivariate nonlinear function that can be expressed in the form:

$$y_k(\mathbf{x}) = \sum_{j=1}^M w_{kj} g \left(\sum_{i=1}^d w_{ji} x_i + w_{j0} \right) + w_{k0}. \quad (9)$$

Here w_{j0} and w_{k0} are *bias* parameters, and the basis functions are called *hidden units*. The function $g(\cdot)$ is the logistic sigmoid function of Eq. 6. This can also be represented as a network diagram as in Figure 3. Such a model is able to take account of the intrinsic dimensionality of the data because the first-layer weights w_{ji} can adapt and hence orient the surfaces along which the basis function response is constant. It has been demonstrated that models of this form can approximate to arbitrary accuracy any continuous function, defined on a compact domain, provided the number M of hidden units is sufficiently large. The MLP model can be extended by considering several successive layers of weights. Note that the use of nonlinear activation functions is crucial, since if $g(\cdot)$ in Eq. 9 were replaced by the identity, the network would reduce to several successive linear transformations which would itself be linear.

The second common network model is obtained by choosing the basis functions $\phi_j(\mathbf{x})$ in Eq. 8 to be functions of the radial variable $\mathbf{x} - \boldsymbol{\mu}_j$ where $\boldsymbol{\mu}_j$ is the *center* of the j th basis function, which gives rise to the *radial basis function* (RBF) network model. The most common example uses Gaussians of the form:

$$\phi_j(\mathbf{x}) = \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\mathbf{x} - \boldsymbol{\mu}_j) \right\}. \quad (10)$$

Here both the mean vector $\boldsymbol{\mu}_j$ and the covariance matrix $\boldsymbol{\Sigma}_j$ are considered to be adaptive parameters. The curse of dimensionality is alleviated because the basis functions can be positioned and oriented in input space such as to overlay the regions of high data density and hence to capture the nonlinear correlations between input variables. Indeed, a common approach to training an RBF network is to use a two-stage procedure [Bishop, 1995]. In the first stage the basis function parameters are determined using the input data alone, which corresponds to a density estimation problem using a mixture model in which the component densities are given by the basis functions $\phi_j(\mathbf{x})$. In the second stage the basis function parameters are frozen and the second-layer weights w_{kj} are found by standard least-squares optimization procedures.

2.4 Decision trees

MLP and RBF networks are often contrasted in terms of the support of the basis functions that compose them. MLP networks are often referred to as “global,” given that linear-logistic basis functions are bounded away from zero over a significant fraction of the input space. Accordingly, in an MLP, each input vector generally gives rise to a distributed pattern over the hidden units. RBF networks, on the other hand, are referred to as “local,” due to the fact that their Gaussian basis functions typically have support over a local region of the input space. It is important to note, however, that local support does not necessarily mean non-overlapping support; indeed, there is nothing in the RBF model that prefers basis functions that have non-overlapping support. A third class of model that does focus on basis functions with non-overlapping support is the *decision tree* model [Breiman, et al., 1984]. A decision tree is a regression or classification model that can be viewed as asking a sequence of questions about the input vector. Each question is implemented as a linear discriminant, and a sequence of questions can be viewed as a recursive partitioning of the input space. All inputs that arrive at a particular leaf of the tree define a polyhedral region in the input space. The collection of such regions can be viewed as a set of basis functions. Associated with each basis function is an output value which (ideally) is close to the average value of the conditional mean (for regression) or discriminant function (for classification; a majority vote is also used). Thus the decision tree output can be written as a weighted sum of basis functions in the same manner as a layered network.

As this discussion suggests, decision trees and MLP/RBF neural networks are best viewed as being different points along the continuum of models having overlapping or non-overlapping basis functions. Indeed, as we show in the following section, decision trees can be treated probabilistically as mixture models, and in the mixture approach the sharp discriminant function boundaries of classical decision trees become smoothed, yielding partially-overlapping basis functions.

There are tradeoffs associated with the continuum of degree-of-overlap—in particular, non-overlapping basis functions are generally viewed as being easier to interpret, and better able to

reject noisy input variables that carry little information about the output. Overlapping basis functions are often viewed as yielding lower variance predictions and as being more robust.

2.5 General mixture models

The use of mixture models is not restricted to density estimation; rather, the mixture approach can be used quite generally to build complex models out of simple parts. To illustrate, let us consider using mixture models to model a conditional density in the context of a regression or classification problem. A mixture model in this setting is referred to as a “mixtures of experts” model [Jacobs, et al., 1991].

Suppose that we have at our disposal an elemental conditional model $p(\mathbf{t}|\mathbf{x}, \mathbf{w})$. Consider a situation in which the conditional mean or discriminant exhibits variation on a local scale that is a good match to our elemental model, but the variation differs in different regions of the input space. We could use a more complex network to try to capture this global variation; alternatively we might wish to combine local variants of our elemental models in some manner. This can be achieved by defining the following probabilistic mixture:

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}) = \sum_{i=1}^M p(i|\mathbf{x}, \mathbf{v}) p(\mathbf{t}|\mathbf{x}, i, \mathbf{w}_i). \quad (11)$$

Comparing this mixture to the unconditional mixture defined earlier (Eq. 2), we see that both the mixing proportions and the component densities are now conditional densities dependent on the input vector \mathbf{x} . The former dependence is particularly important—we now view the mixing proportion $p(i|\mathbf{x}, \mathbf{v})$ as providing a probabilistic device for choosing different elemental models (“experts”) in different regions of the input space. A learning algorithm that chooses values for the parameters \mathbf{v} as well as the values for the parameters \mathbf{w}_i can be viewed as attempting to find both a good partition of the input space and a good fit to the local models within that partition.

This approach can be extended recursively by considering mixtures of models where each model may itself be a mixture model [Jordan and Jacobs, 1994]. Such a recursion can be viewed as providing a probabilistic interpretation for the decision trees discussed in the previous section. We view the decisions in the decision tree as forming a recursive set of probabilistic selections among a set of models. The total probability of a target \mathbf{t} given an input \mathbf{x} is the sum across all paths down the tree:

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}) = \sum_{i=1}^M p(i|\mathbf{x}, \mathbf{u}) \sum_{j=1}^M p(j|\mathbf{x}, i, \mathbf{v}_i) \cdots p(\mathbf{t}|\mathbf{x}, i, j, \dots, \mathbf{w}_{ij\dots}), \quad (12)$$

where i and j are the decisions made at the first level and second level of the tree, respectively, and $p(\mathbf{t}|\mathbf{x}, i, j, \dots, \mathbf{w}_{ij\dots})$ is the elemental model at the leaf of the tree defined by the sequence of decisions. This probabilistic model is a conditional hierarchical mixture. Finding parameter values \mathbf{u} , \mathbf{v}_i , etc. to fit this model to data can be viewed as finding a nested set of partitions of the input space and fitting a set of local models within the partition.

The mixture model approach can be viewed as a special case of a general methodology known as *learning by committee*. Bishop [1995] provides a discussion of committees; we will also meet them in the section on Bayesian methods later in the chapter.

3 Learning from Data

The previous section has provided a selection of models to choose from; we now face the problem of matching these models to data. In principle the problem is straightforward: given a family of models of interest we attempt to find out how probable each of these models is in the light of the data. We can then select the most probable model (a selection rule known as *maximum a posteriori* or *MAP* estimation), or we can select some highly probable subset of models, weighted by their probability (an approach that we discuss below in the section on Bayesian methods). In practice there are a number of problems to solve, beginning with the specification of the family of models of interest. In the simplest case, in which the family can be described as a fixed structure with varying parameters (e.g., the class of feedforward MLP's with a fixed number of hidden units), the learning problem is essentially one of *parameter estimation*. If on the other hand the family is not easily viewed as a fixed parametric family (e.g., feedforward MLP's with variable number of hidden units), then we must solve the *model selection* problem.

In this section we discuss the parameter estimation problem. The goal will be to find MAP estimates of the parameters by maximizing the probability of the parameters given the data \mathcal{D} . We compute this probability using Bayes rule:

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}, \quad (13)$$

where we see that to calculate MAP estimates we must maximize the expression in the numerator (the denominator does not depend on \mathbf{w}). Equivalently we can minimize the negative logarithm of the numerator. We thus define the following *cost function* $J(\mathbf{w})$:

$$J(\mathbf{w}) = -\ln p(\mathcal{D}|\mathbf{w}) - \ln p(\mathbf{w}), \quad (14)$$

which we wish to minimize with respect to the parameters \mathbf{w} . The first term in this cost function is a (negative) log likelihood. If we assume that the elements in the training set \mathcal{D} are conditionally independent of each other given the parameters, then the likelihood factorizes into a product form. For density estimation we have:

$$p(\mathcal{D}|\mathbf{w}) = \prod_{n=1}^N p(\mathbf{x}_n|\mathbf{w}) \quad (15)$$

and for classification and regression we have:

$$p(\mathcal{D}|\mathbf{w}) = \prod_{n=1}^N p(\mathbf{t}_n|\mathbf{x}_n, \mathbf{w}). \quad (16)$$

In both cases this yields a log likelihood which is the sum of the log probabilities for each individual data point. For the remainder of this section we will assume this additive form; moreover, we will assume that the log prior probability of the parameters is uniform across the parameters and drop the second term. Thus we focus on *maximum likelihood* (ML) estimation, where we choose parameter values \mathbf{w}_{ML} that maximize $\ln p(\mathcal{D}|\mathbf{w})$.

3.1 Likelihood-based cost functions

Regression, classification and density estimation make different probabilistic assumptions about the form of the data and therefore require different cost functions.

Eq. 3 defines a probabilistic model for regression. The model is a conditional density for the targets \mathbf{t} in which the targets are distributed as Gaussian random variables (assuming Gaussian errors ϵ) with mean values $f(\mathbf{x})$. We now write the conditional mean as $f(\mathbf{x}, \mathbf{w})$ to make explicit the dependence on the parameters \mathbf{w} . Given the training set $\mathcal{D} = \{\mathbf{x}_n, \mathbf{t}_n\}_{n=1}^N$, and given our assumption that the targets \mathbf{t}_n are sampled independently (given the inputs \mathbf{x}_n and the parameters \mathbf{w}), we obtain:

$$J(\mathbf{w}) = \frac{1}{2} \sum_n \|\mathbf{t}_n - f(\mathbf{x}_n, \mathbf{w})\|^2, \quad (17)$$

where we have assumed an identity covariance matrix and dropped those terms that do not depend on the parameters. This cost function is the standard least squares cost function which is traditionally used in neural network training for real-valued targets. Minimization of this cost function is typically achieved via some form of gradient optimization, as we discuss in the following section.

Classification problems differ from regression problems in the use of discrete-valued targets, and the likelihood accordingly takes a different form. For binary classification the Bernoulli probability model $p(t|\mathbf{x}, \mathbf{w}) = y^t(1-y)^{1-t}$ is natural, where we use y to denote the probability $p(t=1|\mathbf{x}, \mathbf{w})$. This model yields the following log likelihood:

$$J(\mathbf{w}) = - \sum_n [t_n \ln y_n + (1 - t_n) \ln(1 - y_n)], \quad (18)$$

which is known as the *cross entropy* function. It can be minimized using the same generic optimization procedures as are used for least squares.

For multi-way classification problems in which there are C categories, where $C > 2$, the multinomial distribution is natural. Define \mathbf{t}_n such that its elements $t_{n,i}$ are one or zero according to whether the n^{th} data point belongs to the i^{th} category, and define $y_{n,i}$ to be the network's estimate of the posterior probability of category i for data point n ; i.e., $y_{n,i} \equiv p(t_{n,i} = 1|\mathbf{x}_n, \mathbf{w})$. Given these definitions we obtain the following cost function:

$$J(\mathbf{w}) = - \sum_n \sum_i t_{n,i} \ln y_{n,i}, \quad (19)$$

which again has the form of a cross entropy.

We now turn to density estimation as exemplified by Gaussian mixture modeling. The probabilistic model in this case is that given in Eq. 2. Assuming Gaussian component densities with arbitrary covariance matrices, we obtain the following cost function:

$$J(\mathbf{w}) = - \sum_n \ln \sum_i \pi_i \frac{1}{|\boldsymbol{\Sigma}_i|^{1/2}} \exp \left\{ -\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_i) \right\}, \quad (20)$$

where the parameters \mathbf{w} are the collection of mean vectors $\boldsymbol{\mu}_i$, the covariance matrices $\boldsymbol{\Sigma}_i$, and the mixing proportions π_i . A similar cost function arises for the generalized mixture models (cf. Eq. 12).

3.2 Gradients of the cost function

Once we have defined a probabilistic model, obtained a cost function and found an efficient procedure for calculating the gradient of the cost function, the problem can be handed off to an optimization routine. Before discussing optimization procedures, however, it is useful to examine the form that the gradient takes for the examples that we have discussed in the previous two sections.

The i^{th} output unit in a layered network is endowed with a rule for combining the activations of units in earlier layers, yielding a quantity that we denote by z_i , and a function that converts z_i into the output y_i . For regression problems, we assume linear output units such that $y_i = z_i$. For binary classification problems, our earlier discussion showed that a natural output function is the logistic: $y_i = 1/(1 + e^{-z_i})$. For multi-way classification, it is possible to generalize the derivation of the logistic function to obtain an analogous representation for the multi-way posterior probabilities known as the *softmax function* [cf. Bishop, 1995]:

$$y_i = \frac{e^{z_i}}{\sum_k e^{z_k}}, \quad (21)$$

where y_i represents the posterior probability of category i .

If we now consider the gradient of $J(\mathbf{w})$ with respect to z_i , it turns out that we obtain a single canonical expression of the following form:

$$\frac{\partial J}{\partial \mathbf{w}} = \sum_i (t_i - y_i) \frac{\partial z_i}{\partial \mathbf{w}}. \quad (22)$$

As discussed by [Rumelhart, et al. 1995], this form for the gradient is predicted from the theory of Generalized Linear Models [McCullagh and Nelder, 1983], where it is shown that the linear, logistic, and softmax functions are (inverse) *canonical links* for the Gaussian, Bernoulli, and multinomial distributions, respectively. Canonical links can be found for all of the distributions in the exponential family, thus providing a solid statistical foundation for handling a wide variety of data formats at the output layer of a network, including counts, time intervals and rates.

The gradient of the cost function for mixture models has an interesting interpretation. Taking the partial derivative of $J(\mathbf{w})$ in Eq. 20 with respect to $\boldsymbol{\mu}_i$, we find:

$$\frac{\partial J}{\partial \boldsymbol{\mu}_i} = \sum_n h_{n,i} \boldsymbol{\Sigma}_i (\mathbf{x}_n - \boldsymbol{\mu}_i), \quad (23)$$

where $h_{n,i}$ is defined as follows:

$$h_{n,i} = \frac{\pi_i |\boldsymbol{\Sigma}_i|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_i)\}}{\sum_k \pi_k |\boldsymbol{\Sigma}_k|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)\}}. \quad (24)$$

When summed over i , the quantity $h_{n,i}$ sums to one, and is often viewed as the “responsibility” or “credit” assigned to the i^{th} component for the n^{th} data point. Indeed, interpreting Eq. 24 using Bayes rule shows that $h_{n,i}$ is the posterior probability that the n^{th} data point is generated by the

i^{th} component Gaussian. A learning algorithm based on this gradient will move the i^{th} mean μ_i toward the data point \mathbf{x}_n , with the effective step size proportional to $h_{n,i}$.

The gradient for a mixture model will always take the form of a weighted sum of the gradients associated with the component models, where the weights are the posterior probabilities associated with each of the components. The key computational issue is whether these posterior weights can be computed efficiently. For Gaussian mixture models, the calculation (Eq. 24) is clearly efficient. For decision trees there are a set of posterior weights associated with each of the nodes in the tree, and a recursion is available that computes the posterior probabilities in an upward sweep [Jordan and Jacobs, 1994]. Mixture models in the form of a chain are known as hidden Markov models, and the calculation of the relevant posterior probabilities is performed via an efficient algorithm known as the Baum-Welch algorithm.

For general layered network structures, a generic algorithm known as “backpropagation” is available to calculate gradient vectors [Rumelhart, et al., 1986]. Backpropagation is essentially the chain rule of calculus realized as a graphical algorithm. As applied to layered networks it provides a simple and efficient method that calculates a gradient in $O(W)$ time per training pattern, where W is the number of weights.

3.3 Optimization algorithms

By introducing the principle of maximum likelihood in Section 1, we have expressed the problem of learning in neural networks in terms of the minimization of a cost function $J(\mathbf{w})$ which depends on a vector \mathbf{w} of adaptive parameters. An important aspect of this problem is that the gradient vector $\nabla_{\mathbf{w}}J$ can be evaluated efficiently (for example by backpropagation). Gradient-based minimization is a standard problem in unconstrained nonlinear optimization, for which many powerful techniques have been developed over the years. Such algorithms generally start by making an initial guess for the parameter vector \mathbf{w} and then iteratively updating the vector in a sequence of steps:

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} + \Delta \mathbf{w}^{(\tau)} \quad (25)$$

where τ denotes the step number. The initial parameter vector $\mathbf{w}^{(0)}$ is often chosen at random, and the final vector represents a minimum of the cost function at which the gradient vanishes. Due to the nonlinear nature of neural network models, the cost function is generally a highly complicated function of the parameters, and may possess many such minima. Different algorithms differ in how the update $\Delta \mathbf{w}^{(\tau)}$ is computed.

The simplest such algorithm is called *gradient descent* and involves a parameter update which is proportional to the negative of the cost function gradient $\Delta = -\eta \nabla E$ where η is a fixed constant called the learning rate. It should be stressed that gradient descent is a particularly inefficient optimization algorithm. Various modifications have been proposed, such as the inclusion of a *momentum* term, to try to improve its performance. In fact much more powerful algorithms are readily available, as described in standard textbooks such as [Fletcher, 1987]. Two of the best known are called *conjugate gradients* and *quasi-Newton* (or *variable metric*) methods. For the particular case of a sum-of-squares cost function, the *Levenberg-Marquardt* algorithm can also be very effective. Software implementations of these algorithms are widely available.

The algorithms discussed so far are called *batch* since they involve using the whole data set for each evaluation of the cost function or its gradient. There is also a *stochastic* or *on-line* version of gradient descent in which, for each parameter update, the cost function gradient is evaluated using just one of the training vectors at a time (which are then cycled either in order or in a random sequence). While this approach fails to make use of the power of sophisticated methods such as conjugate gradients, it can prove effective for very large data sets, particularly if there is significant redundancy in the data.

3.4 Hessian matrices, error bars and pruning

After a set of weights have been found for a neural network using an optimization procedure, it is often useful to examine second-order properties of the fitted network as captured in the Hessian matrix $H = \partial^2 J / \partial \mathbf{w} \partial \mathbf{w}^T$. Efficient algorithms have been developed to compute the Hessian matrix in time $O(W^2)$ [Bishop, 1995]. As in the case of the calculation of the gradient by backpropagation, these algorithms are based on recursive message passing in the network.

One important use of the Hessian matrix lies in the calculation of error bars on the outputs of a network. If we approximate the cost function locally as a quadratic function of the weights (an approximation which is equivalent to making a Gaussian approximation for the log likelihood), then the estimated variance of the i^{th} output y_i can be shown to be:

$$\hat{\sigma}_{y_i}^2 = \left(\frac{\partial y_i}{\partial \mathbf{w}} \right)^T H^{-1} \left(\frac{\partial y_i}{\partial \mathbf{w}} \right), \quad (26)$$

where the gradient vector $\partial y_i / \partial \mathbf{w}$ can be calculated via backpropagation.

The Hessian matrix is also useful in pruning algorithms. A pruning algorithm deletes weights from a fitted network to yield a simpler network that may outperform a more complex, overfitted network (see below), and may be easier to interpret. In this setting, the Hessian is used to approximate the increase in the cost function due to the deletion of a weight. A variety of such pruning algorithms are available [cf. Bishop, 1995].

3.5 Complexity control

In previous sections we have introduced a variety of models for representing probability distributions, we have shown how the parameters of the models can be optimized by maximizing the likelihood function, and we have outlined a number of powerful algorithms for performing this minimization. Before we can apply this framework in practice there is one more issue we need to address, which is that of model complexity. Consider the case of a mixture model given by Eq. 2. The number of input variables will be determined by the particular problem at hand. However, the number M of component densities has yet to be specified. Clearly if M is too small the model will be insufficiently flexible and we will obtain a poor representation of the true density. What is not so obvious is that if M is too large we can also obtain poor results. This effect is known as *overfitting* and arises because we have a data set of finite size. It is illustrated using a simple example of mixture density estimation in Figure 4. Here a set of 100 data points in one dimension has been generated from a distribution consisting of a mixture of two Gaussians (shown by the

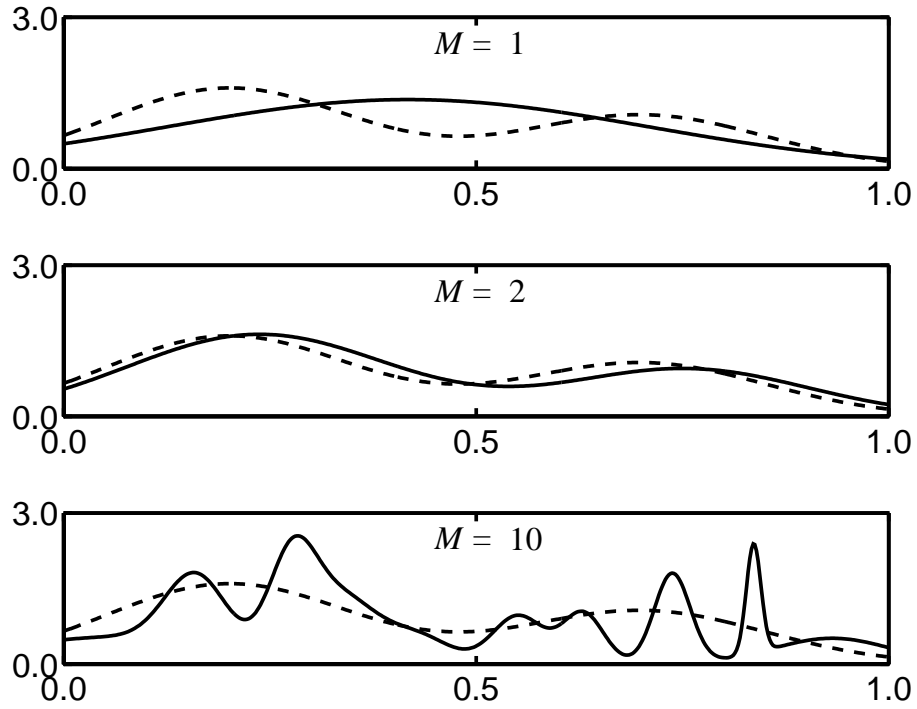


Figure 4: Effects of model complexity illustrated by modeling a mixture of two Gaussians (shown by the dashed curves) using a mixture of M Gaussians (shown by the solid curves). The results are obtained for 20 cycles of EM.

dashed curves). This data set has then been fitted by a mixture of M Gaussians by use of the EM algorithm. We see that a model with 1 component ($M = 1$) gives a poor representation of the true distribution from which the data was generated, and in particular is unable to capture the bimodal aspect. For $M = 2$ the model gives a good fit, as we expect since the data was itself generated from a two-component Gaussian mixture. However, increasing the number of components to $M = 10$ gives a poorer fit, even though this model contains the simpler models as special cases.

The problem is a very fundamental one and is associated with the fact that we are trying to infer an entire distribution function from a finite number of data points, which is necessarily an ill-posed problem. In regression for example there are infinitely many functions which will give a perfect fit to the finite number of data points. If the data are noisy, however, the best generalization will be obtained for a function which does not fit the data perfectly but which captures the underlying function from which the data were generated. By increasing the flexibility of the model we are able to obtain ever better fits to the training data, and this is reflected in a steadily increasing value for the likelihood function at its maximum. Our goal is to model the true underlying density function from which the data was generated since this allows us to make the best predictions for new data. We see that the best approximation to this density occurs for an intermediate value of M .

The same issue arises in connection with nonlinear regression and classification problems. For example, the number M of hidden units in an MLP network controls the model complexity and must be optimized to give the best generalization. In a practical application we can train a variety of different models having different complexity, and compare their generalization performance using an independent validation set, and then select the model with the best generalization. In fact the process of optimizing the complexity using a validation set can lead to some partial overfitting to the validation data itself, and so the final performance of the selected model should be confirmed using a third independent data set called a *test* set.

Some theoretical insight into the problem of overfitting can be obtained by decomposing the error into the sum of *bias* and *variance* terms [Geman, et al., 1992]. A model which is too inflexible is unable to represent the true structure in the underlying density function and this gives rise to a high bias. Conversely a model which is too flexible becomes tuned to the specific details of the particular data set and gives a high variance. The best generalization is obtained from the optimum trade-off of bias against variance.

As we have already remarked, the problem of inferring an entire distribution function from a finite data set is fundamentally ill-posed since there are infinitely many solutions. The problem only becomes well-posed when some additional constraint is imposed. This constraint might be that we model the data using a network having a limited number of hidden units. Within the range of functions which this model can represent there is then a unique function which best fits the data. Implicitly we are assuming that the underlying density function from which the data were drawn is relatively smooth. Instead of limiting the number of parameters in the model, we can encourage smoothness more directly using the technique of *regularization*. This involves adding a penalty term Ω to the original cost function J to give a total cost function \tilde{J} of the form:

$$\tilde{J} = J + \nu\Omega \tag{27}$$

where ν is called a regularization coefficient. The network parameters are determined by minimizing \tilde{J} , and the value of ν controls the degree of influence of the penalty term Ω . In practice Ω is typically

chosen to encourage smooth functions. The simplest example is called *weight decay* and consists of the sum of the squares of all the adaptive parameters in the model:

$$\Omega = \sum_i w_i^2 \quad (28)$$

Consider the effect of such a term on the MLP function (Eq. 9). If the weights take very small values then the network outputs become approximately linear functions of the inputs (since the sigmoidal function is approximately linear for small values of its argument). The value of ν in Eq. 27 controls the effective complexity of the model, so that for large ν the model is over-smoothed (corresponding to high bias) while for small ν the model can overfit (corresponding to high variance). We can therefore consider a network with a relatively large number of hidden units and control the effective complexity by changing ν . In practice, a suitable value for ν can be found by seeking the value which gives the best performance on a validation set.

The weight decay regularizer (Eq. 28) is simple to implement but suffers from a number of limitations. Regularizers used in practice may be more sophisticated and may contain multiple regularization coefficients [Neal, 1994].

Regularization methods can be justified within a general theoretical framework known as *structural risk minimization* [Vapnik, 1995]. Structural risk minimization provides a quantitative measure of complexity known as the *VC dimension*. The theory shows that the VC dimension predicts the difference between performance on a training set and performance on a test set; thus, the sum of log likelihood and (some function of) VC dimension provides a measure of generalization performance. This motivates regularization methods (Eq. 27) and provides some insight into possible forms for the regularizer Ω .

3.6 Bayesian viewpoint

In earlier sections we discussed network training in terms of the minimization of a cost function derived from the principle of maximum a posteriori or maximum likelihood estimation. This approach can be seen as a particular approximation to a more fundamental, and more powerful, framework based on Bayesian statistics. In the maximum likelihood approach the weights \mathbf{w} are set to a specific value \mathbf{w}_{ML} determined by minimization of a cost function. However, we know that there will typically be other minima of the cost function which might give equally good results. Also, weight values close to \mathbf{w}_{ML} should give results which are not too different from those of the maximum likelihood weights themselves.

These effects are handled in a natural way in the Bayesian viewpoint, which describes the weights not in terms of a specific set of values, but in terms of a probability distribution over all possible values. As discussed earlier (cf. Eq. 13), once we observe the training data set \mathcal{D} we can compute the corresponding *posterior* distribution using Bayes' theorem, based on a *prior* distribution function $p(\mathbf{w})$ (which will typically be very broad), and a *likelihood* function $p(\mathcal{D}|\mathbf{w})$:

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}. \quad (29)$$

The likelihood function will typically be very small except for values of \mathbf{w} for which the network function is reasonably consistent with the data. Thus the posterior distribution $p(\mathbf{w}|\mathcal{D})$ will be much more sharply peaked than the prior distribution $p(\mathbf{w})$ (and will typically have multiple maxima). The quantity we are interested in is the predicted distribution of target values \mathbf{t} for a new input vector \mathbf{x} once we have observed the data set \mathcal{D} . This can be expressed as an integration over the posterior distribution of weights of the form:

$$p(\mathbf{t}|\mathbf{x}, \mathcal{D}) = \int p(\mathbf{t}|\mathbf{x}, \mathbf{w})p(\mathbf{w}|\mathcal{D}) d\mathbf{w} \quad (30)$$

where $p(\mathbf{t}|\mathbf{x}, \mathbf{w})$ is the conditional probability model discussed in the introduction.

If we suppose that the posterior distribution $p(\mathbf{w}|\mathcal{D})$ is sharply peaked around a single most-probable value \mathbf{w}_{MP} , then we can write Eq. 30 in the form:

$$p(\mathbf{t}|\mathbf{x}, \mathcal{D}) \simeq p(\mathbf{t}|\mathbf{x}, \mathbf{w}_{\text{MP}}) \int p(\mathbf{w}|\mathcal{D}) d\mathbf{w} \quad (31)$$

$$= p(\mathbf{t}|\mathbf{x}, \mathbf{w}_{\text{MP}}) \quad (32)$$

and so predictions can be made by fixing the weights to their most probable values. We can find the most probable weights by maximizing the posterior distribution, or equivalently by minimizing its negative logarithm. Using Eq. 29, we see that \mathbf{w}_{MP} is determined by minimizing a regularized cost function of the form in Eq. 27 in which the negative log of the prior $-\ln p(\mathbf{w})$ represents the regularizer $\nu\Omega$. For example, if the prior consists of a zero-mean Gaussian with variance ν^{-1} then we obtain the weight-decay regularizer of Eq. 28.

The posterior distribution will become sharply peaked when the size of the data set is large compared to the number of parameters in the network. For data sets of limited size, however, the posterior distribution has a finite width and this adds to the uncertainty in the predictions for \mathbf{t} which can be expressed in terms of error bars. Bayesian error bars can be evaluated using a local Gaussian approximation to the posterior distribution [MacKay, 1992]. The presence of multiple maxima in the posterior distribution also contributes to the uncertainties in predictions. The capability to assess these uncertainties can play a crucial role in practical applications.

The Bayesian approach can also deal with more general problems in complexity control. This can be done by considering the probabilities of a set of alternative models, given the data set:

$$p(\mathcal{H}_i|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{H}_i)p(\mathcal{H}_i)}{p(\mathcal{D})}. \quad (33)$$

Here different models can also be interpreted as different values of regularization parameters as these too control model complexity. If the models are given the same prior probabilities $p(\mathcal{H}_i)$ then they can be ranked by considering the *evidence* $p(\mathcal{D}|\mathcal{H}_i)$ which itself can be evaluated by integration over the model parameters \mathbf{w} . We can simply select the model with the greatest probability. However, a full Bayesian treatment requires that we form a linear combination of the predictions of the models in which the weighting coefficients are given by the model probabilities.

In general, the required integrations, such as that in Eq. 30, are analytically intractable. One approach is to approximate the posterior distribution by a Gaussian centered on \mathbf{w}_{MP} and then

to linearize $p(\mathbf{t}|\mathbf{x}, \mathbf{w})$ about \mathbf{w}_{MP} so that the integration can be performed analytically [MacKay, 1992]. Alternatively, sophisticated Monte Carlo methods can be employed to evaluate the integrals numerically [Neal, 1994]. An important aspect of the Bayesian approach is that there is no need to keep data aside in a validation set as is required when using maximum likelihood. In practical applications for which the quantity of available data are limited, it is found that a Bayesian treatment generally outperforms other approaches.

3.7 Pre-processing, invariances and prior knowledge

We have already seen that neural networks can approximate essentially arbitrary nonlinear functional mappings between sets of variables. In principle we could therefore use a single network to transform the raw input variables into the required final outputs. However, in practice for all but the simplest problems the results of such an approach can be improved upon considerably by incorporating various forms of pre-processing, for reasons which we shall outline below.

One of the simplest and most common forms of pre-processing consists of a simple normalization of the input, and possibly also target, variables. This may take the form of a linear rescaling of each input variable independently to give it zero mean and unit variance over the training set. For some applications the original input variables may span widely different ranges. Although a linear rescaling of the inputs is equivalent to a different choice of first-layer weights, in practice the optimization algorithm may have considerable difficulty in finding a satisfactory solution when typical input values are substantially different. Similar rescaling can be applied to the output values in which case the inverse of the transformation needs to be applied to the network outputs when the network is presented with new inputs. Pre-processing is also used to encode data in a suitable form. For example, if we have categorical variables such as ‘red’, ‘green’ and ‘blue’, these may be encoded using a 1-of-3 binary representation.

Another widely used form of pre-processing involves reducing the dimensionality of the input space. Such transformations may result in loss of information in the data, but the overall effect can be a significant improvement in performance as a consequence of the curse of dimensionality discussed in Section 3.5. The finite data set is better able to specify the required mapping in the lower-dimensional space. Dimensionality reduction may be accomplished by simply selecting a subset of the original variables, but more typically involves the construction of new variables consisting of linear or nonlinear combinations of the original variables called *features*. A standard technique for dimensionality reduction is principal component analysis [Anderson, 1984]. Such methods, however, make use only of the input data and ignore the target values, and can sometimes be significantly sub-optimal.

Yet another form of pre-processing involves correcting deficiencies in the original data. A common occurrence is that some of the input variables are missing for some of the data points. Correction of this problem in a principled way requires that the probability distribution $p(\mathbf{x})$ of input data be modeled.

One of the most important factors determining the performance of real-world applications of neural networks is the use of *prior knowledge* which is information additional to that present in the data. As an example, consider the problem of classifying hand-written digits discussed in Section 1. The most direct approach would be to collect a large training set of digits and to train a feedforward

network to map from the input image to a set of 10 output values representing posterior probabilities for the 10 classes. However, we know that the classification of a digit should be independent of its position within the input image. One way of achieving such *translation invariance* is to make use of the technique of *shared weights*. This involves a network architecture having many hidden layers in which each unit takes inputs only from a small patch, called a *receptive field*, of units in the previous layer. By a process of constraining neighboring units to have common weights, it can be arranged that the output of the network is insensitive to translations of the input image. A further benefit of weight sharing is that the number of independent parameters is much smaller than the number of weights, which assists with the problem of model complexity. This approach is the basis for the highly successful US postal code recognition system of [LeCun, et al., 1989]. An alternative to shared weights is to enlarge the training set artificially by generating “virtual examples” based on applying translations and other transformations to the original training set [Poggio and Vetter, 1992].

4 Graphical models

Neural networks express relationships between variables by utilizing the representational language of graph theory. Variables are associated with nodes in a graph and transformations of variables are based on algorithms that propagate numerical messages along the links of the graph. Moreover, the graphs are often accompanied by probabilistic interpretations of the variables and their interrelationships. As we have seen, such probabilistic interpretations allow a neural network to be understood as a form of probabilistic model, and reduce the problem of learning the weights of a network to a problem in statistics.

Related graphical models have been studied throughout statistics, engineering and AI in recent years. Hidden Markov models, Kalman filters, and path analysis models are all examples of graphical probabilistic models that can be fitted to data and used to make inferences. The relationship between these models and neural networks is rather strong; indeed it is often possible to reduce one kind of model to the other. In this section, we examine these relationships in some detail and provide a broader characterization of neural networks as members of a general family of graphical probabilistic models.

Many interesting relationships have been discovered between graphs and probability distributions [Spiegelhalter, et al., 1993]; [Pearl, 1988]. These relationships derive from the use of graphs to represent conditional independencies among random variables. In an undirected graph, there is a direct correspondence between conditional independence and graph separation—random variables X_i and X_k are conditionally independent given X_j if nodes X_i and X_k are separated by node X_j (we use the symbol “ X_i ” to represent both a random variable and a node in a graph). This statement remains true for sets of nodes (see Figure 5(a)). Directed graphs have a somewhat different semantics, due to the ability of directed graphs to represent “induced dependencies.” An induced dependency is a situation in which two nodes which are marginally independent become conditionally dependent given the value of a third node (see Figure 5(b)). Suppose, for example, that X_i and X_k represent independent coin tosses, and X_j represents the sum of X_i and X_k . Then X_i and X_k are marginally independent but are conditionally dependent given X_j . The semantics of

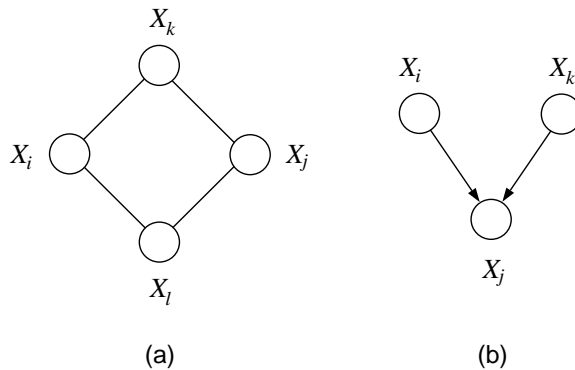


Figure 5: (a) An undirected graph in which X_i is independent of X_j given X_k and X_l , and X_k is independent of X_l given X_i and X_j . (b) A directed graph in which X_i and X_k are marginally independent but are conditionally dependent given X_j .

independence in directed graphs is captured by a graphical criterion known as *d-separation* [Pearl, 1988], which differs from undirected separation only in those cases in which paths have two arrows arriving at the same node (as in Figure 5(b)).

Although the neural network architectures that we have discussed until now have all been based on directed graphs, undirected graphs also play an important role in neural network research. Constraint satisfaction architectures, including the Hopfield network [Hopfield, 1982] and the Boltzmann machine [Hinton and Sejnowski, 1986], are the most prominent examples. A Boltzmann machine is an undirected probabilistic graph that respects the conditional independency semantics described above (cf. Figure 5(a)). Each node in a Boltzmann machine is a binary-valued random variable X_i (or more generally a discrete-valued random variable). A probability distribution on the 2^N possible configurations of such variables is defined via an *energy function* E . Let J_{ij} be the weight on the link between X_i and X_j , let $J_{ij} = J_{ji}$, let α index the configurations, and define the energy of configuration α as follows:

$$E_\alpha = - \sum_{i < j} J_{ij} X_i^\alpha X_j^\alpha. \quad (34)$$

The probability of configuration α is then defined via the Boltzmann distribution:

$$P_\alpha = \frac{e^{-E_\alpha/T}}{\sum_\gamma e^{-E_\gamma/T}}, \quad (35)$$

where the *temperature* T provides a scale for the energy.

An example of a directed probabilistic graph is the hidden Markov model (HMM). An HMM is defined by a set of *state variables* H_i , where i is generally a time or a space index, a set of output variables O_i , a *probability transition matrix* $A = p(H_i|H_{i-1})$, and an *emission matrix* $B = p(O_i|H_i)$. The directed graph for an HMM is shown in Figure 6(a). As can be seen from considering the separatory properties of the graph, the conditional independencies of the HMM are defined by the

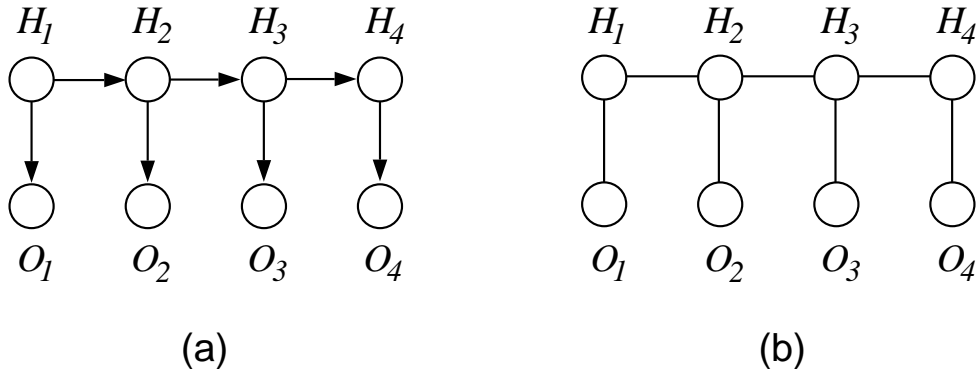


Figure 6: (a) A directed graph representation of an HMM. Each horizontal link is associated with the transition matrix A , and each vertical link is associated with the emission matrix B . (b) An HMM as a Boltzmann machine. The parameters on the horizontal links are logarithms of the entries of the A matrix, and the parameters on the vertical links are logarithms of the entries of the B matrix. The two representations yield the same joint probability distribution.

following Markov conditions:

$$H_i \perp \{H_1, O_1, \dots, H_{i-2}, O_{i-2}, O_{i-1}\} | H_{i-1}, \quad 2 \leq i \leq N \quad (36)$$

and

$$O_i \perp \{H_1, O_1, \dots, H_{i-1}, O_{i-1}\} | H_i, \quad 2 \leq i \leq N, \quad (37)$$

where the symbol \perp is used to denote independence.

Figure 6(b) shows that it is possible to treat an HMM as a special case of a Boltzmann machine [Luttrell, 1989]; [Saul and Jordan, 1995]. The probabilistic structure of the HMM can be captured by defining the weights on the links as the logarithms of the corresponding transition and emission probabilities. The Boltzmann distribution (Eq. 35) then converts the additive energy into the product form of the standard HMM probability distribution. As we will see, this reduction of a directed graph to an undirected graph is a recurring theme in the graphical model formalism.

General mixture models are readily viewed as graphical models [Buntine, 1994]. For example, the unconditional mixture model of Eq. 2 can be represented as a graphical model with two nodes—a multinomial “hidden” node which represents the selected component, a “visible” node representing \mathbf{x} , and a directed link from the hidden node to the visible node (see below for the hidden/visible distinction). Conditional mixture models [Jacobs, et al., 1991] simply require another visible node with directed links to the hidden node and the visible nodes. Hierarchical conditional mixture models [Jordan and Jacobs, 1994] require a chain of hidden nodes, one hidden node for each level of the tree.

Within the general framework of probabilistic graphical models, it is possible to tackle general problems of inference and learning. The key problem that arises in this setting is the problem of computing the probabilities of certain nodes, which we will refer to as *hidden nodes*, given the observed values of other nodes, which we will refer to as *visible nodes*. For example, in an HMM, the

variables O_i are generally treated as visible, and it is desired to calculate a probability distribution on the hidden states H_i . A similar inferential calculation is required in the mixture models and the Boltzmann machine.

Generic algorithms have been developed to solve the inferential problem of the calculation of posterior probabilities in graphs. Although a variety of inference algorithms have been developed, they can all be viewed as essentially the same underlying algorithm [Shachter, Andersen, and Szolovits, 1994]. Let us consider undirected graphs. A special case of an undirected graph is a *triangulated graph* [Spiegelhalter, et al., 1993], in which any cycle having four or more nodes has a chord. For example, the graph in Figure 5(a) is not triangulated, but becomes triangulated when a link is added between nodes X_i and X_j . In a triangulated graph, the cliques of the graph can be arranged in the form of a *junction tree*, which is a tree having the property that any node that appears in two different cliques in the tree also appears in every clique on the path that links the two cliques (the “running intersection property”). This cannot be achieved in non-triangulated graphs. For example, the cliques in Figure 5(a) are $\{X_i, X_k\}$, $\{X_k, X_j\}$, $\{X_j, X_l\}$, and it is not possible to arrange these cliques into a tree that obeys the running intersection property. If a chord is added the resulting cliques are $\{X_i, X_j, X_k\}$ and $\{X_i, X_j, X_l\}$, and these cliques can be arranged as a simple chain that trivially obeys the running intersection property. In general, it turns out that the probability distributions corresponding to triangulated graphs can be characterized as *decomposable*, which implies that they can be factorized into a product of local functions (“potentials”) associated with the cliques in the triangulated graph.¹ The calculation of posterior probabilities in decomposable distributions is straightforward, and can be achieved via a local message-passing algorithm on the junction tree [Spiegelhalter, et al., 1993].

Graphs that are not triangulated can be turned into triangulated graphs by the addition of links. If the potentials on the new graph are defined suitably as products of potentials on the original graph, then the independencies in the original graph are preserved. This implies that the algorithms for triangulated graphs can be used for *all* undirected graphs; an untriangulated graph is first triangulated (see Figure 7). Moreover, it is possible to convert *directed* graphs to undirected graphs in a manner that preserves the probabilistic structure of the original graph [Spiegelhalter, et al., 1993]. This implies that the junction tree algorithm is indeed generic; it can be applied to any graphical model.

The problem of calculating posterior probabilities on graphs is NP-hard; thus, a major issue in the use of the inference algorithms is the identification of cases in which they are efficient. Chain structures such as HMM’s yield efficient algorithms, and indeed the classical forward-backward algorithm for HMM’s is a special, efficient case of the junction tree algorithm [Heckerman, Jordan, and Smyth, 1996]. Decision tree structures such as the hierarchical mixture of experts yield efficient algorithms, and the recursive posterior probability calculation of [Jordan and Jacobs, 1994] described earlier is also a special case of the junction tree algorithm. All of the simpler mixture model calculations described earlier are therefore also special cases. Another interesting special

¹An interesting example is a Boltzmann machine on a triangulated graph. The potentials are products of $\exp(J_{ij})$ factors, where the product is taken over all (i, j) pairs in a particular clique. Given that the product across potentials must be the joint probability, this implies that the partition function (the denominator of Eq. 35) must be unity in this case.

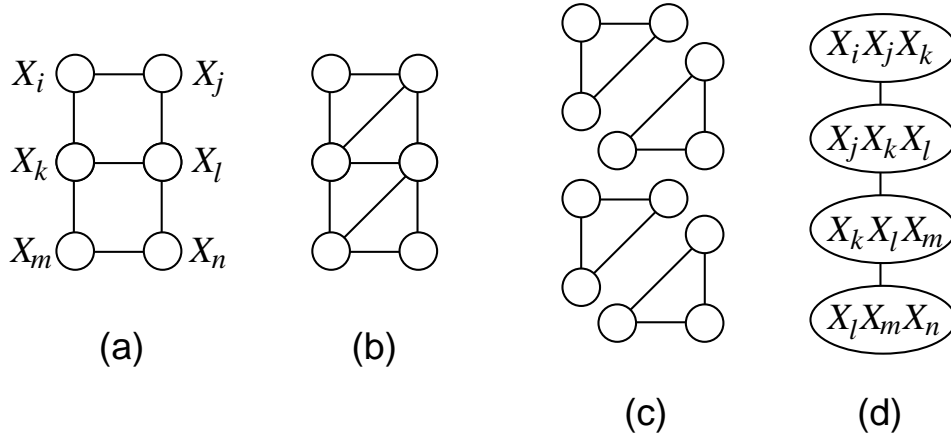


Figure 7: The basic structure of the junction tree algorithm for undirected graphs. The graph in (a) is first triangulated (b), then the cliques are identified (c), and arranged into a tree (d). Products of potential functions on the nodes in (d) yield probability distributions on the nodes in (a).

case is the state estimation algorithm of the Kalman filter [Shachter and Kenley, 1989]. Finally, there are a variety of special cases of the Boltzmann machine which are amenable to the exact calculations of the junction tree algorithm [Saul and Jordan, 1995].

For graphs that are outside of the tractable categories of trees and chains, the junction tree algorithm often performs surprisingly well, but for highly connected graphs the algorithm can be too slow. In such cases, approximate algorithms such as Gibbs sampling are utilized. A virtue of the graphical framework is that Gibbs sampling has a generic form, which is based on the notion of a *Markov boundary* [Pearl, 1988]. A special case of this generic form is the stochastic update rule for general Boltzmann machines.

Our discussion has emphasized the unifying framework of graphical models both for expressing probabilistic dependencies in graphs and for describing algorithms that perform the inferential step of calculating posterior probabilities on these graphs. The unification goes further, however, when we consider learning. A generic methodology known as the Expectation-Maximization (EM) algorithm is available for MAP and Bayesian estimation in graphical models [Dempster, Laird, and Rubin, 1977]. EM is an iterative method, based on two alternating steps: an *E step*, in which the values of hidden variables are estimated, based on the current values of the parameters and the values of visible variables, and an *M step*, in which the parameters are updated based on the estimated values obtained from the E step. Within the framework of the EM algorithm, the junction tree algorithm can readily be viewed as providing a generic E step. Moreover, once the estimated values of the hidden nodes are obtained from the E step, the graph can be viewed as fully observed, and the M step is a standard MAP or ML problem. The standard algorithms for all of the tractable architectures described above (mixtures, trees and chains) are in fact instances of this general graphical EM algorithm, and the learning algorithm for general Boltzmann machines is a special case of a generalization of EM known as GEM [Dempster, et al., 1977].

What about the case of feedforward neural networks such as the multilayer perceptron? It

is in fact possible to associate binary hidden values with the hidden units of such a network (cf. our earlier discussion of the logistic function; see also [Amari, 1995]) and apply the EM algorithm directly. For N hidden units, however, there are 2^N patterns whose probabilities must be calculated in the E step. For large N , this is an intractable computation, and recent research has therefore begun to focus on fast methods for approximating these distributions [Hinton, et al., 1995]; [Saul, et al., 1995].

References

- Amari, S. 1995. The EM algorithm and information geometry in neural network learning. *Neural Computation*, 7(1):13–18.
- Anderson, T. W. 1984. *An Introduction to Multivariate Statistical Analysis*. John Wiley, New York.
- Bengio, Y. 1996. *Neural Networks for Speech and Sequence Recognition*. Thomson Computer Press, London.
- Bishop, C. M. 1995. *Neural Networks for Pattern Recognition*. Oxford University Press.
- Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. 1984. *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA.
- Buntine, W. 1994. Operations for learning with graphical models. *Journal of Artificial Intelligence Research* 2:159–225.
- Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum-likelihood from incomplete data via the EM algorithm. *J. of Royal Statistical Society, B39*:1–38.
- Fletcher, R. 1987. *Practical Methods of Optimization*, 2nd ed. John Wiley, New York.
- Geman, S., Bienenstock, E., and Doursat, R. 1992. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58.
- Heckerman, D., Jordan, M. I., and Smyth, P. 1996. *Conditional independence graphs for hidden Markov probability models*. Tech. Rep., Center for Biological and Computational Learning, Massachusetts Institute of Technology.
- Hertz, J., Krogh, A, and Palmer, R. G. 1991. *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City, CA.
- Hinton, G. E., Dayan, P., Frey, B., and Neal, R. 1995. The wake-sleep algorithm for unsupervised neural networks. *Science*, 268:1158–1161.
- Hinton, G. E. and Sejnowski, T. 1986. Learning and relearning in Boltzmann machines. In *Parallel distributed processing: Volume 1*, ed., D. E. Rumelhart and J. L. McClelland, p. 282–317. MIT Press, Cambridge, MA.

- Hopfield, J. J. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79:2554–2558.
- Jacobs, R.A., Jordan, M.I., Nowlan, S.J., and Hinton, G.E. 1991. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87.
- Jordan, M.I. and Jacobs, R.A. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214.
- Le Cun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. and Jackel, L. D. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551.
- Luttrell, S. 1989. The Gibbs machine applied to hidden Markov model problems. *Royal Signals and Radar Establishment: SP Research Note 99*, Malvern, UK.
- MacKay, D. J. C. 1992. A practical Bayesian framework for back-propagation networks. *Neural Computation*, 4:448–472.
- McCullagh, P. and Nelder, J.A. 1983. *Generalized Linear Models*. Chapman and Hall, London.
- Neal, R. M. 1994. *Bayesian Learning for Neural Networks*. Unpublished PhD thesis, Department of Computer Science, University of Toronto, Canada.
- Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.
- Poggio, T. and Vetter, T. 1992. *Recognition and structure from one 2D model view: Observations on prototypes, object classes and symmetries*. AI Memo 1347, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.
- Rabiner, L.R. 1989. A tutorial on Hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286.
- Rumelhart, D. E., Durbin, R., Golden, R., and Chauvin, Y. 1995. Backpropagation: The basic theory. In *Backpropagation: Theory, Architectures, and Applications*, ed. Y. Chauvin, and D.E. Rumelhart, p. 1–35. Lawrence Erlbaum, Hillsdale, NJ.
- Saul, L. K., Jaakkola, T., and Jordan, M. I. 1995. Mean field learning theory for sigmoid belief networks. Computational Cognitive Science Tech. Rep. 9501, MIT, Cambridge, MA.
- Saul, L. K. and Jordan, M. I. 1995. Boltzmann chains and hidden Markov models. In *Advances in Neural Information Processing Systems 7*, ed., G. Tesauro, D. Touretzky, and T. Leen. Cambridge, MA, MIT Press.
- Sandler, D. G., Barrett, T. K., Palmer, D. A., Fugate, R. Q., and Wild, W. J. 1991. Use of a neural network to control an adaptive optics system for an astronomical telescope, *Nature*, 351:300–302.

- Shachter, R., Andersen, S., and Szolovits, P. 1994. Global conditioning for probabilistic inference in belief networks. In *Uncertainty in Artificial Intelligence: Proceedings of the Tenth Conference*, p. 514–522. Seattle, WA.
- Shachter, R. and Kenley, C. 1989. Gaussian influence diagrams. *Management Science*, 35(5):527–550.
- Spiegelhalter, D., Dawid, A., Lauritzen, S., and Cowell, R. 1993. Bayesian analysis in expert systems. *Statistical Science*, 8(3):219–283.
- Tarassenko, L. 1995. Novelty detection for the identification of masses in mammograms. *Proceedings Fourth IEE International Conference on Artificial Neural Networks*, 4:442–447.
- Vapnik, V. N. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Further information

In this chapter we have emphasized the links between neural networks and statistical pattern recognition. A more extensive treatment from the same perspective can be found in [Bishop, 1995]. For a view of recent research in the field, the proceedings of the annual NIPS (Neural Information Processing Systems; MIT Press) conferences are highly recommended.

Neural computing is now a very broad field and there are many topics which have not been discussed for lack of space. Here we aim to provide a brief overview of some of the more significant omissions, and to give pointers to the literature.

The resurgence of interest in neural networks during the 1980's was due in large part to work on the statistical mechanics of fully connected networks having symmetric connections (i.e. if unit i sends a connection to unit j then there is also a connection from unit j back to unit i with the same weight value). We have briefly discussed such systems; a more extensive introduction to this area can be found in [Hertz, et al., 1991].

The implementation of neural networks in specialist VLSI hardware has been the focus of much research, although by far the majority of work in neural computing is undertaken using software implementations running on standard platforms.

An implicit assumption throughout most of this chapter is that the processes which give rise to the data are stationary in time. The techniques discussed here can readily be applied to problems such as time series forecasting, provided this stationarity assumption is valid. If, however, the generator of the data is itself evolving with time then more sophisticated techniques must be used, and these are the focus of much current research [see Bengio, 1996].

One of the original motivations for neural networks was as models of information processing in biological systems such as the human brain. This remains the subject of considerable research activity, and there is a continuing flow of ideas between the fields of neurobiology and of artificial neural networks. Another historical springboard for neural network concepts was that of adaptive control, and again this remains a subject of great interest.

Defining terms

Classification A learning problem in which the goal is to assign input vectors to one of a number of (usually mutually exclusive) classes.

Boltzmann machine An undirected network of discrete valued random variables, where an energy function is associated with each of the links, and for which a probability distribution is defined by the Boltzmann distribution.

Cost function A function of the adaptive parameters of a model whose minimum is used to define suitable values for those parameters. It may consist of a likelihood function and additional terms.

Decision tree A network that performs a sequence of classificatory decisions on an input vector and produces an output vector that is conditional on the outcome of the decision sequence.

Density estimation The problem of modeling a probability distribution from a finite set of examples drawn from that distribution.

Discriminant function A function of the input vector which can be used to assign inputs to classes in a classification problem.

Hidden Markov model A graphical probabilistic model characterized by a state vector, an output vector, a state transition matrix, an emission matrix and an initial state distribution.

Likelihood function The probability of observing a particular data set under the assumption of a given parametrized model, expressed as a function of the adaptive parameters of the model.

Mixture model A probability model which consists of a linear combination of simpler component probability models.

Multilayer perceptron The most common form of neural network model, consisting of successive linear transformations followed by processing with nonlinear activation functions.

Overfitting The problem in which a model which is too complex captures too much of the noise in the data, leading to poor generalization.

Radial basis function network A common network model consisting of a linear combination of basis functions each of which is a function of the difference between the input vector and a center vector.

Regression A learning problem in which the goal is to map each input vector to a real-valued output vector.

Regularization A technique for controlling model complexity and improving generalization by the addition of a penalty term to the cost function.

VC dimension A measure of the complexity of a model. Knowledge of the VC dimension permits an estimate to be made of the difference between performance on the training set and performance on a test set.

Pattern Recognition approaches to Machine Translation

Francesc Casacuberta and Enrique Vidal

Pattern Recognition and Human Language Technology Group
Instituto Tecnológico de Informática
Departamento de Sistemas Informáticos y Computación
Universisdad Politécnica de Valencia, Spain

January 2005

F. Casacuberta & E. Vidal – ITI-UPV-DSIC

[Pattern Recognition Machine Translation](#)

[General Index](#)

Index

1. Introduction to Machine Translation,
Statistical Framework for Machine Translation (Mon. E.Vidal)
2. Statistical Alignment Models (Mon. F.Casacuberta)
3. Advanced Statistical Alignment Models (Tue. F.Casacuberta)
4. Stochastic Finite-State Translation Models (Tue. E.Vidal)
5. Phrase-based Alignment Models and Alignment Templates (Wed. F.Casacuberta)
6. State-Merging Approaches (Wed. E.Vidal)
7. Finite-State Translation Models based on Alignments (Thu. E.Vidal)
8. Recursive Alignment Models (Thu. F.Casacuberta)
9. Speech-to-Speech Translation (Fri. F.Casacuberta)
10. Computer-Assisted Translation (Fri. E.Vidal)

Pattern Recognition approaches to Machine Translation

F. Casacuberta and E. Vidal

Pattern Recognition and Human Language Technology Group
Instituto Tecnológico de Informática
Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de Valencia, Spain

Introduction to Machine Translation

Enrique Vidal

`evidal@iti.upv.es`

January 2005

E. Vidal – ITI-UPV-DSIC

[Pattern Recognition Machine Translation](#)

[Introduction to Machine Translation](#)

Index

- 1 Objectives of Machine Translation (MT) ▷ [2](#)
- 2 Approaches to MT ▷ [7](#)
- 3 Linguistic Resources ▷ [10](#)
- 4 Assessment ▷ [12](#)
- 5 Limited Domains ▷ [14](#)
- 6 Speech-to-speech MT ▷ [19](#)
- 7 Computer Assisted Translation ▷ [22](#)
- 8 Brief History of MT ▷ [25](#)

Bibliography:

D. Arnold, L. Balkan, R. Lee Humphreys, S. Meijer, L. Sadler:
“*Machine Translation, an introductory guide*”. NCC Blackwell, 1994

Index

- 1 *Objectives of Machine Translation (MT)* ▷ 2
- 2 Approaches to MT ▷ 7
- 3 Linguistic Resources ▷ 10
- 4 Assessment ▷ 12
- 5 Limited Domains ▷ 14
- 6 Speech-to-speech MT ▷ 19
- 7 Computer Assisted Translation ▷ 22
- 8 Brief History of MT ▷ 25

MT objectives: Erroneous conceptions

- MT is a waste of time because a machine never will translate Shakespeare.
- Generally, the quality of translation you can get from an MT system is very low.
- MT threatens the jobs of translators
- There is an MT system that translates what you say into Japanese and translates the other speaker's replies in English.
- There is an amazing South American Indian language with a structure of such logical perfection that it solves the problem of design MT systems.
- MT systems are machines, and buying an MT system should be very much like buying a car.

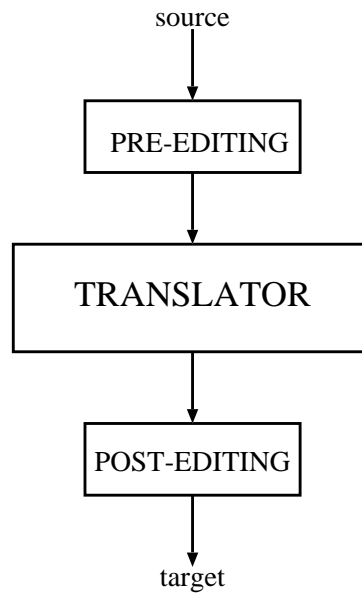
MT objectives: Facts

- MT is useful.
- There are many situations that MT systems produce reliable, if less than perfect, translations at high speed.
- In some circumstances, MT systems can produce good quality outputs.
- MT does not threaten translators' jobs: High demand of translations and too repetitive translation jobs.
- Speech-to-speech MT is still a research topic.
- There are many open research problems in MT.
- Building a traditional MT system is a time consuming job.
- A user will typically have to invest a considerable amount of effort in customizing an MT system.

Need of pre/post-editing

- While the number of errors and bad constructions is high, “post-editing” can make the result useful.
- Many problems could have been avoided by making the source text “simpler”.
- Simplification of the translation problem by using adequate rules to produce “controlled” (i.e., simple and regular) source text.

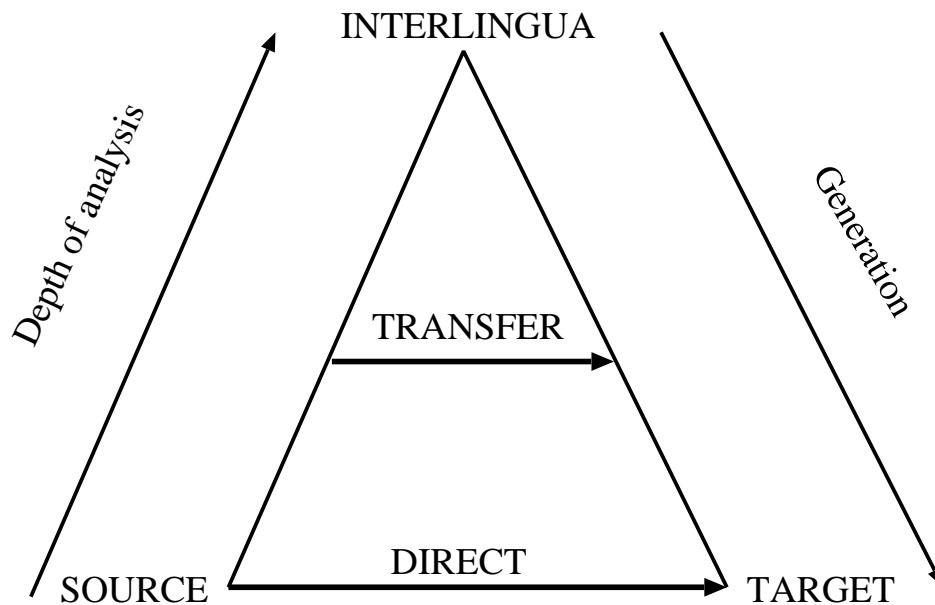
General scheme for MT



Index

- 1 Objectives of Machine Translation (MT) ▷ 2
- 2 *Approaches to MT* ▷ 7
- 3 Linguistic Resources ▷ 10
- 4 Assessment ▷ 12
- 5 Limited Domains ▷ 14
- 6 Speech-to-speech MT ▷ 19
- 7 Computer Assisted Translation ▷ 22
- 8 Brief History of MT ▷ 25

Approaches to MT: Analysis detail



Approaches to MT: Technologies

- (Linguistic) knowledge-based methods
- (Memorized) example-based methods
 - Translation memories
- Statistical models
 - Alignment models
 - Finite-State models
- Hybrid models

Index

- 1 Objectives of Machine Translation (MT) ▷ 2
- 2 Approaches to MT ▷ 7
- 3 *Linguistic Resources* ▷ 10
- 4 Assessment ▷ 12
- 5 Limited Domains ▷ 14
- 6 Speech-to-speech MT ▷ 19
- 7 Computer Assisted Translation ▷ 22
- 8 Brief History of MT ▷ 25

Linguistic resources

- Dictionaries
- Grammars
- Corpora
- Paragraph-aligned and Labeled Corpora

Index

- 1 Objectives of Machine Translation (MT) ▷ 2
- 2 Approaches to MT ▷ 7
- 3 Linguistic Resources ▷ 10
- 4 *Assessment* ▷ 12
- 5 Limited Domains ▷ 14
- 6 Speech-to-speech MT ▷ 19
- 7 Computer Assisted Translation ▷ 22
- 8 Brief History of MT ▷ 25

Assessment

- Test sentences
- Subjective evaluation based on the number of words that need to be corrected or deleted
- Test sentences with reference translation
- Automatic assessment
 - Editing Distances:
Translation Word Error Rate (TWER)
 - Multireference TWER
 - N-Gram based: *BLUE*

Index

- 1 Objectives of Machine Translation (MT) ▷ 2
- 2 Approaches to MT ▷ 7
- 3 Linguistic Resources ▷ 10
- 4 Assessment ▷ 12
- 5 *Limited Domains* ▷ 14
- 6 Speech-to-speech MT ▷ 19
- 7 Computer Assisted Translation ▷ 22
- 8 Brief History of MT ▷ 25

Limited domains or “sublanguages”

- Tasks with *small or medium-sized vocabularies* and *restricted semantic scope*.
- *Robust systems* needed.
- *Manual “post-editing”* should be *avoided* or minimized.
- Only *low development costs* can be afforded.

Limited domain or “sublanguages”: Example

The “Traveler Task” [Vidal et al., 96] (EuTrans ESPRIT project – first-phase)

- Domain: *human-to-human communication* situations in the front-desk of a hotel.
- Three language pairs:
 - *Spanish-English*,
 - *Spanish-German*
 - *Spanish-Italian*

Features of the Spanish-English task (similar for the other language pairs)

Input/output <i>vocabulary sizes</i>	~ 700 / 500
Average input/output <i>sentence lengths</i>	~ 10 / 10
Input/output <i>test-set perplexities</i>	~ 11 / 6

Limited domain or “sublanguages”: Example

The Traveler Task: examples of Spanish-English paired sentences

<i>Spanish:</i>	Reservé una habitación individual y tranquila con televisión hasta pasado mañana.
<i>English:</i>	I booked a quiet, single room with a tv. until the day after tomorrow.
<i>Spanish:</i>	Por favor, prepárenos nuestra cuenta de la habitación dos veintidós.
<i>English:</i>	Could you prepare our bill for room number two two two for us, please?

Language translation and language understanding

- Under the *Limited-Domain* (LD) framework both Language Understanding (LU) and Language Translation (LT) can be properly formulated in a *uniform* way.
- The ultimate goal of a LD LU system is to *drive the actions* associated to the meaning conveyed by the sentences issued by the users.
- Since actions are to be performed by machines, the understanding problem can then be simply formulated as *translating* the *natural language* sentences into *formal sentences* of an adequate (computer) command language in which the actions to be carried out can be specified.
- Thus, LU can be seen as a specific (simpler) case of LT in which the output language is *formal* rather than *natural*.

Index

- 1 Objectives of Machine Translation (MT) ▷ 2
- 2 Approaches to MT ▷ 7
- 3 Linguistic Resources ▷ 10
- 4 Assessment ▷ 12
- 5 Limited Domains ▷ 14
- 6 *Speech-to-speech MT* ▷ 19
- 7 Computer Assisted Translation ▷ 22
- 8 Brief History of MT ▷ 25

Approaches to speech-to-speech translation

- **Traditional** → Serially couple the following (existing) devices:
 1. Conventional continuous word recognition front-end.
 2. Text-to-text, general-purpose, knowledge-based MT system (adapted by experts to the task in hand).
 3. Text-to-speech output language synthesizer.
- **Integrated approach** → Consider language translation as a global input-output *decoding problem*:
 1. Develop an INTEGRATED DEVICE that directly accepts speech (or text) input sentences and outputs corresponding sentences in the target language.
 2. Implement input-output decoding as a global optimization search that takes into account all the information compiled into the integrated recognition/translation device.
 3. Chose a translation model that is *trainable* from input-output translation examples.

Speech translation: Advantages of integration and automatic learning

- *Tight integration* leads to speech-input translation systems which are significantly more *robust*, as compared with other based on the more traditional, *loosely coupled* approach.
- *Trainability* leads to *better adaptation* to specific domains at much *lower development costs*.

Index

- 1 Objectives of Machine Translation (MT) ▷ 2
- 2 Approaches to MT ▷ 7
- 3 Linguistic Resources ▷ 10
- 4 Assessment ▷ 12
- 5 Limited Domains ▷ 14
- 6 Speech-to-speech MT ▷ 19
- 7 *Computer Assisted Translation* ▷ 22
- 8 Brief History of MT ▷ 25

Computer Assited Translation (CAT)

- Do *not* attempt *fully automated* MT
- Aim at *high-quality* results
- Let *the human* translator *fully command* the process
- Allow for *tight human-machine cooperation*
- Aim to *increase* human translator *productivity*
- Ergonomic issues and *multimodality*:
keyboard, mouse, speech, ...

Typical Computer Assisted Translation Scenario

Text prediction based on both the source-language text to be translated *and* preceding text that has been *validated by the user*.

For each source sentence or paragraph to be translated:

1. The system provides its best (or N-best) translation suggestion
2. The user selects a *correct* part (typically a prefix) of this suggestion and starts amending the remaining part or entering new text by him/herself
3. After each user-entered word (or key-stroke), the system recomputes its best suggestion(s), thereby starting a new human-system interaction cycle.

Index

- 1 Objectives of Machine Translation (MT) ▷ 2
- 2 Approaches to MT ▷ 7
- 3 Linguistic Resources ▷ 10
- 4 Assessment ▷ 12
- 5 Limited Domains ▷ 14
- 6 Speech-to-speech MT ▷ 19
- 7 Computer Assisted Translation ▷ 22
- 8 *Brief History of MT* ▷ 25

Brief history of MT

- **1949** Weaver: Information-theory based approach
- **1957** Chomsky: Natural language is not governed by statistics
- **1960** ALPAC (Automatic Language Processing Advisory Committee) report: No useful MT results are foreseen
- **1960-nowadays**
 - SYSTRAN system: based on dictionaries
 - Several (linguistic) knowledge-based approaches
- **1985-95** “Empiricists” methods are introduced: corpus-based and statistical approaches (IBM, 1989)
- **1995-nowadays** “Empiricists” methods are thriving. Speech-to-speech MT in limited domains

Recent history of MT: “Empiricists” methods

- **1989-95** Statistical approach to MT by IBM Yorktown Heights researchers
 - Corpus: Hansards
 - Parallel English/French transcriptions of parliamentary discussions
 - DARPA competitive assessment (1994): Results comparable to those achieved by traditional approaches
- **1990-05** Development of statistical techniques and other empiricists methods
 - Progress of the statistical approach (by IBM and other groups)
 - Other “example-based”, empiricist techniques: Memory-Based, Finite-State, etc.
 - Statistics are applied to other MT-related fields: Lexicography, syntactic labeling of corpora, etc.
 - Progress in Grammars and Syntactic Analysis
 - Computer Assisted Translation

Pattern Recognition approaches to Machine Translation

F. Casacuberta and E. Vidal

Pattern Recognition and Human Language Technology Group
Instituto Tecnológico de Informática
Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de Valencia, Spain

Statistical Framework to Machine Translation

Enrique Vidal

`evidal@iti.upv.es`

January 2005

E. Vidal – ITI-UPV-DSIC

Pattern Recognition Machine Translation

Statistical Framework to Machine Translation

Index

- 1 Notation and background ▷ [2](#)
- 2 Text-input machine translation ▷ [4](#)
- 3 Speech-input machine translation ▷ [10](#)
- 4 Computer-assisted translation ▷ [13](#)
- 5 Bibliography ▷ [18](#)

Index

- 1 *Notation and background* ▷ 2
- 2 Text-input machine translation ▷ 4
- 3 Speech-input machine translation ▷ 10
- 4 Computer-assisted translation ▷ 13
- 5 Bibliography ▷ 18

Notation and Basic Concepts

- x and y will generally denote *source* and *target* texts, respectively
- **CONDITIONAL AND UNCONDITIONAL PROBABILITIES:**
 $\Pr(X = x \mid Y = y) \equiv \Pr(x \mid y), \quad \Pr(X = x) \equiv \Pr(x)$
- **BAYES' RULE:** $\Pr(x \mid y) \cdot \Pr(y) = \Pr(y \mid x) \cdot \Pr(x)$
- **JOINT PROBABILITY:** $\Pr(x, y) = \Pr(x) \cdot \Pr(y \mid x)$
- $\Pr(x) = \sum_y \Pr(x, y)$
- $\max_x \Pr(x) = \Pr(\operatorname{argmax}_x \Pr(x))$
- $\sum_x \Pr(x) \approx \max_x \Pr(x)$

Index

- 1 Notation and background ▷ 2
- 2 *Text-input machine translation* ▷ 4
- 3 Speech-input machine translation ▷ 10
- 4 Computer-assisted translation ▷ 13
- 5 Bibliography ▷ 18

General Framework

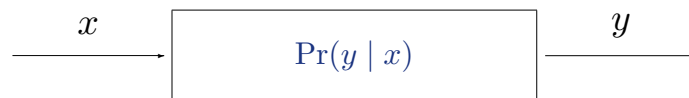
- Every sentence y in a *target* language is considered as a possible translation of any other sentence x in another *source* language.
- For each possible pair of sentences y, x , there is a probability $\Pr(y | x)$.
- $\Pr(y | x)$ should be *low* for pairs (y, x) such as:
(*una habitación con vistas al mar , are all expenses included in the bill ?*)
- $\Pr(y | x)$ should be *high* for pairs such as:
(*¿ hay alguna habitación tranquila libre ? , is there a quiet room available ?*)

A direct approach

Search for a target sentence with maximum *posterior* probability:

$$\hat{y} = \underset{y}{\operatorname{argmax}} \Pr(y \mid x)$$

A “direct model”



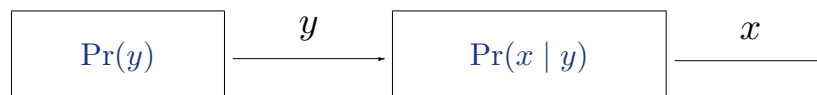
Need: *alignment and lexicon models*

An inverse approach

Decompose $\Pr(y \mid x)$ using Bayes’ rule:

$$\hat{y} = \underset{y}{\operatorname{argmax}} \Pr(y \mid x) = \underset{y}{\operatorname{argmax}} \Pr(x \mid y) \cdot \Pr(y)$$

A “distorted channel model”



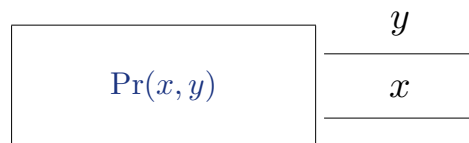
Need: *a target-language model + alignment and lexicon models*

A finite-state approach

The direct probability can be decomposed in a different way:

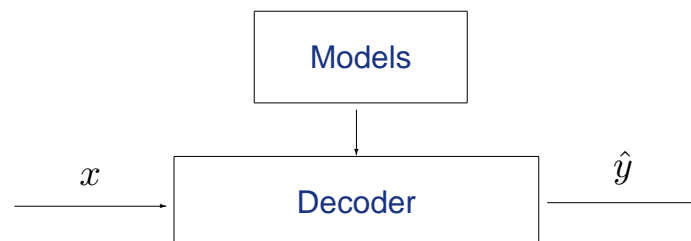
$$\hat{y} = \underset{y}{\operatorname{argmax}} \Pr(y \mid x) = \underset{y}{\operatorname{argmax}} \Pr(x, y)$$

A “joint” model



A stochastic finite-state transducer can model the joint distribution

Translation search



- Direct approach: **alignment and lexicon models**

$$\hat{y} = \underset{y}{\operatorname{argmax}} \Pr(y \mid x)$$

- Inverse approach: **a target-language model + alignment and lexicon models**

$$\hat{y} = \underset{y}{\operatorname{argmax}} \Pr(x \mid y) \cdot \Pr(y)$$

- Joint approach: **stochastic finite-state transducer**

$$\hat{y} = \underset{y}{\operatorname{argmax}} \Pr(x, y)$$

Index

- 1 Notation and background ▷ 2
- 2 Text-input machine translation ▷ 4
- 3 *Speech-input machine translation* ▷ 10
- 4 Computer-assisted translation ▷ 13
- 5 Bibliography ▷ 18

Speech-input translation

Given an input acoustic sequence v , search for a target sentence with maximum posterior probability:

$$\hat{y} = \underset{y}{\operatorname{argmax}} \Pr(y \mid v)$$

But this can be seen as a “*two-step process*”:

$$v \longrightarrow x \longrightarrow y$$

where the “*hidden variable*” x accounts for all possible input decodings of v :

$$\hat{y} = \underset{y}{\operatorname{argmax}} \sum_x \Pr(y, x \mid v) = \underset{y}{\operatorname{argmax}} \sum_x \Pr(x, y) \cdot \Pr(v \mid x)$$

(with the assumption: $\Pr(v \mid x, y)$ does not depend on the target sentence y)

Speech-input translation

$$\begin{aligned}
 \underset{y}{\operatorname{argmax}} \Pr(y \mid v) &\approx \underset{y}{\operatorname{argmax}} \max_x (\Pr(x, y) \cdot \Pr(v \mid x)) \\
 &= \underset{y}{\operatorname{argmax}} \max_x (\Pr(y) \cdot \Pr(x \mid y) \cdot \Pr(v \mid x)) \\
 &= \underset{y}{\operatorname{argmax}} \max_x (\Pr(x) \cdot \Pr(y \mid x) \cdot \Pr(v \mid x))
 \end{aligned}$$

- $\Pr(v \mid x) \approx$ **ACOUSTIC MODELS**
- $\Pr(x, y) \approx$ **FINITE-STATE TRANSDUCERS**
- $\Pr(x \mid y), \Pr(y \mid x) \approx$ **ALIGNMENT AND LEXICON MODELS**
- $\Pr(y) \approx$ **TARGET LANGUAGE MODELS**
- $\Pr(x) \approx$ **SOURCE LANGUAGE MODELS**

Index

- 1 Notation and background ▷ 2
- 2 Text-input machine translation ▷ 4
- 3 Speech-input machine translation ▷ 10
- 4 *Computer-assisted translation* ▷ 13
- 5 Bibliography ▷ 18

Text prediction for Computer-Assisted Translation (CAT)

Given a source text x and a “correct” *prefix* y_p of the target text, search for a *suffix* \hat{y}_s , that maximizes the posterior probability over all possible suffixes:

$$\hat{y}_s = \underset{y_s}{\operatorname{argmax}} \Pr(y_s \mid x, y_p)$$

Taking into account that $\Pr(y_p \mid x)$ does not depend on y_s , we can write:

$$\begin{aligned} \hat{y}_s &= \underset{y_s}{\operatorname{argmax}} \Pr(y_p y_s \mid x) \\ &= \underset{y_s}{\operatorname{argmax}} \Pr(x, y_p y_s) \\ &= \underset{y_s}{\operatorname{argmax}} \Pr(x \mid y_p y_s) \cdot \Pr(y_p y_s) \end{aligned}$$

Main difference with text-input machine translation: **search over the set of suffixes.**

Target language dictation in CAT

A *human* translator *dictates* the translation of a source text, x , producing a *target language* acoustic sequence v .

Given v and x , the system should search for a most likely decoding of v :

$$\hat{y} = \underset{y}{\operatorname{argmax}} \Pr(y \mid x, v)$$

By the assumption that $\Pr(v \mid x, y)$ does not depend on x ,

$$\hat{y} = \underset{y}{\operatorname{argmax}} \Pr(v \mid y) \cdot \Pr(x \mid y) \cdot \Pr(y)$$

- $\Pr(v \mid y) \approx$ **(TARGET LANGUAGE) ACOUSTIC MODELS**
- $\Pr(x \mid y) \approx$ **TRANSLATION MODEL**
- $\Pr(y) \approx$ **TARGET LANGUAGE MODEL**

Similar to plain speech decoding, where: $\hat{y} = \underset{y}{\operatorname{argmax}} \Pr(v \mid y) \cdot \Pr(y)$

Further use of speech recognition in CAT

Let x be the source text and y_p a “correct” prefix of the target sentence.
As in pure text CAT the system suggests an optimal suffix:

$$\hat{y}_s = \underset{y_s}{\operatorname{argmax}} \Pr(y_s \mid x, y_p) . \quad (1)$$

The user is now allowed to *utter some words*, v , generally aimed at amending parts of \hat{y}_s and the system has then to obtain a most probable decoding of v :

$$\hat{d} = \underset{d}{\operatorname{argmax}} \Pr(d \mid x, y_p, \hat{y}_s, v) . \quad (2)$$

Finally, the user can enter additional amendment keystrokes k , to produce a new consolidated prefix, y_p , based on the previous y_p , \hat{d} , k and parts of \hat{y}_s .

Further use of speech recognition in CAT (cont.)

From Eq. (2):

$$\hat{d} = \underset{d}{\operatorname{argmax}} \Pr(d \mid x, y_p, \hat{y}_s) \cdot \Pr(v \mid x, y_p, \hat{y}_s, d)$$

and, by making the assumption that $\Pr(v \mid x, y_p, \hat{y}_s, d)$ only depends on d :

$$\hat{d} = \underset{d}{\operatorname{argmax}} \Pr(d \mid x, y_p, \hat{y}_s) \cdot \Pr(v \mid d)$$

- $\Pr(v \mid d) \approx \text{(TARGET LANGUAGE) ACOUSTIC MODELS}$
- $\Pr(d \mid x, y_p, \hat{y}_s) \approx \text{TARGET LANGUAGE MODEL CONSTRAINED BY THE SOURCE SENTENCE, THE PREFIX AND THE SUFFIX}$

Index

- 1 Notation and background ▷ 2
- 2 Text-input machine translation ▷ 4
- 3 Speech-input machine translation ▷ 10
- 4 Computer-assisted translation ▷ 13
- 5 *Bibliography* ▷ 18

Bibliografy

1. P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, P. S. Roosin: *A statistical approach to machine translation*. Computational Linguistics, vol. 16, pp. 79–85, 1990.
2. P. F. Brown, S. F. Chen, S. A. Della Pietra, V. J. Della Pietra, A. S. Kehler, R. L. Mercer: *Automatic speech recognition in machine aided translation*. Computer Speech and Language, vol. 8 (3), pp. 177–87. 1994.
3. H. Ney, S. Nießen, F. Och, H. Sawaf, C. Tillmann, S. Vogel: *Algorithms for statistical translation of spoken language*. IEEE Transactions on Speech and Audio Processing, vol. 8, pp. 24–36, 2000.
4. F. Casacuberta, E. Vidal, J. M. Vilar: *Architectures for speech-to-speech translation using finite-state 5 models*. ACL-02 Workshop on Speech-to-Speech Translation: Algorithms and Systems, July 11, 2002.
5. F. Casacuberta, E. Vidal, A. Sanchis, and J.-M. Vilar: *Pattern recognition approaches for speech-to-speech translation*. Cybernetic and Systems: an International Journal, 2004.
6. G. Foster, P. Langlais, G. Lapalme: *User-friendly text prediction for translators*, Proc. of the Conf. on Empirical methods in Natural Language Processing. pp. 148-155. July 2002.
7. F. Casacuberta, H. Ney, F. J. Och, E. Vidal, J. M. Vilar, S. Barrachina, I. García-Varea, D. Llorens, C. Martínez, S. Molau, F. Nevado, M. Pastor, D. Picó, A. Sanchis, Ch. Tillmann: *Statistical and finite-state approaches to speech-to-speech translation*. Computer Speech and Language. 2004.
8. E. Vidal, F. Casacuberta, L. Rodríguez, J. Civera and C. Martínez *Computer-Assisted Translation Using Speech Recognition*. 2005 (to be published).

Pattern Recognition Approaches to Machine Translation

E. Vidal and F. Casacuberta

Pattern Recognition and Human Language Technology Group

Departament de Sistemes Informàtics i Computació

Institut Tecnològic d'Informàtica

Universitat Politècnica de València

2: Statistical Alignment Models

Francisco Casacuberta Nolla

`fcn@iti.upv.es`

24-28 January 2005

F. Casacuberta – DSIC-ITI-UPV

[Pattern Recognition approaches to Machine Translation](#)

[Statistical Alignment Models](#)

Index

- 1 Statistical framework to machine translation ▷ [2](#)
- 2 Alignments ▷ [11](#)
- 3 Statistical alignment models ▷ [20](#)
- 4 First-order alignment models ▷ [50](#)
- 5 Categorization in statistical modeling ▷ [66](#)
- 6 Bibliography ▷ [74](#)

Index

- 1 *Statistical framework to machine translation* ▷ 2
- 2 Alignments ▷ 11
- 3 Statistical alignment models ▷ 20
- 4 First-order alignment models ▷ 50
- 5 Categorization in statistical modeling ▷ 66
- 6 Bibliography ▷ 74

General framework

- Every sentence **y** in one language is a translation of any sentence **x** in another language.
- For each possible pair of sentences, **y** and **x**, there is a probability $\text{Pr}(\mathbf{y} \mid \mathbf{x})$.
- The probability of pairs of sentences as
quiero una habitación doble con vistas al mar # are all expenses included in the bill ?
should be low.
- The probability of pairs of sentences as
¿ hay alguna habitación tranquila libre ? # is there a quiet room available ?
should be high.

General framework

Given a source sentence x , search for the sentence \hat{y}

$$\hat{y} = \underset{y}{\operatorname{argmax}} \operatorname{Pr}(y \mid x)$$

Approaches

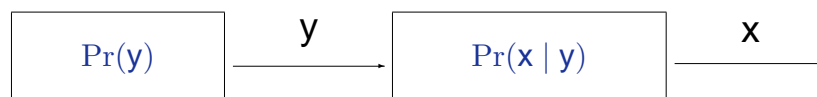
- A direct approach: *maximum entropy models*
- An inverse approach: *channel models*

An inverse approach

Given a source sentence x , search for the sentence \hat{y}

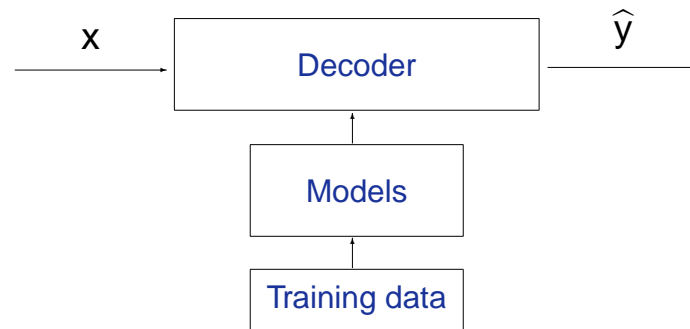
$$\hat{y} = \underset{y}{\operatorname{argmax}} \operatorname{Pr}(y \mid x) = \underset{y}{\operatorname{argmax}} \operatorname{Pr}(x \mid y) \cdot \operatorname{Pr}(y)$$

A channel model



A target-language model + alignment and lexicon models

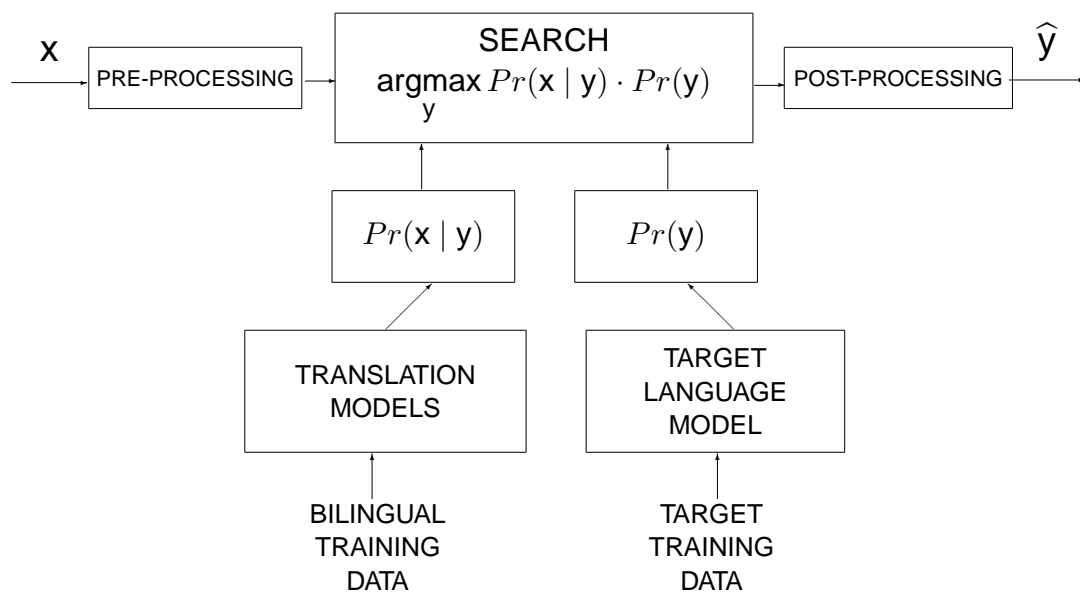
Translation search



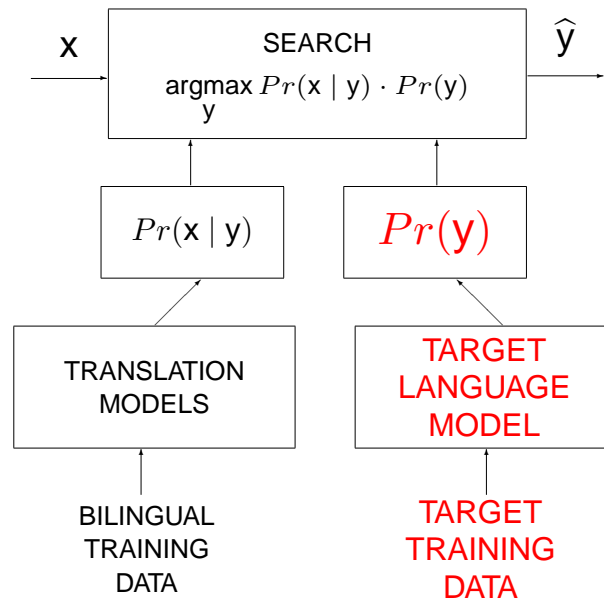
- Inverse approach:

- A target-language model: $Pr(y) \approx Pr(y)$
- Translation models (alignment and lexicon models): $Pr(x | y) \approx Pr(x | y)$
- Search procedure: $\hat{y} = \underset{y}{\operatorname{argmax}} Pr(x | y) \cdot Pr(y)$

An inverse approach



An inverse approach: The target language model



Language models

Word n-grams

$$\Pr(y) = \prod_{i=1}^{|y|} \Pr(y_i | y_1 \dots y_{i-1}) \approx Pr(y) = \prod_{i=1}^{|y|} p_n(y_i | y_{i-n+1} \dots y_{i-1})$$

n-grams of categories

$$\Pr(y) \approx Pr(y) = \prod_{i=1}^{|y|} p_n(C_i | C_{i-n+1} \dots C_{i-1}) \cdot p(y_i | C_i)$$

Regular or context-free grammars

$$\Pr(y) \approx Pr(y) = \sum_{d(y)} p_G(d(y)) \approx \max_{d(y)} p_G(d(y))$$

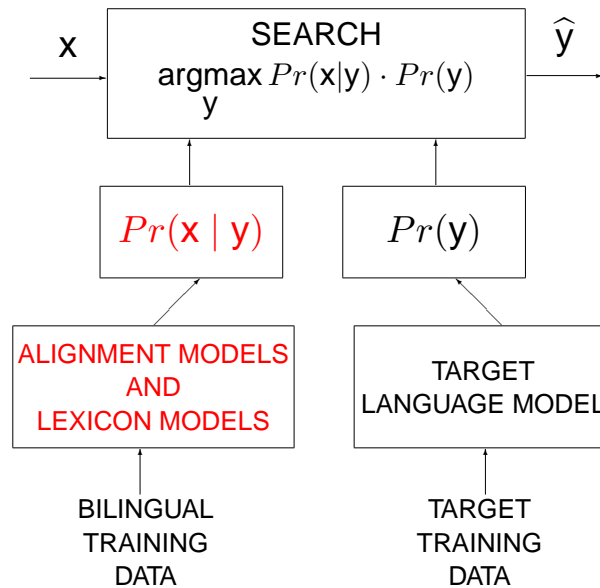
Learning language models

- Probabilistic estimation techniques.
- Grammatical inference techniques.
- **SMOOTHING**.
- Extensions: cache, triggers, categories, etc.
- Widely used toolkits for n -grams:
 - SRILM - The SRI Language Modeling Toolkit
<http://www.speech.sri.com/projects/srilm/>
 - The CMU Statistical Language Modeling (SLM) Toolkit
http://www.speech.cs.cmu.edu/SLM_info.html

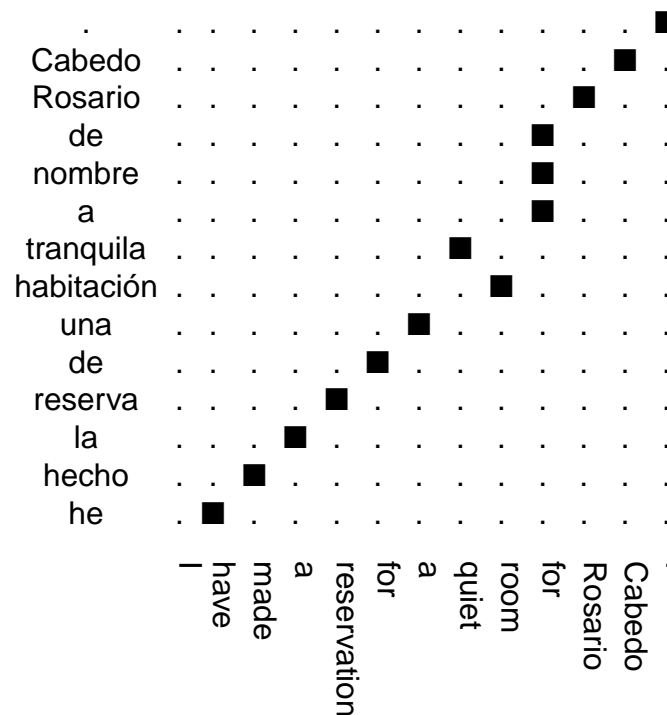
Index

- 1 Statistical framework to machine translation ▷ 2
- 2 **Alignments** ▷ 11
- 3 Statistical alignment models ▷ 20
- 4 First-order alignment models ▷ 50
- 5 Categorization in statistical modeling ▷ 66
- 6 Bibliography ▷ 74

An inverse approach



Example of word alignments



Example of word alignments

taxi	■	.	.	.
un	■
pídame	■	■	■	■
,	■	.	.
favor	■	.
por	■	.
	could	you	ask	for	a	taxi	,	please	?

Example of word alignments

H. Ney, *Statistical Natural Language Processing*, 2003: Canadian Hansards

[illegible]

Example of word alignments

AMETRA corpus

1996	.	.	■	.	.
de	.	.	■	.	.
marzo	.	.	.	■	.
de	.	.	.	■	.
20	■
a	■
,	.	■	.	.	.
Lemoa	■
En	■
		Lemoan	,	1996ko	martxoaren
					20an

Example of word alignments

METEO corpus

sud	■
meitat	■	.
seva	■	.	.
la
en	■	.	.	.
Llevant	.	.	.	■
de	.	.	■
des	.	.	■
sobretot	■	■
	sobre	todo	desde	Levante	en	su	mitad	sur

Alignments

- **Alignments:** (Brown et al. 90) $J = |x|$ y $I = |y|$

$$\mathbf{a} \subseteq \{1, \dots, J\} \times \{1, \dots, I\}$$

– Number of connections: $I \cdot J$

– Number of alignments: $2^{I \cdot J}$

- **Constrain:** $\mathbf{a} : \{1, \dots, J\} \rightarrow \{0, \dots, I\}$, ($a_j = 0 \Rightarrow j$ in x is not aligned with any position in y).

– Number of alignments: $(I + 1)^J$

- Set of possible alignments: $\mathcal{A}(\mathbf{x}, \mathbf{y})$

- The probability of translation \mathbf{y} to \mathbf{x} through an alignment \mathbf{a} is $\Pr(\mathbf{x}, \mathbf{a} \mid \mathbf{y})$

$$\Pr(\mathbf{x} \mid \mathbf{y}) = \sum_{\mathbf{a} \in \mathcal{A}(\mathbf{y}, \mathbf{x})} \Pr(\mathbf{x}, \mathbf{a} \mid \mathbf{y})$$

Alignments

$$\begin{aligned} \Pr(\mathbf{x}, \mathbf{a} \mid \mathbf{y}) &= \Pr(J \mid \mathbf{y}) \cdot \Pr(\mathbf{x}, \mathbf{a} \mid J, \mathbf{y}) \\ &= \Pr(J \mid \mathbf{y}) \cdot \Pr(\mathbf{a} \mid J, \mathbf{y}) \cdot \Pr(\mathbf{x} \mid \mathbf{a}, J, \mathbf{y}) \end{aligned}$$

- **Length probability:** $\Pr(J \mid \mathbf{y})$
- **Alignment probability:** $\Pr(\mathbf{a} \mid J, \mathbf{y})$
- **Lexicon probability:** $\Pr(\mathbf{x} \mid \mathbf{a}, J, \mathbf{y})$

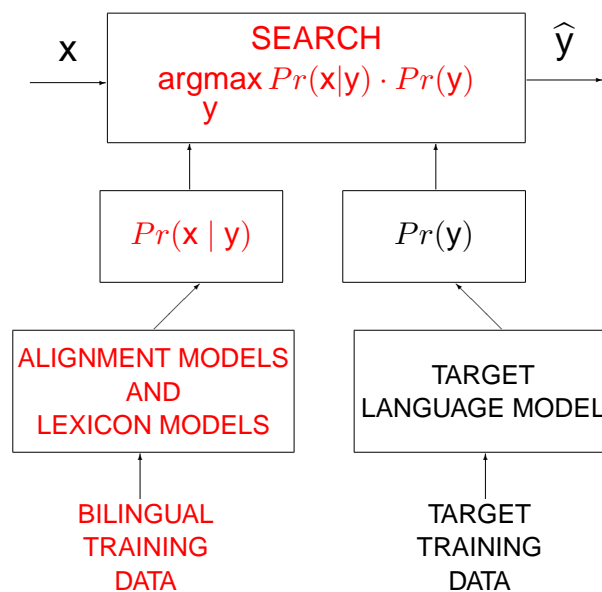
$$\Pr(\mathbf{a} \mid J, \mathbf{y}) = \prod_{j=1}^J \Pr(a_j \mid \mathbf{a}_1^{j-1}, J, \mathbf{y}) \quad \Pr(\mathbf{x} \mid \mathbf{a}, J, \mathbf{y}) = \prod_{j=1}^J \Pr(x_j \mid \mathbf{x}_1^{j-1}, \mathbf{a}, J, \mathbf{y})$$

$$\Pr(\mathbf{x}, \mathbf{a} \mid \mathbf{y}) = \Pr(J \mid \mathbf{y}) \cdot \prod_{j=1}^J \Pr(a_j \mid \mathbf{a}_1^{j-1}, \mathbf{x}_1^{j-1}, J, \mathbf{y}) \cdot \Pr(x_j \mid \mathbf{a}_1^j, \mathbf{x}_1^{j-1}, J, \mathbf{y})$$

Index

- 1 Statistical framework to machine translation ▷ 2
- 2 Alignments ▷ 11
- 3 *Statistical alignment models* ▷ 20
- 4 First-order alignment models ▷ 50
- 5 Categorization in statistical modeling ▷ 66
- 6 Bibliography ▷ 74

An inverse approach



Zero-order models

- Model 1
- Model 2
- The Viterbi approximation
- The search problem

Model 1

$$\Pr(\mathbf{x}, \mathbf{a} \mid \mathbf{y}) = \Pr(J \mid \mathbf{y}) \cdot \prod_{j=1}^J \Pr(\mathbf{a}_j \mid \mathbf{a}_1^{j-1}, \mathbf{x}_1^{j-1}, J, \mathbf{y}) \cdot \Pr(\mathbf{x}_j \mid \mathbf{a}_1^j, \mathbf{x}_1^{j-1}, J, \mathbf{y})$$

- $\Pr(J \mid \mathbf{y}) \approx n(J \mid I)$
- $\Pr(\mathbf{a}_j \mid \mathbf{a}_1^{j-1}, \mathbf{x}_1^{j-1}, J, \mathbf{y}) \approx \frac{1}{(I+1)^J}$
- $\Pr(\mathbf{x}_j \mid \mathbf{a}_1^j, \mathbf{x}_1^{j-1}, J, \mathbf{y}) \approx l(\mathbf{x}_j \mid \mathbf{y}_{\mathbf{a}_j})$

$l(\mathbf{x}_j \mid \mathbf{y}_i)$ defines a **statistical lexicon**

$$\Pr(\mathbf{x} \mid \mathbf{y}) \approx P_{M1}(\mathbf{x} \mid \mathbf{y}) = \frac{n(J \mid I)}{(I+1)^J} \prod_{j=1}^J \sum_{i=0}^I l(\mathbf{x}_j \mid \mathbf{y}_i)$$

Model 1

$$\begin{aligned}
\Pr(\mathbf{x} | \mathbf{y}) &= \sum_{\mathbf{a}} \Pr(J | \mathbf{y}) \cdot \Pr(\mathbf{x}, \mathbf{a} | J, \mathbf{y}) \\
&\approx \sum_{\mathbf{a}} n(J|I) \cdot \prod_{j=1}^J \left[\frac{1}{(I+1)^J} \cdot l(\mathbf{x}_j | \mathbf{y}_{\mathbf{a}_j}) \right] \\
&= \frac{n(J|I)}{(I+1)^J} \sum_{\mathbf{a}_1=0}^I \cdots \sum_{\mathbf{a}_J=0}^I \prod_{j=1}^J l(\mathbf{x}_j | \mathbf{y}_{\mathbf{a}_j}) \\
&= \frac{n(J|I)}{(I+1)^J} \prod_{j=1}^J \sum_{\mathbf{a}_j=0}^I l(\mathbf{x}_j | \mathbf{y}_{\mathbf{a}_j}) \\
&= \frac{n(J|I)}{(I+1)^J} \prod_{j=1}^J \sum_{i=0}^I l(\mathbf{x}_j | \mathbf{y}_i) = P_{M1}(\mathbf{x} | \mathbf{y})
\end{aligned}$$

Model 1

- $\Pr(J | \mathbf{y}) \approx n(J|I)$
- $\Pr(\mathbf{a}_j | \mathbf{a}_1^{j-1}, \mathbf{x}_1^{j-1}, J, \mathbf{y}) \approx \frac{1}{(I+1)^J}$
- $\Pr(\mathbf{x}_j | \mathbf{a}_1^j, \mathbf{x}_1^{j-1}, J, \mathbf{y}) \approx l(\mathbf{x}_j | \mathbf{y}_{\mathbf{a}_j})$

Generative process: Given a target sentence \mathbf{y} of length I ,

1. Choose the length of the source sentence J according to $n(J|I)$
2. For each $1 \leq j \leq J$, choose a position \mathbf{a}_j in the target sentence according to an uniform distribution.
3. For each $1 \leq j \leq J$ choose a source word \mathbf{x}_j according to $l(\mathbf{x}_j | \mathbf{y}_{\mathbf{a}_j})$

An example

Given y :	a	$double$	$room$	$(I = 3)$	
Choose J ($n(J 3)$): ($J = 5$)	1	2	3	4	5
Choose a_j (uniform)	1	3	2	2	2
	a	$room$	$double$	$double$	$double$
Choose x_j ($l(x_j y_i)$)	Una	habitación	con	dos	camas

Parameter estimation with Model 1

- Training sample: $A = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(K)}, y^{(K)})\}$
- Function to be maximized: likelihood

$$\mathcal{L}_A(l) = \prod_{k=1}^K P_{M1}(x^{(k)} | y^{(k)}) = \prod_{k=1}^K \frac{n(J^{(k)} | I^{(k)})}{(I^{(k)} + 1)^{J^{(k)}}} \cdot \prod_{j=1}^{J^{(k)}} \sum_{i=0}^{I^{(k)}} l(x_j^{(k)} | y_i^{(k)})$$

or the log-likelihood

$$\mathcal{L}_A(l) = \sum_{k=1}^K \sum_{j=1}^{J^{(k)}} \log \sum_{i=0}^{I^{(k)}} l(x_j^{(k)} | y_i^{(k)})$$

- Procedure: **Expectation-maximization** or **growth transformations** ($\mathcal{T}_1 : \theta \rightarrow \theta$):

$$\mathcal{T}_1(l(x | y)) = \frac{l(x | y) \cdot \left(\frac{\partial \mathcal{L}_A(l)}{\partial l(x | y)} \right)}{\sum_{x'} l(x' | y) \cdot \left(\frac{\partial \mathcal{L}_A(l)}{\partial l(x' | y)} \right)}$$

Parameter estimation with Model 1

$$\begin{aligned}
l(x | y) \frac{\partial \mathcal{L}_A(l)}{\partial l(x|y)} &= l(x|y) \sum_{k=1}^K \prod_{\substack{k'=1 \\ k' \neq k}}^K P_{M1}(\mathbf{x}^{(k')} | \mathbf{y}^{(k')}) \frac{\partial P_{M1}(\mathbf{x}^{(k)} | \mathbf{y}^{(k)})}{\partial l(x|y)} = \mathcal{L}_A(l) \sum_{k=1}^K \frac{l(x|y) \cdot \frac{\partial P_{M1}(\mathbf{x}^{(k)} | \mathbf{y}^{(k)})}{\partial l(x|y)}}{P_{M1}(\mathbf{x}^{(k)} | \mathbf{y}^{(k)})} \\
l(x | y) \cdot \frac{\partial P_{M1}(\mathbf{x}^{(k)} | \mathbf{y}^{(k)})}{\partial l(x | y)} &= l(x | y) \cdot \frac{n(J^{(k)} | I^{(k)})}{(I^{(k)} + 1)^{J^{(k)}}} \cdot \frac{\partial}{\partial l(x | y)} \prod_{j=1}^{J^{(k)}} \sum_{i=0}^{I^{(k)}} l(\mathbf{x}_j^{(k)} | \mathbf{y}_i^{(k)}) \\
&= l(x | y) \cdot \frac{n(J^{(k)} | I^{(k)})}{(I^{(k)} + 1)^{J^{(k)}}} \cdot \sum_{j=1}^{J^{(k)}} \left(\prod_{j'=1; j' \neq j}^{J^{(k)}} \sum_{i=0}^{I^{(k)}} l(\mathbf{x}_{j'}^{(k)} | \mathbf{y}_i^{(k)}) \right) \cdot \frac{\partial}{\partial l(x | y)} \sum_{i=0}^{I^{(k)}} l(\mathbf{x}_j^{(k)} | \mathbf{y}_i^{(k)}) \\
&= l(x | y) \cdot P_{M1}(\mathbf{x}^{(k)} | \mathbf{y}^{(k)}) \cdot \sum_{j=1}^{J^{(k)}} \left(\frac{1}{\sum_{i=0}^{I^{(k)}} l(\mathbf{x}_j^{(k)} | \mathbf{y}_i^{(k)})} \cdot \sum_{i=0}^{I^{(k)}} \delta(\mathbf{x}_j^{(k)}, x) \cdot \delta(\mathbf{y}_i^{(k)}, y) \right) \\
&= l(x | y) \cdot P_{M1}(\mathbf{x}^{(k)} | \mathbf{y}^{(k)}) \cdot \sum_{j=1}^{J^{(k)}} \left(\frac{\delta(\mathbf{x}_j^{(k)}, x)}{\sum_{i=0}^{I^{(k)}} l(\mathbf{x}_j^{(k)} | \mathbf{y}_i^{(k)})} \right) \cdot \#(y, \mathbf{y}^{(k)}) \\
&= P_{M1}(\mathbf{x}^{(k)} | \mathbf{y}^{(k)}) \cdot l(x | y) \cdot \frac{\#(x, \mathbf{x}^{(k)}) \cdot \#(y, \mathbf{y}^{(k)})}{\sum_{i=0}^{I^{(k)}} l(x | \mathbf{y}_i^{(k)})}
\end{aligned}$$

Parameter estimation in Model 1

Iterative *E-M* procedure:

Expectation step:

$$c(x | y; \mathbf{x}^{(k)}, \mathbf{y}^{(k)}) = \frac{l(x | y)}{\sum_{i=0}^{I^{(k)}} l(x | \mathbf{y}_i^{(k)})} \cdot \#(y, \mathbf{y}^{(k)}) \cdot \#(x, \mathbf{x}^{(k)})$$

Maximization step:

$$\mathcal{T}_1(l(x | y)) = \frac{\sum_{k=1}^K c(x | y; \mathbf{x}^{(k)}, \mathbf{y}^{(k)})}{\sum_{x'} \sum_{k=1}^K c(x' | y; \mathbf{x}^{(k)}, \mathbf{y}^{(k)})}$$

Parameter estimation in Model 1

- PROPERTY: the increase of the likelihood of the training set in each iteration:

$$\prod_{k=1}^K P_{M1}(\mathbf{x}^{(k)} \mid \mathbf{y}^{(k)}) \leq \prod_{k=1}^K P_{T_1(M1)}(\mathbf{x}^{(k)} \mid \mathbf{y}^{(k)})$$

- PROPERTY: Eventually an **absolute maximum** is achieved!
- COMPUTATIONAL COST OF T_1 : If $I_M = \max_k I^{(k)}$ y $J_M = \max_k J^{(k)}$
 - time: $O(K \times (I_M + J_M))$
 - space: $O(|\Sigma| \times |\Delta|)$

Model 2

$$\Pr(\mathbf{x}, \mathbf{a} \mid \mathbf{y}) = \Pr(J \mid \mathbf{y}) \cdot \prod_{j=1}^J \Pr(\mathbf{a}_j \mid \mathbf{a}_1^{j-1}, \mathbf{x}_1^{j-1}, J, \mathbf{y}) \cdot \Pr(\mathbf{x}_j \mid \mathbf{a}_1^j, \mathbf{x}_1^{j-1}, J, \mathbf{y})$$

- $\Pr(J \mid \mathbf{y}) \approx n(J|I)$
- $\Pr(\mathbf{a}_j \mid \mathbf{a}_1^{j-1}, \mathbf{x}_1^{j-1}, J, \mathbf{y}) \approx a(\mathbf{a}_j \mid j, J, I)$
- $\Pr(\mathbf{x}_j \mid \mathbf{a}_1^j, \mathbf{x}_1^{j-1}, J, \mathbf{y}) \approx l(\mathbf{x}_j \mid \mathbf{y}_{\mathbf{a}_j})$

$l(\mathbf{x}_j \mid \mathbf{y}_i)$ defines a **statistical lexicon**

$a(i \mid j, J, I)$ defines **statistical alignments**

$$\Pr(\mathbf{x} \mid \mathbf{y}) \approx P_{M2}(\mathbf{x} \mid \mathbf{y}) = n(J|I) \cdot \prod_{j=1}^J \sum_{i=0}^I a(i \mid j, J, I) \cdot l(\mathbf{x}_j \mid \mathbf{y}_i)$$

Model 2

$$\begin{aligned}
\Pr(\mathbf{x} | \mathbf{y}) &= \sum_{\mathbf{a}} \Pr(J | \mathbf{y}) \cdot \Pr(\mathbf{x}, \mathbf{a} | J, \mathbf{y}) \\
&= \sum_{\mathbf{a}} n(J | I) \cdot \prod_{j=1}^J \left[a(\mathbf{a}_j | j, J, I) \cdot l(\mathbf{x}_j | \mathbf{y}_{\mathbf{a}_j}) \right] \\
&= n(J | I) \cdot \sum_{\mathbf{a}_1=0}^I \cdots \sum_{\mathbf{a}_J=0}^I \prod_{j=1}^J \left[a(\mathbf{a}_j | j, J, I) \cdot l(\mathbf{x}_j | \mathbf{y}_{\mathbf{a}_j}) \right] \\
&= n(J | I) \cdot \prod_{j=1}^J \sum_{\mathbf{a}_j=0}^I a(\mathbf{a}_j | j, J, I) \cdot l(\mathbf{x}_j | \mathbf{y}_{\mathbf{a}_j}) \\
&= n(J | I) \cdot \prod_{j=1}^J \sum_{i=0}^I a(i | j, J, I) \cdot l(\mathbf{x}_j | \mathbf{y}_i) = P_{M2}(\mathbf{x} | \mathbf{y})
\end{aligned}$$

Model 2

- $\Pr(J | \mathbf{y}) \approx n(J|I)$
- $\Pr(\mathbf{a}_j | \mathbf{a}_1^{j-1}, \mathbf{x}_1^{j-1}, J, \mathbf{y}) \approx a(\mathbf{a}_j | j, J, I)$
- $\Pr(\mathbf{x}_j | \mathbf{a}_1^j, \mathbf{x}_1^{j-1}, J, \mathbf{y}) \approx l(\mathbf{x}_j | \mathbf{y}_{\mathbf{a}_j})$

Generative process: Given a target sentence \mathbf{y} of length I ,

1. Choose the length of the source sentence J according to $n(J|I)$.
2. For each $1 \leq j \leq J$, choose a position \mathbf{a}_j in the target sentence according to $a(\mathbf{a}_j | j, J, I)$.
3. For each $1 \leq j \leq J$ choose a source word \mathbf{x}_j according to $l(\mathbf{x}_j | \mathbf{y}_{\mathbf{a}_j})$.

An example

Given y :	a	$double$	$room$	$(I = 3)$	
Choose J ($n(J 3)$): ($J = 5$)	1	2	3	4	5
Choose a_j ($a(a_j j, I, J)$)	1 a	3 room	2 double	2 double	2 double
Choose x_j ($l(x_j y_i)$)	Una	habitación	con	dos	camas

Parameter estimation in Model 2

- Training sample: $A = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(K)}, y^{(K)})\}$
- Function to be maximized: likelihood

$$\begin{aligned}
 \mathcal{L}_A(a, l) &= \prod_{k=1}^K P_{M2}(x^{(k)} | y^{(k)}) \\
 &= \prod_{k=1}^K n(J^{(k)} | I^{(k)}) \cdot \prod_{j=1}^{J^{(k)}} \sum_{i=0}^{I^{(k)}} a(i | j, J^{(k)}, I^{(k)}) \cdot l(x_j^{(k)} | y_i^{(k)})
 \end{aligned}$$

or the log-likelihood:

$$\mathcal{L}_A(a, l) = \sum_{k=1}^K \sum_{j=1}^{J^{(k)}} \log \sum_{i=0}^{I^{(k)}} a(i | j, J^{(k)}, I^{(k)}) \cdot l(x_j^{(k)} | y_i^{(k)})$$

- Procedure: **Expectation-maximization** or **growth transformations** ($\mathcal{T}_2 : \theta \rightarrow \theta$)

Parameter estimation in Model 2

Iterative *E-M* procedure:

Expectation step:

$$c(x | y; \mathbf{x}^{(k)}, \mathbf{y}^{(k)}) = \sum_{j=1}^{J^{(k)}} \sum_{i=0}^{I^{(k)}} \frac{l(x | y) \cdot a(i | j, J^{(k)}, I^{(k)}) \cdot \delta(x, \mathbf{x}_j^{(k)}) \cdot \delta(y, \mathbf{y}_i^{(k)})}{\sum_{n=0}^{I^{(k)}} l(x | \mathbf{y}_n) \cdot a(n | j, J^{(k)}, I^{(k)})}$$

$$c(i | j; J, I, \mathbf{x}^{(k)}, \mathbf{y}^{(k)}) = \begin{cases} \frac{l(\mathbf{x}_j^{(k)} | \mathbf{y}_i^{(k)}) \cdot a(i | j, J^{(k)}, I^{(k)})}{\sum_{i'=0}^{I^{(k)}} l(\mathbf{x}_j^{(k)} | \mathbf{y}_{i'}^{(k)}) \cdot a(i' | j, J^{(k)}, I^{(k)})} & \text{if } I = I^{(k)} \\ & \text{and } J = J^{(k)} \\ 0 & \text{otherwise} \end{cases}$$

Maximization step:

$$\mathcal{T}_2(l(x | y)) = \frac{\sum_{k=1}^K c(x | y; \mathbf{x}^{(k)}, \mathbf{y}^{(k)})}{\sum_{x'} \sum_{k=1}^K c(x' | y; \mathbf{x}^{(k)}, \mathbf{y}^{(k)})}$$

$$\mathcal{T}_2(a(i | j, J, I)) = \frac{\sum_{k=1}^K c(i | j; J, I, \mathbf{x}^{(k)}, \mathbf{y}^{(k)})}{\sum_{i'} \sum_{k=1}^K c(i' | j; J, I, \mathbf{x}^{(k)}, \mathbf{y}^{(k)})}$$

Parameter estimation in Model 2

- PROPERTY: the increase of the likelihood of the training set in each iteration.

$$\prod_{k=1}^K P_{M2}(\mathbf{x}^{(k)} | \mathbf{y}^{(k)}) \leq \prod_{k=1}^K P_{\mathcal{T}_2(M2)}(\mathbf{x}^{(k)} | \mathbf{y}^{(k)})$$

- PROPERTY: Eventually an **local maximum** is achieved.
- COMPUTATIONAL COST OF \mathcal{T}_2 : If $I_M = \max_k I^{(k)}$ y $J_M = \max_k J^{(k)}$
 - time: $O(K \times I_M \times J_M)$
 - space: $O((|\Sigma| \times |\Delta|) + I_M + J_M)$

Optimal alignment with Model 2

$$P_{M2}(\mathbf{x} | \mathbf{y}) = n(J|I) \cdot \prod_{j=1}^J \sum_{i=0}^I a(i | j, J, I) \cdot l(\mathbf{x}_j | \mathbf{y}_i) \approx$$

$$\hat{P}_{M2}(\mathbf{x} | \mathbf{y}) = n(J | I) \cdot \prod_{j=1}^J \max_{0 \leq i \leq I} a(i | j, I, J) \cdot l(\mathbf{x}_j | \mathbf{y}_i)$$

Algorithm Viterbi ($\mathbf{x}, \mathbf{y}, l, a$)

Input: A pair \mathbf{x}, \mathbf{y} and the parameters l and a of Model 2

Output: An optimal alignment A between \mathbf{x} and \mathbf{y} .

For $j := 1$ **until** J

$A[j] := \operatorname{argmax}_{0 \leq i \leq I} a(i | j, J, I) \cdot l(\mathbf{x}_j | \mathbf{y}_i)$

End-for

Return: A

The computational cost of this algorithm is $O(J \times I)$.

Examples of alignments

EUTRANS-I corpus: Spanish-English

- **Vocabulary:** 680 Spanish words, and 513 English words.
- **Training:** 10,000 pairs (97,000/99,000 words).

An example

1	2	3	4	5	6	7	8	9	10
por	favor	,	¿	podría	ver	alguna	habitación	tranquila	?

- MODEL 1, ITERATION 5
could (5) I (6) see (6) a (7) quiet (9) room (8) , (3) please (2) ? (4)
- MODEL 2, ITERATION 2
could (5) I (6) see (6) a (7) quiet (9) room (8) , (3) please (3) ? (10)

Examples of alignments

MODEL 2 ITERATION 2

por favor , he hecho una reserva a nombre de Federico Redondo .

I (4) have (4) made (5) a (6) reservation (5) for (9) Federico (11) Redondo (12) . (0)

por favor , ¿ podría pedir nuestro taxi ?

could (5) you (4) ask (6) for (6) our (7) taxi (8) , (3) please (3) ? (9)

¿ les importaría despertarnos mañana a las siete y cuarto , por favor ?

would (2) you (1) mind (3) waking (4) us (4) up (6) tomorrow (5) at (7) a (9) quarter (10) past (9) seven (8) , (13) please (13) ? (1)

me voy a ir el jueves tres de junio a la una y media de la tarde .

I (2) am (2) leaving (2) on (5) Thursday (6) June (9) the (5) third (9) at (10) half (14) past (13) one (11) in (4) the (11) afternoon (17) . (18)

Viterbi estimation

- Training sample: $A = \{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), (\mathbf{x}^{(2)}, \mathbf{y}^{(2)}), \dots, (\mathbf{x}^{(K)}, \mathbf{y}^{(K)})\}$
- Function to be maximized: Viterbi score
- Procedure:

$$\text{VITERBI}(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}, l, a) \forall k : 1 \leq k \leq K$$

$$\hat{c}(x | y; \mathbf{x}^{(k)}, \mathbf{y}^{(k)}) = \#(x, \mathbf{x}^{(k)}) \times \#(y, \mathbf{y}^{(k)})$$

$$\hat{c}(i | j; J, I, \mathbf{x}^{(k)}, \mathbf{y}^{(k)}) = \begin{cases} 1 & \text{if } i = a_j \text{ in } \text{VITERBI}(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}, l, a) \text{ and if} \\ & J^{(k)} = J \text{ and } I^{(k)} = I \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{T}_v(l(x | y)) = \frac{\sum_{k=1}^K \hat{c}(x | y; \mathbf{x}^{(k)}, \mathbf{y}^{(k)})}{\sum_{x'} \sum_{k=1}^K \hat{c}(x' | y; \mathbf{x}^{(k)}, \mathbf{y}^{(k)})}$$

$$\mathcal{T}_v(a(i | j, J, I)) = \frac{\sum_{k=1}^K \hat{c}(i | j; J, I, \mathbf{x}^{(k)}, \mathbf{y}^{(k)})}{\sum_{i'} \sum_{k=1}^K \hat{c}(i' | j; J, I, \mathbf{x}^{(k)}, \mathbf{y}^{(k)})}$$

Viterbi estimation

- PROPERTY: the increasing of the Viterbi score in each iteration:

$$\prod_{k=1}^K \hat{P}_{M2}(\mathbf{x}^{(k)} | \mathbf{y}^{(k)}) \leq \prod_{k=1}^K \hat{P}_{\mathcal{T}_v(M2)}(\mathbf{x}^{(k)} | \mathbf{y}^{(k)})$$

- PROPERTY: Eventually a **local maximum** is achieved.
- COMPUTATIONAL COST OF \mathcal{T}_v : If $I_M = \max_k I^{(k)}$ y $J_M = \max_k J^{(k)}$
 - time: $O(K \times I_M \times J_M)$
 - space: $O((|\Sigma| \times |\Delta|) + I_M \times J_M)$

Simplified version of Model 2 (Ney et al. 2000)

$$\Pr(\mathbf{x}, \mathbf{a} | \mathbf{y}) = \Pr(J | \mathbf{y}) \cdot \prod_{j=1}^J \Pr(\mathbf{a}_j | \mathbf{a}_1^{j-1}, \mathbf{x}_1^{j-1}, J, \mathbf{y}) \cdot \Pr(\mathbf{x}_j | \mathbf{a}_1^j, \mathbf{x}_1^{j-1}, J, \mathbf{y})$$

- $\Pr(J | \mathbf{y}) \approx n(J | I)$
- $\Pr(\mathbf{a}_j | \mathbf{a}_1^{j-1}, \mathbf{x}_1^{j-1}, J, \mathbf{y}) \approx a(\mathbf{a}_j | j, I)$
- $\Pr(\mathbf{x}_j | \mathbf{a}_1^j, \mathbf{x}_1^{j-1}, J, \mathbf{y}) \approx l(\mathbf{x}_j | \mathbf{y}_{\mathbf{a}_j})$

$$\Pr(\mathbf{x} | \mathbf{y}) \approx P^A(\mathbf{x} | \mathbf{y}) = n(J | I) \cdot \prod_{j=1}^J \sum_{i=1}^I a(i | j, I) \cdot l(\mathbf{x}_j | \mathbf{y}_i)$$

$$\mathcal{T}_A(t(x | \mathbf{y})) = \frac{\sum_{k=1}^K c(x | \mathbf{y}; \mathbf{x}^{(k)}, \mathbf{y}^{(k)})}{\sum_{x'} \sum_{k=1}^K c(x' | \mathbf{y}; \mathbf{x}^{(k)}, \mathbf{y}^{(k)})} \text{ and } \mathcal{T}_A(a(i | j, J)) = \frac{r(i - j \frac{I}{J})}{\sum_{i'=1}^I r(i' - j \frac{I}{J})}$$

It is assumed that the diagonal of the plain (j,i) is the dominant factor.

The translation process: searching

$$\operatorname{argmax}_y Pr(x | y) \cdot Pr(y)$$

A computational difficult problem

(K.Knight *Decoding complexity in word-replacement translation models*. Comp. Ling. 1999)

ALGORITHMIC SOLUTIONS:

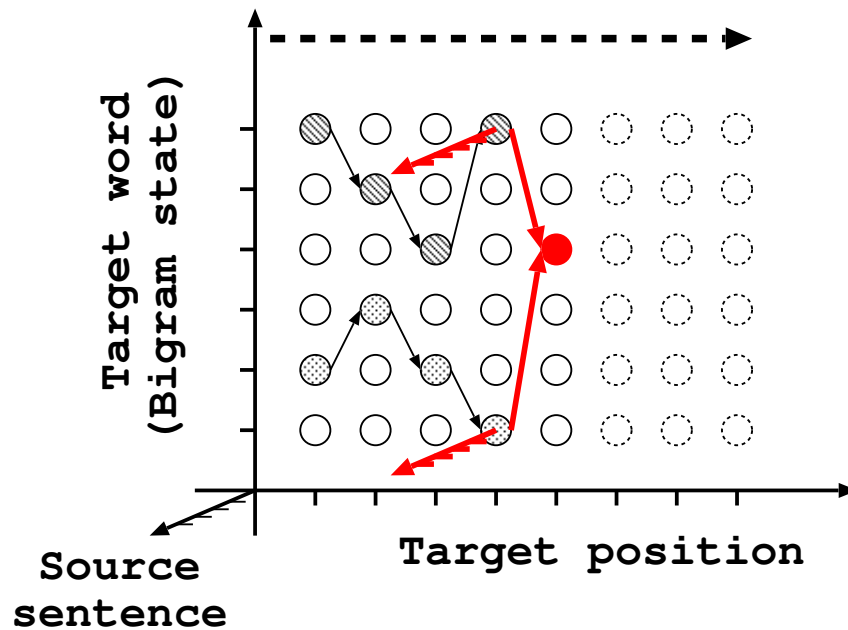
- **Dynamic Programming like** (Garcia-Varea, 1998) (Ney, 2000)
- **Stack-Decoding**, A* or Branch & Bound(Brown, 1990)(Wang, 1997)
- **Greedy** (Germann, 2001)
- **Using finite-state transducers** (Kumar, 2004)

Dynamic-programming like search

An approximate solution: DPSearchM2

- Characteristics:
 - Building partial hypothesis $(1, 2, \dots, I)$
 - Search graph: a $|\mathcal{E}| \times I$ *trellis*.
- Assumptions:
 - Language model: n -grams (bigrams)
 - The length of target sentence is known: I .

Dynamic-programming like search: *DPSearchM2*



Dynamic-programming like search: *DPSearchM2*

- Search criterium:

$$\max_y \left\{ \prod_{i=1}^I p_2(y_i | y_{i-1}) \cdot \prod_{j=1}^J \sum_{i=0}^I l(\mathbf{x}_j | y_i) \cdot a(i|j, J, I) \right\}$$

- Auxiliar variables:

- $Q(e, i, j)$: Contribution of the translation models for each j :

$$Q(t, i, j) = l(\mathbf{x}_j | t) \cdot a(i|j, J, I) + \sum_{k=0}^{i-1} l(\mathbf{x}_j | y_k) \cdot a(k|j, J, I)$$

- $T(t, i)$: Contribution of the language model:

$$T(t, i) = p_2(t | y_{i-1}) \prod_{k=1}^{i-1} p(y_k | y_{k-1})$$

Dynamic-programming like search: *DPSearchM2*

- Recursion:

$$Q(t, i, j) = Q(\hat{t}, i-1, j) + l(\mathbf{x}_j | \hat{t}) \cdot a(i|j, J, I)$$

$$T(t, i) = T(\hat{t}, i-1) \cdot p_2(t | \hat{t})$$

where \hat{t} is the optimal state in $i-1$

$$\hat{t} = \operatorname{argmax}_{t'} T(t', i-1) \cdot p_2(t | t') \cdot \prod_{j=1}^J (Q(t', i-1, j) + l(\mathbf{x}_j | t') \cdot a(i|j, J, I))$$

- Basis of the recursion $\forall t \forall j : 1 \leq j \leq J$:

$$Q(t, 1, j) = l(\mathbf{x}_j | y_0) \cdot a(0|j, J, I)$$

$$T(t, 1) = 1.0$$

- An approximation to the optimal solution is:

$$\hat{y} = \operatorname{argmax}_y \left\{ T(y_I | I) \cdot \prod_{j=1}^J Q(y_I, I, j) \right\}$$

Dynamic-programming like search: *DPSearchM2*

- Problem: in i , y_{i+1}^I is unknown:

$$Q(t, i, j) = l(\mathbf{x}_j | t) \cdot a(i|j, J, I) + \sum_{k=0; k \neq i}^I l(\mathbf{x}_j | y_k) \cdot a(k|j, J, I)$$

- Solution: Iterative search

$$Q(t, i, j) = Q(\hat{t}(t, i), i-1, j) + l(\mathbf{x}_j | \hat{t}(t, i)) \cdot a(i|j, J, I) + R(j, i+1)$$

$$R(j, i) = \sum_{k=i}^I l(\mathbf{x}_j | \tilde{y}_k) a(k|j, J, I); \quad \tilde{y}_1^I \text{ is the last optimal solution}$$

- Initialization: $R(j, i) = 0$
- But

$$R(i, j) = \sum_{k=i}^I \max_t \{ l(\mathbf{x}_j | t) \cdot a(k|j, J, I) \} \rightarrow \text{Heuristic initialization}$$

Index

- 1 Statistical framework to machine translation ▷ 2
- 2 Alignments ▷ 11
- 3 Statistical alignment models ▷ 20
- 4 *First-order alignment models* ▷ 50
- 5 Categorization in statistical modeling ▷ 66
- 6 Bibliography ▷ 74

First-order models

- Homogeneous HMM model (HMM)
- Search: Quasi-monotone alignments
- Search: Inverted alignments
- Results
- Other search solution

Homogeneous HMM alignment

(H. Ney et al. *Algorithms for statistical translation of spoken language*. IEEE TSAP. 2000.)

$$\Pr(\mathbf{x}, \mathbf{a} \mid \mathbf{y}) = \Pr(J \mid \mathbf{y}) \cdot \prod_{j=1}^J \Pr(\mathbf{a}_j \mid \mathbf{a}_1^{j-1}, \mathbf{x}_1^{j-1}, J, \mathbf{y}) \cdot \Pr(\mathbf{x}_j \mid \mathbf{a}_1^j, \mathbf{x}_1^{j-1}, J, \mathbf{y})$$

- $\Pr(J \mid \mathbf{y}) \approx n(J \mid I)$
- $\Pr(\mathbf{a}_j \mid \mathbf{a}_1^{j-1}, \mathbf{x}_1^{j-1}, J, \mathbf{y}) \approx h(\mathbf{a}_j \mid \mathbf{a}_{j-1}, J, I)$
- $\Pr(\mathbf{x}_j \mid \mathbf{a}_1^j, \mathbf{x}_1^{j-1}, J, \mathbf{y}) \approx l(\mathbf{x}_j \mid \mathbf{y}_{\mathbf{a}_j})$

$h(\mathbf{a}_j \mid \mathbf{a}_{j-1}, J, I)$ defines **statistical alignment with first-order dependencies**

$$P_{HMM}(\mathbf{x} \mid \mathbf{y}) = n(J \mid I) \cdot \sum_{\mathbf{a}} \prod_{j=1}^J h(\mathbf{a}_j \mid \mathbf{a}_{j-1}, J, I) \cdot l(\mathbf{x}_j \mid \mathbf{y}_{\mathbf{a}_j})$$

Homogeneous HMM alignment

Forward computation of $P_{HMM}(\mathbf{x} \mid \mathbf{y})$

$$P_{HMM}(\mathbf{x} \mid \mathbf{y}) = n(J \mid I) \cdot \sum_{\mathbf{a}} \prod_{j=1}^J h(\mathbf{a}_j \mid \mathbf{a}_{j-1}, J, I) \cdot l(\mathbf{x}_j \mid \mathbf{y}_{\mathbf{a}_j}) = n(J \mid I) \cdot Q(I, J)$$

with

$$Q(i, j) = l(\mathbf{x}_j \mid \mathbf{y}_i) \cdot \sum_{i'} h(i \mid i', I, J) \cdot Q(i', j-1)$$

MAXIMUM APPROACH

$$P_{HMM}(\mathbf{x} \mid \mathbf{y}) \approx n(J \mid I) \cdot \max_{\mathbf{a}} \prod_{j=1}^J h(\mathbf{a}_j \mid \mathbf{a}_{j-1}, J, I) \cdot l(\mathbf{x}_j \mid \mathbf{y}_{\mathbf{a}_j})$$

Viterbi computation in the maximum approach

$$\hat{Q}(i, j) = l(\mathbf{x}_j \mid \mathbf{y}_i) \cdot \max_{i'} \left(h(i \mid i', I, J) \cdot \hat{Q}(i', j-1) \right)$$

Homogeneous HMM alignment

ALIGNMENT PROBABILITY DISTRIBUTION:

$$h(i|i', I, J) = \frac{q(i - i')}{\sum_{i''=1}^I q(i'' - i')}$$

TRAINING WITH THE MAXIMUM APPROACH

- Position alignment by computing $\hat{Q}(i, j)$
- Parameter estimation (relative frequencies)

Searching with homogeneous HMM alignments

$$\max_y \Pr(y) \cdot \Pr(x | y) \approx$$

$$\max_I \left\{ n(J | I) \cdot \max_{y \in \Delta^I} \left(\prod_{i=1}^I p_2(y_i | y_{i-1}) \cdot \max_a \prod_{j=1}^J \left[h(a_j | a_{j-1}, J) \cdot l(x_j | y_{a_j}) \right] \right) \right\}$$

- Quasi-monotone alignments and quasi-monotone search.

$$p_2(y_i | y_{i-1}) \Rightarrow p[a_j - a_{j-1}](y_{a_j} | y_{a_{j-1}})$$

- Inverted alignments and search.

$$h(a_j | a_{j-1}, J) \cdot l(x_j | y_{a_j}) \Rightarrow q(i | b_i, J, I) \cdot t(x_{b_i} | y_i)$$

Quasi-monotone alignments and quasi-monotone search

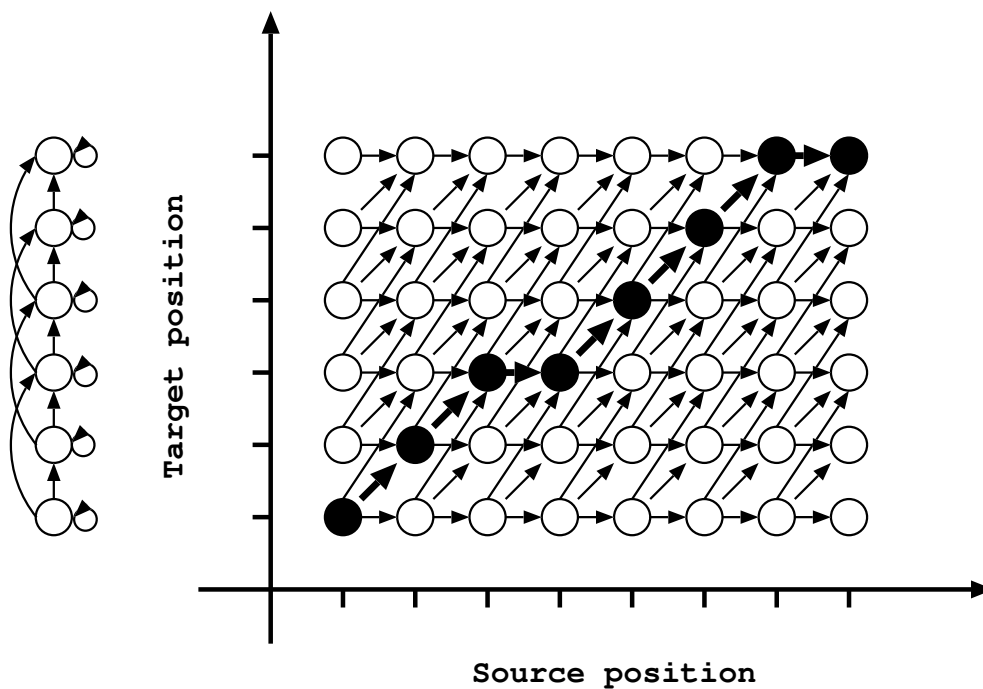
$$\delta \equiv a_j - a_{j-1} \in \{0, 1, 2\}$$

Modification of the target language model

- If $\delta = 0$, $p_{[\delta]}(y \mid y') = \begin{cases} 1 & y = y' \\ 0 & y \neq y' \end{cases}$
- If $\delta = 1$, $p_{[\delta]}(y \mid y') = p_2(y \mid y')$
- If $\delta = 2$, $p_{[\delta]}(y \mid y') = \max_{y''} (p_2(y \mid y'') \cdot p_2(y'' \mid y'))$

$$\max_I \left\{ n(J \mid I) \cdot \max_{y, a} \left\{ \prod_{j=1}^J h(a_j \mid a_{j-1}, J) \cdot p_{[a_j - a_{j-1}]}(y_{a_j} \mid y_{a_{j-1}}) \cdot l(x_j \mid y_{a_j}) \right\} \right\}$$

Quasi-monotone alignments and quasi-monotone search



Quasi-monotone alignments and quasi-monotone search

(C.Tillmann. *Word re-ordering and DP based search algorithm for SMT*. Ph.D. Thesis. 2001.)

$$\max_I \left\{ n(J | I) \cdot \max_{y, a} \left\{ \prod_{j=1}^J h(a_j | a_{j-1}, J) \cdot p_{[a_j - a_{j-1}]}(y_{a_j} | y_{a_{j-1}}) \cdot l(x_j | y_{a_j}) \right\} \right\}$$

$Q(i, j, s)$ = the probability of the best partial hypothesis (y_1^i, a_1^j) with $y_i = y$ and $a_j = i$.

$$Q(i, j, s) = t(x_j | s) \cdot \max_{\delta, e'} (h(i | i - \delta, I) \cdot p_{[\delta]}(s | s') \cdot Q(i - \delta, j - 1, s'))$$

Solution

$$\max_{I, \hat{s}} (n(J | I) \cdot Q(I, J, \hat{s}))$$

Computational cost: $O(I_{max} \cdot J \cdot |\Delta|^2)$

Quasi-monotone alignments and quasi-monotone search

(H. Ney et al. *Algorithms for statistical translation of spoken language*. IEEE TSAP. 2000.)

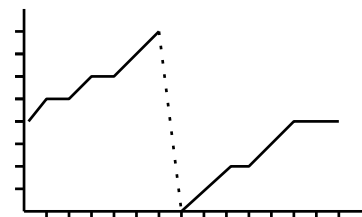
- Problem with the monotone models: assumption of similar syntactic structures in both languages.

– First solution: Re-ordering and monotone models.

– Second solution: Two-level monotone alignments:

– Third solution: a new alignment model

- * Concept of **inverted alignment**: $b_i = j \Rightarrow y_i \Leftrightarrow x_j$
- * Associated distribution: $q(i | b_i, J, I)$
- * (+ optional) **Fertility**



Inverted alignments

H. Ney, *Statistical Natural Language Processing*, STC Doctorate Program, UPC. 2003

An **inverted alignment** is: $i \rightarrow B_i \subset \{1, \dots, j, \dots, J\}$

$$\begin{aligned}
 \Pr(\mathbf{x}, \mathbf{b} \mid \mathbf{y}) &= \Pr(J \mid \mathbf{y}) \cdot \Pr(\mathbf{x}, \mathbf{b} \mid J, \mathbf{y}) \\
 &= \Pr(J \mid \mathbf{y}) \cdot \prod_{i=1}^I \Pr(\mathbf{x}_{\mathbf{b}_i}, \mathbf{b}_i \mid \mathbf{x}_{\mathbf{b}_1}^{\mathbf{b}_1^{i-1}}, \mathbf{b}_1^{i-1}, J, \mathbf{y}) \\
 &= \Pr(J \mid \mathbf{y}) \cdot \prod_{i=1}^I \left(\Pr(\mathbf{b}_i \mid \mathbf{x}_{\mathbf{b}_1}^{\mathbf{b}_1^{i-1}}, \mathbf{b}_1^{i-1}, J, \mathbf{y}) \cdot \Pr(\mathbf{x}_{\mathbf{b}_i} \mid \mathbf{x}_{\mathbf{b}_1}^{\mathbf{b}_1^{i-1}}, \mathbf{b}_1^{i-1}, J, \mathbf{y}) \right) \\
 &\approx n(J \mid I) \cdot \prod_{i=1}^I \left(q(\mathbf{b}_i \mid \mathbf{b}_1^{i-1}; \mathbf{x}_{\mathbf{b}_i}, \mathbf{y}_{i-1}) \cdot l(\mathbf{x}_{\mathbf{b}_i} \mid \mathbf{y}_i) \right) \\
 &= n(J \mid I) \cdot \prod_{i=1}^I \left(q(\mathbf{b}_i \mid \mathbf{b}_1^{i-1}; \mathbf{x}_{\mathbf{b}_i}, \mathbf{y}_{i-1}) \cdot \prod_{j \in \mathbf{b}_i} l(\mathbf{x}_j \mid \mathbf{y}_i) \right)
 \end{aligned}$$

Inverted alignments and search

$$\max_I \left\{ n(J \mid I) \cdot \max_{\mathbf{y}, \mathbf{b}} \left\{ \prod_{i=1}^I [p_2(\mathbf{y}_i \mid \mathbf{y}_{i-1}) \cdot q(i \mid \mathbf{b}_i, J, I) \cdot l(\mathbf{x}_{\mathbf{b}_i} \mid \mathbf{y}_i)] \right\} \right\}$$

$Q_I(i, j, y)$ = probability of the best partial hypothesis $(\mathbf{y}_1^i, \mathbf{b}_1^i)$ con $\mathbf{y}_i = y$ y $\mathbf{b}_i = j$.

General recursion:

$$Q_I(i, j, y) = l(\mathbf{x}_j \mid y) \cdot q(i \mid j, I, J) \cdot \max_{j', y'} (p_2(y \mid y') \cdot Q_I(i-1, j', y'))$$

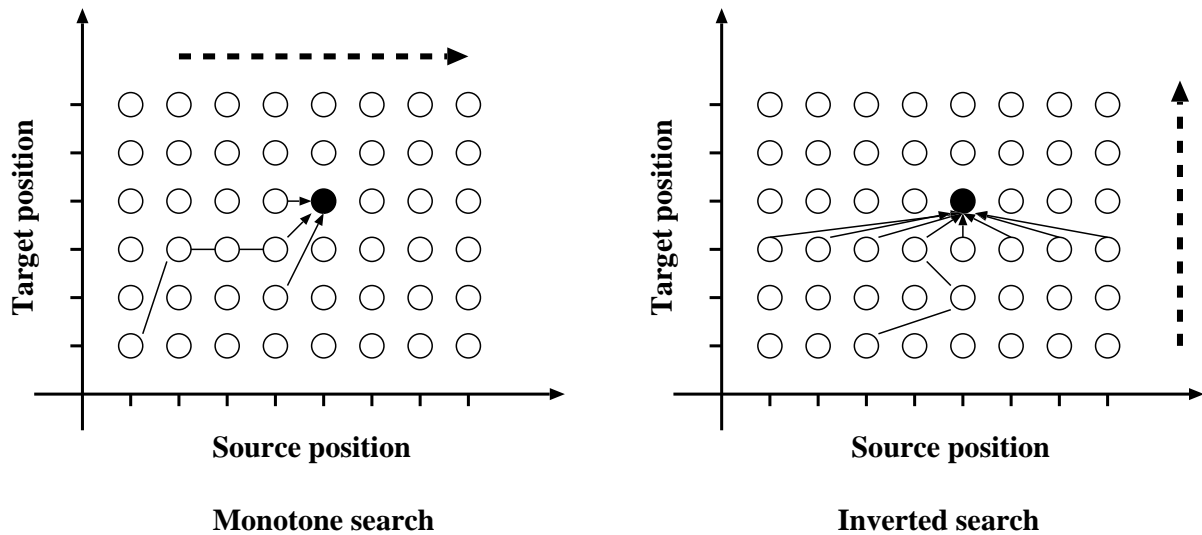
Solution:

$$\max_{I, \hat{y}} (n(J \mid I) \cdot Q_I(I, J, \hat{y}))$$

Computational cost: $O(I_{max}^2 \cdot J^2 \cdot |\Delta|^2)$

Inverted alignments

(H. Ney et al. *Algorithms for statistical translation of spoken language*. IEEE TSAP. 2000.)



Results

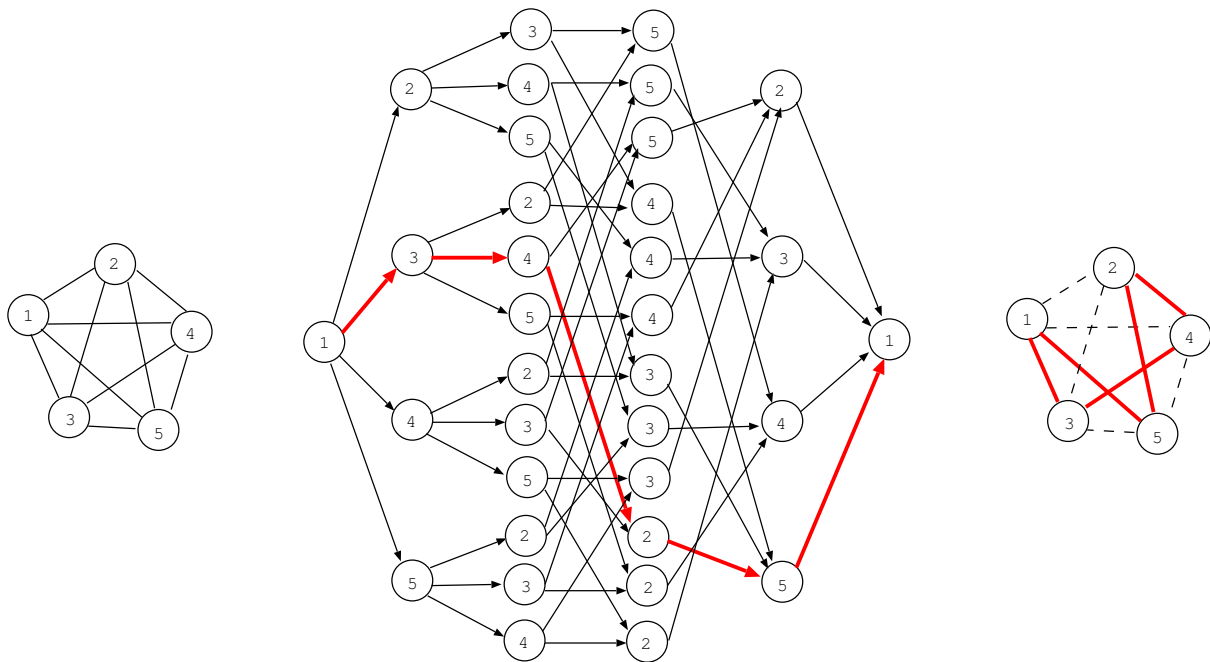
EuTrans-I corpus (Spanish-English)

- Vocabulary: 680 Spanish words, and 513 English words.
- Training: 10,000 pairs (97,000/99,000 words).
- Test: 2,996 pairs (PP=8.6/5.2) (35,000/35,590 words).
- Manual categories: 7.

Model	WER
Quasi-monotone search	10.8
DP-search with M2	13.9

Word error rate (WER): The minimum number of substitution, insertion and deletion operations needed to convert the word string hypothesized by the translation system into a given single reference word string.

Dynamic programming approach: The traveling salesman problem



Cities \equiv source positions

Dynamic programming approach: The traveling salesman problem

(H. Ney, *Statistical Natural Language Processing*, STC Doctorate Program, UPC. 2003)

DP Algorithm for SMT (\mathbf{x}, l, a)

Input: A source sentence \mathbf{x} and the parameters l and a .

Output: A target sentence y .

Initialization

For $c := 1$ **until** J **do**

For each (C, j) with $C \subset \{1, \dots, J\}$, $j \in C$ and $|C| = c$ **do**

For each pair of target words y, y' **do**

$$Q_{yy'}(C, j) = l(\mathbf{x}_j | y) \cdot \max_{y'', j' \in C - \{j\}} (a(j | j') \cdot p(y | y', y'') \cdot Q_{y'y''}(C - \{j\}, j'))$$

End-for

End-for

Return: $\operatorname{argmax}_{y, y', j} p(\# | y, y') \cdot Q_{yy'}(\{1, \dots, J\}, j)$ and traceback

The set C is constraint to be a maximum number of positions.

Index

- 1 Statistical framework to machine translation ▷ 2
- 2 Alignments ▷ 11
- 3 Statistical alignment models ▷ 20
- 4 First-order alignment models ▷ 50
- 5 *Categorization in statistical modeling* ▷ 66
- 6 Bibliography ▷ 74

Categorization

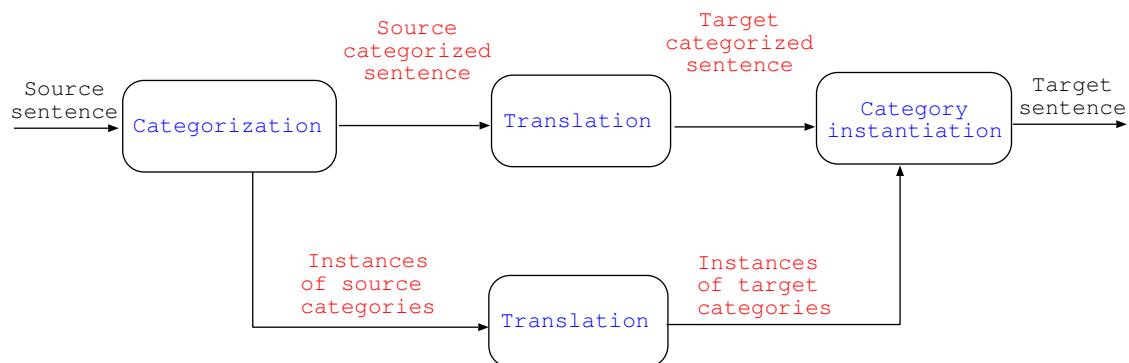
- Too many parameters to be estimated
- Many words play the same role: names, dates, etc.
- Substitution of words by categories:
 - The vocabulary size decreases.
 - Easy word addition to the vocabulary.
- Examples:
 - `mi nombre es $NAME.masc $SURNAME . # my name is $NAME.masc $SURNAME .`
 - `nos vamos a ir el $DATE a $HOUR . # we are leaving on $DATE at $HOUR .`
- Given a bilingual corpus:
 - Automatic extraction of bilingual categories.
 - Manual extraction of bilingual categories.

Categorization and learning

- Given a bilingual corpus:
 - CATEGORIZED TRANSLATOR:** Training a statistical translator (a translation model plus a target language model) from a corpus of categorized pairs.
 - A TRANSLATOR FOR EACH CATEGORY:** Training a statistical translator (translation model plus target language model) from the set of pairs of segments associated to each category.
 - A SOURCE CATEGORIZER:** Training a statistical translator (translation model plus language model) from the set of source no-categorized/categorized sentences.

An approach

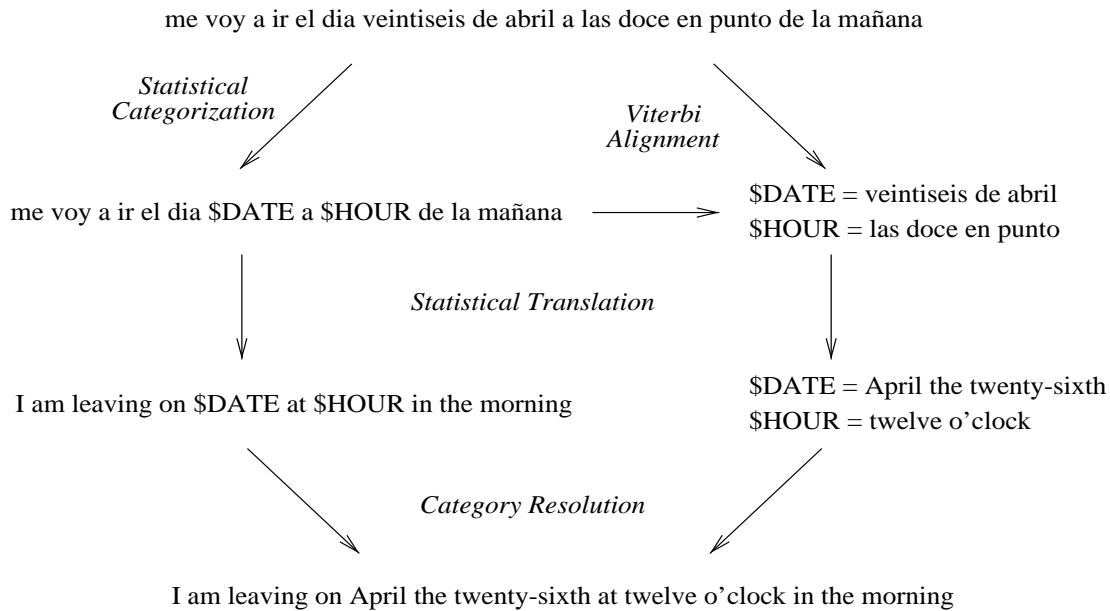
(I.Garcia-Varea, F.Casacuberta. *An iterative, DP-based search algorithm for statistical machine translation*. ICSLP. 1998.)



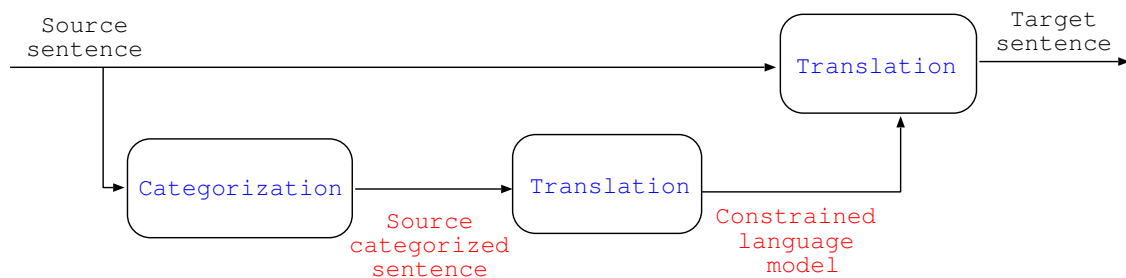
- CATEGORIZATION:** Translating the source sentence into an source categorized sentence and Obtaining the source instances of each category.
- CATEGORIZED TRANSLATION:** Translating the source categorized sentence into a target categorized sentence.
- TRANSLATION OF EACH CATEGORY:** Translating the source instances of each category detected.
- CATEGORY RESOLUTION:** Substitution of each target category by the corresponding instance translation.

An example

(I.Garcia-Varea, F.Casacuberta. *An iterative, DP-based search algorithm for statistical machine translation*. ICSLP. 1998.)



Another approach



1. **CATEGORIZATION:** Translating the source sentence into an source categorized sentence.
2. **CATEGORIZED TRANSLATION:** Translating the source categorized sentence into a target categorized sentence.
3. **DETAILED TRANSLATION:** Translating the source non-categorized sentence using the target categorized sentence as a restricted target language

Results

EuTrans-I corpus (Spanish-English)

- Vocabulary: 680 Spanish words, and 513 English words.
- Training: 10,000 pairs (97,000/99,000 words).
- Test: 2,996 pairs (PP=8.6/5.2) (35,000/35,590 words).
- Manual categories: 7.

Model	categorization	WER
Quasi-monotone search	manual	6.7
DP-search with M2	manual	9.8
Quasi-monotone search	no	10.8
DP-search with M2	no	13.9

Automatic categorization

- *Extended word categories*
(Barrachina & Vilar. *Bilingual clustering using monolingual algorithms*. TMI. 1999.)
 1. Align a bilingual corpus
 2. Build extended words using the alignments
 3. Apply a clustering algorithm to the corpus of extended word sentences
- *Statistical bilingual categories*
(Och. *An Efficient method for determining bilingual word classes*. ECACL. 1999.)
 1. Align a bilingual corpus
 2. Apply a clustering algorithm to the target corpus.
 3. Apply a clustering algorithm to the source corpus taking into account the categories of target words aligned to the source words.

Index

- 1 Statistical framework to machine translation ▷ 2
- 2 Alignments ▷ 11
- 3 Statistical alignment models ▷ 20
- 4 First-order alignment models ▷ 50
- 5 Categorization in statistical modeling ▷ 66
- 6 *Bibliography* ▷ 74

Bibliography

1. P. F. Brown et al. *A statistical approach to machine translation*. Computational Linguistics, vol. 16, pp. 79–85, 1990.
2. P. F. Brown et al. *The mathematics of statistical machine translation: parameter estimation*. Computational Linguistics, vol. 19 (2), 263–310, 1993.
3. I. Garcia-Varea, F. Casacuberta. *An iterative, DP-based search algorithm for statistical machine translation*. Proceedings of the ICSLP. 1998.
4. S. Barrachina and J. Vilar. *Bilingual clustering using monolingual algorithms*. TMI. 1999.
5. F. Och. *An Efficient method for determining bilingual word classes*. ECACL. 1999.
6. H. Ney, S. Nießen, F. Och, H. Sawaf, C. Tillmann, S. Vogel: *Algorithms for statistical translation of spoken language*. IEEE Transactions on Speech and Audio Processing, vol. 8 (1), 24–36, 2000.
7. F. J. Och, H. Ney: *Improved statistical alignment models*. Proc. of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 440–447, Hongkong, China, October 2000.
www-i6.Informatik.RWTH-Aachen.de/Colleagues/och/software/GIZA++.html
8. H. Ney: *Statistical Natural Language Processing*. Signal Theory and Communications Doctorate Program. UPC. Barcelona, 2003.
9. R. C. Moore *Improving IBM Word Alignment Model 1* Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics: ACL 2004. 518–525

Pattern Recognition Approaches to Machine Translation

E. Vidal y F. Casacuberta

Pattern Recognition and Human Language Technology Group

Departament de Sistemes Informàtics i Computació

Institut Tecnològic d'Informàtica

Universitat Politècnica de València

3: Advanced Statistical Alignment Models

Francisco Casacuberta Nolla

`fcn@iti.upv.es`

24-28 January 2005

F. Casacuberta – DSIC-ITI-UPV

[Pattern Recognition approaches to Machine Translation](#)

[Advanced Statistical Alignment Models](#)

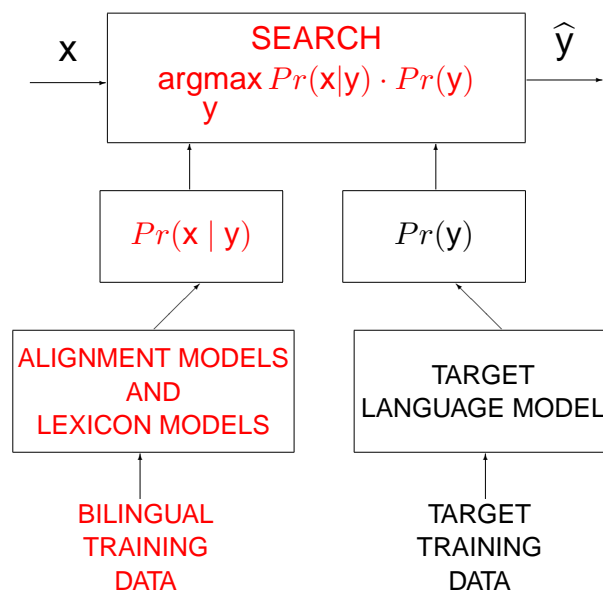
Index

- 1 Statistical framework to machine translation ▷ [2](#)
- 2 Fertility-based models ▷ [7](#)
- 3 The search problem ▷ [27](#)
- 4 Maximum entropy models ▷ [42](#)
- 5 Using linguistic knowledge ▷ [56](#)
- 6 Bibliography ▷ [65](#)

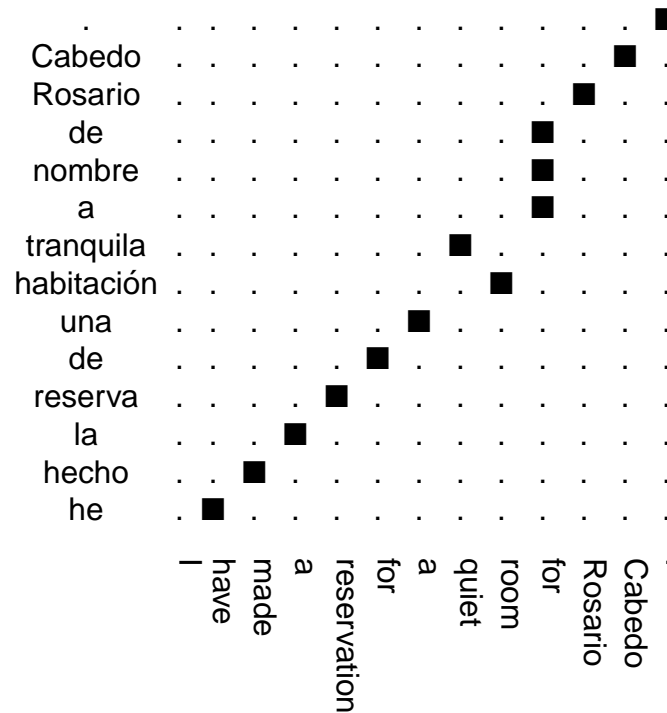
Index

- 1 *Statistical framework to machine translation* ▷ 2
- 2 Fertility-based models ▷ 7
- 3 The search problem ▷ 27
- 4 Maximum entropy models ▷ 42
- 5 Using linguistic knowledge ▷ 56
- 6 Bibliography ▷ 65

An inverse approach



An example of word alignments



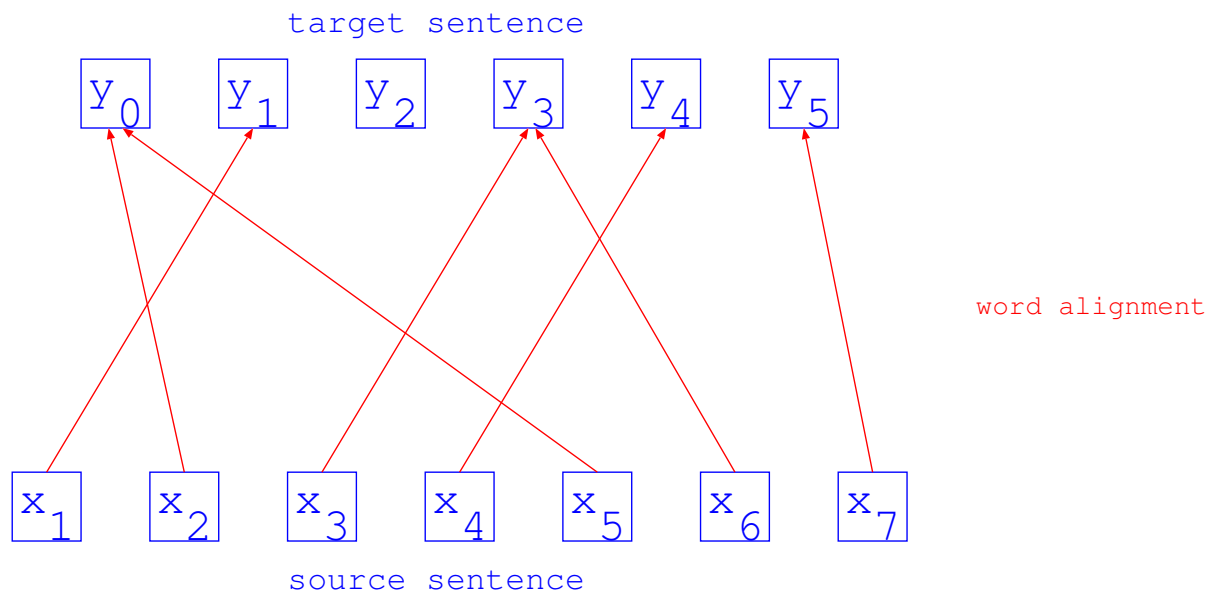
Alignments

$$\Pr(\mathbf{x} \mid \mathbf{y}) = \sum_{\mathbf{a} \in \mathcal{A}(\mathbf{y}, \mathbf{x})} \Pr(\mathbf{x}, \mathbf{a} \mid \mathbf{y}) = \Pr(J \mid \mathbf{y}) \cdot \sum_{\mathbf{a} \in \mathcal{A}(\mathbf{y}, \mathbf{x})} \Pr(\mathbf{x}, \mathbf{a} \mid J, \mathbf{y})$$

Alignment probabilities and lexicon probabilities

- Model 1
- Model 2
- Hidden Markov model

Models 1, 2 or HMM



Index

- 1 Statistical framework to machine translation ▷ 2
- 2 *Fertility-based models* ▷ 7
- 3 The search problem ▷ 27
- 4 Maximum entropy models ▷ 42
- 5 Using linguistic knowledge ▷ 56
- 6 Bibliography ▷ 65

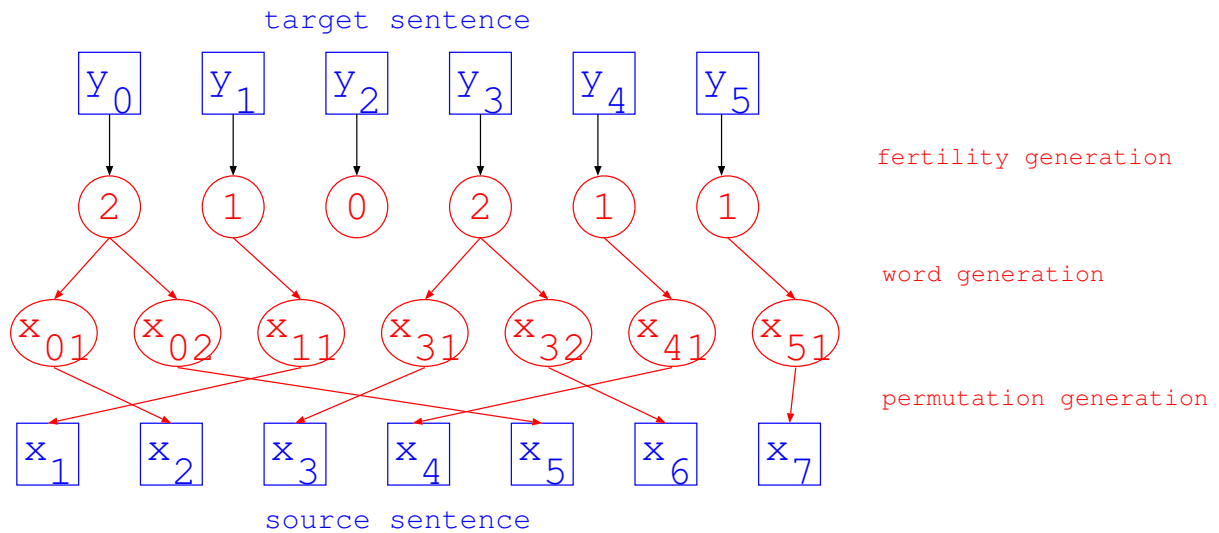
Fertility-based models

- Fertility
- Model 3
- Model 4
- Model 5
- Model 6
- The training process

Models 3, 4, 5 and 6

- Model 3 is a zero-order model: Lexicon, fertility and distortion models.
- Model 4 is a refined version (first-order) of distortion distribution in Model 3.
- Model 5 is a consistent version of distortion distribution in Model 4.
- Model 6 is a log-linear combination of HMM and Model 4.

Fertility



Fertility

Fertility ϕ of $y_i \in \Delta$: number of the source words connected to an target word y_i .

1. Choose how many source words are connected to a target word y_i : **fertility** of y_i
 $(\Phi = \phi(y_i))$
2. Choose a set of the source words, a **tablet** τ_i , that is connected to i -th target word
 $(\Gamma_{i,k} = \tau_{i,k} \in \Sigma \text{ for } 1 \leq k \leq \phi(y_i))$
3. Choose the **position** $\pi_{i,k}$ in the source sentence of the k -th word $\tau_{i,k}$ that is connected to the i -th target word
 $(\Pi_{i,k} = \pi_{i,k}, 1 \leq \pi_{i,k} \leq J)$

An example

Given y : $a \quad double \quad room \quad (I = 3)$

i	1	2	3	
Choose $\phi(y_i) = \phi$	1	3	1	
Choose $\tau_{i,k} = x$	{una}	{con, camas, dos}	{habitación}	
Choose $\pi_{i,k} = j$	1	3	5	4
				2

j	1	2	3	4	5
x	una	habitación	con	dos	camas

Model 3

$$\Pr(\mathbf{x} \mid \mathbf{y}) = \sum_{\mathbf{a}} \Pr(\mathbf{x}, \mathbf{a} \mid \mathbf{y}) = \sum_{\mathbf{a}} \sum_{(\tau, \pi) \in \mathcal{F}(\mathbf{x}, \mathbf{a})} \Pr(\phi, \tau, \pi \mid \mathbf{y})$$

The probability for a tablet τ and a permutation π is:

$$\Pr(\phi, \tau, \pi \mid \mathbf{y}) = \prod_{i=1}^I \Pr(\phi_i \mid \phi_1^{i-1}, \mathbf{y}) \Pr(\phi_0 \mid \phi_1^I, \mathbf{y}) \times \prod_{i=0}^I \prod_{k=1}^{\phi_i} \Pr(\tau_{i,k} \mid \tau_{i,1}^{k-1}, \tau_0^{i-1}, \phi_0^I, \mathbf{y}) \times \\ \prod_{i=1}^I \prod_{k=1}^{\phi_i} \Pr(\pi_{i,k} \mid \pi_{i,1}^{k-1}, \pi_1^{i-1}, \tau_0^I, \phi_0^I, \mathbf{y}) \times \prod_{k=1}^{\phi_i} \Pr(\pi_{0,k} \mid \pi_{0,1}^{k-1}, \pi_1^I, \tau_0^I, \phi_0^I, \mathbf{y})$$

- $\Pr(\phi_i \mid \phi_1^{i-1}, \mathbf{y}) \approx f(\phi_i \mid \mathbf{y}_i)$ *fertility probability*
- $\Pr(\tau_{i,k} = x \mid \tau_{i,1}^{k-1}, \tau_0^{i-1}, \phi_0^I, \mathbf{y}) \approx l(x \mid \mathbf{y}_i)$ *lexicon probability*
- $\Pr(\pi_{i,k} = j \mid \pi_{i,1}^{k-1}, \pi_1^{i-1}, \tau_0^I, \phi_0^I, \mathbf{y}) \approx d(j \mid i, J, I)$ *distortion probability*

Model 3

- $\Pr(\phi_i \mid \phi_1^{i-1}, \mathbf{y}) \approx f(\phi_i \mid \mathbf{y}_i)$ *fertility probability*
- $\Pr(\Gamma_{i,k} = x \mid \tau_{i,1}^{k-1}, \tau_0^{i-1}, \phi_0^I, \mathbf{y}) \approx l(x \mid \mathbf{y}_i)$ *lexicon probability*
- $\Pr(\Pi_{i,k} = j \mid \pi_{i,1}^{k-1}, \pi_1^{i-1}, \tau_0^I, \phi_0^I, \mathbf{y}) \approx d(j \mid i, J, I)$ *distortion probability*

$$P_{M3}(\mathbf{x} \mid \mathbf{y}) = \sum_{\mathbf{a}} \sum_{(\tau, \pi) \in \mathcal{F}(\mathbf{x}, \mathbf{a})} P_{M3}(\phi, \tau, \pi \mid \mathbf{y}) =$$

$$\sum_{a_1=0}^I \cdots \sum_{a_J=0}^I \binom{J - \phi_0}{\phi_0} p_0^{J-2\phi_0} p_1^{\phi_0} \prod_{i=1}^I \phi_i! \cdot f(\phi_i \mid \mathbf{y}_i) \prod_{j=1}^J l(\mathbf{x}_j \mid \mathbf{y}_{\mathbf{a}_j}) \cdot d(j \mid \mathbf{a}_j, J, I)$$

Model 3

- $\Pr(\phi_i \mid \phi_1^{i-1}, \mathbf{y}) \approx f(\phi_i \mid \mathbf{y}_i)$ *fertility probability*
- $\Pr(\Gamma_{i,k} = x \mid \tau_{i,1}^{k-1}, \tau_0^{i-1}, \phi_0^I, \mathbf{y}) \approx l(x \mid \mathbf{y}_i)$ *lexicon probability*
- $\Pr(\Pi_{i,k} = j \mid \pi_{i,1}^{k-1}, \pi_1^{i-1}, \tau_0^I, \phi_0^I, \mathbf{y}) \approx d(j \mid i, J, I)$ *distortion probability*

Given a target sentence \mathbf{y} of length I ,

1. For each $1 \leq i \leq I$ choose a length ϕ_i according to $f(\phi_i \mid \mathbf{y}_i)$.
2. Choose a length ϕ_0 according to $f_0(\phi_0 \mid \sum_{i=1}^I \phi_i)$.
3. $J = \sum_{i=0}^I \phi_i$.
4. For each $1 \leq i \leq I$ and $1 \leq k \leq \phi_i$, choose a source word $\tau_{i,k} \in \Sigma$ according to $l(\tau_{i,k} \mid \mathbf{y}_i)$.
5. For each $1 \leq i \leq I$ and $1 \leq k \leq \phi_i$, choose a position $\pi_{i,k}$ ($1 \leq \pi_{i,k} \leq J$) in the source sentence according to $d(\pi_{i,k} \mid i, J, I)$.
6. If any position has been chosen then **error** (*inconsistent model*).
7. For each $1 \leq k \leq \phi_0$ choose a position $\pi_{0,k}$ from the vacant positions according to a uniform distribution.

An example

Given y : $a \quad double \quad room \quad (I = 3)$

	i								
Choose $\phi(y_i) = \phi$ using $f(\phi y_i)$		1		2		3			
Choose $\tau_{i,k} = x$ using $l(x y_i)$		1		3		1			
Choose $\pi_{i,k} = j$ using $d(j i, I, J)$		{una}	{con,	camas,	dos}	{habitación}			
		1	3	5	4	2			

	j								
x		1	2	3	4	5			
		una	habitación	con	dos	camas			

Examples of alignments

Corpus EUTRANS-I: Spanish-English

1 2 3 4 5 6 7 8 9 10
 por favor , ¿ podría ver alguna habitación tranquila ?

- MODEL 1, ITERATION 5
 could (5) I (6) see (6) a (7) quiet (9) room (8) , (3) please (2) ? (4)
- MODEL 2, ITERATION 2
 could (5) I (6) see (6) a (7) quiet (9) room (8) , (3) please (3) ? (10)
- MODEL 3, ITERATION 2
 could (5) I (5) see (6) a (7) quiet (9) room (8) , (3) please (2) ? (10)

Model 4

For a target word y_i :

- The center of y_i , $c(i) = \frac{\sum_k \pi_{i,k}}{\phi_i}$

- $\Pr(\phi_i \mid \phi_1^{i-1}, \mathbf{y}) \approx f(\phi_i \mid \mathbf{y}_i)$ *fertility probability*

- $\Pr(\Gamma_{i,k} = x \mid \tau_{i,1}^{k-1}, \tau_0^{i-1}, \phi_0^I, \mathbf{y}) \approx l(x \mid \mathbf{y}_i)$ *lexicon probability*

- $\Pr(\Pi_{i,1} = j \mid \pi_1^{i-1}, \tau_0^I, \phi_0^I, \mathbf{y}) \approx d_{=1}(j - c(i-1) \mid \mathcal{C}_Y(\mathbf{y}_{i-1}), \mathcal{C}_X(\mathbf{x}_j))$

distortion probability for the first position in a tablet

- $\Pr(\Pi_{i,k} = j \mid \pi_{i,1}^{k-1}, \pi_1^{i-1}, \tau_0^I, \phi_0^I, \mathbf{y}) \approx d_{>1}(j - \pi_{i,k-1} \mid \mathcal{C}_X(\mathbf{x}_j))$

distortion probability for the rest of positions in a tablet

Model 4

$$\begin{array}{ccccccc}
 \mathbf{y}_1 & \cdots & \mathbf{y}_{i-1} & & \mathbf{y}_i & \cdots & \mathbf{y}_I \\
 \phi_1 & \cdots & \phi_{i-1} & & \phi_i & \cdots & \phi_I \\
 \{x_{1,1}, \dots, x_{1,\phi_1}\} & \cdots & \{x_{i-1,1}, \dots, x_{i-1,\phi_{i-1}}\} & & \{x_{i,1}, \dots, x_{i,\phi_i}\} & \cdots & \{x_{I,1}, \dots, x_{I,\phi_I}\} \\
 \{\pi_{1,1} < \dots < \pi_{1,\phi_1}\} & \cdots & \{\pi_{i-1,1} < \dots < \pi_{i-1,\phi_{i-1}}\} & & & & \\
 c(1) = \frac{\sum_{t=1}^{\phi_1} \pi_{1,t}}{\phi_1} & \cdots & c(i-1) = \frac{\sum_{t=1}^{\phi_{i-1}} \pi_{i-1,t}}{\phi_{i-1}} & & & &
 \end{array}$$

$$\pi_{i,1} = j \text{ according to } d_{=1}(j - c(i-1) \mid \mathcal{C}_Y(\mathbf{y}_{i-1}), \mathcal{C}_X(x_{i,1}))$$

$$\pi_{i,k} = j, \text{ for } 1 < k \leq \phi_i, \text{ according to } d_{>1}(j - \pi_{i,k-1} \mid \mathcal{C}_X(x_{i,k}))$$

$$\pi_{i,1} < \cdots < \pi_{i,\phi_i}$$

Model 4

- $f(\phi_i | y_i)$ *fertility probability*
- $l(x | y_i)$ *lexicon probability*
- $d_{=1}(j - c(i - 1) | \mathcal{C}_Y(y_{i-1}), \mathcal{C}_X(x_j))$ *distortion probability for the first position in a tablet*
- $d_{>1}(j - \pi_{i,k-1} | \mathcal{C}_X(x_j))$ *distortion probability for the rest of positions in a tablet*

Given a target sentence y of length I ,

1. For each $1 \leq i \leq I$ choose a length ϕ_i according to $f(\phi_i | y_i)$.
2. Choose a length ϕ_0 according to $f_0(\phi_0 | \sum_{i=1}^I \phi_i)$.
3. $J = \sum_{i=0}^I \phi_i$.
4. For each $1 \leq i \leq I$ and $1 \leq k \leq \phi_i$, choose a source word $\tau_{i,k}$ according to $l(\tau_{i,k} | y_i)$.
5. For each $1 \leq i \leq I$ and $1 \leq k \leq \phi_i$, choose a position $\pi_{i,k}$
 - if $k = 1$ according to d_1
 - if $k > 1$ according to $d_{>1}$ but greater than $\pi_{i,k-1}$
6. If any position has been chosen then **error**. (*inconsistent model*)
7. For each $1 \leq k \leq \phi_0$ choose a position $\pi_{0,k}$ from the vacant positions according to a uniform distribution.

Examples of alignments

Corpus EUTRANS-I: Spanish-English

1	2	3	4	5	6	7	8	9	10	11	12	13
por	favor	,	he	hecho	una	reserva	a	nombre	de	Federico	Redondo	.

[2-2] I (4) have (4) made (5) a (6) **reservation** (5) for (9) Federico (11) Redondo (12) . (0)
 [4-2] I (4) **have** (5) made (5) a (6) reservation (7) for (9) Federico (11) Redondo (12) . (13)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
me	voy	a	ir	el	jueves	tres	de	junio	a	la	una	y	media	de	la	tarde	.

[2-2] I (2) am (2) leaving (2) on (5) Thursday (6) June (9) the (5) **third** (9) at (10) half (14)
 past (13) **one** (11) in (4) **the** (11) afternoon (17) . (18)
 [4-2] I (2) am (2) leaving (2) on (5) Thursday (6) June (9) **the** (0) third (7) at (10) half (14)
 past (13) **one** (11) in (15) the (16) afternoon (17) . (18)
 [4-5] I (2) am (2) leaving (2) on (5) Thursday (6) June (9) **the** (0) third (7) at (10) half (14)
 past (13) one (12) in (15) the (16) afternoon (17) . (18)

Model 5

For a target word y_i :

- Number of vacant positions up to and including position j just before $\tau_{i,k}$ is placed, $v(j, \tau_1^{i-1}, \tau_{i,1}^{k-1}) \equiv v_j$.

- $\Pr(\phi_i \mid \phi_1^{i-1}, y) \approx f(\phi_i \mid y_i)$ *fertility probability*

- $\Pr(\Gamma_{i,k} = x \mid \tau_{i,1}^{k-1}, \tau_0^{i-1}, \phi_0^I, y) \approx l(x \mid y_i)$ *lexicon probability*

- $\Pr(\Pi_{i,1} = j \mid \pi_1^{i-1}, \tau_0^I, \phi_0^I, y) \approx d_{=1}(v_j \mid \mathcal{C}_X(\mathbf{x}_j), v_{c(i-1)}, v_J - \phi_i + 1) \cdot (1 - \delta(v_j, v_{j-1}))$

distortion probability for the first position in a tablet

- $\Pr(\Pi_{i,k} = j \mid \pi_{i,1}^{k-1}, \pi_1^{i-1}, \tau_0^I, \phi_0^I, y) \approx d_{>1}(v_j - v_{\pi_{i,k-1}} \mid \mathcal{C}_X(\mathbf{x}_j), v_J - v_{\pi_{i,k-1}} - \phi_i + k) \cdot (1 - \delta(v_j, v_{j-1}))$

distortion probability for the rest of positions in a tablet

Model 6

A linear combination of Model 4 and Homogeneous hidden Markov model.

$$Pr_{M6}(\mathbf{x} \mid \mathbf{y}) = \alpha \cdot Pr_{M4}(\mathbf{x} \mid \mathbf{y}) + (1 - \alpha) \cdot Pr_{HM}(\mathbf{x} \mid \mathbf{y})$$

The training process

H. Ney, *Statistical Natural Language Processing*, STC Doctorate Program, UPC. 2003

- Maximum likelihood by EM estimation.
- The counts in the reestimation are multiplied by $Pr_M(x, a | y)$ and are added for all possible alignment.
- No efficient method is computing these estimated counts.
- The estimated counts are approximate by:
 - Computing the (approximate) most probable alignment (Model 2)
 - Apply modifications: moves and swaps
 - Sum the estimated counts for all alignments whose probability is larger than the probability of the probable alignment times a given constant.
 - More details: P. F. Brown et al. *The mathematics of statistical machine translation: parameter estimation*. Computational Linguistics, vol. 19 (2), 263–310, 1993.

Conventional IBM Models Training

- Every model has a specific set of free parameters.
- For example for IBM Model 4: $\theta = \{ \{l(x|y)\}, \{p_{=1}(\Delta_j)\}, \{p_{>1}(\Delta_j)\}, \{p(\phi|x)\}, p_1 \}$
- To train the model parameters θ : A maximum likelihood criterium, using a parallel training corpus consisting of S sentence pairs $\{(x^{(n)}, y^{(n)}) : n = 1, \dots, N\}$:

$$\hat{\theta} = \operatorname{argmax}_{\theta} \prod_{n=1}^N \sum_{\mathbf{a}} p_{\theta}(x^{(n)}, \mathbf{a} | y^{(n)}) \quad .$$

- The training is carried out using the Expectation-Maximization (EM) algorithm.

The EM algorithm

Given a set of pairs $(\mathbf{x}^n, \mathbf{y}^n)$, for $n = 1, \dots, N$,

- Initialize parameters $\theta = \{l(x|y), \dots\}$
- Iterate (EM-procedure)
 - In the E-step, the lexicon parameter counts for every sentence pair (\mathbf{y}, \mathbf{x}) are calculated:

$$c(x|y; \mathbf{y}, \mathbf{x}) = N(\mathbf{y}, \mathbf{x}) \cdot \sum_{\mathbf{a}} Pr(\mathbf{a}|\mathbf{y}, \mathbf{x}) \sum_j \delta(x, \mathbf{x}_j) \delta(y, \mathbf{y}_{\mathbf{a}_j})$$

- In the M-step, the lexicon parameters $\hat{l}(s|t)$ that maximize the likelihood on the training corpus are computed:

$$\hat{l}(x|y) = \frac{\sum_n c(x|y; \mathbf{x}^{(n)}, \mathbf{y}^{(n)})}{\sum_{n,s} c(x|y; \mathbf{x}^{(n)}, \mathbf{y}^{(n)})}$$

Similarly, the alignment/distortion and fertility parameters can be estimated for all other alignment models.

- Compute Viterbi alignments

The output is a set of aligned sentence pairs $V(\mathbf{x}^n, \mathbf{y}^n); \hat{\theta}$

Index

- 1 Statistical framework to machine translation ▷ 2
- 2 Fertility-based models ▷ 7
- 3 *The search problem* ▷ 27
- 4 Maximum entropy models ▷ 42
- 5 Using linguistic knowledge ▷ 56
- 6 Bibliography ▷ 65

The search problem in statistical machine translation

$$\hat{y} = \operatorname{argmax}_y Pr(x | y) \cdot Pr(y)$$

- Search is a **NP-Hard problem**. (Knight, 1999)
- **Algorithmic solutions**: (+ heuristics for efficient suboptimal solutions)
 - *Dynamic Programming* (Garcia-Varea, 2003) (Tillman, 2003)
 - *Stack-decoding, A* or Branch & Bound* (Ortiz, 2003)
 - *Greedy strategies* (Germann, 2001)
 - *Using finite-state transducers* (Kumar, 2004)

Dynamic programming approach: DPSearchM2 for the models 3, 4 and 5*

- Using DPSearchM2 and Viterbi alignments.
- The Viterbi alignments for models 3, 4 and 5, are based on model 2.
- Solution:
 - In the final states in DPSearchM2
 - To choose the hypothesis with the best Viterbi score for a model M .
 - To iterate the process.
- Computational complexity: $O(J \cdot I_{max} \cdot L \cdot |\mathcal{E}|^2)$

*The slides on searching are modified versions of some material supplied by Ismael García-Varea

Some stack-decoding proposals

- Candide systems from IBM [Berger et al. 96]: Multiple stacks, model 3.
- Multiple stack-decoding [Wang and Waibel 98]: Model 2.
- Algorithm A^* [Ueffing et al. 01]: model 4.
- Algorithm A^* [Och and Ney 03]: model 6.
- Basic stack-decoding strategy:
 - Origin of the *stack decoding* or A^* : ASR
 - Optimal solution to the search problem (Jelinek, 1976)
 - Incremental development of practical hypothesis
 - The hypothesis are stored in a priority queue (a type of 'stack')
 - Selection and expansion of the top of the stack(s).

A taxonomy of the stack-decoding algorithms

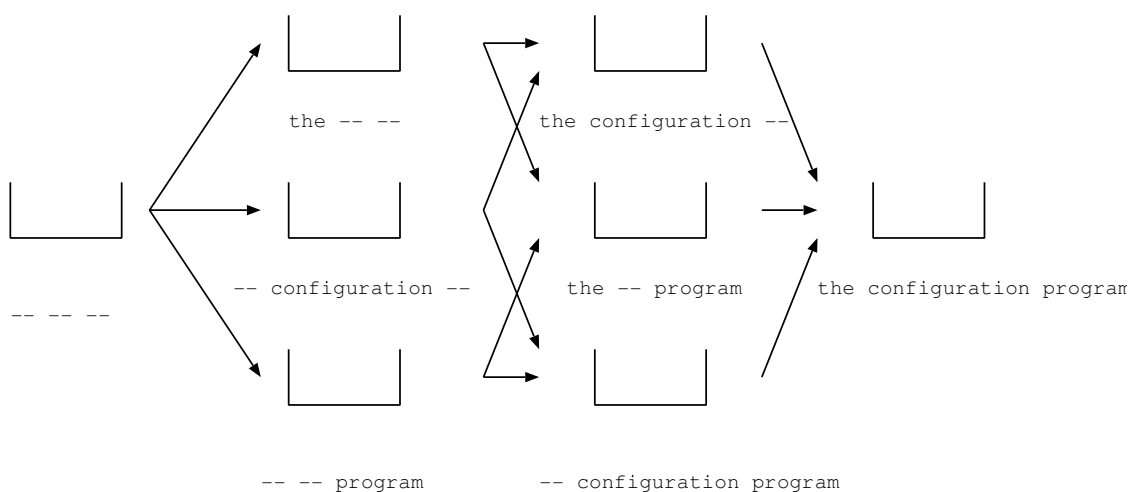
- Basic stack-decoding algorithm:
 - All the hypothesis are stored in a one stack
 - A hypothesis is selected in each iteration: the hypothesis with higher score in the stack
- Problem: hypothesis with a high number of aligned words are discarded.
- Possible solutions:
 - Use of heuristics: an estimation of the contribution to the set of the optimal score.
 - Multiple stacks.
- Taxonomy:
 - Single stack algorithms A^*
 - Multiple stack algorithms

Basic multiple stack decoding *StackDecoding*

- A hypothesis in a stack:
 - A prefix of the target sentence (y_1^i)
 - A coverage subset of source positions (\mathcal{C})
 - A score (S).
- There is one stack for each possible subset of source positions which words has already been translated.
- The possible number of stacks can be very high.
- In each iteration, the best hypothesis from each available stack is selected to generate new extended hypothesis.
- The new target prefix is the concatenation of the target prefix of the selected hypothesis and each possible target word.
- The new source positions are selected from the complementary set of \mathcal{C} (assuming some constraints).
- The new score is computed using the new ngram and the new source positions.
- The new hypothesis is stored in the corresponding stack.

Basic multiple stack decoding *StackDecoding*

Source sentence: "the configuration program"



Basic greedy algorithm *GreedySearch*

- Previous works: [Germann et al. 01] for models 3 and 4.
- Characteristics:
 - Local optimization
 - No incremental building of hypothesis
 - Dependence on the initialization
 - Approximation to the search problem.
 - Fast.
 - They can be used to refine other algorithms.

Basic greedy algorithm *GreedySearch*

- Algorithm:
 - An initial complete hypothesis and the corresponding alignment are required.
 - The hypothesis is modified iteratively until no improvements are achieved.
 - Hillclimbing algorithm
 - Building a neighbourhood from $\langle y, a \rangle$
- Temporal cost: $O(J^2 \cdot |\mathcal{E}|^2 \cdot I^2)$

Experiments

- EUTRANS-I corpus:
 - Training: 10,000 pairs
 - Test: 2,636 sentences (length ≤ 15)
- HANSARDS corpus:
 - Training: 128.000 pairs
 - Test: 500 sentences of 4, 6, 8, 10, 12 words
- Translation models: IBM+HMM, $1^5 2^5 3^5 4^5 5^5$
- Language models: 3-grams + smoothing *Good Turing*
- Assessment:
 - **Word Error Rate (WER)**: The minimum number of substitution, insertion and deletion operations needed to convert the word string hypothesized by the translation system into a given single reference word string.
 - **Position Independent word error Rate (PER)**: Similar to WER but the order is not taken into account.

The HANSARD corpus

- Task definition:
 - Proceedings of the Canadian parliament. (French \rightarrow English)
 - Vocabulary sizes (more than two occurrences): 58.016 (French), 42.055 (English).
 - Training set: $1,7 \times 10^6$ pairs (sentence length less than 30)
 - Test set: 73 sentences.
- First results in (Brown et al. 1993)
 - Models:
 - * 12 training iterations (1 IBM1 + 6 IBM2 + 1 IBM3 + 4 IBM5)
 - * Language model: trigrams.
 - * Search: stack-decoding.
 - Results:
 - * 48% of sentences were successfully translated.

The EUTRANS-I corpus

- **Vocabulary:** 680 Spanish words, and 513 English words.
- **Training:** 10,000 pairs (97,000/99,000 words).
- **Test:** 2,996 pairs (PP=3.3) (35,000/35,590 words).

Experimental results

- EUTRANS-I corpus:

Strategy	sec.	SerErr	ModErr	Accuracy	WER	PER
DPSearch-M2	55.7	5.5	55.2	39.3	12.7	10.5
DPSearch-M4	69.5	12.2	45.1	42.7	10.2	9.4
StackDecoding-M4	87.1	18.4	44.1	37.5	14.2	11.1
GreedySearch-M3	18.7	61.3	20.5	18.2	24.8	18.6
GreedySearch-M4	165.9	53.0	23.3	23.7	20.0	16.2

- HANSARDS corpus:

Strategy	seg.	SerErrs	ModErr	Accuracy	WER	PER
DPSearch-M2	102.9	2.6	81.2	16.2	50.5	46.8
StackDecoding-M4	163.1	12.0	78.6	9.4	54.2	51.3
GreedySearch-M3	17.0	15.0	75.0	10.0	55.9	51.0

Comparasion results

	EUTRANS-I task		HANSARDS task	
Search strategy	WER	PER	WER	PER
DPSearch-M2	12.7	10.5	50.5	46.8
DPSearch-M4	10.2	9.4		
StackDecoding-M4	14.2	11.1	54.2	51.3
GreedySearch-M3	24.8	18.6		
GreedySearch-M4	20.0	16.2	55.9	51.0
SWB ⁽¹⁾	10.8	10.0	64.9	51.4
SWB+IBM ⁽¹⁾			64.9	51.4
AT ⁽¹⁾	4.4	2.9	61.5	49.2
A [*] ⁽²⁾			68.7	61.5
A [*] ⁽³⁾ -M4	5.5			

⁽¹⁾ In [Och 02] (Ph.D.)

⁽²⁾ In [Ueffing et al. 02] for sentences (≤ 12) and computational time of 127 sec.

⁽³⁾ In [Prat 02].

Results with categories

EUTRANS-I task		
Model	categorization	WER
Alignment templates	manual	2.5
STM category translation + A [*] with M4	automatic	3.8
Alignment templates	automatic	4.4
A [*] with M4	no	5.3
Quasi-monotone search	manual	6.7
STM	no	7.0
DP-search with M2	manual	9.8
Quasi-monotone search	no	10.8
DP-search with M2	no	13.9

Index

- 1 Statistical framework to machine translation ▷ 2
- 2 Fertility-based models ▷ 7
- 3 The search problem ▷ 27
- 4 *Maximum entropy models* ▷ 42
- 5 Using linguistic knowledge ▷ 56
- 6 Bibliography ▷ 65

Context-dependent lexicon models*

- The performance of a statistical machine translation system depends on the quality of lexicon and alignment models used.
- Typically, these statistical alignment models are based on single-word dependencies → lack of useful context information that can lead to inadequate alignments.
- A possible solution would be to include more dependencies in the lexicon model i.e. $l(x_j|y_{a_{j-1}}, y_{a_j}) \Rightarrow$ problem: significant data sparseness.
- A possible solution: Use maximum entropy to build context-dependent lexicon models.
- Some advantages of using maximum entropy
 - Easy to integrate additional knowledge sources
 - No problem with overlapping features
 - Well-founded mathematical theory
 - Efficient training algorithms
 - ...

*The slides on searching are modified versions of some material supplied by Ismael García-Varea

Maximum entropy principle

- A model that takes a **context** w into account $\Rightarrow p_y(x|w)$ instead of $l(x|y)$.
- The properties that can be useful: by **feature functions** $\phi_{y,k}(w, x), k = 1, \dots, K_y$.
 - For example, to model the existence or absence of a specific target word y' in the context of a target word y , which can be translated by the source word x' .
 - This dependence using the following indicator function (*feature*):

$$\phi_{y,1}(w, x) = \begin{cases} 1 & \text{if } x = x' \text{ and } y' \in w \\ 0 & \text{otherwise} \end{cases}$$

Consequently the first feature for word y has associated the pair (y', x') .

Maximum entropy principle

- The entropy maximum principle suggests that the optimal parametric form of a model $p_y(x|w)$ taking into account the feature functions $\phi_{y,k}, k = 1, \dots, K_y$ is given by:

$$p_y(x|w) = \frac{1}{Z_{\Lambda_y}(w)} \cdot \exp \left\{ \sum_{k=1}^{K_y} \lambda_{y,k} \cdot \phi_{y,k}(w, x) \right\}$$

- The resulting model has an exponential form with free parameters:

$$\Lambda_y \equiv \{\lambda_{y,k}, k = 1, \dots, K_y\}$$

- The parameter values that maximize the likelihood for a given training corpus can be computed using the so-called GIS algorithm.

Contextual information and features definition

- A model $p_y(x|w)$ and a sample training for each target word y are needed.
- In a pair of sentences (x, y) , contextual information (easily extended):
 - Target context: $y_{i-3} \dots y_i \dots y_{i+3}$
 - Source context: x_j
 - Word classes: syntactic and semantic information $(\mathcal{T}(y_i), \mathcal{S}(x_j))$.
- Feature categories:

Category	$\phi_{y_i,k}(w, x_j) = 1$ if and only if ...							
1	$x_j = \diamond$ and $\square \in$ <table><tr><td></td><td></td><td></td><td>y_i</td><td></td><td></td><td></td></tr></table>				y_i			
			y_i					
2	$x_j = \diamond$ and $\square \in$ <table><tr><td></td><td></td><td>•</td><td>y_i</td><td></td><td></td><td></td></tr></table>			•	y_i			
		•	y_i					
3	$x_j = \diamond$ and $\square \in$ <table><tr><td></td><td></td><td></td><td>y_i</td><td>•</td><td></td><td></td></tr></table>				y_i	•		
			y_i	•				
4	$x_j = \diamond$ and $\square \in$ <table><tr><td>•</td><td>•</td><td>•</td><td>y_i</td><td></td><td></td><td></td></tr></table>	•	•	•	y_i			
•	•	•	y_i					
5	$x_j = \diamond$ and $\square \in$ <table><tr><td></td><td></td><td></td><td>y_i</td><td>•</td><td>•</td><td>•</td></tr></table>				y_i	•	•	•
			y_i	•	•	•		

In all cases the k -th feature has associated the pair (\diamond, \square) .

Maximum entropy models training integration

- Model parameters to be learnt: $\Lambda_t \equiv \{\lambda_{t,k} : k = 1, \dots, K\}$
- In the E-step, a refined count collection for the lexicon parameters is performed

$$c(x|y, w; \mathbf{x}, \mathbf{y}) = N(\mathbf{x}, \mathbf{y}) \cdot \sum_{\mathbf{a}} Pr(\mathbf{a}|\mathbf{x}, \mathbf{y}) \sum_j \delta(x, x_j) \delta(y, y_{a_j}) \delta(w, w_{j,a_j})$$

$w_{j,a_j} \equiv$ the maximum entropy context that surrounds x_j and y_{a_j}

- In the M-step, the new lexicon parameters are computed:

$$\hat{\Lambda}_y = \operatorname{argmax}_{\Lambda_y} \prod_{x,w} c(s|y, w; \mathbf{x}, \mathbf{y}) \cdot \log p_y(x|w)$$

$c(x|y, w; \mathbf{x}, \mathbf{y}) \equiv$ weights of the training samples (x, y, w) used to train the maximum entropy model (number of times that (x, y, w) occurs).

- The re-estimation of the alignment/distortion and fertility probabilities does not change if we use a maximum entropy lexicon model.

The EM-ME algorithm

Given a set of pairs $(\mathbf{x}^n, \mathbf{y}^n)$, for $n = 1, \dots, N$,

- Initialize parameters $\theta = \{l(x|y), \dots\}$
- Iterate (EM-procedure)
 - In the E-step:
 1. Collect counts for alignment/distortion and fertility parameters.
 2. Collect refined lexicon counts (Overhead on space and computation time).
 - In the M-step:
 1. Reestimate alignment/distortion and fertility parameters.
 2. Perform GIS training for lexicon parameters (Overhead on space and computation time).
- Compute Viterbi alignments

The output is a set of aligned sentence pairs $V(\mathbf{x}^n, \mathbf{y}^n); \hat{\theta}; \hat{\Lambda}_y$

Potential problems of the ME-EM integration

- Computation overhead:
 - In the k -th iteration of the E-step
 - In the M-step the computation of the GIS training for each word
- Space overhead:

We have to store every possible maximum entropy training event (s, t, x) , that is, every possible combination of $t \in \mathcal{V}_T$, $s \in \mathcal{V}_S$ and $x \Rightarrow$ requires a huge quantity of memory.

Experimental results

- Efficiency: time consumption of different approaches.
- Performance: comparison of the alignment quality (of 500 randomly selected pairs) obtained with all the IBM models (1 to 5) with and without using maximum entropy modeling.
- Tasks: Verbmobil and Hansards

		Verbmobil		Hansards	
		German	English	French	English
Training	Sentences	34,446		1,470K	
	Words	329,625	343,076	24.33M	22.16M
	Vocabulary	5,936	3,505	100,269	78,332

Experimental results

Time consumption results

- Time consumption in seconds of different approaches per EM iteration (on average for the five IBM models) for different sizes of training corpus.

Task	Size of train.	# of e	Conventional train	ME-train	Simplified ME-train
Verbmobil	0.5K	29	1	29	1.5
	8K	84	18	235	68
	35K	209	60	2290	675
Hansards	0.5K	15	2.5	29	3
	8K	80	35	1180	100
	128K	1214	655	16890	6870

Alignment quality results: evaluation methodology

Example of a manual alignment

- An annotation scheme that explicitly allows for ambiguous alignments.
- Two different kinds of alignments: a $S(ure)$ alignment (unambiguous ■) and a $P(ossible)$ alignment (ambiguous □).
- The P labels are used specially to align words within idiomatic expressions, free translations, and missing function words ($S \subseteq P$).
- Reference alignment: many-to-one and one-to-many relationships.

[illegible]

Alignment quality results: evaluation criterion

- The quality of an alignment $A = \{(j, a_j) | a_j > 0\}$ is then computed by appropriately redefined precision and recall measures:

$$recall = \frac{|A \cap S|}{|S|}, \quad precision = \frac{|A \cap P|}{|A|}$$

and using the following alignment error rate, which is derived from the well known F-measure:

$$AER(S, P; A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$

- In such a way, a recall error can only occur if a $S(ure)$ alignment is not found and a precision error can only occur if the found alignment is not even $P(ossible)$.

Alignment quality results: AER

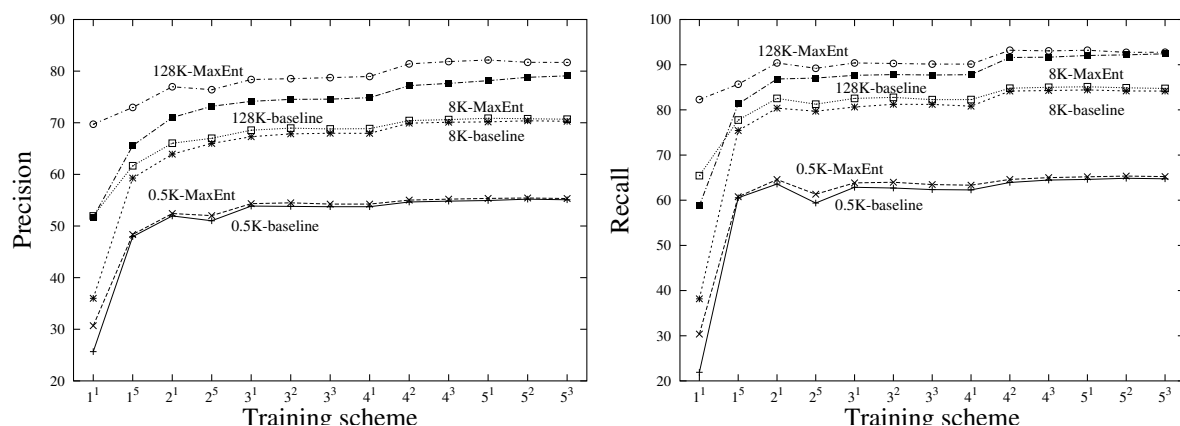
- AER of 500 randomly selected sentence pairs.

Training Scheme	Model	Hansards task			Verbmobil task		
		Size of train corpus			Size of train corpus		
		0.5K	8K	128K	0.5K	8K	34K
1^5	1	48.0	35.1	29.2	27.7	19.2	17.6
	1+ME	47.7	32.7	22.5	24.6	16.6	13.7
$1^5 2^5$	2	46.0	29.2	21.9	26.8	15.7	13.5
	2+ME	44.7	28.0	19.0	25.3	14.1	10.8
$1^5 2^5 3^3$	3	43.2	27.3	20.8	25.6	13.7	10.8
	3+ME	42.5	26.4	17.2	24.1	11.6	8.8
$1^5 2^5 3^3 4^3$	4	41.8	24.9	17.4	23.6	10.0	7.7
	4+ME	41.5	24.3	14.1	22.8	9.3	7.0
$1^5 2^5 3^3 4^3 5^3$	5	41.5	24.8	16.2	22.6	9.9	7.2
	5+ME	41.5	24.5	14.3	22.3	9.6	6.8

- The alignment error rate improves using the context-dependent lexicon models.
- For the Verbmobil task, the improvements are smaller than for the Hansards task, which might be due to the fact that already the baseline alignment quality is very good.

Alignment quality results: precision and recall

- Precision and recall [%] results for Hansards task for different corpus sizes in every iteration of the training:



Index

- 1 Statistical framework to machine translation ▷ 2
- 2 Fertility-based models ▷ 7
- 3 The search problem ▷ 27
- 4 Maximum entropy models ▷ 42
- 5 *Using linguistic knowledge* ▷ 56
- 6 Bibliography ▷ 65

Is the linguistic knowledge needed for statistical machine translation?

- YES?
 - There are many linguistic knowledge available.
 - The bilingual training data can be better exploited.
- NOT?
 - Many linguistic knowledge is hard to formalize.
 - The generation of new linguistic knowledge requires great human effort.

Linguistic knowledge that has been used in statistical machine translation

- Morpho-syntactic knowledge: lexicon, Part-of-Speech, etc... (Nießen and Ney, 2004)

Hybrid linguistic-statistical approaches have been used with success (i.e. *hidden markov models*)

- Others: Cognates (Kondrak, Marcu and Knight, 2003), named entities (Huang, Vogel and Waibel, 2003), ...
- Syntactic information: next talk!

Morpho-syntactic knowledge in statistical machine translation

Nießen and Ney, 2004. *Statistical machine translation with scarce resources using morpho-syntactic information*. Computational Linguistics.

- Present statistical machine translation systems often treat different inflected forms of the same lemma as if they were independent of one another.
- The bilingual data can be better exploited by explicitly taking into account the interdependencies of related inflected forms.

A possible proposal: **HIERARCHICAL LEXICON MODELS**

Morpho-syntactic knowledge in statistical machine translation

yo **como** pan

- Morphological and syntactic tags (POS, tense, person, ...)
- The *base form*

$T = t_1^6 = \text{comer}$ verb indicative present singular 1

Morpho-syntactic knowledge in statistical machine translation

$$\Pr(\mathbf{x} \mid \mathbf{y}) = \sum_{\mathbf{a} \in \mathcal{A}(\mathbf{y}, \mathbf{x})} \Pr(J \mid \mathbf{y}) \cdot \Pr(\mathbf{a} \mid J, \mathbf{y}) \cdot \Pr(\mathbf{x} \mid \mathbf{a}, J, \mathbf{y})$$

$$(t_1^n)_j \equiv T_j$$

$$\begin{aligned} \Pr(\mathbf{x} \mid \mathbf{a}, J, \mathbf{y}) &= \sum_{T_1^J} \Pr(\mathbf{x}, T_1^J \mid \mathbf{a}, J, \mathbf{y}) \\ &= \sum_{T_1^J} \prod_{j=1}^J \Pr(\mathbf{x}_j, T_j \mid \mathbf{x}_1^j, T_1^j, \mathbf{a}, J, \mathbf{y}) \\ &\approx \sum_{T_1^J} \prod_{j=1}^J l(\mathbf{x}_j, T_j \mid \mathbf{y}_{\mathbf{a}_j}) \end{aligned}$$

A lemma-tag lexicon: $l(\mathbf{x}_j, T_j \mid \mathbf{y}_{\mathbf{a}_j})$

Estimation of the lemma-tag lexicon

Maximum entropy modelling

$$l(s, T \mid t) \equiv l_{\Lambda}(s, T \mid t) = \frac{\exp [\sum_m \lambda_m h_m(t, s, t_1^n)]}{\sum_{\bar{s}, \bar{t}_1^n \exp [\sum_m \lambda_m h_m(t, \bar{s}, \bar{t}_1^n)]}$$

$$\Lambda = \{\lambda_m\}$$

- During training, the sum on \bar{s} and \bar{t}_1^n is restricted to the reading of word forms having the same base form and partial reading as a word forms aligned at least once with t .
- Three types of feature functions for maximum entropy modeling: 1. Base forms; 2. Subsets of tags and 3. Fully inflected words.

Experiments

Nießen and Ney, 2004. *Statistical machine translation with scarce resources using morpho-syntactic information*. Computational Linguistics.

- Verbmobil task:
 - Automatic translation of spontaneously spoken dialogs (English → German)
 - Vocabulary sizes: 1. 4,674 word forms (English) and 7,940 word forms (German).
2. 3,639 base forms (English) and 6,063 base forms (German)
 - Training set: 58,073 pairs (549Kw/519Kw).
 - Test set: 527 sentences.
- Results (m-WER):
 - 31.8% (34,1% in the baseline).

Open problems

- Automatically induction of the morphology of inflectional languages using only text corpora and no human input: Using prefix trees (Schone and Jurafsky, 2000) or pairs of hidden Markov models (Clark, 2000)
- Using “conventional” dictionaries (collections of word or phrase pairs collected by hand) (Nießen and Ney, 2004)
- Unknown words by some semantic information of the context words (Widdows, 2003).
- Extracting named entity translingual equivalences from bilingual parallel corpora (Huang, Vogel and Waibel, 2003)
- Using cognates (Kondrak, Marcu and Knight, 2003)

Index

- 1 Statistical framework to machine translation ▷ 2
- 2 Fertility-based models ▷ 7
- 3 The search problem ▷ 27
- 4 Maximum entropy models ▷ 42
- 5 Using linguistic knowledge ▷ 56
- 6 *Bibliography* ▷ 65

Bibliography

1. P. F. Brown et al. *The mathematics of statistical machine translation: parameter estimation*. Computational Linguistics, 19(2):263–310, 1993.
2. Y.Y. Wang, A. Waibel: *Decoding algorithms in statistical machine translation*, Proc. of the 37th Annual Meeting of the ACL and 8th Conf. of the European Chapter of the ACL, pp. 366-372, 1997.
3. K. Knight. *Decoding complexity in word-replacement translation models* Computational Linguistics. 25(4):607-615. 1999.
4. Clark. *Learning morphology with pair hidden markov models*. Student Workshop at ACL. 2000.
5. Schone and Jurafsky. *Knowledge-free induction of inflectional morphologies*. North American chapter of the Association for Computational Linguistics NAACL. 2000.
6. D. Ortiz, I. García-Varea, F. Casacuberta: *An empirical comparison of stack-based decoding algorithms for statistical machine translation*. Proc. of the IbPRIA, 2003.
7. H. Ney, *Statistical Natural Language Processing*, Signal Theory and Communications Doctorate Program, UPC. March 17-21, 2003
8. I. García-Varea. *Traducción automática estadística: modelos de traducción basados en máxima entropía y algoritmos de búsqueda*. PhD thesis, DSIC, Universidad Politécnica de Valencia, 2003.
9. Huang, Vogel and Waibel. *Extracting named entity translingual equivalence with limited resources*. ACM Transactions on Asian Language Information Processing. 2(2).2003.

Bibliography

10. Kondrak, Marcu and Knight. *Cognates can improve statistical translation models*. Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics HLT-NAACL), 46-48, Edmonton, May 2003.
11. F. J. Och, H. Ney. *A systematic comparison of various statistical alignment models*. Computational Linguistics. 2003.
12. C. Tillmann, H. Ney. *Word Reordering and a Dynamic Programming Beam Search Algorithm for Statistical Machine Translation*. Computational Linguistics, 29(1):97-133, 2003.
13. Widdows. *Unsupervised methods for developing taxonomies by combining syntactic and statistical information*. HLT-NAACL. 197–204, Edmonton, June 2003.
14. Nießen and Ney. *Statistical machine translation with scarce resources using morpho-syntactic information*. Computational Linguistics, 30(2):181–204. 2004.
15. I. García-Varea, F. Casacuberta. *Maximum Entropy Modeling: A Suitable Framework to Learn Context-Dependent Lexicon Models for Statistical Machine Translation*. Machine Learning. 2005.

Pattern Recognition approaches to Machine Translation

F. Casacuberta and E. Vidal

Pattern Recognition and Human Language Technology Group
Instituto Tecnológico de Informática
Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de Valencia, Spain

Stochastic Finite-State Translation Models

Enrique Vidal

`evidal@iti.upv.es`

January 2005

E. Vidal – ITI-UPV-DSIC

[Pattern Recognition Machine Translation](#)

[Stochastic Finite-State Transducers](#)

Index

- 1 Introduction ▷ [2](#)
- 2 Rational or Finite-State Transduction ▷ [6](#)
- 3 Stochastic Finite-State Transducers ▷ [11](#)
- 4 Error Correcting ▷ [20](#)
- 5 Sequential Transduction ▷ [26](#)
- 6 Subsequential Transduction:
Introduction to the “OSTI” Algorithm ▷ [32](#)
- 7 Bibliography ▷ [36](#)

Index

- 1 *Introduction* ▷ 2
- 2 Rational or Finite-State Transduction ▷ 6
- 3 Stochastic Finite-State Transducers ▷ 11
- 4 Error Correcting ▷ 20
- 5 Sequential Transduction ▷ 26
- 6 Subsequential Transduction:
Introduction to the “OSTI” Algorithm ▷ 32
- 7 Bibliography ▷ 36

Pattern Recognition, Natural Language Processing and Finite-State Transduction

- (Stochastic) Grammars and Automata are adequate models for *Classification tasks*. But there are many Pattern Recognition problems which are better framed within the most general *Interpretation* paradigm
- Interpretation tasks and be conceptually (and practically) tackled through *Formal Transduction*.
- E.g., many *Continuous Speech Recognition and Understanding* tasks can be seen as (simple) *transductions* from certain acoustic, phonetic or lexical input sequences into output sequences of higher-level linguistic categories
- Many *direct* applications such as *Language Translation* and *Semantic Decoding*
- *Simple transducers* are often powerful enough to deal with useful mappings between *complex languages*

Probabilistic problem statement

Given a source text x , its most probable translation is given by:

$$\hat{y} = \underset{y}{\operatorname{argmax}} \Pr(y | x) = \underset{y}{\operatorname{argmax}} \Pr(x, y)$$

The joint probability $\Pr(x, y)$ can be adequately modelled by means of a *stochastic finite-state transducer* \mathcal{T} :

$$\Pr(x, y) \approx P_{\mathcal{T}}(x, y)$$

However, *not all the transduction tasks are equally difficult. . .*

Not all the transduction tasks are equally difficult: examples

- **1... Spanish to English, word by word**

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

- **2... Division by 7**

3 5 7 6 8 1 8 0 3 1 (: 7 =)

- **3... English to Decimal**

NINEHUNDREDANDNINETEENTHOUSANDANDNINE

- **4... Roman to Decimal**

III XIX XLII LXXIV CDII CMLXXXIX

- **5... ATIS: English to "Pseudo English"**

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- **6... Spanish to English**

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- **1... Spanish to English, word by word**

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what time

- **2... Division by 7**

3 5 7 6 8 1 8 0 3 1 (: 7 =)

- **3... English to Decimal**

NINEHUNDREDANDNINETEENTHOUSANDANDNINE

- **4... Roman to Decimal**

III XIX XLII LXXIV CDII CMLXXXIX

- **5... ATIS: English to "Pseudo English"**

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- **6... Spanish to English**

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- **1... Spanish to English, word by word**

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what time departs

- **2... Division by 7**

3 5 7 6 8 1 8 0 3 1 (: 7 =)

- **3... English to Decimal**

NINEHUNDREDANDNINETEENTHOUSANDANDNINE

- **4... Roman to Decimal**

III XIX XLII LXXIV CDII CMLXXXIX

- **5... ATIS: English to "Pseudo English"**

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- **6... Spanish to English**

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

to what time departs the flight more early of Boston to Denver in TWA?

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

to what time departs the flight more early of Boston to Denver in TWA?

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)

0

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what time departs the flight more early of Boston to Denver in TWA?

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)
0 5

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what time departs the flight more early of Boston to Denver in TWA?

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)
0 5 1

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what time departs the flight more early of Boston to Denver in TWA?

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)
0 5 1 0

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what time departs the flight more early of Boston to Denver in TWA?

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)
0 5 1 0 9

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what time departs the flight more early of Boston to Denver in TWA?

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)
0 5 1 0 9 7 4 0 0 4

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what time departs the flight more early of Boston to Denver in TWA?

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)
0 5 1 0 9 7 4 0 0 4

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what time departs the flight more early of Boston to Denver in TWA?

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)
0 5 1 0 9 7 4 0 0 4

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE
9

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what time departs the flight more early of Boston to Denver in TWA?

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)
0 5 1 0 9 7 4 0 0 4

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE
9

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what time departs the flight more early of Boston to Denver in TWA?

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)
0 5 1 0 9 7 4 0 0 4

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE
9

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what time departs the flight more early of Boston to Denver in TWA?

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)
0 5 1 0 9 7 4 0 0 4

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE
9

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what time departs the flight more early of Boston to Denver in TWA?

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)
0 5 1 0 9 7 4 0 0 4

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE
9 19

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what time departs the flight more early of Boston to Denver in TWA?

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)
0 5 1 0 9 7 4 0 0 4

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE
9 19

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what time departs the flight more early of Boston to Denver in TWA?

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)
0 5 1 0 9 7 4 0 0 4

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE
9 19

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what time departs the flight more early of Boston to Denver in TWA?

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)
0 5 1 0 9 7 4 0 0 4

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE
9 19 0

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what time departs the flight more early of Boston to Denver in TWA?

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)
0 5 1 0 9 7 4 0 0 4

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE
9 19 0

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what time departs the flight more early of Boston to Denver in TWA?

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)
0 5 1 0 9 7 4 0 0 4

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE
9 19 0

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what time departs the flight more early of Boston to Denver in TWA?

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)
0 5 1 0 9 7 4 0 0 4

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE
9 19 0 09

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what time departs the flight more early of Boston to Denver in TWA?

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)
0 5 1 0 9 7 4 0 0 4

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE
9 19 0 09

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX
3 19 4 2 74 4 02 9 8 9

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what time departs the flight more early of Boston to Denver in TWA?

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)
0 5 1 0 9 7 4 0 0 4

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE
9 19 0 09

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX
3 19 4 2 74 4 02 9 8 9

- 5... ATIS: English to "Pseudo English"

WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?
List departure time of earliest morning TWA flights from Boston and to Denver

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?

Not all the transduction tasks are equally difficult: examples

- 1... Spanish to English, word by word

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
to what time departs the flight more early of Boston to Denver in TWA?

- 2... Division by 7

3 5 7 6 8 1 8 0 3 1 (: 7 =)
0 5 1 0 9 7 4 0 0 4

- 3... English to Decimal

NINEHUNDREDANDNINETEENTHOUSANDANDNINE
9 19 0 09

- 4... Roman to Decimal

III XIX XLII LXXIV CDII CMLXXXIX
3 19 4 2 74 4 02 9 8 9

- 5... ATIS: English to "Pseudo English"

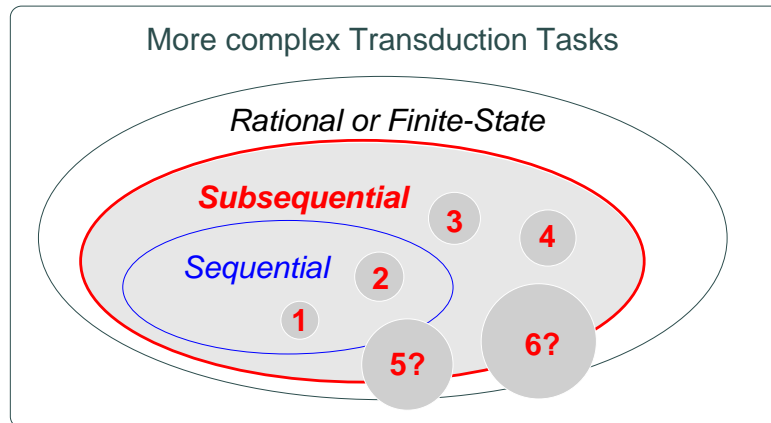
WHAT IS THE DEPARTURE TIME OF TWA EARLIEST FLIGHT FROM BOSTON TO DENVER?
List departure time of earliest morning TWA flights from Boston and to Denver

- 6... Spanish to English

¿A QUE HORA SALE EL VUELO MAS TEMPRANO DE BOSTON A DENVER EN TWA?
What is the departure time of TWA earliest flight from Boston to Denver?

Not all the transduction tasks are equally difficult

- | | |
|-------------------------------------|--------------------------------------|
| 1. Spanish to English, word by word | 4. Roman to Decimal |
| 2. Division by 7 | 5. ATIS: English to "Pseudo English" |
| 3. English to Decimal | 6. Spanish to English |



THE MAIN CONCERN IS THE REQUIRED **degree of “sequentiality”** OR **position monotonicity** BETWEEN INPUT-OUTPUT SUBSEQUENCES

Index

1 Introduction ▷ 2

◦ 2 *Rational or Finite-State Transduction* ▷ 6

3 Stochastic Finite-State Transducers ▷ 11

4 Error Correcting ▷ 20

5 Sequential Transduction ▷ 26

6 Subsequential Transduction:
Introduction to the “OSTI” Algorithm ▷ 32

7 Bibliography ▷ 36

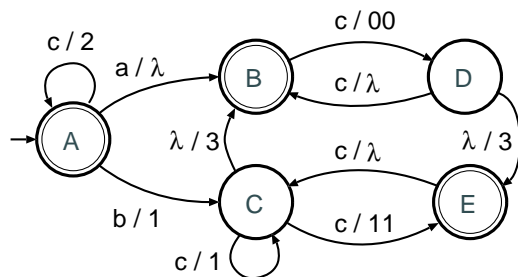
Finite State Transducers (FST): formal definition

A *Finite State or Rational Transducer* τ is a 6-tuple $\tau = (Q, X, Y, q_0, Q_F, E)$:

Q :	Finite set of <i>States</i>
X, Y :	Input and output <i>Alphabets</i>
$q_0 \in Q$:	<i>Initial State</i>
$Q_F \subset Q$:	<i>Set of Final States</i>
$E \subset Q \times X^* \times Y^* \times Q$:	<i>"Edges" or Transitions</i>

Transitions can equivalently defined as $E \subset Q \times (X \cup \lambda) \times Y^* \times Q$.

EXAMPLE

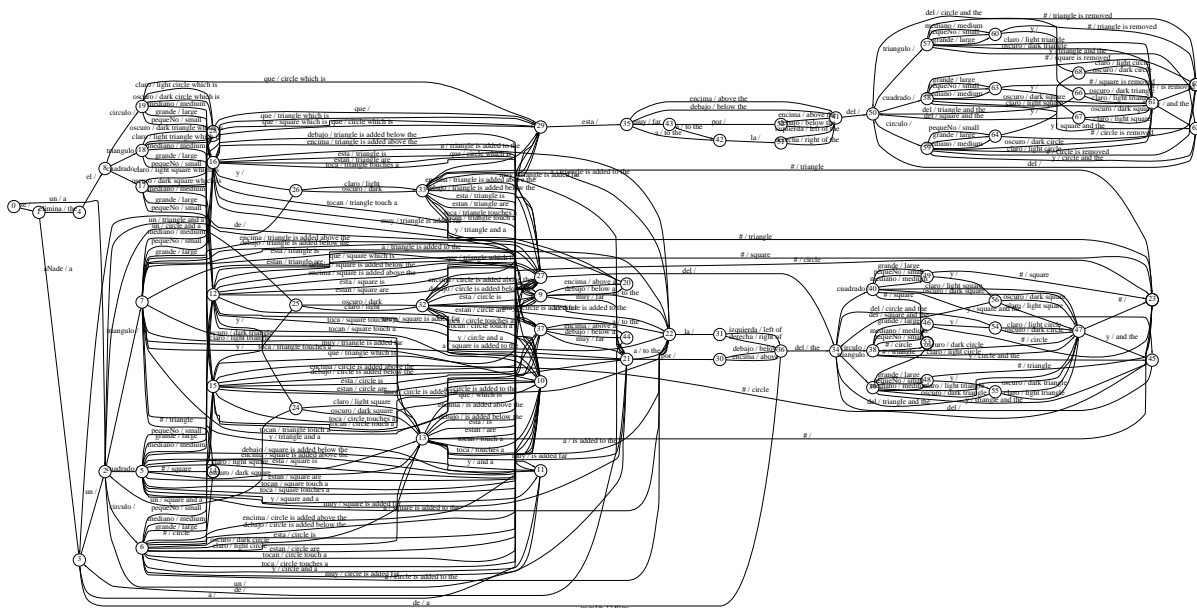


$T_\tau = \{ (\lambda, \lambda), (cb, 213), (ccb, 2213), (a, \lambda), (ac, 003), (cac, 2003), (c, 2), (bc, 111), (cbc, 2111), (b, 13), (bc, 113), (cbc, 213003), (ca, 2), (bc, 13003), (bcc, 1113), (cc, 22), (cca, 22), \dots \}$

Three possible types of ambiguity: **input**, **output** and **path**

Another example of a (very small) FST for a toy, but *real* task

(Learned from MLA Spanish-English training sentences with OSTIADR using input and output 4-Gram constraints)



Finite State Transducers: Paths and Translations

- A **path** \mathcal{P} of a Finite State transducer $\tau = (Q, X, Y, q_0, Q_F, E)$ is a sequence of transitions of E
- A **translation** of τ is pair of strings $(x, y) \in X^* \times Y^*$ such that there is a **path** \mathcal{P} in τ which “matches” x and y ; that is:

$$\mathcal{P} = (q'_1, u_1, v_1, q_1), (q'_2, u_2, v_2, q_2), \dots, (q'_m, u_m, v_m, q_m)$$

$$q'_1 = q_0, \quad q_i = q'_{i+1} \quad 1 \leq i < m, \quad q_m \in Q_F$$

$$x = u_1 \cdots u_m, \quad y = v_1 \cdots v_m$$

- $T_\tau \subset X^* \times Y^* : T_\tau = \{(x, y) \text{ which are translations of } \tau\}$
- Let $\mathcal{P}(\tau, x, y)$ be the set of matching paths of x, y in τ .
 τ is **ambiguous** if $\exists x', y'$ such that $|\mathcal{P}(\tau, x', y')| > 1$

Example: $\mathcal{P}(\tau, bcc, 1113) =$
 $\{(A, b, 1, C)(C, c, 1, C)(C, \lambda, 3, B), (A, b, 1, c)(C, c, 11, E)(C, \lambda, 3, B)\}$

Finite State Transducer Learning and Grammatical Inference

- A Finite State (regular) Grammar (FSG), G , can be seen as a particular case of Finite State Transducer (FST), T which, for each input string x , produces an output string y , such that $y = \text{YES}$ if x belongs to the language of G and $y = \text{NO}$ otherwise.
- Any algorithm that would learn any FST could also learn any FSG and, therefore, learning Finite State Transducers (FST) is at least as hard as learning Finite State (regular) Grammars (FSG).
- Transducer Learning can be properly framed within the paradigm of *Grammatical Inference*

Transducer Identification in the Limit:

Let $f : X^* \rightarrow Y^*$ be a transduction function. A transducer learning algorithm \mathcal{A} is said to *identify f in the limit* if, for any positive presentation S of input-output pairs of f , \mathcal{A} converges to a transduction $g : X^* \rightarrow Y^*$ such that $\forall x \in \text{Dom}(f), g(x) = f(x)$, when the number of pairs in S tends to infinity.

Index

- 1 Introduction ▷ 2
- 2 Rational or Finite-State Transduction ▷ 6
- 3 *Stochastic Finite-State Transducers* ▷ 11
- 4 Error Correcting ▷ 20
- 5 Sequential Transduction ▷ 26
- 6 Subsequential Transduction:
Introduction to the “OSTI” Algorithm ▷ 32
- 7 Bibliography ▷ 36

Stochastic Finite State Transducers

A *Stochastic* Finite State transducer \mathcal{T} is defined by (τ, P, P_F) , where:

- $\tau = (Q, X, Y, q_0, Q_F, E)$ is a Finite State transducer
- $P : E \rightarrow \mathbb{R}^+$ and $P_F : Q_F \rightarrow \mathbb{R}^+$ are functions such that:

$$\sum_{(q', u, v, q) \in E} P(q', u, v, q) + P_F(q') = 1 \quad \forall q' \in Q$$

- Probability of a *path*, \mathcal{P}_m , ending at the state q_m :

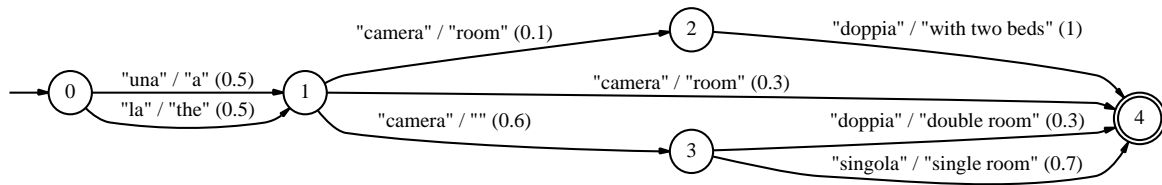
$$Pr(\mathcal{P}_m) = \prod_{(q', u, v, q) \in \mathcal{P}_m} P(q', u, v, q) P_F(q_m)$$

- Probability of a translation (x, y) of τ :

$$P_{\mathcal{T}}(x, y) = \sum_{\mathcal{P}_m \in \mathcal{P}(\tau, x, y)} Pr(\mathcal{P}_m) = \sum_{\mathcal{P}_m \in \mathcal{P}(\tau, x, y)} \prod_{(q', u, v, q) \in \mathcal{P}_m} P(q', u, v, q) P_F(q_m)$$

$P_{\mathcal{T}}(x, y)$ defines a joint distribution in X^*, Y^*

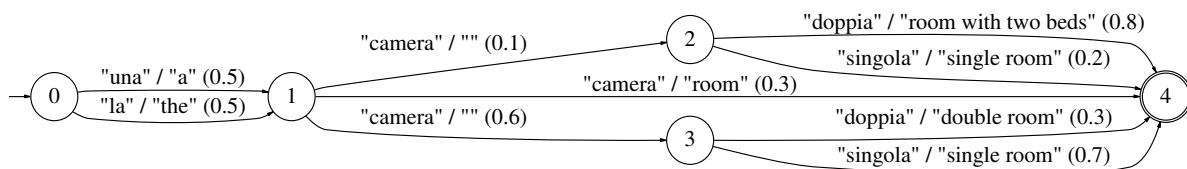
Example of a Stochastic Finite-State Transducer



$$Pr(\text{una camera doppia} , \text{a double room}) = 0.5 \cdot 0.6 \cdot 0.3 = \mathbf{0.09}$$

$$Pr(\text{una camera doppia} , \text{a room with two beds}) = 0.5 \cdot 0.1 \cdot 1.0 = \mathbf{0.05}$$

Stochastic Finite-State Transducer: another example



$$Pr(\text{una camera singola} , \text{a single room}) =$$

$$0.5 \cdot 0.1 \cdot 0.2 + 0.5 \cdot 0.6 \cdot 0.7 = 0.01 + 0.21 = \mathbf{0.22}$$

Stochastic Finite State Transducers: embedded language models

The marginals of the joint probability distribution $P_{\mathcal{T}}(x, y)$ defined by a stochastic finite-state transducer \mathcal{T} are stochastic *regular* languages:

$$P_i(x) = \sum_{y \in Y^*} P_{\mathcal{T}}(x, y), \quad P_o(y) = \sum_{x \in X^*} P_{\mathcal{T}}(x, y).$$

These languages can be properly considered as *input* and *output Language Models* corresponding to \mathcal{T} .

In practice, these Language Models are simply the regular languages associated to the automata obtained by dropping the input and output symbols of each transition of the finite-state transducer, respectively.

Stochastic Finite State Transducers: search problems

- **Most probable path:** given \mathcal{T} , $x \in X^*$, $y \in Y^*$, find

$$\hat{\mathcal{P}} = \underset{\mathcal{P} \in \mathcal{P}(\tau, x', y'); x'=x, y'=y}{\operatorname{argmax}} Pr(\mathcal{P})$$

Efficient solution by Dynamic Programming

- **Most probable translation:** given $x \in X^*$, find

$$\hat{y} = \underset{y \in Y^*}{\operatorname{argmax}} P_{\mathcal{T}}(x, y)$$

No efficient solution (shown to be NP-Hard!).

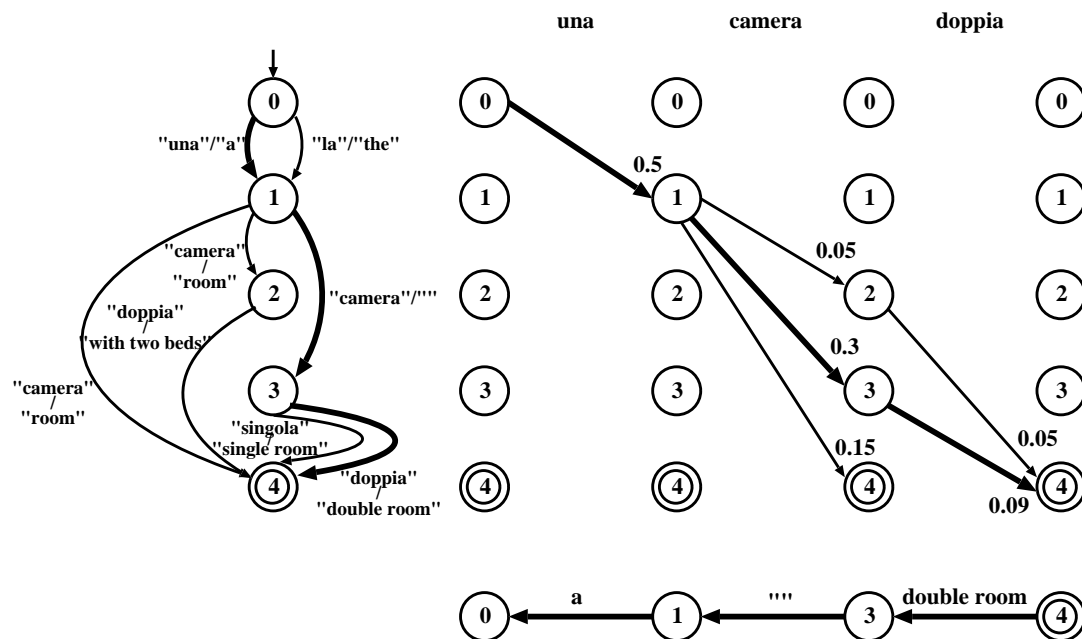
Approximation:

$$\tilde{y} = \underset{\mathcal{P} \in \mathcal{P}(\tau, x', y'); x'=x, y' \in Y^*}{\operatorname{argmax}} Pr(\mathcal{P})$$

Efficient solution by Viterbi search

Both problems are easy if τ is un-ambiguous – trivial if τ is deterministic

Example of Viterbi translation



$$\operatorname{argmax}_y \Pr(\text{"una camera doppia", } y) \approx \text{"a double room"}$$

Learning Stochastic Finite State Transducers

Three main families of techniques to learn a SFST from a parallel corpus of source-target sentences:

- **Traditional syntactic pattern recognition paradigm:**

- Learn the SFST “topology” (the *states and transitions*)
- Estimate the probabilities from the same data

Problem: The class of finite-state transducers as a whole is at least as hard to learn as the class of finite-state automata!

⇒ Try to learn adequate subclasses and/or use heuristics!

- **Hybrid methods:** Under the *traditional* paradigm, use statistical methods to guide the structure learning

- **Pure statistical approach (new):**

- Adequately parameterize the SFST structure and consider it as a hidden variable
- Estimate everything by Expectation Maximization (EM)

Estimating probabilities of Stochastic Finite State Transducers

- **Estimating transition and final-state probabilities:**
 - *Un-ambiguous transducers:*
Maximum likelihood estimation from the frequency of use of transition and states in the paths matching the training pairs
 - *Ambiguous transducers:*
EM re-estimation based on a *forward-backward*-like algorithm or a Viterbi-like approximation [Picó & Casacuberta, 01]
- **Modeling of unseen events – smoothing:**
 - *Back-off and interpolation*
Adapted from techniques used in language modeling [Llorens 01]
(so far fully developed only for techniques based on N-Grams)
 - *Stochastic error-correcting parsing*
Given a source sentence, x , find a path in the transducer that error-correcting matches x with maximum probability

Index

- 1 Introduction ▷ 2
- 2 Rational or Finite-State Transduction ▷ 6
- 3 Stochastic Finite-State Transducers ▷ 11
- 4 *Error Correcting* ▷ 20
- 5 Sequential Transduction ▷ 26
- 6 Subsequential Transduction:
Introduction to the “OSTI” Algorithm ▷ 32
- 7 Bibliography ▷ 36

Error Correcting techniques: Motivation

Often needed to allow parsing unseen input sentences through learned Finite State models that do not completely “cover” the input language:

- Can be understood as a kind of *smoothing* that can be applied to most types of Finite State devices.
- Explicitly copes with “*imperfect*” *input sentences* (i.e., sentences inappropriately modeled by the trained models).
- Also copes with *insufficiently trained models*. In an extreme view, this is similar to Memory Based techniques, where only the (raw) training data is considered (no generalization).

Finite State Error Correcting Parsing

- Each input sentence, x , is considered as a *corrupted* version of some sentence $x' \in L$, where L is the language (domain) of the FSM.
- The corruption process is modelled by means of an *Error Model* E , that accounts for (single-word) substitutions, insertions and deletions.
- The parsing of x consists in finding a string \hat{x} in L which has *maximum posterior probability* of having been distorted into x ; that is,

$$\hat{x} = \underset{x' \in L}{\operatorname{argmax}} P(x'|x) = \underset{x' \in L}{\operatorname{argmax}} P_L(x') \cdot P_E(x|x')$$

where $P_L(x')$ is the probability of x' in L , given by the (input part of the) FSM, and $P_E(x|x')$ is the probability of x being a corrupted version of x' according to E .

The resulting translation, y' , is the string associated to \hat{x} through the SST.

Finite State Error Correcting Parsing: training and search

- The parameters of $P_E(x|x')$ are estimated from a set of “distorted” sentences; i.e., sentences that can not be exactly parsed through the given FST.
- If both $P_E(x|x')$ and $P_L(x)$ are given by a Finite-State models, the Error Correcting search

$$\hat{x} = \underset{x' \in L}{\operatorname{argmax}} P_L(x') \cdot P_E(x|x')$$

can be efficiently performed through appropriate extensions of the *Viterbi Algorithm* [Amengual & Vidal, PAMI-1999]

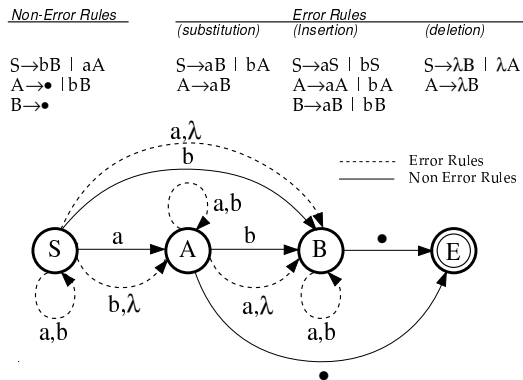
Efficient Finite State Error Correcting Parsing

The required search can be efficiently performed through appropriate extensions of the *Viterbi Algorithm*:

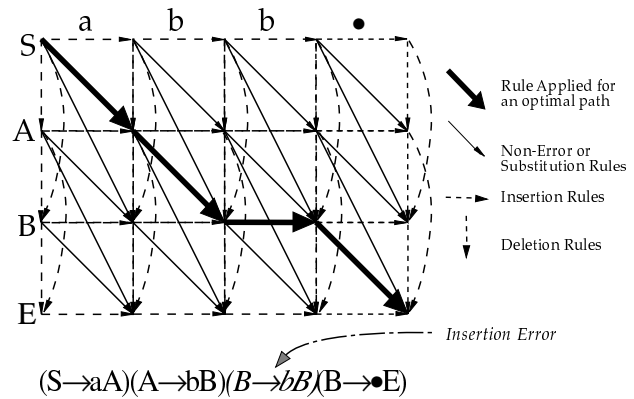
- For every input language word, a *loop* transition is added to each *state* of the FSM to account for *insertion-errors*.
- Each *transition* is expanded with the appropriate *substitution-error* transitions plus an *empty transition* to account for a *deletion-error*.
- The standard Viterbi trellis is extended with “horizontal” arcs for the insertion-errors and “vertical” arcs for the deletion-errors.
- Actual expansion of the trellis is not necessary (nor generally possible!). Virtual expansion is one of the key issues for efficiency.
- Efficient techniques can be used to process the scores in each stage of the trellis, overcoming the trouble raised by vertical arcs [Amengual & Vidal, PAMI-1999].

Error Correcting Parsing: example

Grammar, equivalent Automaton and
(virtual) Error-Correcting extensions



Ins-Sub-Del-Extended trellies and
Error-Correcting Parsing process

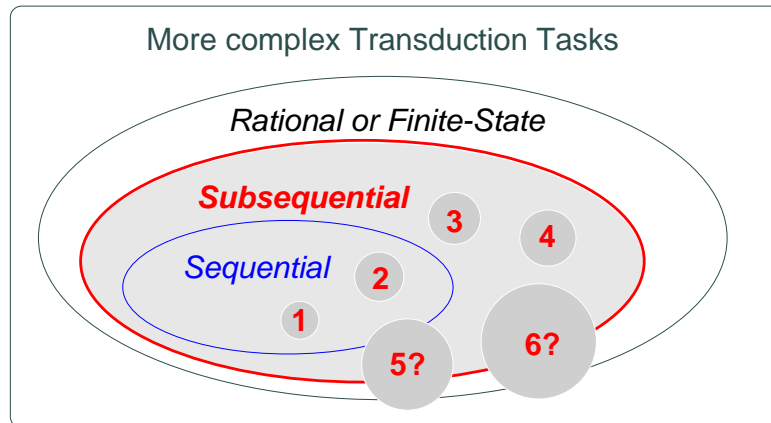


Index

- 1 Introduction ▷ 2
- 2 Rational or Finite-State Transduction ▷ 6
- 3 Stochastic Finite-State Transducers ▷ 11
- 4 Error Correcting ▷ 20
- 5 *Sequential Transduction* ▷ 26
- 6 Subsequential Transduction:
Introduction to the “OSTI” Algorithm ▷ 32
- 7 Bibliography ▷ 36

Not all the transduction tasks are equally difficult

- | | |
|-------------------------------------|--------------------------------------|
| 1. Spanish to English, word by word | 4. Roman to Decimal |
| 2. Division by 7 | 5. ATIS: English to "Pseudo English" |
| 3. English to Decimal | 6. Spanish to English |



THE MAIN CONCERN IS THE REQUIRED **degree of “sequentiality”** OR **position monotonicity** BETWEEN INPUT-OUTPUT SUBSEQUENCES

Sequential Transducers

A *Sequential Transducer* (ST) τ is a 5-tuple $\tau = (Q, X, Y, q_0, E)$:

Q :	Finite set of <i>States</i>
X, Y :	Input and output <i>Alphabets</i>
$q_0 \in Q$:	<i>Initial State</i>
$E \subset Q \times X \times Y^* \times Q$:	<i>“Edges” or Transitions</i>

- All the states are *accepting*
- Edges are *deterministic*:
 $(q, a, u, r), (q, a, v, s) \in E \Rightarrow (u = v \wedge r = s)$

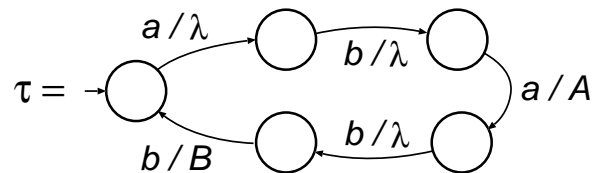
PROPERTIES:

1. T_τ is a *function*: $X^* \rightarrow Y^*$
2. STs \equiv *Generalized Sequential Machines* \supset (Mealy and Moore machines)
3. STs *preserve prefixes*: $T_\tau(\lambda) = \lambda$; $T_\tau(uv) \in T_\tau(u)Y^*$

“Property” 2 entails *strict sequentiality*,
which can hardly be adequate in many cases of interest

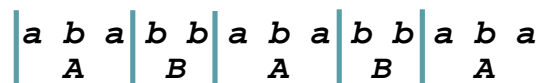
An example of Sequential Transduction; sequential segmentation

$$X = \{a, b\}; Y = \{A, B\}$$

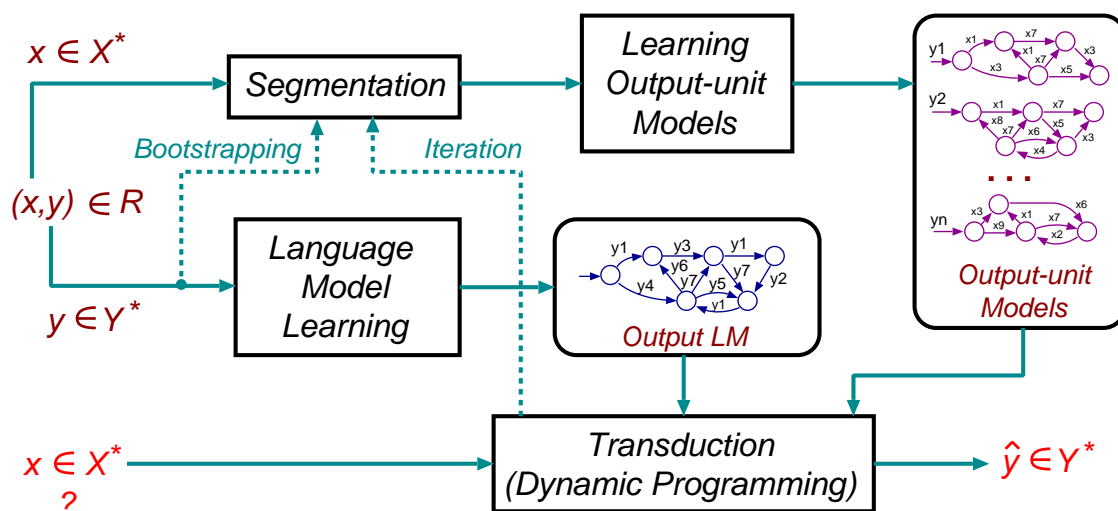


$$T_\tau = \{(\lambda, \lambda), (aba, A), (ababbaba, ABA), (ababbababb), (ABAB), \dots\}$$

Sequential segmentation of the input string "ababbababbaba"

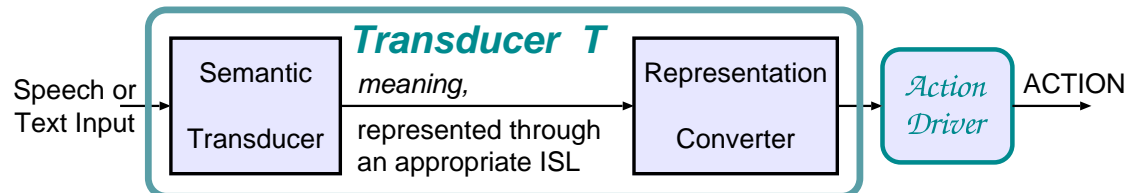


Learning Sequential Transducers using language learning (GI) techniques



Applications of Sequential Transduction: Language Understanding using “intermediate” semantic languages

Basic idea: Split the single-block transducer T into two blocks, using an **Intermediate Semantic Language** (ISL) which is *sequential* with the input



Example (Spanish numbers to Decimal)

<i>Input:</i>	dos		cientos		do		ce		mil		dieci		seis
<i>ISL:</i>	+2		*100		+2		+10) * 1000 (+10		+6

Example (From the ATIS task)

<i>Input:</i>	I'd like to fly		from Boston		to Denver		with American Airlines		on Tuesday
<i>ISL:</i>	REQ=FLIGHTS		ORG=BBOS		DST=DDEN		AIRLINE=AA		WEEKDAY=TU

Index

- 1 Introduction ▷ [2](#)
- 2 Rational or Finite-State Transduction ▷ [6](#)
- 3 Stochastic Finite-State Transducers ▷ [11](#)
- 4 Error Correcting ▷ [20](#)
- 5 Sequential Transduction ▷ [26](#)
- 6 *Subsequential Transduction: Introduction to the “OSTI” Algorithm* ▷ [32](#)
- 7 Bibliography ▷ [36](#)

Subsequential Transduction

[Berstel, 79]

A *Subsequential Transducer* (SST) τ is a 6-tuple $\tau = (Q, X, Y, q_0, E, \sigma)$, where:

- $\tau' = (Q, X, Y, q_0, E)$ is a Sequential Transducer
- $\sigma : Q \rightarrow Y^*$ is a *state output* (partial) *function*
- For each input string x , the output string y is obtained by concatenating $\sigma(q)$ to $\tau'(x)$, where q is the last state reached through the analysis of x by τ' ; i.e.:

$$y = \tau(x) = \tau'(x)\sigma(q)$$

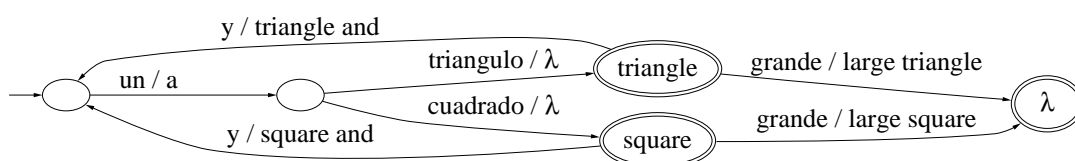
PROPERTIES:

1. T_τ is a *function*: $X^* \rightarrow Y^*$
2. Sequential \subset **Subsequential Transduction** \subset Finite State.
3. Input-output monotonicity (sequentiality) needs *not* be as strict as in STs.

Subsequential Transducers (intuitive concept)

- **Deterministic Finite State Networks** which accept sentences from an *input* language and produce sentences of an *output* language.
- In addition to input symbols, output strings are assigned to the edges.
- Output strings are also assigned to final states.
- **SST operation relies on “delaying” the production of output symbols** until enough of the input sentence has been seen to guarantee a correct output.

An example of SST:



Learning SSTs: the OSTI Algorithm

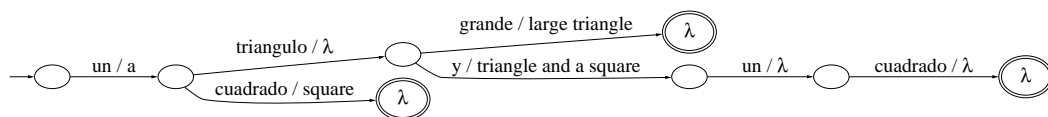
[Oncina, 91-93]

SSTs can be learned from training examples using the **Onward Subsequential Transducer Inference Algorithm (OSTIA)**.

1. Build an **“onward” tree representation** of the training data (a tree in which output strings are as close as possible to the root – called “OTST”)

Example:

(*un triángulo y un cuadrado* , *a triangle and a square*),
 (*un triángulo grande* , *a large triangle*),
 (*un cuadrado* , *a square*)



2. Orderly traverse the tree, while **merging states** in order to get, hopefully, adequate generalizations.

Index

- 1 Introduction ▷ 2
- 2 Rational or Finite-State Transduction ▷ 6
- 3 Stochastic Finite-State Transducers ▷ 11
- 4 Error Correcting ▷ 20
- 5 Sequential Transduction ▷ 26
- 6 Subsequential Transduction:
Introduction to the “OSTI” Algorithm ▷ 32
- 7 Bibliography ▷ 36

Bibliography

- E.Vidal, P.García, E.Segarra: "Inductive Learning of Finite-State Transducers for Interpretation of Unidimensional Objects". In "Structural Pattern Analysis," pp.17-35. R.Mohr, T.Pavlidis, A.Sanfeliu, eds., World Scient. Pub., Series in Computer Science, 19, 1990.
- J.Oncina, P.García, E.Vidal: "Learning Subsequential Transducers for Pattern Recognition Interpretation Tasks". IEEE Trans. on Pattern Analysis and Machine Intelligence. Vol.PAMI-15, No.5, pp.448-458, 1993.
- E.Vidal: "Language Learning, Understanding and Translation". En "Progress and Prospects of Speech Research and Technology", pp.131-140. H.Niemann, R.de Mori, G.Hanrieder (Eds.). Infix, 1994.
- J.C.Amengual, E.Vidal: "Efficient Error-Correcting Viterbi Parsing". IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.20, no. 10, 1998.
- A.Castellanos, E.Vidal, A.Varó, J.Oncina: "Language Understanding and Subsequential Transducer Learning". Computer Speech and Language, No.12, pp.193-228. 1998.
- E.Vidal, F.Thollard, C.de la Higuera, F.Casacuberta and R.C.Carrasco: "Probabilistic Finite-State Machines – Parts I and II" IEEE Trans on Pattern Analysis and Machine Intelligence (PAMI), 2005. To appear.

Pattern Recognition approaches to Machine Translation

E. Vidal and F. Casacuberta

Pattern Recognition and Human Language Technology Group

Departament de Sistemes Informàtics i Computació

Institut Tecnològic d'Informàtica

Universitat Politècnica de València

5: Phrase-based Models and Alignment Templates

Francisco Casacuberta Nolla

`fcn@iti.upv.es`

24-28 January 2005

F. Casacuberta – DSIC-ITI-UPV

[Pattern Recognition approaches to Machine Translation](#)

[Statistical Alignment Models](#)

Index

- 1 Beyond word models ▶ [2](#)
- 2 Phrase-based models ▶ [9](#)
- 3 Alignment templates ▶ [47](#)
- 4 Phrases and finite-state transducers ▶ [58](#)
- 5 Using linguistic knowledge ▶ [66](#)
- 6 Bibliography ▶ [72](#)

Index

- 1 *Beyond word models* ▷ 2
- 2 Phrase-based models ▷ 9
- 3 Alignment templates ▷ 47
- 4 Phrases and finite-state transducers ▷ 58
- 5 Using linguistic knowledge ▷ 66
- 6 Bibliography ▷ 72

Exemple of word alignments

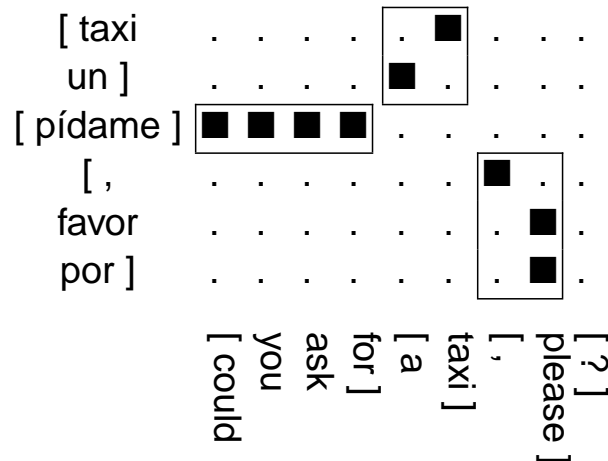
taxi	■	.	.	.
un	■
pídame	■	■	■	■
,	■	.	.
favor	■	.
por	■	.
	could	you	ask	for	a	taxi	,	please	?

Segment alignment

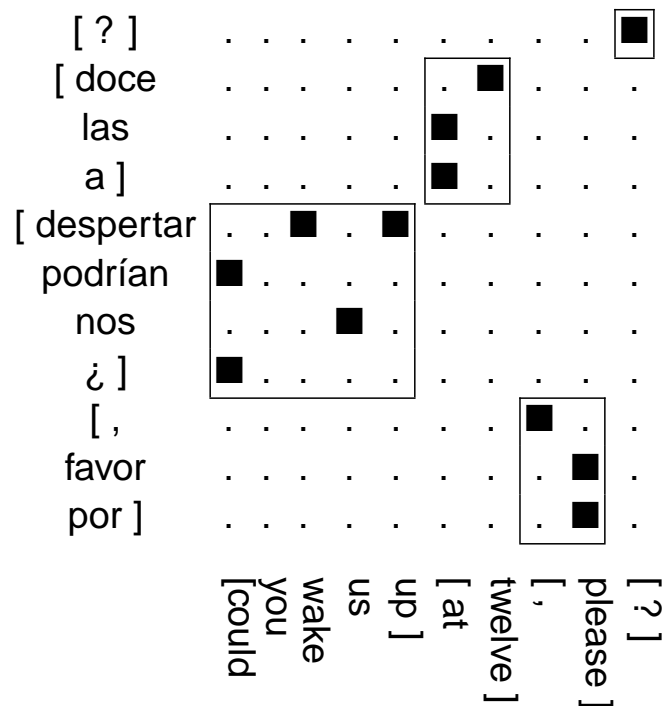
SINGLE-WORD ALIGNMENTS: only model the correspondence between words.

Alternative:

SEGMENT ALIGNMENTS: modelling the correspondences between word segments.



Segment alignment

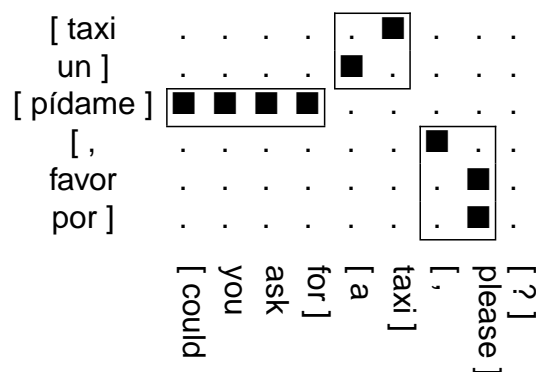


Beyond word-based models

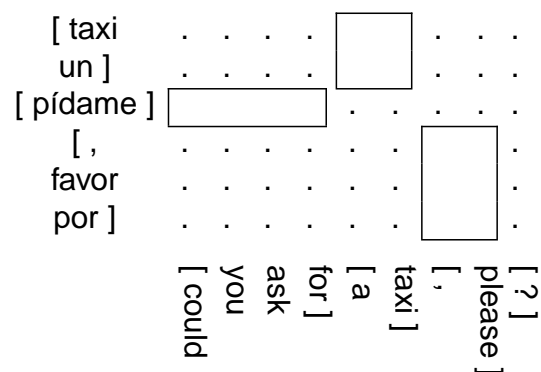
- The basic assumption in the current word-based models: Each source word is generated by only one target word.
- This assumption does not correspond to the nature of natural language. In some cases, it is necessary to know the context.
- Solutions:
 - *Context-dependent dictionaries* (previous talk). The basic unit is the word.
 - *Word sequences*:
 - * *Alignment templates*: A sequence of source (classes of) words is aligned with a sequence of target (classes of) words. Inside the templates there are word-to-word correspondences. The basic unit is the word.
 - * *Phrase-based models*:¹ A sequence of source words is aligned with a sequence of target words. The basic unit is the phrase.

¹By “phrase” we will mean a possible word sequence.

Word sequences



Alignment templates



Bilingual phrases

Phrase-based models

The statistical dictionaries of single word pairs are substituted by statistical dictionaries of *bilingual phrases*.

Bilingual phrases are related with a bilingual segmentation.

- Problem: The generalisation capability, since only sequences of segments that have been seen in the training corpus are accepted.
- Problem: The selection of adequate bilingual phrases.

Index

- 1 Beyond word models ▷ 2
- 2 *Phrase-based models* ▷ 9
- 3 Alignment templates ▷ 47
- 4 Phrases and finite-state transducers ▷ 58
- 5 Using linguistic knowledge ▷ 66
- 6 Bibliography ▷ 72

An example

y: could you ask for a taxi , please ?									
Segmentation	y	could	you	ask	for	a	taxi	,	please ?
	i	1	2	3	4	5	6	7	8 9=I
	μ				μ_1		μ_2		μ_3
Permutation	α		$\alpha_1 = 2$			$\alpha_2 = 3$		$\alpha_3 = 1$	
		,	please	?	could you ask for		a	taxi	
Translation	x	por	favor	,		pídame	un	taxi	.
	j	1	2	3		4	5	6	7
Segmentation	γ			γ_{α_3}		γ_{α_1}			γ_{α_2}
x: por favor , pídame un taxi .									

General framework

- Let K be the number of segments in x and in y,
- Segmentation of the target sentence

$$\mu : \{1, \dots, K\} \rightarrow \{1, \dots, I\} : \mu_k \geq \mu_{k-1} \quad 1 < k \leq K \quad \& \quad \mu_K = I \quad (\mu_0 = 0)$$

- Segmentation of the source sentence

$$\gamma : \{1, \dots, K\} \rightarrow \{1, \dots, J\} : \gamma_k \geq \gamma_{k-1} \quad 1 < k \leq K \quad \& \quad \gamma_K = J \quad (\gamma_0 = 0)$$

- Segment alignment (Permutation):

$$\alpha : \{1, \dots, K\} \rightarrow \{1, \dots, K\} : \alpha(k) = \alpha(k') \quad \text{iff} \quad k = k'$$

Monotone vs. no monotone alignments

NO MONOTONE ALIGNMENT

$$\Pr(\mathbf{x}|\mathbf{y}) \approx P(\mathbf{x}|\mathbf{y}) = p(J|I) \cdot \sum_K \sum_{\mu_1^K} \sum_{\alpha_1^K} \sum_{\gamma_1^K} \prod_{k=1}^K p(\alpha_k|\alpha_{k-1}) \cdot p(\mathbf{x}_{\gamma_{\alpha_k-1}+1}^{\gamma_{\alpha_k}} | \mathbf{y}_{\mu_{k-1}+1}^{\mu_k})$$

$$\text{MONOTONE ALIGNMENT} \Rightarrow \alpha_k = k$$

$$\Pr(\mathbf{x}|\mathbf{y}) \approx P(\mathbf{x}|\mathbf{y}) = p(J|I) \cdot \sum_K \sum_{\mu_1^K} \sum_{\gamma_1^K} \prod_{k=1}^K p(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k} | \mathbf{y}_{\mu_{k-1}+1}^{\mu_k})$$

Maximum approaches

NO MONOTONE ALIGNMENT

$$\Pr(\mathbf{x}|\mathbf{y}) \approx \hat{P}(\mathbf{x}|\mathbf{y}) = p(J|I) \cdot \max_K \max_{\mu_1^K} \max_{\alpha_1^K} \max_{\gamma_1^K} \prod_{k=1}^K p(\alpha_k|\alpha_{k-1}) \cdot p(\mathbf{x}_{\gamma_{\alpha_k-1}+1}^{\gamma_{\alpha_k}} | \mathbf{y}_{\mu_{k-1}+1}^{\mu_k})$$

$$\text{MONOTONE ALIGNMENT} \Rightarrow \alpha_k = k$$

$$\Pr(\mathbf{x}|\mathbf{y}) \approx \hat{P}(\mathbf{x}|\mathbf{y}) = p(J|I) \cdot \max_K \max_{\mu_1^K} \max_{\gamma_1^K} \prod_{k=1}^K p(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k} | \mathbf{y}_{\mu_{k-1}+1}^{\mu_k})$$

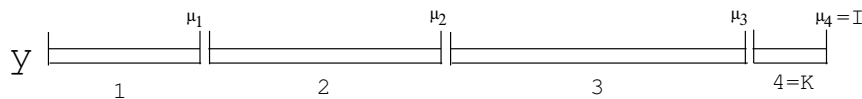
Formal derivation (I)

$$\hat{y} = \operatorname{argmax}_y \Pr(y|x) = \operatorname{argmax}_y \Pr(y) \cdot \Pr(x|y)$$

$$\begin{aligned} \Pr(x | y_1^I) &= \Pr(J | y_1^I) \cdot \Pr(x | y_1^I, J) \\ &= \Pr(J | y_1^I) \cdot \sum_K \Pr(K | y_1^I, J) \cdot \Pr(x | y_1^I, J, K) \end{aligned}$$

Segmentation of target sentences

$$\mu : \{1, \dots, K\} \rightarrow \{1, \dots, I\} : \mu_k \geq \mu_{k-1} \quad 1 < k \leq K \quad \& \quad \mu_K = I \quad (\mu_0 = 0)$$



$$\Pr(x | y_1^I) = \Pr(J | y_1^I) \cdot \sum_K \Pr(K | y_1^I, J) \cdot \sum_{\mu_1^K} \Pr(x_1^I, \mu_1^K | y_1^I, J, K)$$

An example (I)

x: por favor , pícame un taxi

y: could you ask for a taxi , please ?

y	could	you	ask	for	a	taxi	,	please	?
1	1	2	3	4	5	6	7	8	9=I

Number of segments in y: $K = 3$

Segmentation of y: μ

y	could	you	ask	for	a	taxi	,	please	?
i	1	2	3	4	5	6	7	8	9=I
μ				μ_1		μ_2			μ_3

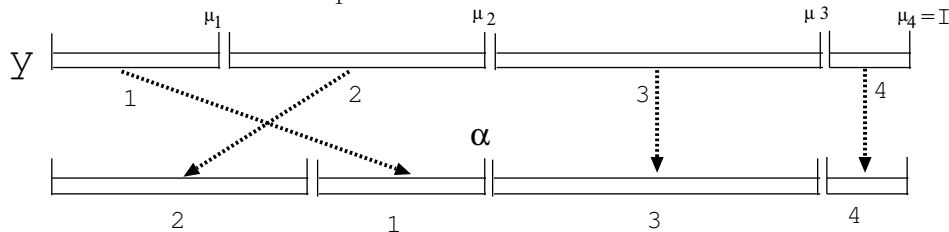
Formal derivation (II)

$$\Pr(\mathbf{x} | \mathbf{y}_1^I) = \Pr(J | \mathbf{y}_1^I) \cdot \sum_K \sum_{\mu_1^K} \Pr(K | \mathbf{y}_1^I, J) \cdot \Pr(\mu_1^K | \mathbf{y}_1^I, J, K) \cdot \Pr(\mathbf{x}_1^J | \mathbf{y}_1^I, J, K, \mu_1^K)$$

Permutation of target segments:

$$\alpha : \{1, \dots, K\} \rightarrow \{1, \dots, K\} : \alpha(k) = \alpha(k') \text{ iff } k = k'$$

$$\begin{aligned} \Pr(\mathbf{x}_1^J | \mathbf{y}_1^I, J, K, \mu_1^K) &= \sum_{\alpha_1^K} \Pr(\mathbf{x}_1^J, \alpha_1^K | \mathbf{y}_1^I, J, K, \mu_1^K) \\ &= \sum_{\alpha_1^K} \Pr(\alpha_1^K | \mathbf{y}_1^I, J, K, \mu_1^K) \cdot \Pr(\mathbf{x}_1^J | \mathbf{y}_1^I, J, K, \mu_1^K, \alpha_1^K) \end{aligned}$$



An example (II)

x: por favor , pídame un taxi

y: could you ask for a taxi , please ?

Segmentation of y: μ

y	could	you	ask	for	a	taxi	,	please	?
i	1	2	3	4	5	6	7	8	9=I
μ				μ_1		μ_2			μ_3
α		$\alpha_1 = 2$			$\alpha_2 = 3$			$\alpha_3 = 1$	

Permutation of segments in y: α

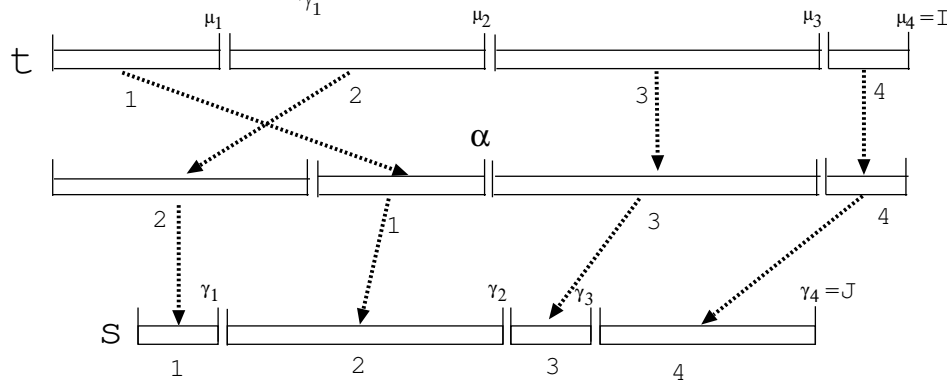
y	,	please	?	could	you	ask	for	a	taxi
i	7	8	9	1	2	3	4	5	6

Formal derivation (III)

Segmentation of source segments:

$$\gamma : \{1, \dots, K\} \rightarrow \{1, \dots, J\} : \gamma_k \geq \gamma_{k-1} \quad 1 < k \leq K \quad \& \quad \gamma_K = J \quad (\gamma_0 = 0)$$

$$\begin{aligned} \Pr(\mathbf{x}_1^J \mid \mathbf{y}_1^I, J, K, \mu_1^K, \alpha_1^K) &= \sum_{\gamma_1^K} \Pr(\mathbf{x}_1^J, \gamma_1^K \mid \mathbf{y}_1^I, J, K, \mu_1^K, \alpha_1^K) \\ &= \sum_{\gamma_1^K} \Pr(\gamma_1^K \mid \mathbf{y}_1^I, J, K, \mu_1^K, \alpha_1^K) \cdot \Pr(\mathbf{x}_1^J \mid \mathbf{y}_1^I, J, K, \mu_1^K, \alpha_1^K, \gamma_1^K) \end{aligned}$$



An example (III)

x: por favor , pícame un taxi y: could you ask for a taxi , please ?

Segmentation of y: μ

y	could	you	ask	for	a	taxi	,	please	?
i	1	2	3	4	5	6	7	8	9=I
μ				μ_1	μ_2				μ_3
α		$\alpha_1 = 2$			$\alpha_2 = 3$		$\alpha_3 = 1$		

Permutation of segments in y: α

y	,	please	?	could	you	ask	for	a	taxi
i	7	8	9	1	2	3	4	5	6

Segmentation of x: γ

j	1	2	3	4	5	6=J
γ		γ_1	γ_2	γ_3		

Segments of x

x | por favor , | pícame | un taxi |

Summary

$$\begin{aligned} \Pr(\mathbf{x}|\mathbf{y}_1^I) &= \Pr(J | \mathbf{y}_1^I) \cdot \sum_K \sum_{\mu_1^K} \sum_{\alpha_1^K} \sum_{\gamma_1^K} \Pr(K | \mathbf{y}_1^I, J) \cdot \Pr(\mu_1^K | \mathbf{y}_1^I, J, K) \cdot \\ &\quad \Pr(\alpha_1^K | \mathbf{y}_1^I, J, K, \mu_1^K) \cdot \Pr(\gamma_1^K | \mathbf{y}_1^I, J, K, \mu_1^K, \alpha_1^K) \cdot \Pr(\mathbf{x}_1^J | \mathbf{y}_1^I, J, K, \mu_1^K, \alpha_1^K, \gamma_1^K) \end{aligned}$$

$$\Pr(\mu_1^K | \mathbf{y}_1^I, J, K) = \prod_{k=1}^K \Pr(\mu_k | \mathbf{y}_1^I, J, K, \mu_1^{k-1})$$

$$\Pr(\alpha_1^K | \mathbf{y}_1^I, J, K, \mu_1^K) = \prod_{k=1}^K \Pr(\alpha_k | \mathbf{y}_1^I, J, K, \mu_1^K, \alpha_1^{k-1})$$

$$\Pr(\gamma_1^K | \mathbf{y}_1^I, J, K, \mu_1^K, \alpha_1^K) = \prod_{k=1}^K \Pr(\gamma_{\alpha_k} | \mathbf{y}_1^I, J, K, \mu_1^K, \alpha_1^K, \gamma_{\alpha_1}, \dots, \gamma_{\alpha_{k-1}})$$

$$\Pr(\mathbf{x}_1^J | \mathbf{y}_1^I, J, K, \mu_1^K, \alpha_1^K, \gamma_1^K) = \prod_{k=1}^K \Pr(\mathbf{x}_{\gamma_{\alpha_{k-1}+1}}^{\gamma_{\alpha_k}} | \mathbf{y}_1^I, J, K, \mu_1^K, \alpha_1^K, \gamma_1^K, \mathbf{x}_{\gamma_{\alpha_1-1}+1}^{\gamma_{\alpha_1}}, \dots, \mathbf{x}_{\gamma_{\alpha_{k-1}-1}+1}^{\gamma_{\alpha_{k-1}}})$$

An approach

$$\Pr(J | \mathbf{y}_1^I) \approx p(J|I)$$

$$\Pr(K | \mathbf{y}_1^I, J) \approx p(K|I, J)$$

$$\Pr(\mu_k | \mathbf{y}_1^I, J, K, \mu_1^{k-1}) \approx p(\mu_k|I)$$

$$\Pr(\alpha_k | \mathbf{y}_1^I, J, K, \mu_1^K, \alpha_1^{k-1}) \approx p(\alpha_k|\alpha_{k-1})$$

$$\Pr(\gamma_{\alpha_k} | \mathbf{y}_1^I, J, K, \mu_1^K, \alpha_1^K, \gamma_{\alpha_1}, \dots, \gamma_{\alpha_{k-1}}) \approx p(\gamma_{\alpha_k}|I)$$

$$\Pr(\mathbf{x}_{\gamma_{\alpha_{k-1}+1}}^{\gamma_{\alpha_k}} | \mathbf{y}_1^I, J, K, \mu_1^K, \alpha_1^K, \gamma_1^K, \mathbf{x}_{\gamma_{\alpha_1-1}+1}^{\gamma_{\alpha_1}}, \dots, \mathbf{x}_{\gamma_{\alpha_{k-1}-1}+1}^{\gamma_{\alpha_{k-1}}}) \approx p(\mathbf{x}_{\gamma_{\alpha_{k-1}+1}}^{\gamma_{\alpha_k}} | \mathbf{y}_{\mu_{k-1}+1}^{\mu_k})$$

Another approach

$$\Pr(J \mid \mathbf{y}_1^I) \approx p(J|I)$$

$$\Pr(K \mid \mathbf{y}_1^I, J) \approx p(K|I, J)$$

$$\Pr(\mu_k \mid \mathbf{y}_1^I, J, K, \mu_1^{k-1}) \approx p(\mu_k|I, \mu_{k-1})$$

$$\Pr(\alpha_k \mid \mathbf{y}_1^I, J, K, \mu_1^K, \alpha_1^{k-1}) \approx p(\alpha_k|\alpha_{k-1})$$

$$\Pr(\gamma_{\alpha_k} \mid \mathbf{y}_1^I, J, K, \mu_1^K, \alpha_1^K, \gamma_{\alpha_1}, \dots, \gamma_{\alpha_{k-1}}) \approx p(\gamma_{\alpha_k}|J, \mu_k, \gamma_{\alpha_{k-1}})$$

$$\Pr(\mathbf{x}_{\gamma_{\alpha_k-1}+1}^{\gamma_{\alpha_k}} \mid \mathbf{y}_1^I, J, K, \mu_1^K, \alpha_1^K, \gamma_1^K, \mathbf{x}_{\gamma_{\alpha_1-1}+1}^{\gamma_{\alpha_1}}, \dots, \mathbf{x}_{\gamma_{\alpha_{k-1}-1}+1}^{\gamma_{\alpha_{k-1}}}) \approx p(\mathbf{x}_{\gamma_{\alpha_k-1}+1}^{\gamma_{\alpha_k}} \mid \mathbf{y}_{\mu_{k-1}+1}^{\mu_k})$$

An extra approach

$$\Pr(J \mid \mathbf{y}_1^I) \approx p(J|I)$$

$$\Pr(K \mid \mathbf{y}_1^I, J) \approx p(K|I, J)$$

$$\Pr(\mu_k \mid \mathbf{y}_1^I, J, K, \mu_1^{k-1}) \approx p(\mu_k|I, \mathbf{y}_{\mu_{k-1}}^{\mu_k})$$

$$\Pr(\alpha_k \mid \mathbf{y}_1^I, J, K, \mu_1^K, \alpha_1^{k-1}) \approx p(\alpha_k|\alpha_{k-1})$$

$$\Pr(\gamma_{\alpha_k} \mid \mathbf{y}_1^I, J, K, \mu_1^K, \alpha_1^K, \gamma_{\alpha_1}, \dots, \gamma_{\alpha_{k-1}}) \approx \frac{p(\gamma_{\alpha_k}, \mathbf{x}_{\gamma_{\alpha_k}-1}, \mathbf{x}_{\gamma_{\alpha_k}})}{\sum_{s, s'} p(\gamma_{\alpha_k}, s, s')}$$

$$\Pr(\mathbf{x}_{\gamma_{\alpha_k-1}+1}^{\gamma_{\alpha_k}} \mid \mathbf{y}_1^I, J, K, \mu_1^K, \alpha_1^K, \gamma_1^K, \mathbf{x}_{\gamma_{\alpha_1-1}+1}^{\gamma_{\alpha_1}}, \dots, \mathbf{x}_{\gamma_{\alpha_{k-1}-1}+1}^{\gamma_{\alpha_{k-1}}}) \approx p(\mathbf{x}_{\gamma_{\alpha_k-1}+1}^{\gamma_{\alpha_k}} \mid \mathbf{y}_{\mu_{k-1}+1}^{\mu_k})$$

Encore un plus

$$\Pr(J \mid y_1^I) \approx p(J|I)$$

$$\Pr(K \mid y_1^I, J) \approx p(K|I, J)$$

$$\Pr(\mu_k \mid y_1^I, J, K, \mu_1^{k-1}) \approx p(\mu_k|I, \mu_{k-1}, y_{\mu_{k-2}+1}^{\mu_{k-1}})$$

$$\Pr(\alpha_k \mid y_1^I, J, K, \mu_1^K, \alpha_1^{k-1}) \approx p(\alpha_k|k, K) \cdot \prod_{l=1}^k (1 - \delta(\alpha_k, \alpha_l))$$

$$\Pr(\gamma_{\alpha_k} \mid y_1^I, J, K, \mu_1^K, \alpha_1^K, \gamma_{\alpha_1}, \dots, \gamma_{\alpha_{k-1}}) \approx p(\gamma_{\alpha_k} - \gamma_{\alpha_{k-1}} \mid K, \mu_k - \mu_{k-1})$$

$$\Pr(\mathbf{x}_{\gamma_{\alpha_{k-1}+1}^{\alpha_k}}^{\gamma_{\alpha_k}} \mid y_1^I, J, K, \mu_1^K, \alpha_1^K, \gamma_1^K, \mathbf{x}_{\gamma_{\alpha_1-1}+1}^{\gamma_{\alpha_1}}, \dots, \mathbf{x}_{\gamma_{\alpha_{k-1}-1}+1}^{\gamma_{\alpha_{k-1}}}) \approx p(\mathbf{x}_{\gamma_{\alpha_{k-1}+1}^{\alpha_k}}^{\gamma_{\alpha_k}} \mid y_{\mu_{k-1}+1}^{\mu_k})$$

Monotone vs. no monotone alignments (I)

NO MONOTONE ALIGNMENT

$$\Pr(\mathbf{x} \mid y_1^I) \approx P(\mathbf{x} \mid y_1^I) = p(J|I) \cdot \sum_K \sum_{\mu_1^K} \sum_{\alpha_1^K} \sum_{\gamma_1^K} p(K|I, J) \cdot$$

$$\prod_{k=1}^K p(\mu_k|I) \cdot p(\alpha_k|\alpha_{k-1}) \cdot p(\gamma_{\alpha_k}|I) \cdot p(\mathbf{x}_{\gamma_{\alpha_{k-1}+1}^{\alpha_k}}^{\gamma_{\alpha_k}} \mid y_{\mu_{k-1}+1}^{\mu_k})$$

MONOTONE ALIGNMENT $\Rightarrow \alpha_k = k$

$$\Pr(\mathbf{x} \mid y_1^I) \approx P(\mathbf{x} \mid y_1^I) = p(J|I) \cdot \sum_K \sum_{\mu_1^K} \sum_{\gamma_1^K} p(K|I, J) \cdot \prod_{k=1}^K p(\mu_k|I) \cdot p(\gamma_k|J) \cdot p(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k} \mid y_{\mu_{k-1}+1}^{\mu_k})$$

Monotone vs. no monotone alignments (II)

NO MONOTONE ALIGNMENT

$$\Pr(\mathbf{x}|\mathbf{y}_1^I) \approx P(\mathbf{x}|\mathbf{y}_1^I) = p(J|I) \cdot \sum_K \sum_{\mu_1^K} \sum_{\alpha_1^K} \sum_{\gamma_1^K} p(K|I, J) \cdot \prod_{k=1}^K p(\mu_k|I, \mu_{k-1}, \mathbf{y}_{\mu_{k-2}+1}^{\mu_{k-1}}).$$

$$p(\alpha_k|\alpha_{k-1}, k, K) \cdot p(\mathbf{x}_{\gamma_{\alpha_{k-1}+1}}^{\gamma_{\alpha_k}} | \mathbf{y}_{\mu_{k-1}+1}^{\mu_k}) \cdot \prod_{l=1}^k (1 - \delta(\alpha_k, \alpha_l)) \cdot p(\gamma_{\alpha_k} - \gamma_{\alpha_{k-1}} | K, \mu_k - \mu_{k-1})$$

$$\text{MONOTONE ALIGNMENT} \Rightarrow \alpha_k = k$$

$$\Pr(\mathbf{x} | \mathbf{y}_1^I) \approx P(\mathbf{x} | \mathbf{y}_1^I) = p(J | I) \cdot \sum_K \sum_{\mu_1^K} \sum_{\gamma_1^K} p(K | I, J) \cdot$$

$$\prod_{k=1}^K p(\mu_k | I, \mu_{k-1}, \mathbf{y}_{\mu_{k-2}+1}^{\mu_{k-1}}) \cdot p(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k} | \mathbf{y}_{\mu_{k-1}+1}^{\mu_k}) \cdot \prod_{l=1}^k p(\gamma_k - \gamma_{k-1} | K, \mu_k - \mu_{k-1})$$

Monotone phrase-based models

- Uniform distributions are assumed for $p(J | I)$, $p(K | I, J)$, $p(\mu_k | I)$ and $p(\gamma_k | J)$,

$$P(\mathbf{x} | \mathbf{y}_1^I) \propto \sum_K \sum_{\mu_1^K} \sum_{\gamma_1^K} \prod_{k=1}^K p(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k} | \mathbf{y}_{\mu_{k-1}+1}^{\mu_k})$$

- The sums are approximate by maximizations.

$$P(\mathbf{x} | \mathbf{y}_1^I) \approx \max_K \max_{\mu_1^K} \max_{\gamma_1^K} \prod_{k=1}^K p(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k} | \mathbf{y}_{\mu_{k-1}+1}^{\mu_k})$$

Learning phrase-based models

- Models
 - Learning monotone phrase-based models
 - Learning nonmonotone phrase-based models
- Phrase-based units
 - Training with a sentence-aligned corpus.
 - Training with a word-aligned corpus.

Learning monotone phrase-based models*

Training with a sentence-aligned corpus.

Given a sentence-aligned corpus \mathcal{T} of pairs of sentences (\mathbf{x}, \mathbf{y}) , the maximum likelihood criterion tries to estimate the parameters $p(\tilde{x} \mid \tilde{y})$ that maximize:

$$\prod_{(\mathbf{x}, \mathbf{y}) \in \mathcal{T}} P(\mathbf{x} \mid \mathbf{y})$$

$$\text{subject to: } \sum_{\tilde{x}} p(\tilde{x} \mid \tilde{y}) = 1 \text{ for each target segment } \tilde{y}$$

By applying an EM procedure:

$$p(\tilde{x} \mid \tilde{y}) = \lambda_{\tilde{y}} \cdot \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{T}} \sum_{K, \mu_1^K, \gamma_1^K} \left(\prod_{k=1}^K p(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k} \mid \mathbf{y}_{\mu_{k-1}+1}^{\mu_k}) \cdot \sum_{l=1}^K \delta(\tilde{x} = \mathbf{x}_{\gamma_{l-1}+1}^{\gamma_l}) \cdot \delta(\tilde{y} = \mathbf{y}_{\mu_{l-1}+1}^{\mu_l}) \right)$$

where $\lambda_{\tilde{y}}$ is a normalization factor and $\delta: \delta(true) = 1$ y $\delta(false) = 0$.

*The slides on phrase-based models are modified versions of some material supplied by Jesús Tomás.

Learning monotone phrase-based models

Training with a word-aligned corpus.

Given a sentence-aligned corpus \mathcal{T} ,

- A word-aligned corpus is generated using the GIZA++ toolkit with \mathcal{T}
<http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html>
- A set of bilingual word sequences from the word aligned corpus is extracted.
- The parameters of the phrase-model are estimated.

Learning monotone phrase-based models

Extracting bilingual word sequences.

For each $\mathbf{x}, \mathbf{y} \in \mathcal{T}$, aligned by \mathbf{a} ,

$$BP_1(\mathbf{x}, \mathbf{y}, \mathbf{a}) = \left\{ (\mathbf{x}_{j_1}^{j_2}, \mathbf{y}_{i_1}^{i_2}) : \begin{array}{l} \forall j : j_1 \leq j \leq j_2; \exists i : i_1 \leq i \leq i_2 : \mathbf{a}(j) = i \\ \forall i : i_1 \leq i \leq i_2; \exists j : j_1 \leq j \leq j_2 : \mathbf{a}(j) = i \end{array} \right\}$$

$$BP_2(\mathbf{x}, \mathbf{y}, \mathbf{a}) = \left\{ (\mathbf{x}_{j_1}^{j_2}, \mathbf{y}_{i_1}^{i_2}) : \begin{array}{l} \forall j : j_1 \leq j \leq j_2; (i_1 \leq \mathbf{a}(j) \leq i_2) \vee (\mathbf{a}(j) = 0) \\ \forall j : (j < j_1) \vee (j_2 < j) : (\mathbf{a}(j) < i_1) \vee (i_2 < \mathbf{a}(j)) \end{array} \right\}$$

$$BP_3(\mathbf{x}, \mathbf{y}, \mathbf{a}) = \left\{ (\mathbf{x}_{j_1}^{j_2}, \mathbf{y}_{i_1}^{i_2}) : \begin{array}{l} \forall j : j_1 \leq j \leq j_2; (i_1 \leq \mathbf{a}_{\mathbf{x}, \mathbf{y}}(j) \leq i_2) \vee (\mathbf{a}(j) = 0) \\ \forall j : j < j_1; \mathbf{a}_{\mathbf{x}, \mathbf{y}}(j) < i_1 \quad \forall j : j > j_2; \mathbf{a}(j) > i_2 \end{array} \right\}$$

Learning monotone phrase-based models

Extracting bilingual multiword sequences: an example

x:	configuration	program	
y:	programa	de	configuración
a:	2	0	1

- $BP_1 = \{\text{configuration-configuration, program-programa}\}$
- $BP_2 = \{\text{configuration-configuration, program-programa, configuration-de configuración, program-programa de, configuration program-programa de configuración}\}$
- $BP_3 = \{\text{configuration program-programa de configuración}\}$

Learning monotone phrase-based models

Estimating the parameters.

By relative frequencies, for each pair of segments (x, y) :

$$p(\tilde{x} \mid \tilde{y}) = \frac{N(\tilde{x}, \tilde{y})}{N(\tilde{y})}$$

where $N(\tilde{y})$ denotes the number of times that phrase \tilde{y} has appeared, and $N(\tilde{x}, \tilde{y})$ is the number of times that the bilingual phrase (\tilde{x}, \tilde{y}) has appeared.

A refinement: the combination of the method based on a sentence-aligned corpus and one of the techniques for the bilingual multiword sequences:

$$p(\tilde{x} \mid \tilde{y}) = \lambda_{\tilde{y}} \cdot \sum_{(x,y) \in \mathcal{T}} p_I \cdot \sum_{K, \mu_1^K, \gamma_1^K} \left(\prod_{k=1}^K p(x_{\gamma_{k-1}+1}^{\gamma_k} \mid y_{\mu_{k-1}+1}^{\mu_k}) \cdot \sum_{l=1}^K \delta(\tilde{x} = x_{\gamma_{l-1}+1}^{\gamma_l}) \cdot \delta(\tilde{y} = y_{\mu_{l-1}+1}^{\mu_l}) \cdot \delta((\tilde{x}, \tilde{y}) \in BP(\mathcal{T})) \right)$$

Learning no-monotone phrase-based models

- The procedures for estimating the models parameters are similar to the ones for monotone models.
- For the distortion model, $p(\alpha_k \mid \alpha_{k-1})^*$:

$$p(\alpha_k \mid \alpha_{k-1}) = p_0^{|\gamma_{\alpha_k} - \gamma_{\alpha_{k-1}}|},$$

where p_0 is a parameter to be adjusted using a validation set.

(F.Och, H. Ney *The Alignment Template Approach to Statistical Machine Translation*. Computational Linguistics, 30(4), 2004.)

Search algorithms for monotone phrase-based models

- Basic idea is to generate partial hypothesis about the target sentence in an incremental way.
- Each of these hypothesis is composed by a prefix of the target sentence, a subset of source positions that have been aligned with the positions of the prefix of the target sentence and a score.
- New hypothesis can be generated for a previous hypothesis by adding a target word to the prefix of the target sentence that is the translation of a source(s) word(s) that is (are) not translated yet.

The adopted search algorithm for phrase-based models is the *multi-stack-decoding algorithm*.

Search algorithms for monotone phrase-based models

Given a source sentence x , a hypothesis is a tuple

$$(x_1^j, y_1^i, S(x_1^j, y_1^i) = P(y_1^i) \cdot P(x_1^j | y_1^i))$$

where

$$P(y_1^i) = \prod_{i'=1}^i p(y_{i'} | y_{i'-n+1}^{i'-1})$$

and

$$P(x_1^j | y_1^i) = \max_K \max_{\mu_1^K} \max_{\gamma_1^K} \prod_{k=1}^K p(x_{\gamma_{k-1}+1}^{\gamma_k} | y_{\mu_{k-1}+1}^{\mu_k})$$

with $\gamma_k = j$ and $\mu_k = i$

Search algorithms for monotone phrase-based models

- The initialization consists on building a hypothesis for the empty target and empty source prefixes and a score of 1.0.
- The algorithm selects a hypothesis $(x_1^j, y_1^i, S(x_1^j, y_1^i))$ of each stack and for each bilingual segment (\tilde{x}, \tilde{y}) with $x_{j+1}^{j+|\tilde{x}|} \equiv \tilde{x}$, a new hypothesis is created $(x_1^j \tilde{x}, y_1^i \tilde{y}, S(x_1^j \tilde{x}, y_1^i \tilde{y}))$

$$S(x_1^j \tilde{x}, y_1^i \tilde{y}) = S(x_1^j, y_1^i) \cdot \prod_{l=i+1}^{i+|\tilde{y}|} p(y_l | y_{l-n+1}^{l-1}) \cdot p(\tilde{x} | \tilde{y}) \quad (1)$$

Each new hypothesis, $(x_1^j \tilde{x}, y_1^i \tilde{y}, S(x_1^j \tilde{x}, y_1^i \tilde{y}))$ will be stored in the stack associated to the source prefixes of length $j + |\tilde{x}|$.

Search algorithms for no-monotone search algorithm.

- The procedure is quite similar to the monotone search algorithm,
- A hypothesis consists on a prefix of the target sentence, a subset of source positions and a score with the partial contributions of the target language model and translation model.
- The implementation requires a stack for each possible subset of source positions and consequently, the computational cost can be very high.

Experimental results

Corpora

“*El Periódico*”: From a bilingual newspaper (Spanish to Catalan)

	Spanish	Catalan
Train: Sentence pairs	643,961	
Running words (Kwords)	7,180	7,435
Vocabulary (Kwords)	129	128
Test: Sentences	240	
Running words	4,316	4,389

Experimental results

Corpora

XRCE: From Xerox printer manuals (English to and from Spanish, French and German)

	En	Sp	En	Ge	En	Fr
Train: Sentence pairs	56K		49K		53K	
Running words	665K	753K	633K	696K	587K	534K
Vocabulary	8K	11K	8K	10K	8K	19K
Test: Sentence pairs	1,125		984		996	
Running words	8K	10K	11K	12K	12K	12K
Test perplexity	48	33	51	87	73	52

Experimental results

Corpora

EU: Bulletin of the European Union (English to and from Spanish, French and German)

	En	Sp	En	Ge	En	Fr
Train: Sentence pairs	214K		223K		215K	
Running words	5.9M	6.6M	6.5M	6.1M	6.0M	6.6M
Vocabulary	84K	97K	87K	152K	85K	91K
Test: Sentence pairs	800		800		800	
Running words	2K	25K	22K	21K	22K	24K
Test perplexity	47	39	47	71	48	38

Experimental results

Corpora

Hansards: Proceedings of Canadian Parliament (French to English)

	English	French
Train: Sentence pairs	137,381	
Running words (Kwords)	1,941	2,130
Vocabulary (Kwords)	29.5	37.5
Test: Sentences	250	
Running words	2,633	2,805

Assessment

- **Word error rate (WER)**: The minimum number of substitution, insertion and deletion operations needed to convert the word string hypothesized by the translation system into a given single reference word string.
- **Multi reference WER (mWER)**: Similar to WER, but for each source test sentence there are more than one target sentences as references.
- **BiLingual Evaluation Understudy (BLEU)**: it is based on the n -grams of the hypothesized translation that occur in the reference translations. The BLEU metric ranges from 0.0 (worst score) to 1.0 (best score).

Effect of the size of the segment length (MSL)

“El Periódico” task.

MSL	2	3	4
WER	12.1	10.6	10.5
Parameters	2.0M	7.0M	14,5M

XRCE task.

	MSL	2	4	6	8	10	12	14	16
English to Spanish	WER	50.6	36.6	29.5	27.4	26.1	25.6	25.4	25.4
	Params.	0.1M	0.4M	0.8M	1,1M	1,4M	1,6M	1,8M	1,9M
Spanish to English	WER	48.1	35.2	29.9	28.2	27.6	27.5	27.3	27.2
	Params.	0,1M	0.5M	0.9M	1,3M	1,6M	1,8M	2.0M	2,2M
English to French	WER	59.0	55.9	54.7	54.2	54.2	54.2	54.0	53.9
	Params.	0.1M	0.5M	0.8M	1,0M	1,3M	1,5M	1,6M	1,8M
French to English	WER	52.8	52,3	52.9	52,8	52.9	52.7	52.6	52.3
	Params.	0.1M	0.5M	0.9M	1,2M	1,4M	1,6M	1,8M	1,9M
English to German	WER	68.5	66.0	65.8	65.5	65.1	65.1	65.0	64.8
	Params.	0.1M	0.5M	0.8M	1.0M	1.3M	1.4M	1.5M	1.6M
German to English	WER	59.7	56.5	55.2	54.5	54.2	54.2	54.0	54.0
	Params.	0.1M	0.4M	0.7M	0.8M	1.0M	1.1M	1.2M	1.3M

Some results

Different methods for building segments. XRCE task.

Procedure	En-Es	Es-En	En-Fr	Fr-En	En-De	De-En
BP1	45.7	28.6	54.2	52.4	65.1	55.6
BP2	26.4	27.4	53.6	52.2	64.1	54.3
BP3	25.4	27.3	54.0	52.5	64.9	53.9

Monotone vs. non monotone search. XRCE task.

Search	En-Es	Es-En	En-Fr	Fr-En	En-De	De-En
Monotone	28.5	30.9	51.4	51.6	66.4	54.1
No monotone	28.0	31.6	52.0	51.3	66.4	54.0

Effect of the training set on the system performance. “El Periódico” task.

Corpus size	5K	10K	20K	40K	80K	160K	320K	640K
WER	20.3	17.3	15.2	13.4	12.4	11.7	11.1	10.7
Parameters	0.1	0.2M	0.4M	0.7M	1.2M	2.1M	3.6M	7.0M

Comparison with other machine translation systems.

“El Periódico” task.

- **Salt** a knowledge-based machine translation systems supported by the Government of the Generalitat Valenciana (<http://www.cultgva.es>).
- **Incyta**, a knowledge-based commercial systems (<http://www.incyta.com>).
- **InterNOSTRUM**, a hybrid knowledge-based and finite-state translation system (<http://www.internostrum.com>).

MT system	WER (%)	mWER (%)	BLEU
Salt	9.9	6.6	0.866
Incyta	10.0	7.6	0.855
Phrase-based	10.7	7.8	0.857
InterNOSTRUM	11.9	8.5	0.837

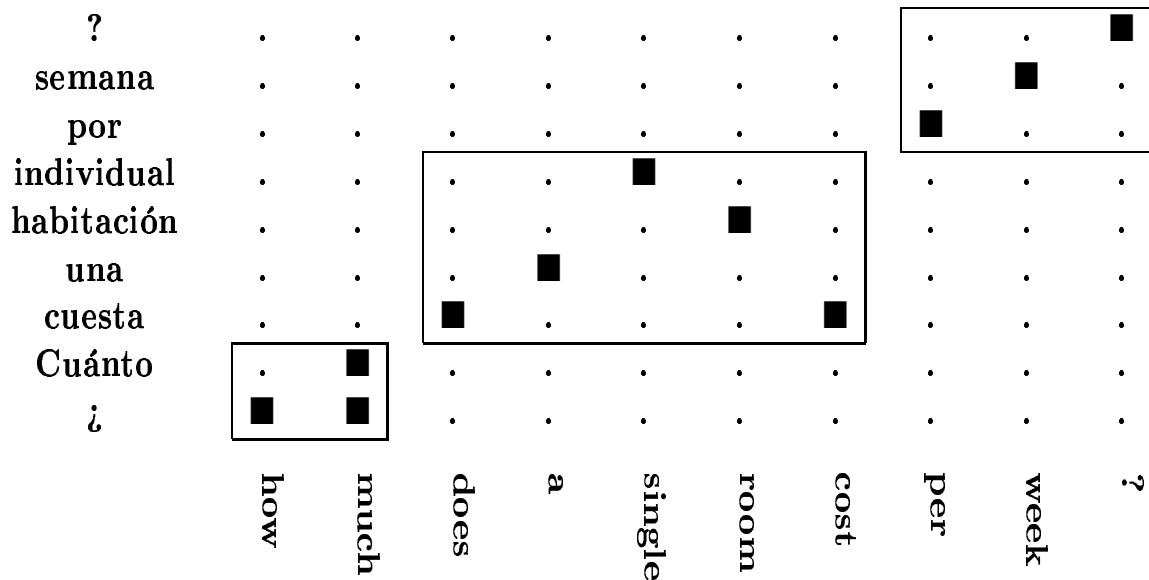
Finally, a simple experiment was carried out with the HANSARD task. The result obtained was 64.9% of WER.

Index

- 1 Beyond word models ▷ 2
- 2 Phrase-based models ▷ 9
- 3 *Alignment templates* ▷ 47
- 4 Phrases and finite-state transducers ▷ 58
- 5 Using linguistic knowledge ▷ 66
- 6 Bibliography ▷ 72

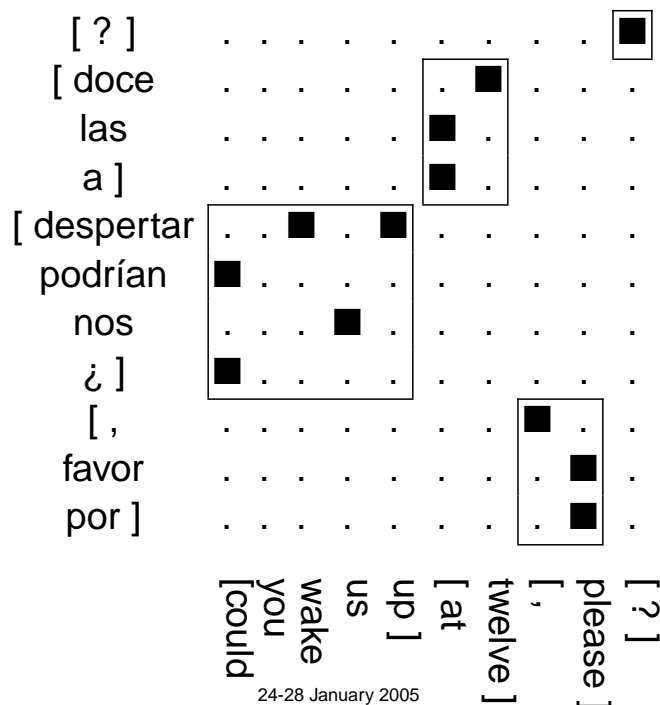
Alignment templates

(H.Ney. *Stochastic modelling: from pattern recognition to language translation*. VIII SNRFAI. 1999.)



Alignment templates

(H. Ney, *Statistical Natural Language Processing*, STC Doctorate Program, UPC. 2003)



Alignment templates

(H. Ney et al. *Algorithms for statistical translation of spoken language*. IEEE TSAP. 2000.)

- Let K be the number of segments in x and in y ,
- Segmentation of the target sentence

$$\mu : \{1, \dots, K\} \rightarrow \{1, \dots, I\} : \mu_k \geq \mu_{k-1} \quad 1 < k \leq K \quad \& \quad \mu_K = I \quad (\mu_0 = 0)$$

$$y_1^I \Rightarrow \tilde{y}_1^K; \tilde{y}_k \equiv y_{\mu_{k-1}+1}^{\mu_k} = y_{\mu_{k-1}+1}, \dots, y_{\mu_k}; 1 \leq k \leq K$$

- Segmentation of the source sentence

$$\gamma : \{1, \dots, K\} \rightarrow \{1, \dots, J\} : \gamma_k \geq \gamma_{k-1} \quad 1 < k \leq K \quad \& \quad \gamma_K = J \quad (\gamma_0 = 0)$$

$$x_1^J \Rightarrow \tilde{x}_1^K; \tilde{x}_k \equiv x_{\gamma_{k-1}+1}^{\gamma_k} = x_{\gamma_{k-1}+1}, \dots, x_{\gamma_k}; 1 \leq k \leq K$$

- Segment alignment (Permutation):

$$\alpha : \{1, \dots, K\} \rightarrow \{1, \dots, K\} : \alpha(k) = \alpha(k') \quad \text{iff} \quad k = k'$$

Alignment templates

ALIGNMENT BETWEEN WORD GROUPS

$$\Pr(\tilde{x} \mid \tilde{y}) = \sum_{\alpha} \Pr(\alpha, \tilde{x} \mid \tilde{y}) \approx \sum_{\alpha} \prod_{k=1}^K p(\alpha_k \mid \alpha_{k-1}) \cdot P(\tilde{x}_k \mid \tilde{y}_{\alpha_k})$$

ALIGNMENT WITHIN WORD GROUPS

$$P(\tilde{x}_k \mid \tilde{y}_l) = \sum_{\mathbf{z}} p(\mathbf{z} \mid \tilde{y}_l) \cdot p(\tilde{x}_k \mid \mathbf{z}, \tilde{y}_l)$$

An **ALIGNMENT TEMPLATE** \mathbf{z} is a binary matrix with I' rows and J' columns: $z_{i,j} = 1$ if the pair y_i and x_j are aligned.

[despertar	.	.	■	.	■
podrían	■
nos	.	.	.	■	.
¿]	■
[could					
you					
wake					
us					
] up]					

Alignment templates

(H. Ney et al. *Algorithms for statistical translation of spoken language*. IEEE TSAP. 2000.)

$$P(\tilde{\mathbf{x}}_k | \tilde{\mathbf{y}}_l) \approx \sum_{\mathbf{z}} p(\mathbf{z} | \tilde{\mathbf{y}}_l) \cdot \prod_{j=1}^{J'} \sum_{i=1}^{I'} a(i | j, z) \cdot l(\tilde{\mathbf{x}}_{k_j} | \tilde{\mathbf{y}}_{l_i})$$

$$= \sum_{\mathbf{z}} p(\mathbf{z} | \tilde{\mathbf{y}}_l) \cdot \prod_{j=1}^{J'} \sum_{i=1}^{I'} \frac{z_{ij}}{\sum_{i'} z_{i'j}} \cdot l(\tilde{\mathbf{x}}_{k_j} | \tilde{\mathbf{y}}_{l_i})$$

Alignment templates: training and search

(H. Ney et al. *Algorithms for statistical translation of spoken language*. IEEE TSAP. 2000.)

- **Training:**
 - Viterbi alignment $\mathbf{x} \rightarrow \mathbf{y}$ and $\mathbf{y} \rightarrow \mathbf{x}$.
 - Obtaining all template templates by considering all possible source-target word groups under the constraint that the words within the source/target word group are only aligned to words within the target/source word group.
- **Translation:**
 - Computing all possible segmentations of the source sentence into word groups.
 - Computing all possible alignments between word groups.
 - Computing all possible word alignments within the word group.

Results

(EUTRANS consortium *Example-Based Language Translation Systems. Final Report.* Deliverable D0.1c. 2000.)

EuTrans-I corpus (Spanish-English)

- **Vocabulary:** 680 Spanish words, and 513 English words.
- **Training:** 10,000 pairs (97,000/99,000 words).
- **Test:** 2,996 pairs (PP=3.3) (35,000/35,590 words).

Model	WER
Alignment templates (with manual categories)	2.5
Quasi-Monotone search	10.8
DP-search M2	13.9

Results

(EUTRANS consortium *Example-Based Language Translation Systems. Final Report.* Deliverable D0.1c. 2000.)

FUB corpus (Italian-English)

- **Vocabulary:** 2,458 Italian words, and 1,701 English words.
- **Training:** 3,338 pairs (61,423/72,689 words).
- **Test:** 278 pairs (PP=31).

Model	WER
Alignment templates	23.8
Monotone search	29.3

Results

(H. Ney et al. *Algorithms for statistical translation of spoken language*. IEEE TSAP. 2000.)

Vermobil corpus (German-English)

- **Vocabulary:** 5,936 German words, and 3,505 English words.
- **Training:** 30,556 pairs (329,000/343,000 words).
- **Test:** 47 pairs (PP=43.7) (701/792 words).

Model	WER
Alignment templates	28.8
Inverted search	41.0
Monotone search	36.5

Results

(H. Ney, *Statistical Natural Language Processing*, STC Doctorate Program, UPC. 2003)

Vermobil corpus (German-English)

- **Vocabulary:** 7,940 German words, and 4,673 English words.
- **Training:** 58,332 pairs (519,523/549,921 words).
- **Test:** 5,069 (German → English) and 4,136 (English → German) sentences.

Model	SER
Semantic Transfer	62
Dialog Act Based	60
Example Based	51
Statistical	29

Index

- 1 Beyond word models ▷ 2
- 2 Phrase-based models ▷ 9
- 3 Alignment templates ▷ 47
- 4 *Phrases and finite-state transducers* ▷ 58
- 5 Using linguistic knowledge ▷ 66
- 6 Bibliography ▷ 72

Joint distributions (I)

$$\hat{y} = \operatorname{argmax}_y \Pr(y \mid x) = \operatorname{argmax}_y \Pr(x, y)$$

Assuming monotone constraints:

$$\begin{aligned}
 \Pr(\mathbf{x}, \mathbf{y}) &= \Pr(J, I) \cdot \Pr(\mathbf{x}_1^J, \mathbf{y}_1^I \mid J, I) \\
 &= \Pr(J, I) \cdot \sum_K \Pr(K \mid J, I) \cdot \Pr(\mathbf{x}_1^J, \mathbf{y}_1^I \mid J, I, K) \\
 &= \Pr(J, I) \cdot \sum_K \Pr(K \mid J, I) \cdot \sum_{\gamma_1^K, \mu_1^K} \Pr(\mathbf{x}_1^J, \mathbf{y}_1^I, \gamma_1^K, \mu_1^K \mid J, I, K)
 \end{aligned}$$

Joint distributions (II)

$$\begin{aligned}
\Pr(\mathbf{x}, \mathbf{y}) &= \Pr(J, I) \cdot \sum_K \Pr(K \mid J, I) \cdot \sum_{\gamma_1^K, \mu_1^K} \Pr(\mathbf{x}_1^J, \mathbf{y}_1^I, \gamma_1^K, \mu_1^K \mid J, I, K) \\
&= \Pr(J, I) \cdot \sum_K \Pr(K \mid J, I) \cdot \sum_{\gamma_1^K, \mu_1^K} \Pr(\gamma_1^K, \mu_1^K \mid J, I, K) \cdot \Pr(\mathbf{x}_1^J, \mathbf{y}_1^I \mid J, I, K, \gamma_1^K, \mu_1^K) \\
&= \Pr(J, I) \cdot \sum_K \Pr(K \mid J, I) \cdot \sum_{\gamma_1^K, \mu_1^K} \prod_{k=1}^K \Pr(\gamma_k, \mu_k \mid J, I, K, \gamma_1^{k-1}, \mu_1^{k-1}) \cdot \\
&\quad \Pr(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k}, \mathbf{y}_{\mu_{k-1}+1}^{\mu_k} \mid J, I, K, \mathbf{x}_1^{\gamma_{k-1}}, \mathbf{y}_1^{\mu_{k-1}}, \gamma_1^K, \mu_1^K)
\end{aligned}$$

Joint distributions (III)

$$\begin{aligned}
\Pr(\mathbf{x}_1^J, \mathbf{y}_1^I, \gamma_1^K, \mu_1^K \mid J, I, K) &= \prod_{k=1}^K \Pr(\gamma_k, \mu_k \mid J, I, K, \gamma_1^{k-1}, \mu_1^{k-1}) \cdot \\
&\quad \Pr(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k}, \mathbf{y}_{\mu_{k-1}+1}^{\mu_k} \mid J, I, K, \mathbf{x}_1^{\gamma_{k-1}}, \mathbf{y}_1^{\mu_{k-1}}, \gamma_1^K, \mu_1^K)
\end{aligned}$$

Assuming,

- $\Pr(\gamma_k, \mu_k \mid J, I, K, \gamma_1^{k-1}, \mu_1^{k-1}) \approx \rho$
- $\Pr(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k}, \mathbf{y}_{\mu_{k-1}+1}^{\mu_k} \mid J, I, K, \mathbf{x}_1^{\gamma_{k-1}}, \mathbf{y}_1^{\mu_{k-1}}, \gamma_1^K, \mu_1^K) \approx \Pr(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k}, \mathbf{y}_{\mu_{k-1}+1}^{\mu_k} \mid \mathbf{x}_1^{\gamma_{k-1}}, \mathbf{y}_1^{\mu_{k-1}})$

$$\Pr(\mathbf{x}_1^J, \mathbf{y}_1^I, \gamma_1^K, \mu_1^K \mid J, I, K) \approx \rho \cdot \prod_{k=1}^K \Pr(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k}, \mathbf{y}_{\mu_{k-1}+1}^{\mu_k} \mid \mathbf{x}_1^{\gamma_{k-1}}, \mathbf{y}_1^{\mu_{k-1}})$$

Assuming n -grams,

$$\Pr(\mathbf{x}_1^J, \mathbf{y}_1^I, \gamma_1^K, \mu_1^K \mid J, I, K) \approx \rho \cdot \prod_{k=1}^K \Pr(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k}, \mathbf{y}_{\mu_{k-1}+1}^{\mu_k} \mid \mathbf{x}_{\gamma_{k-n}+1}^{\gamma_{k-1}}, \mathbf{y}_{\mu_{k-n}+1}^{\mu_{k-1}})$$

An example (IV)

x: he hecho una reserva de una habitación doble .

y: I have made a reservation of a double room .

The lengths of x and y

x	he	hecho	una	reserva	de	una	habitación	doble	.	
j	1	2	3	4	5	6	7	8	9=J	
y	I	have	made	a	reservation	of	a	double	room	.
i	1	2	3	4	5	6	7	8	9	10=I

Number of segments: $K = 3$

An example (V)

x: he hecho una reserva de una habitación doble .

y: I have made a reservation of a double room .

Segmentation of x and y

x	he	hecho	una	reserva	de	una	habitación	doble	.	
j	1	2	3	4	5	6	7	8	9=J	
γ	γ_1		γ_2			γ_3				
y	I	have	made	a	reservation	of	a	double	room	.
i	1	2	3	4	5	6	7	8	9	10=I
μ	μ_1			μ_2			μ_3			

Phrases of x and y

x	he	hecho	una	reserva	de	una	habitación	doble	.	
y	I	have	made	a	reservation	of	a	double	room	.

Joint distributions (IV)

$$\begin{aligned}
 \Pr(\mathbf{x}_1^J, \mathbf{y}_1^I \mid J, I) &\approx \rho \cdot \sum_K \Pr(K \mid J, I) \cdot \sum_{\gamma_1^K, \mu_1^K} \prod_{k=1}^K \Pr(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k}, \mathbf{y}_{\mu_{k-1}+1}^{\mu_k} \mid \mathbf{x}_{\gamma_{k-n}+1}^{\gamma_{k-1}}, \mathbf{y}_{\mu_{k-n}+1}^{\mu_{k-1}}) \\
 &\approx \rho \cdot \max_K \Pr(K \mid J, I) \cdot \max_{\gamma_1^K, \mu_1^K} \prod_{k=1}^K \Pr(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k}, \mathbf{y}_{\mu_{k-1}+1}^{\mu_k} \mid \mathbf{x}_{\gamma_{k-n}+1}^{\gamma_{k-1}}, \mathbf{y}_{\mu_{k-n}+1}^{\mu_{k-1}})
 \end{aligned}$$

A simple case: GIATI (\mathcal{L}_1): $K = J$, $\gamma_j = j$ and the target segments can be empty

Remark: There can be test segmentations that are not in the training corpus!

Possible smoothings:

$$\Pr(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k}, \mathbf{y}_{\mu_{k-1}+1}^{\mu_k} \mid \mathbf{x}_{\gamma_{k-n}+1}^{\gamma_{k-1}}, \mathbf{y}_{\mu_{k-n}+1}^{\mu_{k-1}}) = \begin{cases} \Pr(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k}, \mathbf{y}_{\mu_{k-1}+1}^{\mu_k} \mid -, \mathbf{y}_{\mu_{k-n}+1}^{\mu_{k-1}}) \\ \Pr(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k}, \mathbf{y}_{\mu_{k-1}+1}^{\mu_k} \mid \mathbf{x}_{\gamma_{k-n}+1}^{\gamma_{k-1}}, -) \\ \Pr(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k}, \mathbf{y}_{\mu_{k-1}+1}^{\mu_k} \mid -, -) \end{cases}$$

Joint distributions and HMMs

$$\begin{aligned}
 \Pr(\mathbf{x}_1^J, \mathbf{y}_1^I \mid J, I) &\approx \rho' \cdot \sum_{K, \gamma_1^K, \mu_1^K} \prod_{k=1}^K \Pr(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k}, \mathbf{y}_{\mu_{k-1}+1}^{\mu_k} \mid \mathbf{x}_1^{\gamma_{k-1}}, \mathbf{y}_1^{\mu_{k-1}}) \\
 &= \rho' \cdot \sum_{K, \gamma_1^K, \mu_1^K} \prod_{k=1}^K \Pr(\mathbf{y}_{\mu_{k-1}+1}^{\mu_k} \mid \mathbf{x}_1^{\gamma_{k-1}}, \mathbf{y}_1^{\mu_{k-1}}) \cdot \Pr(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k} \mid \mathbf{x}_1^{\gamma_{k-1}}, \mathbf{y}_1^{\mu_k})
 \end{aligned}$$

$$\begin{aligned}
 \Pr(\mathbf{y}_{\mu_{k-1}+1}^{\mu_k} \mid \mathbf{x}_1^{\gamma_{k-1}}, \mathbf{y}_1^{\mu_{k-1}}) &\approx \Pr(\mathbf{y}_{\mu_{k-1}+1}^{\mu_k} \mid \mathbf{y}_1^{\mu_{k-1}}) \approx \prod_{l=\mu_{k-1}+1}^{\mu_k} \Pr(y_l \mid \mathbf{y}_{l-n+1}^l) \\
 \Pr(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k} \mid \mathbf{x}_1^{\gamma_{k-1}}, \mathbf{y}_1^{\mu_k}) &\approx \Pr(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k} \mid \mathbf{y}_{\mu_{k-1}+1}^{\mu_k})
 \end{aligned}$$

$$\Pr(\mathbf{x}_1^J, \mathbf{y}_1^I \mid J, I) \approx \rho' \cdot \sum_{K, \gamma_1^K, \mu_1^K} \prod_{k=1}^K \left(\Pr(\mathbf{x}_{\gamma_{k-1}+1}^{\gamma_k} \mid \mathbf{y}_{\mu_{k-1}+1}^{\mu_k}) \cdot \prod_{l=\mu_{k-1}+1}^{\mu_k} \Pr(y_l \mid \mathbf{y}_{l-n+1}^l) \right)$$

Index

- 1 Beyond word models ▷ 2
- 2 Phrase-based models ▷ 9
- 3 Alignment templates ▷ 47
- 4 Phrases and finite-state transducers ▷ 58
- 5 *Using linguistic knowledge* ▷ 66
- 6 Bibliography ▷ 72

Chunking in statistical machine translation

Koehn and Knight. *ChunkMT: Statistical machine translation with richer linguistic knowledge*. Draft, Unpublished. 2002.

1. Chunking the source sentence: generating sequence of chunks (with the corresponding source POS)
2. Reordering the source chunks.
3. Chunk mapping: generating the target POS of each chunk
4. Word translations: generating the target words.

Marginal improvements on a corpus of the European Parliament proceedings
(German to English)

Bilingual chunking in statistical machine translation

Wang, Zhou, Huang and Huang. *Structure alignment using bilingual chunking*. 17th International Conference on Computational Linguistics, 2002.

$$\operatorname{argmax}_{\sigma, \tau, \alpha} \Pr(y, \sigma, \tau, \alpha \mid \mathbf{x})$$

- σ = sequence of source chunks;
- τ = sequence of target chunks;
- α = alignment between chunks.

By introducing POS tagging of the source sentence π_x and POS tagging of the target sentence π_y

$$\operatorname{argmax}_{\sigma, \tau, \alpha, \pi_y, \pi_x} \Pr(\pi_x \mid \mathbf{x}) \cdot \Pr(\sigma \mid \pi_x, \mathbf{x}) \cdot \Pr(\pi_y \mid \sigma, \pi_x, \mathbf{x}) \\ \cdot \Pr(y \mid \pi_y, \sigma, \pi_x, \mathbf{x}) \cdot \Pr(\tau \mid y, \pi_y, \sigma, \pi_x, \mathbf{x}) \cdot \Pr(\alpha \mid \tau, y, \pi_y, \sigma, \pi_x, \mathbf{x})$$

Bilingual chunking in statistical machine translation

Wang, Zhou, Huang and Huang. *Structure alignment using bilingual chunking*. 17th International Conference on Computational Linguistics, 2002.

$$\operatorname{argmax}_{\sigma, \tau, \alpha, \pi_y, \pi_x} \Pr(\pi_x \mid \mathbf{x}) \cdot \Pr(\sigma \mid \pi_x, \mathbf{x}) \cdot \Pr(\pi_y \mid \sigma, \pi_x, \mathbf{x})$$

$$\cdot \Pr(y \mid \pi_y, \sigma, \pi_x, \mathbf{x}) \cdot \Pr(\tau \mid y, \pi_y, \sigma, \pi_x, \mathbf{x}) \cdot \Pr(\alpha \mid \tau, y, \pi_y, \sigma, \pi_x, \mathbf{x})$$

- $\Pr(\pi_x \mid \mathbf{x})$: POS tagging of source sentence
- $\Pr(\sigma \mid \pi_x, \mathbf{x})$: chunking the source sentence
- $\Pr(\pi_y \mid \sigma, \pi_x, \mathbf{x})$: target POS tags.
- $\Pr(y \mid \pi_y, \sigma, \pi_x, \mathbf{x})$: target words.
- $\Pr(\tau \mid y, \pi_y, \sigma, \pi_x, \mathbf{x})$: target chunks.
- $\Pr(\alpha \mid \tau, y, \pi_y, \sigma, \pi_x, \mathbf{x})$: alignment between chunks.

Only results on chunk alignments.

Parsing in statistical machine translation

Charniak, Knight and Yamada. *Syntax-based Language Models for Machine Translation*. MT Summit IX. 2003

$$\Pr(y \mid x) = \sum_{\pi_y} \Pr(y, \pi_y) \cdot \Pr(x \mid y, \pi_y) = \sum_{\pi_y} \Pr(\pi_y) \cdot \Pr(x \mid \pi_y)$$

π_y is a parse tree of the target sentence y

Three operations for $\Pr(x \mid \pi_y)$:

- Reordering some the child nodes
- Inserting optional words
- Translating each target word by the corresponding source word

Some improvements on a Chinese to English newspaper task.

Other approaches

- Wang and Waibel. *Modeling with structures in statistical machine translation*. 17th Int. Conf. on Computational Linguistics Montreal, (Coling) 1998.
Shallow parsing to define structures to be aligned.
- Wang, Zhou, Huang and Huang. *Structure alignment using bilingual chunking*. 17th International Conference of Computational Linguistics (COLING) 2002.
- Koehn and Knight, 2003. *Feature-rich statistical translation of noun phrases*. 41nd Annual Meeting of the ACL 2003.
A subsystem for translating noun phrases.

Index

- 1 Beyond word models ▷ 2
- 2 Phrase-based models ▷ 9
- 3 Alignment templates ▷ 47
- 4 Phrases and finite-state transducers ▷ 58
- 5 Using linguistic knowledge ▷ 66
- 6 *Bibliography* ▷ 72

Bibliography

1. Wang and Waibel, 1998. *Modeling with structures in statistical machine translation*. 17th Int. Conf. on Computational linguistics. Montreal, Quebec, Canada. 1357 - 1363, 1998.
2. H. Ney, S. Nießen, F. Och, H. Sawaf, C. Tillmann, S. Vogel: *Algorithms for statistical translation of spoken language*. IEEE Transactions on Speech and Audio Processing, 8(1): 24–36, 2000.
3. F. J. Och, H. Ney: *Improved statistical alignment models*. Proc. of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 440-447, Hongkong, China, October 2000.
www-i6.Informatik.RWTH-Aachen.de/Colleagues/och/software/GIZA++.html
4. Koehn and Knight, 2002. *ChunkMT: Statistical machine translation with richer linguistic knowledge*. Draft, Unpublished.
5. M. Zhou, J.-X. Huang and Ch.-N. Huang. *Structure alignment using bilingual chunking*. 17th International Conference on Computational Linguistics COLING. 2002.
6. Taro Watanabe, Kenji Imamura and Eiichiro Sumita. *Statistical machine translation based on hierarchical phrase alignment*. Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation March 13-17, 2002 Keihanna,
7. Koehn and Knight. *Feature-rich statistical translation of noun phrases*. 41nd Annual Meeting of the ACL. 311-318, July, 2003.
8. Charniak, Knight and Yamada. *Syntax-based Language Models for Machine Translation*. MT Summit IX. 2003.

Bibliography

9. H. Ney, *Statistical Natural Language Processing*, Signal Theory and Communications Doctorate Program, UPC. March 17-21, 2003
10. J. Tomás, F. Casacuberta: *Combining phrase-based and template-based alignment models in statistical translation*. Proc. of the IbPRIA, 2003.
11. Koehn and Knight. *Feature-rich statistical translation of noun phrases*. 41nd Annual Meeting of the ACL Sapporo, JAPAN. 311-318, July, 2003.
12. Watanabe, Sumita and Okuno. *Chunk-based statistical translation*. Proc of HLT-NAAC., 303-310, Edmonton, June 2003.
13. Widdows. *Unsupervised methods for developing taxonomies by combining syntactic and statistical information*. HLT-NAACL. 197–204, Edmonton, June 2003.
14. Widdows. *Unsupervised methods for developing taxonomies by combining syntactic and statistical information*. HLT-NAACL. 197–204, Edmonton, June 2003.
15. Melamed. *Statistical machine translation by parsing*. 42nd Annual Conference of the Association for Computational Linguistics 2004.
16. Zhang and Gildea. *Syntax-based alignment: supervised or unsupervised?*. 20th International Conference on Computational Linguistics COLING. 2004.
17. O. Bender, R. Zens, E. Matusov, and H. Ney: *Alignment Templates: the RWTH SMT System*. In Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2004), Kyoto, Japan, 79-84. September, 2004.

Bibliografy

18. F. J. Och and H. Ney: *The Alignment Template Approach to Statistical Machine Translation*. *Computational Linguistics*, 30(4), 2004.
19. Melamed. *Statistical machine translation by parsing*. 42nd Annual Conference of the Association for Computational Linguistics 42nd ACL. 2004.
20. Zhang and Gildea. *Syntax-based alignment: supervised or unsupervised?*. 20th International Conference on Computational Linguistics COLING.

Pattern Recognition approaches to Machine Translation

F. Casacuberta and E. Vidal

Pattern Recognition and Human Language Technology Group
Instituto Tecnológico de Informática
Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia, Spain

State-Merging Approaches

Enrique Vidal

`evidal@iti.upv.es`

January 2005

E. Vidal – ITI-UPV-DSIC

[Pattern Recognition Machine Translation](#)

[State-Merging Approaches](#)

Index

- 1 Subsequential Transduction: “OSTI” Algorithm ▷ [2](#)
- 2 Using input/output syntactic constraints: OSTIA-DR ▷ [29](#)
- 3 OSTIA-DR: improving scalability ▷ [45](#)
- 4 Bibliography ▷ [70](#)

Index

- 1 *Subsequential Transduction: “OSTI” Algorithm* ▷ 2
- 2 Using input/output syntactic constraints: OSTIA-DR ▷ 29
- 3 OSTIA-DR: improving scalability ▷ 45
- 4 Bibliography ▷ 70

Sequential Transducers

A *Sequential Transducer* (ST) τ is a 5-tuple $\tau = (Q, X, Y, q_0, E)$:

Q :	Finite set of <i>States</i>
X, Y :	Input and output <i>Alphabets</i>
$q_0 \in Q$:	<i>Initial State</i>
$E \subset Q \times X \times Y^* \times Q$:	<i>“Edges” or Transitions</i>

- All the states are *accepting*
- Edges are *deterministic*:
 $(q, a, u, r), (q, a, v, s) \in E \Rightarrow (u = v \wedge r = s)$

PROPERTIES:

1. T_τ is a *function*: $X^* \rightarrow Y^*$
2. STs \equiv *Generalized Sequential Machines* \supset (Mealy and Moore machines)
3. STs *preserve prefixes*: $T_\tau(\lambda) = \lambda$; $T_\tau(uv) \in T_\tau(u)Y^*$

“Property” 2 entails *strict sequentiality*,
 which can hardly be adequate in many cases of interest

Subsequential Transduction

[Berstel, 79]

A *Subsequential Transducer* (SST) τ is a 6-tuple $\tau = (Q, X, Y, q_0, E, \sigma)$, where:

- $\tau' = (Q, X, Y, q_0, E)$ is a Sequential Transducer
- $\sigma : Q \rightarrow Y^*$ is a *state output* (partial) *function*
- For each input string x , the output string y is obtained by concatenating $\sigma(q)$ to $\tau'(x)$, where q is the last state reached through the analysis of x by τ' ; i.e.:

$$y = \tau(x) = \tau'(x)\sigma(q)$$

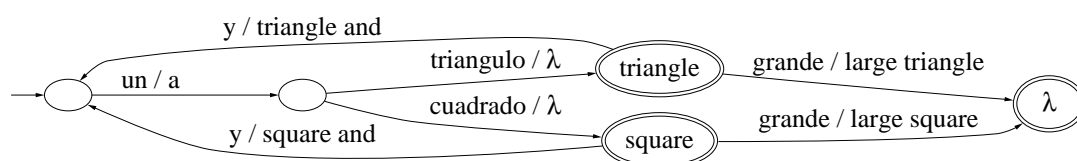
PROPERTIES:

1. T_τ is a *function*: $X^* \rightarrow Y^*$
2. Sequential \subset **Subsequential Transduction** \subset Finite State.
3. Input-output monotonicity (sequentiality) needs *not* be as strict as in STs.

Subsequential Transducers (intuitive concept)

- **Deterministic Finite State Networks** which accept sentences from an *input* language and produce sentences of an *output* language.
- In addition to input symbols, output strings are assigned to the edges.
- Output strings are also assigned to final states.
- **SST operation relies on “delaying” the production of output symbols** until enough of the input sentence has been seen to guarantee a correct output.

An example of SST:



Learning SSTs: the OSTI Algorithm

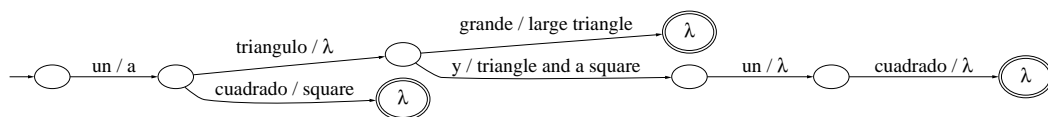
[Oncina, 91-93]

SSTs can be learned from training examples using the **Onward Subsequential Transducer Inference Algorithm (OSTIA)**.

1. Build an **“onward” tree representation** of the training data (a tree in which output strings are as close as possible to the root – called “OTST”)

Example:

(*un triángulo y un cuadrado* , *a triangle and a square*),
 (*un triángulo grande* , *a large triangle*),
 (*un cuadrado* , *a square*)



2. Orderly traverse the tree, while **merging states** in order to get, hopefully, adequate generalizations.

OSTIA State-Merging learning procedure

- The traversal of the tree follows a **level by level order**, typically using the lexicographic order of state names.
- Two kinds of State Merging:
 - Merging based on **local conditions**: involve only the two states under consideration. The most basic idea [Oncina, 91-93]:
If both candidate states have the same output, or at least one has no output, merging is allowed.
 - **Derived merges**: once two states are merged, others may also need to be recursively merged in order to *preserve determinism*.

This process may require to “Push-back” certain output substrings.

- If a cascade of derived merges *fails* preserving determinism, the original and all the derived merges are discarded.

Outline of the OSTIA [Oncina,91]

Algorithm OSTIA ("Onward Subsequential Transducer Inference Algorithm")

Input: Finite set of (non ambiguous) input output pairs $T \subset (X^* \times Y^*)$

Output: Onward Subsequential Transducer τ compatible with T

$\tau' = OTST(T)$; (let $Q(\tau')$ denote the set of states of τ')

$\forall q \in Q(\tau') - \{q_0\}$ in a *level-by-level order*, **do**

$\forall p < q$ **do**

$\tau = merge(\tau', p, q)$

while $\exists q', q'' \in Q(\tau)$ that violate *subsequential conditions*, **do**

- try to restore subsequentiality by *Derived Merging*, possibly requiring to “push-back” some output substrings of the edges incoming to q', q'' towards the leaves of τ
- **if** “*Derived Merging*” possible **then** $\tau = merge(\tau, q', q'')$

end while

if *subsequential*(τ) **then** $\tau' = \tau$

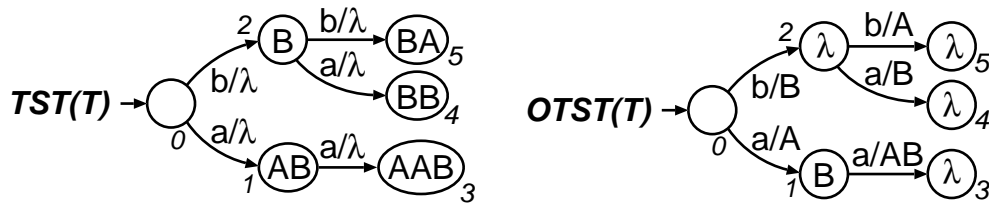
end $\forall p$

end $\forall q$

end OSTIA

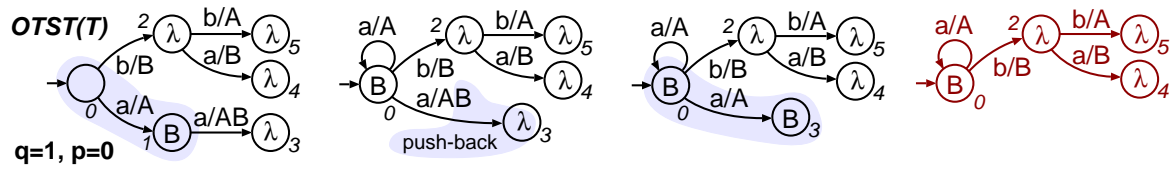
An Example of OSTIA state-merging process

$X=\{a,b\}$; $Y=\{A,B\}$; $T=\{(b,B), (a,AB), (bb,BA), (ba,BB), (aa,AAB)\}$



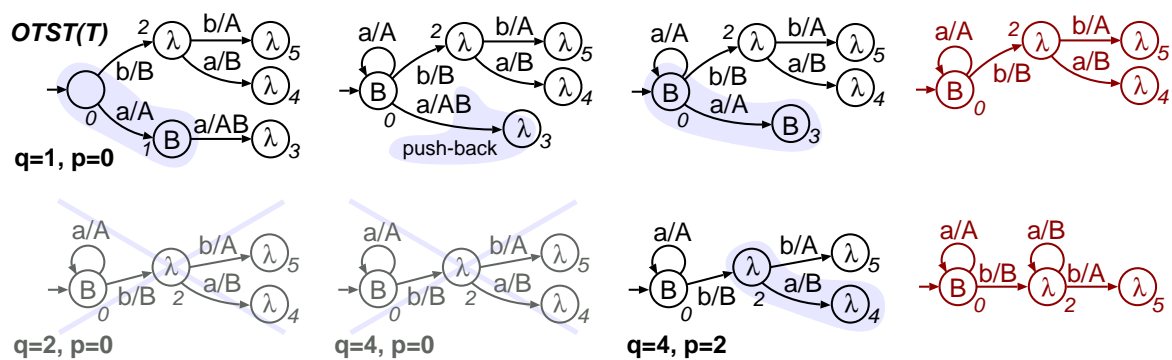
An Example of OSTIA state-merging process

$X=\{a,b\}$; $Y=\{A,B\}$; $T=\{(b,B), (a,AB), (bb,BA), (ba,BB), (aa,AAB)\}$



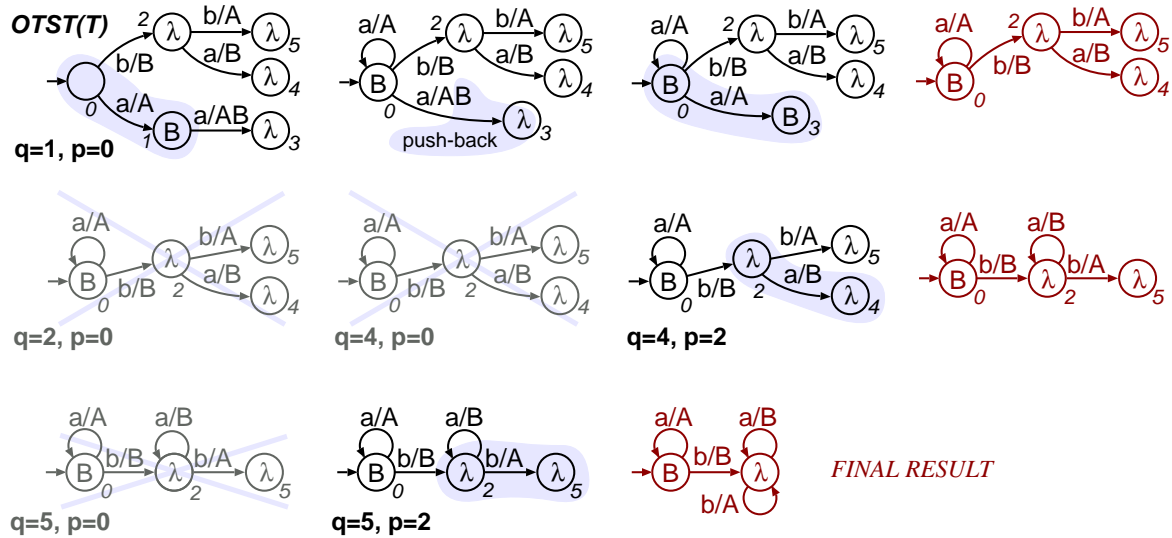
An Example of OSTIA state-merging process

$X=\{a,b\}$; $Y=\{A,B\}$; $T=\{(b,B), (a,AB), (bb,BA), (ba,BB), (aa,AAB)\}$



An Example of OSTIA state-merging process

$X=\{a,b\}$; $Y=\{A,B\}$; $T=\{(b,B), (a,AB), (bb,BA), (ba,BB), (aa,AAB)\}$



The Onward Subsequential Transducer Inference Algorithm (OSTIA)

INPUT: input-output pairs $T \subset (X^* \times Y^*)$; OUTPUT: OST τ consistent with T

```

 $\tau := \text{OTST}(T)$ ;  $q := \text{first}(\tau)$ ;
while  $q < \text{last}(\tau)$  do {
   $q := \text{next}(\tau, q)$ ;  $q' := \text{first}(\tau)$ ;
  while  $q' < q$  do {
    if  $\sigma(q') = \sigma(q)$  or  $\sigma(q') = \emptyset$  or  $\sigma(q) = \emptyset$  then {
       $\tau' := \tau$ ;  $\text{merge}(\tau, q', q)$ ;
      while  $\neg \text{subsequential}(\tau)$  do {
        let  $(r, a, v, s), (r, a, v', s')$  be two edges of  $\tau$  that
          violate the subsequential condition, with  $s' < s$ ;
        if  $s' < q$  and  $v' \notin \text{Pr}(v)$  then exitwhile;
         $u := \text{lcp}(v', v)$ ;
         $\text{push\_back}(\tau, u^{-1}v', (r, a, v', s'))$ ;
         $\text{push\_back}(\tau, u^{-1}v, (r, a, v, s))$ ;
        if  $\sigma(s') = \sigma(s)$  or  $\sigma(s') = \emptyset$  or  $\sigma(s) = \emptyset$ 
          then  $\text{merge}(\tau, s', s)$  else exitwhile;
      } // while  $\neg \text{subsequential}(\tau)$ 
      if  $\neg \text{subsequential}(\tau)$  then  $\tau := \tau'$  else exitwhile;
    } // if  $\sigma(q') = \sigma(q)$ 
     $q' := \text{next}(\tau, q')$ ;
  } // while  $q' < q$ 
} // while  $q < \text{last}(\tau)$ 

```

Properties of OSTIA learning

[Oncina, García & Vidal, 93]

- *Correctness*: the resulting transducer is *subsequential* and is a (state-merging) *generalization* of the set of training pairs T .
- *Convergence*: Using OSTIA the class of *total* Subsequential Transductions can be *identified in the limit*.
- *Efficiency*: OSTIA average running time is observed to be $O(n(m+k))$, where
 - $n = \sum_{(x,y) \in T} |x|$, (overall length of input strings)
 - $m = \max_{(x,y) \in T} |x|$ (longest output string)
 - $k = |X|$ (size of input alphabet).

\Rightarrow *huge sets of training examples can be easily handled.*

Applications of SSTs and OSTIA learning

- *Learning several toy but not trivial transduction tasks* [Oncina, 91-93].
 - Simple Arithmetic (e.g., decimal division by a fixed number).
 - Conversion of (large) English Numbers into Decimal notation.
 - Translation of (large) English Numbers into Spanish (and vice versa).
 - Conversion of Roman Numbers into Decimal.
 - etc.
- *Semantic Decoding*:
 - MLA [Castellanos et al.,98]
 - (Subset of) ATIS [Vidal,94]
- *Language Translation*:
 - MLA [Castellanos et al.,94]
 - Traveler Task [Amengual et al., 95-99]

Language Understanding through semantic decoding

Given a speech or text input sentence, produce an output which can be used to *drive the actions* specified in this sentence.

TYPICAL EXAMPLES:

- ATIS (Air Travel Information Systems):
 - **input:** Spontaneous English Sentences
 - **output:** Formal Query commands to the ATIS Data Base
- BDGEO (Spanish Geographic Quest):
 - **input:** Natural Language Spanish Sentences
 - **output:** Formal Query commands to BDGEO
- MLA ("Miniature Language Acquisition [Feldman et al., 90]):
 - **input:** Quasi-natural English Sentences
 - **output:** First-Order Predicate Logic Formulae

A simple experimental language understanding task: MLA

[Feldman et al., 90]

- Involves description and manipulation of simple visual scenes.
- Originally introduced as a challenging Language Learning task with a fairly simple syntax and small lexicon (about 30 words).
- Extended, as required, to study the impact of *increasing complexity, vocabulary size*, etc.

Examples:

*a medium light square and a circle are far above a light circle and a medium square
a large dark triangle is added far to the left of the square and the medium circle
the large circle which is above the square and the medium triangle is removed*

MLA: language understanding through semantic decoding

[Castellanos et al., 94-98]

- Visual scenes of MLA “understood” in terms of (*first-order*) logic formulae.
- Objects = Variables: x, y, z, w (allow up to *four* objects in a scene).
- 8 unary predicates on variables for *shape*, *shade* and *size*
- 9 (0-ary or binary) predicates for object relative positions (*above*, *below*, *far below*, *to the right*, *touch*, etc).
- Three increasingly non-monotone representations for object relations: L1, L2, L3. Translation into L1 is purely *sequential*; *subsequential* for L2 and L3

Examples:

a small triangle touches a medium light circle and a large square

L1: (Sm(x) & T(x)) **To** (M(z) & Li(z) & C(z) & La(w) & S(w))
L2: Sm(x) & T(x) & **To(x,z)** & M(z) & Li(z) & C(z) & **To(x,w)** & La(w) & S(w)
L3: Sm(x) & T(x) & M(z) & Li(z) & C(z) & La(w) & S(w) & **To(x,z)** & **To(x,w)**

Air Travel Information System (ATIS): semantic decoding

Translate English sentences into a semantic representation
in terms of “Pseudo English” (PE) formal queries.

Examples:

*show all flights and fares from <city> to <*city>*

LIST FLIGHTS FROM <CITY> AND TO <*CITY> ALONG WITH FARES

*I'd like information on <airline> flight from <city> to <*city>*

LIST FLIGHTS FROM <CITY> AND TO <*CITY> AND <AIRLINE>

*I'd like to find cheapest fare one-way fare from <city> to <*city>*

LIST CHEAPEST ONE-WAY FARES CHARGED FOR FLIGHTS FROM <CITY> AND TO <*CITY>

*please tell me about ground transportation from <city> airport to downtown <*city>*

LIST GROUND SERVICES PROVIDED FOR <AIRPORT> AND PROVIDED FOR <*CITY>

what airline is <airline> abbreviation for

LIST AIRLINES WHOSE AIRLINE CODE IS <AIRLINE>

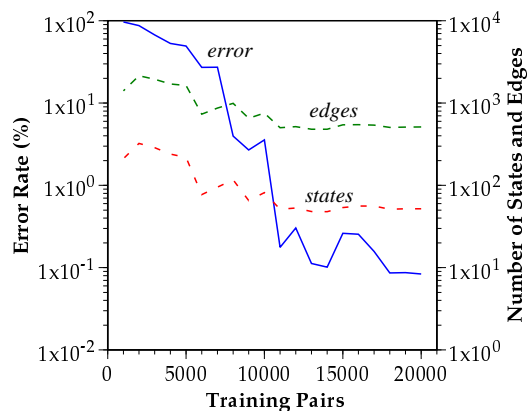
English sentences in lowercase, Pseudo-English commands in capitals.

Tokens within angular brackets are “generic non-terminals” or *bilingual categories*.

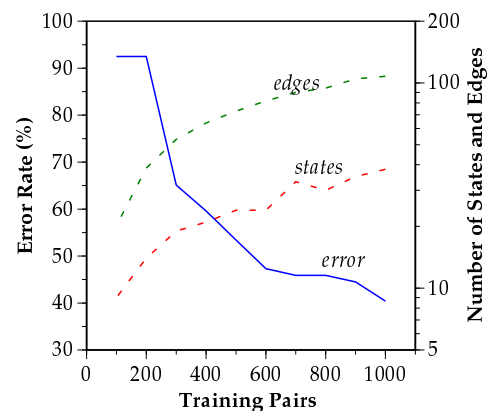
Semantic decoding: OSTIA learning results

Evolution of test-set semantic error and size of the OSTIA learned transducers for increasing amounts of training data.

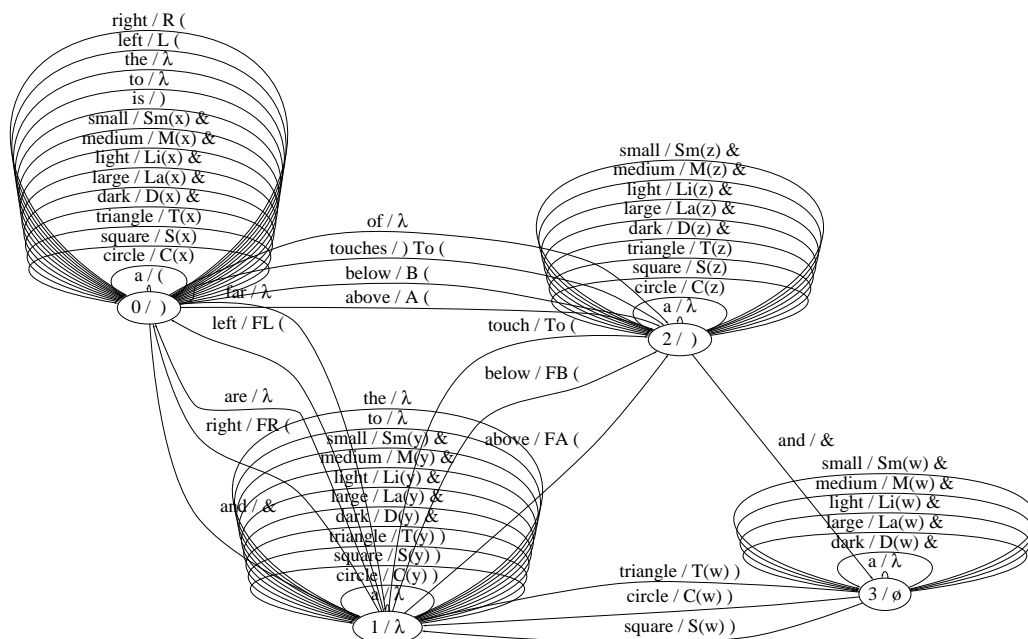
MLA-L3 (10k test sentences)
(similar results for L2;
slightly better for L1)



ATIS (146 test sentences)
(small subset of short,
class A sentences)



OSTIA-learned SST for the MLA language understanding task (L1)



Machine Translation (MT) and Subsequential Transduction

- Translation between languages can be modeled by Finite State (FS) mappings
- An important advantage of FS Translation Models is their great adequacy to be used for speech-input MT
- Theoretically speaking, most language pairs involve only subsequential mappings (*output text can be produced without having to wait until the end of the input discourse!*)
- In practice, many language pairs do involve only short-term *input/output asynchronies*
- **Subsequential Transducers** can be appropriate for **Limited Domain MT applications**

A simple experimental Machine Translation task: MTA

[Feldman et al., 90] [Castellanos et al., 94]

- Based on MLA (description and manipulation of simple visual scenes), which was originally introduced as a challenging Language Learning task with a fairly simple syntax and small lexicon (about 30 words).
- Reformulated for Machine Translation and *extended*, as required, to study the impact of increasing degree of input-output *non-monotonicity*, *vocabulary size*, etc.

Examples (Spanish-English):

un cuadrado mediano y claro y un círculo tocan a un círculo claro y un cuadrado mediano
 a medium light square and a circle touch a light circle and a medium square

se añade un triángulo grande y oscuro muy a la izquierda del cuadrado y del círculo
 a large dark triangle is added far to the left of the square and the circle

se elimina el círculo grande que esta encima del cuadrado y del triángulo mediano
 the large circle which is above the square and the medium triangle is removed

MTA translation results using OSTIA

[Castellanos, Galiano and Vidal, ICGI-94], [Oncina et al., ICSNLP-94]

Spanish-English Translation Word Error Rates for the Extended MTA Task, as a function of the Training Set size supplied to OSTIA.
Test Set: 10,000 independent text input sentences.

Train. Size	WER	States	Edges
1,000	58.8%	412	1652
2,000	57.0%	846	3197
4,000	51.8%	1598	5970
8,000	3.4%	186	891
16,000	0.0%	17	206

- Convergence starts from 4,000–8,000 training pairs (decreasing size of the learned transducers).
- Good results achieved with very compact transducers learned from reasonably small training sets.

▷ **Bad news:** These SSTs perform very poorly with *imperfect text* or *speech* input.

“Good” basic SSTs can accept incorrect input producing even more incorrect output!

OSTIA learning generalizes the training pairs as much as possible, while preserving the input-output mapping represented by these pairs. This may lead to *compact* and accurate transducers but they generally involve excessive *over-generalization* of the input and output sentences.

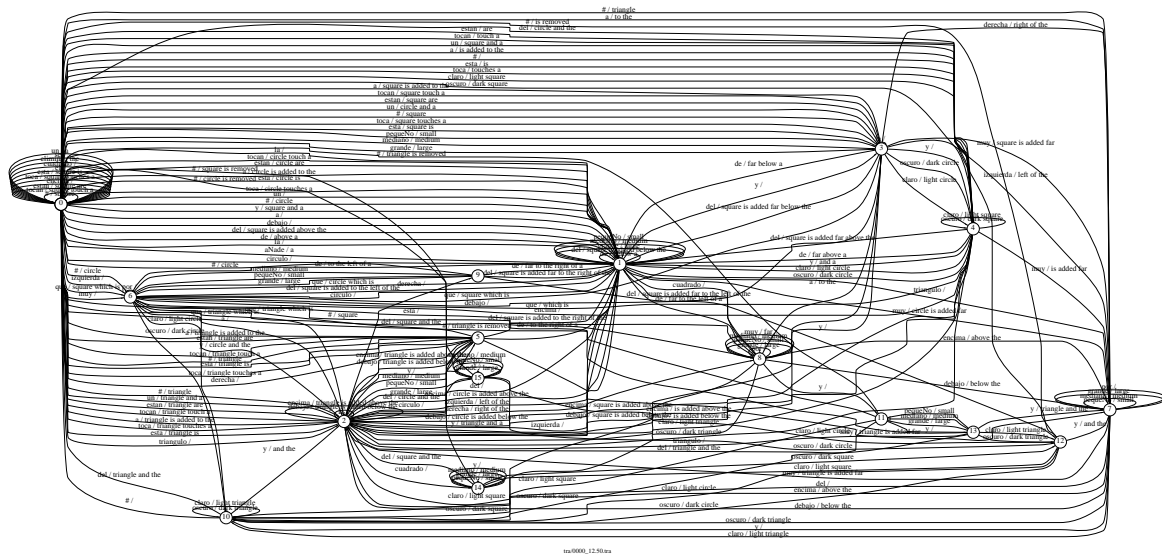
<i>debajo izquierda esta por</i>	→	square is removed
<i>elimina un y</i>	→	the a
<i>a y y claro que</i>	→	light square triangle which is
<i>muy esta oscuro</i>	→	dark square which is square

Examples of Spanish sentences accepted (and translated) by a “good” transducer learned by OSTIA (0.0% translation WER for *clean text* input).



This is *not* a problem for translating *clean text* but it leads to very large translation errors for corrupted text or for *speech input*!

Basic OSTIA–learned SST for Spanish-English MTA

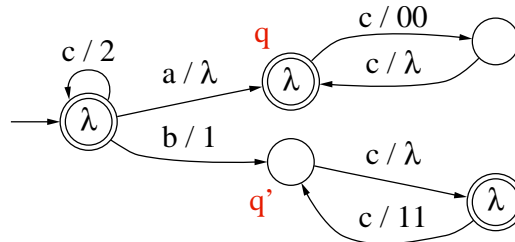


A Difficult-to-learn (partial) Subsequential Transduction

Let $t : \{a, b, c\}^* \rightarrow \{0, 1, 2\}^*$ be a *partial* Subsequential function defined as:

$$t = \{(c^m, 2^m) | m \geq 0\} \cup \{(c^m a c^{2n}, 2^m 0^{2n}) | m, n \geq 0\} \cup \{(c^m b c^{2n+1}, 2^m 1^{2n+1}) | m, n \geq 0\}$$

A Subsequential
Transducer realizing t :

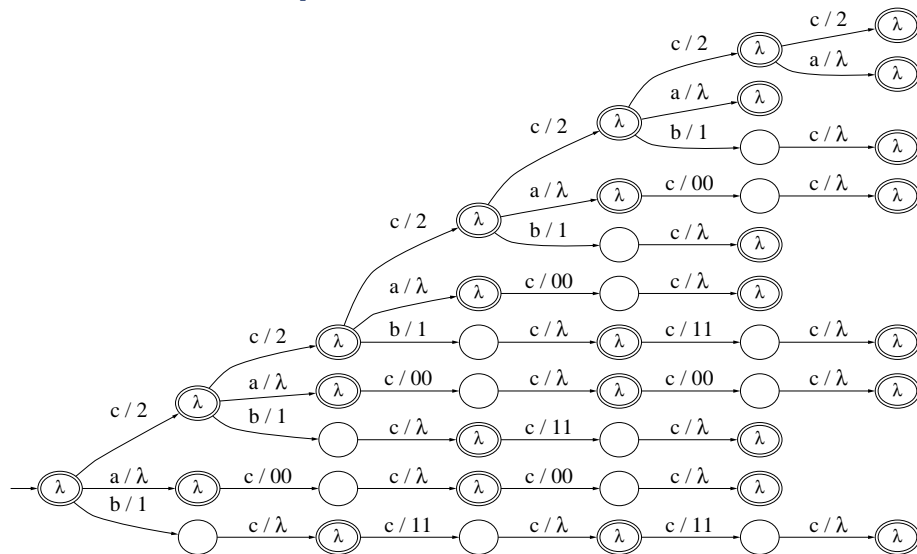


Samples of t , up
to input length 6:

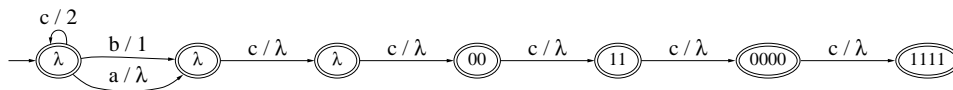
(,)	(cbc, 21)	(cccc, 2222)	(bcccc, 11111)
(a,)	(cca, 22)	(acccc, 0000)	(cacc, 20000)
(c, 2)	(ccc, 222)	(cbccc, 2111)	(ccbccc, 22111)
(bc, 1)	(bcc, 111)	(ccacc, 2200)	(cccacc, 22200)
(ca, 2)	(cacc, 200)	(cccbc, 2221)	(ccccbc, 22221)
(cc, 22)	(ccbc, 221)	(cccca, 2222)	(cccca, 22222)
(acc, 00)	(ccca, 222)	(ccccc, 22222)	(ccccc, 222222)

No transduction example can help distinguish the states q and q' .

Onward Tree Subsequential Transducer and OSTIA result



OTST of a sample of t consisting of all the input-output pairs up to an input length of 6.



Transducer yield by OSTIA from this OTST.

Index

- 1 Subsequential Transduction: “OSTI” Algorithm ▷ 2
- 2 Using input/output syntactic constraints: OSTIA-DR ▷ 29
- 3 OSTIA-DR: improving scalability ▷ 45
- 4 Bibliography ▷ 70

Helping OSTIA with input/output syntactic constraints

Two kind of conditions for OSTIA state merging:

- *Local conditions*: involve only the two states under consideration.

Basic OSTIA allows merging two candidate states if both have the same output or at least one has no output [Oncina, 91-93].

- *Derived merges*: once two states have been merged, others may also need to be merged (while possibly “pushing-back” some output substrings) in order to preserve determinism.

New Local Conditions:

Use *Finite-State Models* of the Input (or Domain) and/or the Output (or Range) to enforce *Input and/or Output Syntactic Constraints*

Idea [Oncina, 93-94]: **disallow the merging of two states if they correspond to different states of the Input or Output models**.

The resulting algorithm is known as OSTIA-DR

OSTIA-DR

[Oncina,93]

- The use of Domain (and Range) information can be accomplished by labeling each state of the initial Onward Tree Subsequential Transducer (OTST) with the name of the state of the Domain (or Range) FS Model that would be reached with the corresponding strings.
- The local compatibility rules then include the condition of disallowing the merging of two states if their labels are distinct.
- **The resulting SSTs only accept input sentences and only produce output sentences compatible with the syntactic constraints represented by the FSMs used**
 - ▷ *This becomes essential for imperfect text or speech input.*
- Using OSTIA-DR, the class of *partial* Subsequential Functions can be *identified in the limit*.

Using input/output syntactic constraints:

outline of OSTIA-DR [Oncina et al.,94]

Algorithm OSTIA-DR ("OSTIA assisted by DOMAIN/RANGE constraints")

Input: Finite set of (non ambiguous) input output pairs $T \subset (X^* \times Y^*)$

Finite-State models, G_D, G_R , of the Domain (X^*) and Range (Y^*)

Output: Onward Subsequential Transducer τ' compatible with T

Method:

$\tau' = OTST(T)$; (let $Q(\tau')$ denote the set of states of τ')

$\forall q \in Q(\tau') - \{q_0\}$ in a *level-by-level order*, **do**

$\forall p < q$ **if** p, q are compatible with G_D and/or G_R **do**

$\tau = merge(\tau', p, q)$

while $\exists q', q'' \in Q(\tau)$ that violate *subsequential conditions*, **do**

– try to restore subsequentiality by *Derived Merging*, possibly requiring to “push-back” some output substrings of the edges incoming to q', q'' towards the leaves of τ'
– **if** “*Derived Merging*” possible **then** $\tau = merge(\tau, q', q'')$

end while

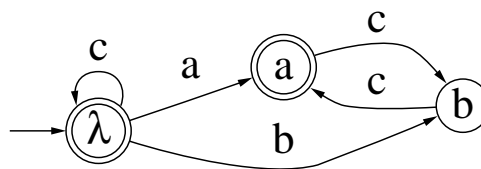
if *subsequential*(τ) **then** $\tau' = \tau$

end $\forall p$

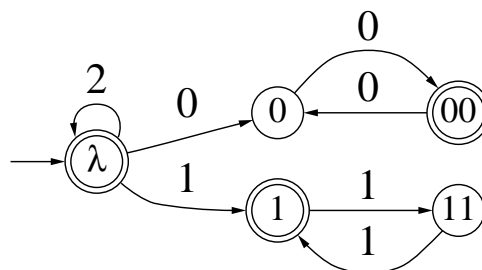
end $\forall q$

end OSTIA

FS input and output models for the “difficult transduction”



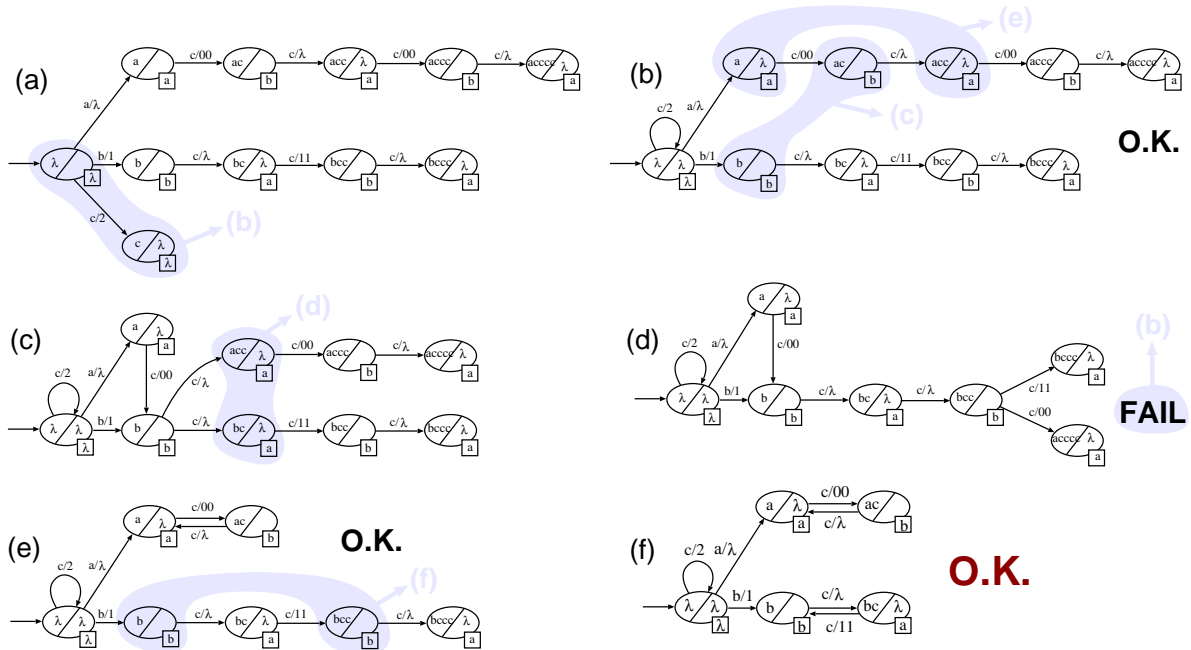
Finite State Domain (input) model



Finite State Range (output) model

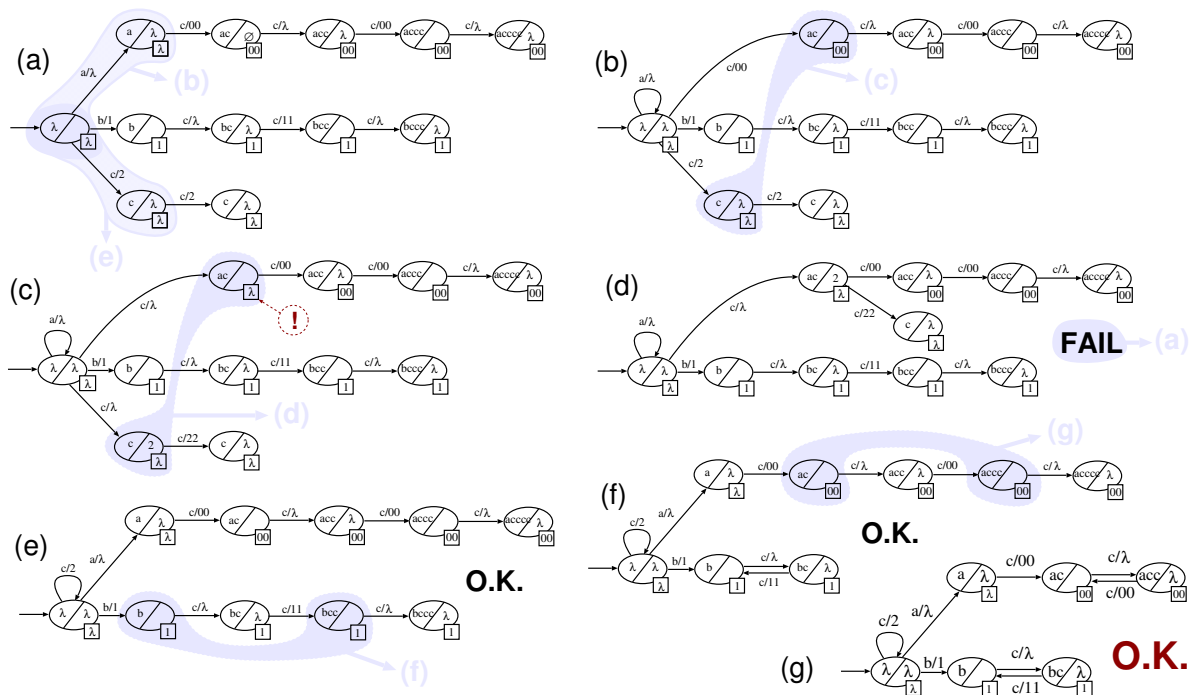
OSTIA-D learning

Training set: $T = \{(a, \lambda), (acc, 00), (acccc, 0000), (bc, 1), (bccc, 111), (c, 2)\}$



OSTIA-R learning

Training set: $T = \{(a, \lambda), (acc, 00), (acccc, 0000), (bc, 1), (bccc, 111), (c, 2), (cc, 22)\}$.



Using input-language constraints: OSTIA-D

INPUT: input-output pairs, $T \subset (X^* \times Y^*)$, Finite-State model, G_D , of the Domain (X^*)

OUTPUT: OST τ consistent with T and G_D

```

 $\tau := \text{OTST}(T); \quad q := \text{first}(\tau);$ 
while  $q < \text{last}(\tau)$  {
   $q := \text{next}(\tau, q); \quad q' := \text{first}(\tau);$ 
  while  $q' < q$  {
    if  $(\sigma(q') = \sigma(q) \text{ or } \sigma(q') = \emptyset \text{ or } \sigma(q) = \emptyset) \text{ and}$ 
       $\delta_D(p_0, \text{input\_prefix}(q')) = \delta_D(p_0, \text{input\_prefix}(q))$  then {
       $\tau' := \tau; \quad \text{merge}(\tau, q', q);$ 
      while  $\neg \text{subsequential}(\tau)$  {
        let  $(r, a, v, s), (r, a, v', s')$  be two edges of  $\tau$  that
          violate the subsequential condition, with  $s' < s$ ;
        if  $s' < q$  and  $v' \notin \text{Pr}(v)$  then exitwhile
         $u := \text{lcp}(v', v);$ 
         $\text{push\_back}(\tau, u^{-1}v', (r, a, v', s'));$   $\text{push\_back}(\tau, u^{-1}v, (r, a, v, s));$ 
        if  $\sigma(s') = \sigma(s) \text{ or } \sigma(s') = \emptyset \text{ or } \sigma(s) = \emptyset$ 
          then merge}(\tau, s', s) \text{ else exitwhile}
      } // while  $\neg \text{subsequential}(\tau)$ 
      if  $\neg \text{subsequential}(\tau)$  then  $\tau := \tau'$  else exitwhile
    } // if  $\sigma(q') = \sigma(q)$ 
     $q' := \text{next}(\tau, q');$ 
  } // while  $q' < q$ 
} // while  $q < \text{last}(\tau)$ 

```

Using output-language constraints: OSTIA-R

INPUT: input-output pairs, $T \subset (X^* \times Y^*)$, Finite-State model, G_R , of the Range (X^*)

OUTPUT: OST τ consistent with T and G_R

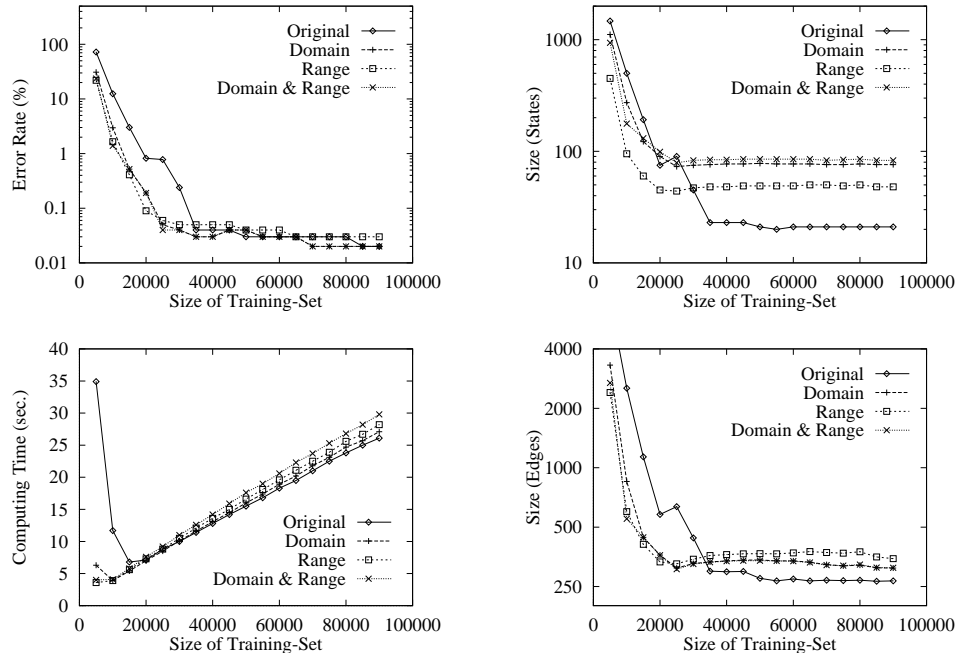
```

 $\tau := \text{OTST}(T); \quad q := \text{first}(\tau);$ 
while  $q < \text{last}(\tau)$  {
   $q := \text{next}(\tau, q); \quad q' := \text{first}(\tau);$ 
  while  $q' < q$  {
    if  $(\sigma(q') = \sigma(q) \text{ or } \sigma(q') = \emptyset \text{ or } \sigma(q) = \emptyset) \text{ and}$ 
       $\delta_R(p_0, \text{output\_prefix}(q')) = \delta_R(p_0, \text{output\_prefix}(q))$  then {
       $\tau' := \tau; \quad \text{merge}(\tau, q', q);$ 
      while  $\neg \text{subsequential}(\tau)$  {
        let  $(r, a, v, s), (r, a, v', s')$  be two edges of  $\tau$  that
          violate the subsequential condition, with  $s' < s$ ;
        if  $s' < q$  and  $v' \notin \text{Pr}(v)$  then exitwhile
         $u := \text{lcp}(v', v);$ 
         $\text{push\_back}(\tau, u^{-1}v', (r, a, v', s'));$   $\text{push\_back}(\tau, u^{-1}v, (r, a, v, s));$ 
        if  $\sigma(s') = \sigma(s) \text{ or } \sigma(s') = \emptyset \text{ or } \sigma(s) = \emptyset$ 
          then merge}(\tau, s', s) \text{ else exitwhile}
      } // while  $\neg \text{subsequential}(\tau)$ 
      if  $\neg \text{subsequential}(\tau)$  then  $\tau := \tau'$  else exitwhile
    } // if  $\sigma(q') = \sigma(q)$ 
     $q' := \text{next}(\tau, q');$ 
  } // while  $q' < q$ 
} // while  $q < \text{last}(\tau)$ 

```

MTA: OSTIA and OSTIA-DR learning performance

Spanish-English Extended MTA Learning performance as a function of training-set size. Domain and/or range Language Models: 3-TSS (3-Gram); Test Set: 100,000 independent input sentences.



Spanish-English MTA: OSTIA and OSTIA-DR learning results

Translation Word Error Rates for the Extended MTA Feldman's Task, as a function of the Training Set size supplied to OSTIA and OSTIA-DR (with 4-Gram Language Models)

Test Set: 10,000 independent input sentences.

Training Set Size	OSTIA			OSTIA-DR		
	WER	States	Edges	WER	States	Edges
1,000	58.8%	412	1652	55.1%	813	2023
2,000	57.0%	846	3197	47.1%	1406	3353
4,000	51.8%	1598	5970	30.1%	1686	4051
8,000	3.4%	186	891	1.4%	244	719
16,000	0.0%	17	206	0.0%	100	363

Using Input/Output syntactic constraints, translation errors can be reduced by a factor of two.

MTA OSTIA and OSTIA-DR learning: impact of noisy text input and input–output language syntactic constraints

Spanish-English Translation Word Error Rates of distorted test sentences for the Extended MTA Task, as a function of the Training Set size supplied to OSTIA and OSTIA-DR (with 4-Gram Input and Output Language Models). Noisy input Translations obtained using Error-Correcting Parsing.

Test Set: 10,000 **clean** and **5%-distorted** independent input sentences.

Train.Set Size	OSTIA Clean	OSTIA 5%Dist	OSTIA-DR Clean	OSTIA-DR 5%Dist
8,000	3.4%	15.0%	1.4%	2.7%
16,000	0.0%	11.7%	0.0%	1.7%

Using Input/Output syntactic constraints increases robustness dramatically

MTA OSTIA and OSTIA-DR Learning: examples of distorted input sentences and the obtained translations

I=Original Input; D=5% Distorted Input; T=System Translation.

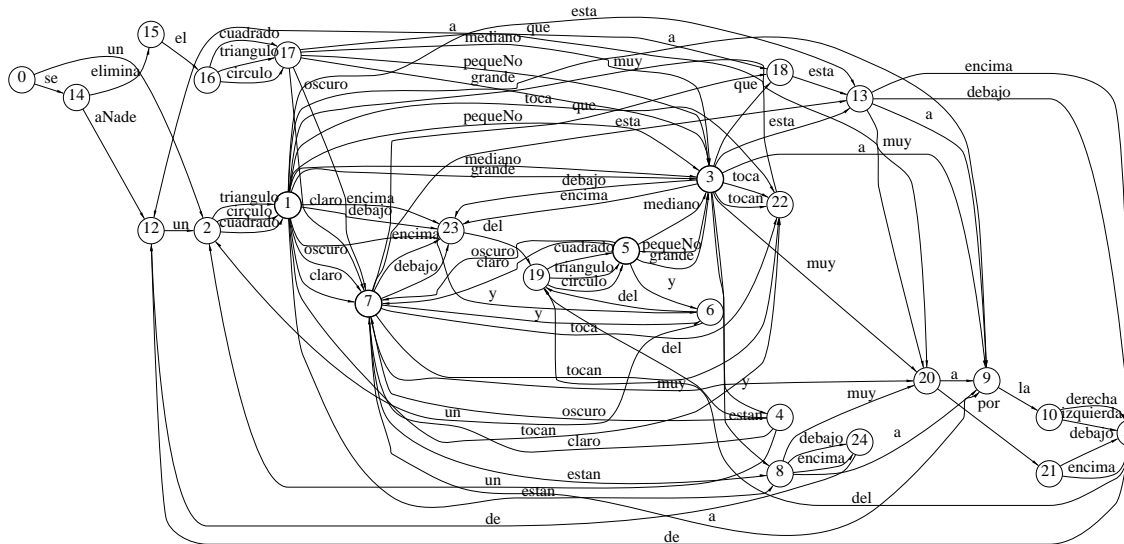
Correctly Translated:

I: *se elimina el círculo grande y claro que está muy por encima del triángulo oscuro y del cuadrado • mediano*
D: *se elimina y círculo grande y claro • está muy por encima • triángulo oscuro y del cuadrado un mediano*
T: *the large light circle which is far above the dark triangle and the medium square is removed*
:
I: *un • círculo mediano y claro está debajo de un cuadrado pequeño y claro y un triángulo pequeño y oscuro*
D: *un tocan círculo mediano y claro • debajo de un cuadrado pequeño claro y se triángulo pequeño y oscuro*
T: *a medium light circle is below a small light square and a small dark triangle*

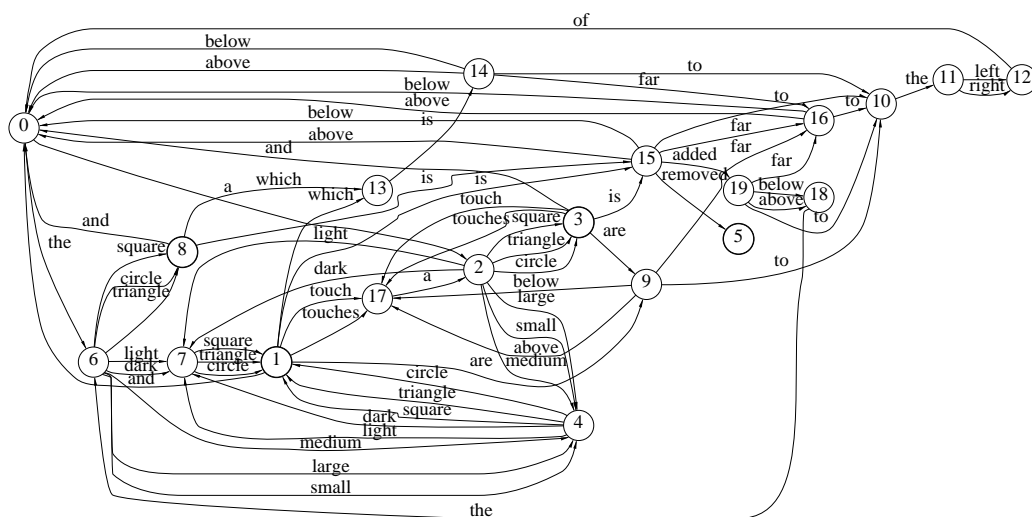
Translation Errors:

I: *se • elimina el círculo que está muy a la izquierda del círculo oscuro y del triángulo mediano y oscuro*
D: *se de de el • que está muy a la izquierda del círculo oscuro y del triángulo mediano y oscuro*
T: *the square which is far to the left of the dark circle and the medium dark triangle is removed*
:
I: *se añade un triángulo mediano y claro muy a la derecha del cuadrado mediano y oscuro y del círculo pequeño y oscuro*
D: *se añade un triángulo la y claro muy a la derecha del cuadrado mediano y oscuro oscuro claro círculo pequeño y oscuro*
T: *a small light triangle is added far to the right of the medium dark square and the small dark circle*

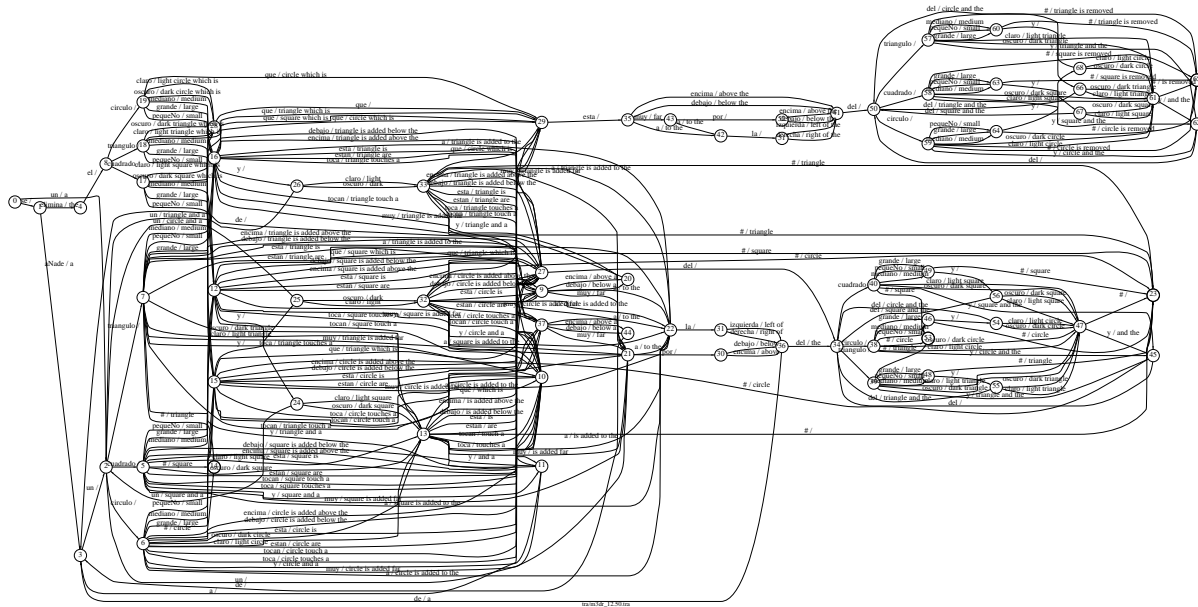
3-TSS Automaton (entailing 3-Gram constraints)



3-TSS Automaton (entailing 3-Gram constraints)



(using both *Domain* and *Range* 3-Gram constraints)



1 Subsequential Transduction: “OSTI” Algorithm ▶ 2

2 Using input/output syntactic constraints: OSTIA-DR 29

- 3 OSTIA-DR: improving scalability ▸ 45

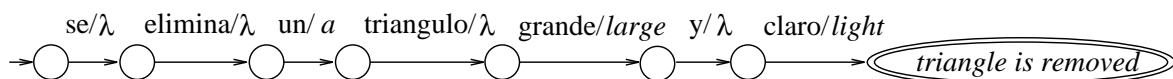
4 Bibliography ▸ 70

Scalability issues

Subsequential Transduction copes with Input-Output non-monotonicity by *delaying the decision for output (sub)strings*.

A training pair and a corresponding SST:

(*se elimina un triángulo grande y claro*, a large light triangle is removed)



Problem:

The number of states can grow as much as $O(n^k)$, where n is the *number of functionally equivalent input words* and k is the *number of word-positions to be delayed*.

The required amount of training data can become prohibitive.

Dealing with increasing vocabulary size (n) and degree of non-monotonicity (k)

Approaches:

$n \Rightarrow$ ***Bilingual Categorization***

[Vilar, Marzal, Vidal, Eurospeech-95]:

While the direct approach degrades rapidly with increasing vocabulary sizes, categorization largely prevents accuracy degradation.

$k \Rightarrow$ ***Partial Alignment and Word Reordering***

[Vilar, Vidal, Amengual, Llorens, ECAI-96, SPECOM-96]:

Training-data requirements can be reduced dramatically.

Cutting down the impact of increasing vocabulary size through Bilingual Categorization

- Substitute words or groups of words by labels representing their syntactic (or semantic) *categories* within a limited rank of options.
- *Learn* a transducer with the *categorized sentences*, which entails a (much) smaller effective vocabulary.
- *Expand* each *category-labeled edge* of the learned transducer with a (small) *transducer for this category*.

Expansion leads to a single, perhaps large transducer which encompasses all the required information.

Categorization helps achieving adequate generalizations and proves very effective to prevent degradation of results with increasing vocabulary sizes.

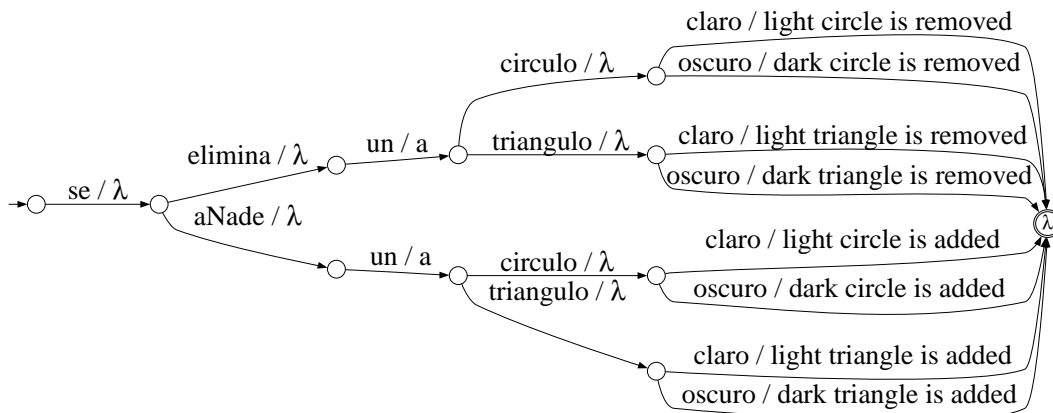
A very small MTA Spanish-English training set

<i>se añade un triángulo claro</i>	↔	a light triangle is added
<i>se añade un círculo claro</i>	↔	a light circle is added
<i>se añade un triángulo oscuro</i>	↔	a dark triangle is added
<i>se añade un círculo oscuro</i>	↔	a dark circle is added
<i>se elimina un triángulo claro</i>	↔	a light triangle is removed
<i>se elimina un círculo claro</i>	↔	a light circle is removed
<i>se elimina un triángulo oscuro</i>	↔	a dark triangle is removed
<i>se elimina un círculo oscuro</i>	↔	a dark circle is removed

A Categorized version of this Training Set

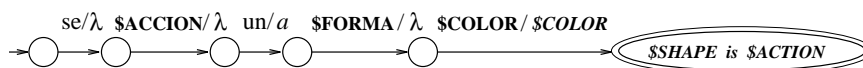
se \$ACCION un \$FORMA \$COLOR ↔ a \$COLOR \$SHAPE is \$ACTION

Subsequential Transducer for the very small MTA Spanish-English training set



Size grows very fast with the number of words in each category.

Categorized Transducer



Size no longer depends on the number of words in each category.

MTA extensions for experimentation with Bilingual Categories

Four extensions to the (extended) Feldman's MTA task:

- EXT1: 6 shapes, 3 sizes, 2 shades (Voc.: 37/28 Spanish/English words)
- EXT2: 12 shapes, 5 sizes, 4 shades/colors (Voc.: 50/36 words)
- EXT3: 18 shapes, 7 sizes, 6 shades/colors (Voc.: 63/48 words)
- EXT4: 118 shapes, 57 sizes, 56 shades/colors (Voc.: 363/248 words)

MTA: cutting down the impact of increasing vocabulary using Bilingual Categories

[Vilar, Marzal and Vidal, Eurospeech-95]

Translation Sentence Error Rate (in %) for two training-set sizes and increasing vocabulary sizes (3 categories: NOUN, ADJ, ADV). Test set: 10,000 independent sentences.

Inp/Out Voc.Sizes	8,000 Train. Pairs		32,000 Train. Pairs	
	Direct	Categ.	Direct	Categ.
37/28	3.1	0.9	0.5	0.2
50/38	42.1	1.5	5.7	0.3
63/48	62.5	3.0	26.5	0.6
363/248	91.3	3.4	98.0	0.7

While the direct approach degrades rapidly with increasing vocabulary sizes, categorization keeps the accuracy essentially unchanged.

A more complex and practical application: the “Traveler Task”

- Domain: *human-to-human communication* situations in the front-desk of a hotel.
- Data produced semi-automatically on the base of a small “seed corpus” obtained from several traveler-oriented booklets.
- Three language pairs: *Spanish-English*, *Spanish-German* and *Spanish-Italian* (only Spanish-English results reported here; similar results for the other languages).

The Traveler Task: features and examples

[Vidal et al., 96] (EuTrans ESPRIT project – first-phase)

Different sentence pairs in the corpus	171,481
Input/output vocabulary sizes	689 / 514
Average input/output sentence lengths	9.5 / 9.8
Input/output (2-Gram) test-set perplexities	12.8 / 7.0

(Similar features for Spanish-German and Spanish-Italian corpora)

Examples (Spanish-English):

Reservé una habitación individual y tranquila con televisión hasta pasado mañana.

I booked a quiet, single room with a tv. until the day after tomorrow.

Despiértenos mañana a las ocho menos cuarto, por favor.

Wake us up tomorrow at a quarter to eight, please.

Por favor, prepárenos nuestra cuenta de la habitación dos veintidós.

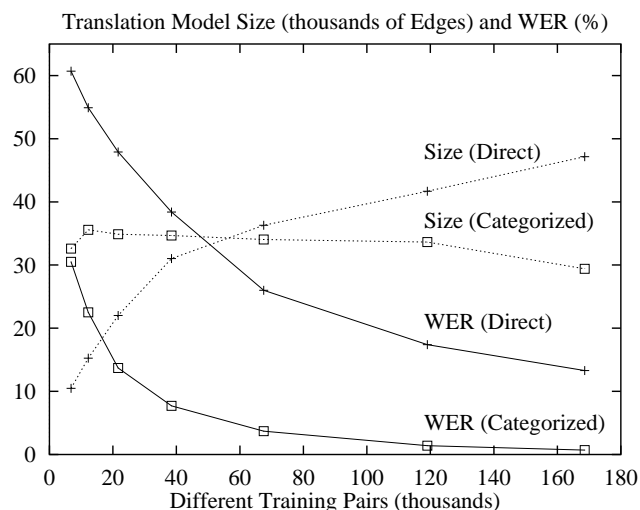
Could you prepare our bill for room number two two two for us, please?

Traveler Task text-input experiments

[Vidal et al., 96] (EuTrans – first-phase Final Report)

OSTIA–DR learning using Input and Output 3–Gram LM Constraints, *with* and *without* Categorization into 7 categories: *dates, times-of-day, room-numbers, etc.*

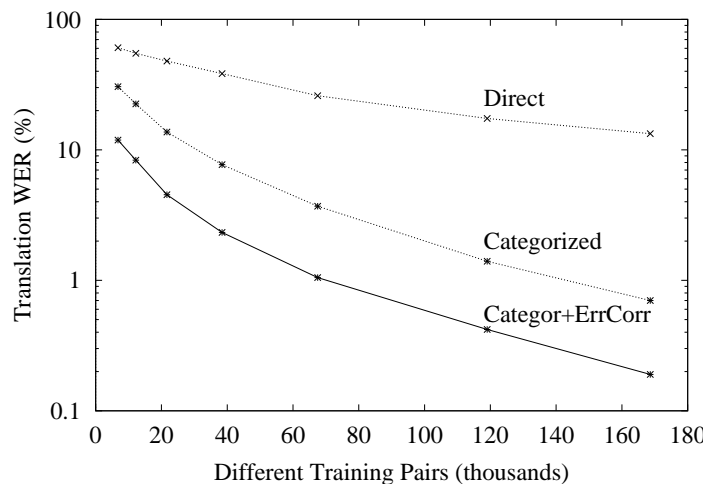
Test-Set:
2,730 different sentences.



► *Categorization leads to **useful accuracy** using moderate amounts of training data.*

Traveler Task Error-Correcting experiments

- OSTIA-DR learning using Input/Output 3-Gram LMs,
- Error model parameters estimated from artificially distorted input sentences, through Expectation-Maximisation and Viterbi re-estimation.



▷ **Training-data demands can be reduced by a factor of 2-3.**

Traveler Task: summary of text-input results

Impact of using *Categories* and *Error Correcting Parsing*

- OSTIA-DR learned Subsequential Transducers
- Training based on the largest training sets available
- Error model parameters estimated from artificially distorted input text
- Test-set: clean (undistorted) independent input text

	OSTIA-DR (baseline)	OSTIA-DR + Categories	OSTIA-DR + Categories+ECP
Spanish-English	13.33 %	0.74 %	0.18 %
Spanish-German	29.86 %	1.23 %	0.54 %
Spanish-Italian	17.60 %	2.54 %	0.51 %

Traveler Task: human subjective assessment results

[Vidal et al., 96] (EuTrans ESPRIT project – first-phase)

- Comparison of EuTrans results with translations provided by low-cost commercial translation packages, adapted to the Traveler Task.
- Human subjective results based on three experts.

	Spanish-to-German	Spanish-to-English		
	EUTRANS	EUTRANS	Power Translator	Spanish Assistant
PCT	81.7%	87.3%	49.0%	
PCIT	93.3%	90.3%	79.7%	75.3%
UM	+0.86	+0.81	+0.64	+0.57

- **PCT**: Percentage of correct translations
- **PCIT**: Percentage of correctly intelligible translations
- **UM**: An approximate usefulness measure

Automatic bilingual word clustering

As task complexity and diversity increase, automated methods are required to *discover* the bilingual categories which are actually relevant in a given corpus of the task.

A basic idea:

- Modify well-known, monolingual, K -means style word clustering techniques, by including translation information.
- Derive this information from an initial bilingual (probabilistic) dictionary.
- This dictionary can be obtained manually and/or using simple statistical techniques such as the IBM-1 translation model.

Preliminary experiments show that techniques based on this idea often supply very adequate bilingual clusters of (individual) words.

Cutting down the impact of increasing vocabulary size (n) and degree of non-monotonicity (k)

Approaches:

$n \Rightarrow$ **Bilingual Categorization**

[Vilar, Marzal, Vidal, Eurospeech-95]:

While the direct approach degrades rapidly with increasing vocabulary sizes, categorization largely prevents accuracy degradation.

$k \Rightarrow$ **Partial Alignment and Word Reordering**

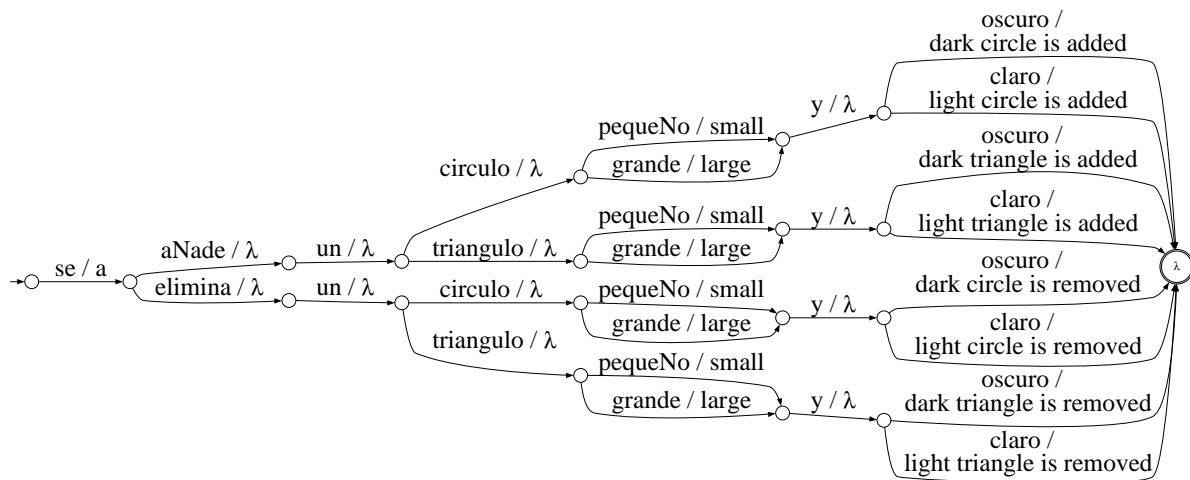
[Vilar, Vidal, Amengual, Llorens, ECAI-96, SPECOM-96]:

Training-data requirements can be reduced dramatically.

A small training set from the MTA task

<i>se elimina un triángulo grande y claro</i>	\leftrightarrow	<i>a large light triangle is removed</i>
<i>se elimina un triángulo pequeño y claro</i>	\leftrightarrow	<i>a small light triangle is removed</i>
<i>se elimina un círculo grande y claro</i>	\leftrightarrow	<i>a large light circle is removed</i>
<i>se elimina un círculo pequeño y claro</i>	\leftrightarrow	<i>a small light circle is removed</i>
<i>se elimina un triángulo grande y oscuro</i>	\leftrightarrow	<i>a large dark triangle is removed</i>
<i>se elimina un triángulo pequeño y oscuro</i>	\leftrightarrow	<i>a small dark triangle is removed</i>
<i>se elimina un círculo grande y oscuro</i>	\leftrightarrow	<i>a large dark circle is removed</i>
<i>se elimina un círculo pequeño y oscuro</i>	\leftrightarrow	<i>a small dark circle is removed</i>
<i>se añade un triángulo grande y claro</i>	\leftrightarrow	<i>a large light triangle is added</i>
<i>se añade un triángulo pequeño y claro</i>	\leftrightarrow	<i>a small light triangle is added</i>
<i>se añade un círculo grande y claro</i>	\leftrightarrow	<i>a large light circle is added</i>
<i>se añade un círculo pequeño y claro</i>	\leftrightarrow	<i>a small light circle is added</i>
<i>se añade un triángulo grande y oscuro</i>	\leftrightarrow	<i>a large dark triangle is added</i>
<i>se añade un triángulo pequeño y oscuro</i>	\leftrightarrow	<i>a small dark triangle is added</i>
<i>se añade un círculo grande y oscuro</i>	\leftrightarrow	<i>a large dark circle is added</i>
<i>se añade un círculo pequeño y oscuro</i>	\leftrightarrow	<i>a small dark circle is added</i>

Transducer for the small MTA training set

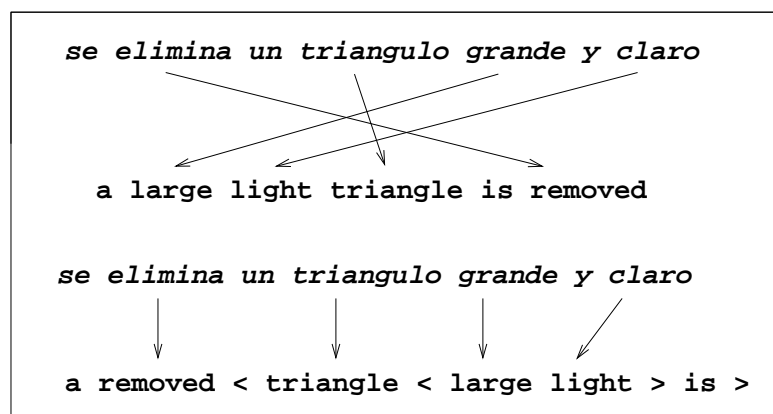


Size grows exponentially with the number of words to be delayed.

Coping with increasing input/output non-monotonicity

[Vilar et al., 1996]

Words of the (training) output sentences can be easily *reordered* on the base of *partial alignments*, which can be obtained, e.g., using a probabilistic bilingual dictionary such as the one obtained by training an IBM-1 translation model.

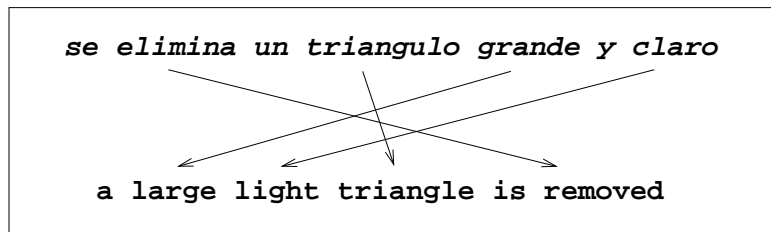


Original pair and a partial alignment (above). Reordering-Bracketing results (below).

Reordering is performed along with a *bracketing* scheme which allows recovering the correct word order.

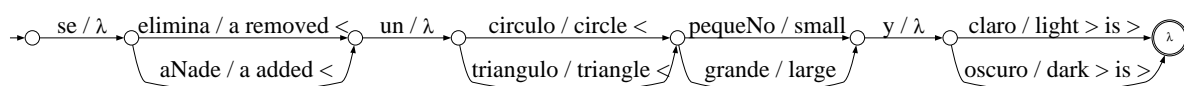
The Reordering Algorithm

Reordering is done by scanning the output sentence from left to right and creating a new reordered and bracketed sentence along the way.



Step	Word	Result (reordered and bracketed sentence)
1	a	a
2	large	a large
3	light	a large light
4	triangle	a triangle < large light >
5	is	a triangle < large light > is
6	removed	a removed < triangle < large light > is >

Transducer for the reordered small MTA training set



▷ **The number of states no longer grows exponentially**

▷ **Learning can be achieved with far less training data**

Recovering the correct word order

“Un-reordering” can be easily done
with the help of the embedded brackets and a stack:

Reordered sentence:

“a removed < triangle < large light > is > ”

Step	Word	Stack	Output
1	a	\emptyset	a
2	removed <	removed	a
3	triangle <	removed, triangle	a
4	large	removed, triangle	a large
5	light	removed, triangle	a large light
6	>	removed	a large light triangle
7	is	removed	a large light triangle is
8	>	\emptyset	a large light triangle is removed

Result: “a large light triangle is removed”

Reordering-based training and translation procedures

[Vilar et al., 1996]

Training: Given a training set S of pairs of input/output sentences (x, y) , the proposed training approach proceeds as follows:

1. Train IBM Model-1 on S and obtain a probabilistic dictionary D .
2. Prune from D those pairs of words with probability below a threshold.
3. Partially align the pairs of sentences in S using the pruned D .
4. Reorder and bracket the output sentences of S to produce S' .
5. Using OSTIA, learn a SST T from S' .

Translation: Given a new test input sentence x the trained system produces a translation y through the following simple steps:

1. Using T , obtain the translation y' of x
2. “Un-reorder” y' with the help of its embedded brackets to obtain y

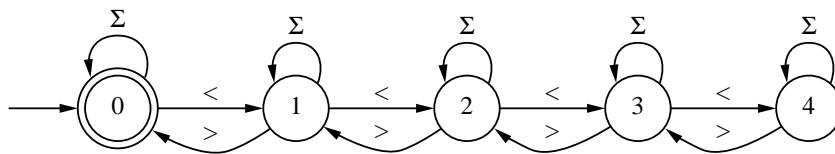
Balancing the brackets

Possible problem:

Transducers learned by OSTIA with output-reordered training data may not perfectly generalise a balanced bracketing for new unseen input test sentences. This becomes even more problematic with noisy (or speech) input.

A simple solution:

Limit the depth of the brackets and perform OSTIA-DR learning using an output finite-state “Language Model” that enforces correct bracketing.



(Σ represents an edge for each word in the output language vocabulary)

The number of states should match the maximum level of embedding allowed.

This can be combined with conventional (e.g., 3-Gram) output Language Models.

MTA OSTIA-DR/Word-Reordering results

[Vilar, Vidal, Amengual, ECAI-96]

Spanish-English Translation Word Error Rates for the Extended Feldman's MTA Task, as a function of the Training Set size.

Test Set: 10,000 5%-distorted independent input sentences.

Train. size	Direct	Reordered
1,000	44.0% (813 / 2023)	17.6% (532 / 1338)
2,000	37.8% (1406 / 3353)	6.2% (358 / 979)
4,000	25.2% (1686 / 4051)	2.2% (144 / 440)
8,000	2.7% (244 / 719)	1.7% (109 / 344)
16,000	1.7% (100 / 363)	1.7% (63 / 183)

In brackets, model sizes (states/edges).

▷ **Reordering can reduce the demand for training data by a factor of four**

Index

- 1 Subsequential Transduction: “OSTI” Algorithm ▷ 2
- 2 Using input/output syntactic constraints: OSTIA-DR ▷ 29
- 3 OSTIA-DR: improving scalability ▷ 45
- 4 *Bibliography* ▷ 70

Bibliography

- J.Oncina, P.García, E.Vidal: “Learning Subsequential Transducers for Pattern Recognition Interpretation Tasks”. IEEE Trans. on Pattern Analysis and Machine Intelligence. Vol.PAMI-15, No.5, pp.448-458, 1993.
- J.M.Vilar, E.Vidal, J.C.Amengual: “Learning Extended Finite-State Models for Language Translation”. Proc. of Extended Finite State Models Workshop (of ECAI’96), pp.92-96. Budapest, Agosto 1996.
- J.M.Vilar, V.M.Jiménez, J.C.Amengual, A.Castellanos, D.Llorens, E.Vidal: “Text and Speech Translation by means of Subsequential Transducers”. Natural Language Engineering, Vol.2, No.4, pp.351-354, 1996.
- E.Vidal: “Finite-State Speech-to-Speech Translation”. Int. Conf. on Acoustics Speech and Signal Processing (ICASSP-97), proc., Vol.1, pp.111-114. Munich, 1997.
- A.Castellanos, E.Vidal, A.Varó, J.Oncina: “Language Understanding and Subsequential Transducer Learning”. Computer Speech and Language, No.12, pp.193-228. 1998.
- J.C.Amengual, J.M.Benedí, F.Casacuberta, A.Castaño, A.Castellanos, V.Jiménez, D.Llorens, A.Marzal, M.Pastor, F.Prat, E.Vidal, J.M.Vilar: “The EuTrans-I Speech Translation System”. Machine Translation. Vol.15, pp.75-103, 2000.

Pattern Recognition approaches to Machine Translation

F. Casacuberta and E. Vidal

Pattern Recognition and Human Language Technology Group
Instituto Tecnológico de Informática
Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de Valencia, Spain

Finite-State Translation Models based on Alignments

Enrique Vidal

`evidal@iti.upv.es`

January 2005

E. Vidal – ITI-UPV-DSIC

Pattern Recognition Machine Translation

Techniques based on Alignments

Index

- 1 Statistical Alignment Models and Finite-State Transducers ▷ [2](#)
- 2 Alignment-controlled state merging: OMEGA ▷ [5](#)
- 3 Alignments and bilingual segmentation: GIATI ▷ [12](#)
- 4 GIATI revisited: pure statistical learning ▷ [24](#)
- 5 Bibliography ▷ [28](#)

Index

- 1 *Statistical Alignment Models and Finite-State Transducers* ▷ 2
- 2 Alignment-controlled state merging: OMEGA ▷ 5
- 3 Alignments and bilingual segmentation: GIATI ▷ 12
- 4 GIATI revisited: pure statistical learning ▷ 24
- 5 Bibliography ▷ 28

Statistical alignments and finite-state models

- Finite state transducer learning techniques seem to require large amounts of training data to produce acceptable results
- Some byproducts of statistical alignment model training can be useful to improve the learning capabilities of finite state methods:
 - *Sentence-to-sentence word alignments*
 - *Word-to-word mappings (statistical dictionaries)*

[Brown et al. *Computational Linguistics*, 1990] : Decomposing $\Pr(x \mid y)$ using bilingual word-position mappings or “alignments” as hidden variables:

$$\Pr(x \mid y) = \sum_{a \in \mathcal{A}(y, x)} \Pr(x, a \mid y)$$

where, $\Pr(x, a \mid y)$ is mainly modeled by means of *position alignment probabilities*, e.g.: $\Pr(i \mid j, I, J)$, and a *statistical dictionary*: $\Pr(x_j \mid y_i)$

Statistical alignment models

- **Alignments:** $a \subseteq \{1, \dots, I\} \times \{1, \dots, J\}$, $I = |x|$, $J = |y|$
- **Restriction:** $a : \{1, \dots, J\} \rightarrow \{0, \dots, I\}$,

where $a_j = 0$ states that the j -th. position in y is not aligned with any position in x

Example:

	1	2	3	4	5	6	
	per	favore	vorrei	una	camera	doppia	
I (0)	would (3)	like (3)	a (4)	double (6)	room (5)	please (2)	
$a_1 = 0$	$a_2 = 3$	$a_3 = 3$	$a_4 = 4$	$a_5 = 6$	$a_6 = 5$	$a_7 = 2$	

per	favore	vorrei	una	camera	doppia	
I	would	like	a	double	room	please

Index

- 1 Statistical Alignment Models and Finite-State Transducers ▷ 2
- 2 *Alignment-controlled state merging: OMEGA* ▷ 5
- 3 Alignments and bilingual segmentation: GIATI ▷ 12
- 4 GIATI revisited: pure statistical learning ▷ 24
- 5 Bibliography ▷ 28

Review of OSTIA State-Merging Learning Procedures

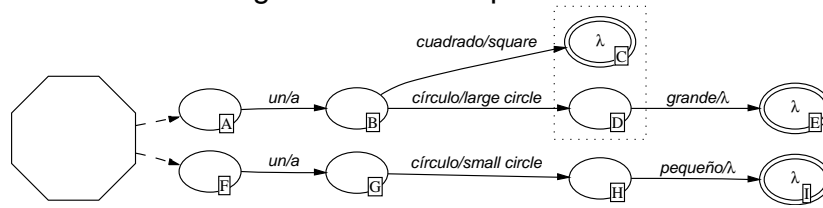
- Build an “*onward*” tree representation of the training data (a tree in which output strings are as close as possible to the root)
- The traversal of the tree goes in a level by level manner, typically by using a lexicographical order of state names.
- Two kinds of State Merging:
 - Merging based on *Local Conditions*: involve only the two states under consideration. ***Different Local Conditions lead to different algorithms.***
 - *Derived merges*: once two states are merged, others may also need to be recursively merged (with the help of possible output substring “*Pushing-back*”) in order to *preserve determinism*.
- If a cascade of derived merges *fails* preserving determinism, the original and all the derived *merges are discarded*

Local Conditions for State Merging

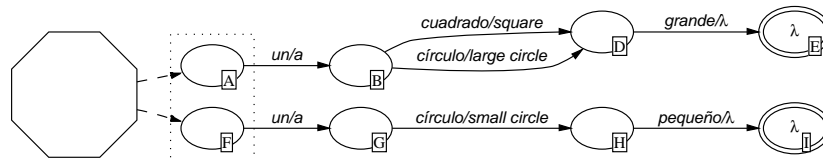
- OSTIA: only considers the output of the states: if both outputs are the same or at least one has no output, the join is possible [*Oncina, 91-93*].
- OSTIA-DR: also takes into account two *Language Models* (LM), one for the Input (or Domain) and one for the Output (or Range): two states cannot be joined if they correspond to different states of the Input or Output LMs [*Oncina, 94-96*].
- ***OMEGA [Vilar, 98]***: also takes into account *alignments* and word to word *dictionaries*.

The Problem of Premature Output

Assume the following situation in the process of OSTIA learning:



OSTIA would join states C and D, yielding:

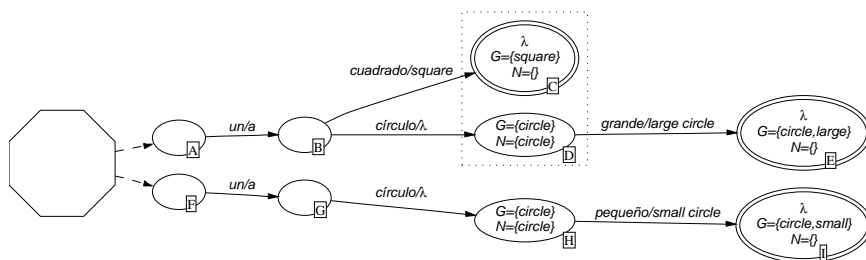


This entails a bad generalisation (un cuadrado grande, a square), and moreover now A and F could not be joined. This problem can be solved with a new extension to OSTIA called “**OMEGA**”¹

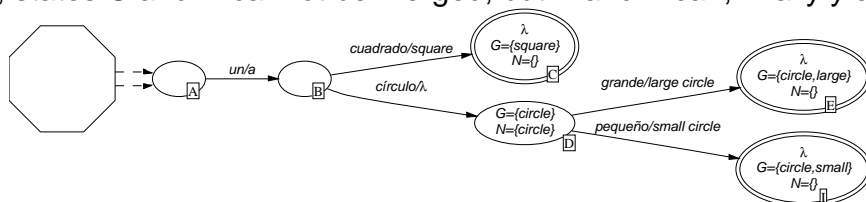
¹For the Spanish “**OSTIA Modificado Empleando Garantías y Alineamientos**” [Vilar,98].

State Labeling with the help of a Dictionary and/or Alignment

Suposse that a known dicctionary or alignment hints that the translation of *grande*, *pequeño* and *círculo* should be *large*, *small* and *circle*, respectively. This can be used for state labelling as follows:



Now, states C and D cannot be merged, but A and F can, finally yielding:



The OMEGA extension to OSTIA

[Vilar, 1998]

- The initial tree is built taking alignments and/or dictionaries into account to avoid premature output. Each state p is labelled with two sets:
 - $G(p)$ representing those words which are “*guaranteed*”, i.e., they will appear in the output of any path passing through p .
 - $N(p)$ representing those words that “*need*” to be seen, i.e., those which have not appeared so far, but which should appear in the translation of at least one of the paths departing from p .
- Local compatibility rules of OSTIA-DR now further include avoiding the join of two states p and q if $N(p) \cup N(q) \not\subseteq G(p) \cap G(q)$.
- N and G can be derived from (probabilistic) dictionaries and/or alignments.
- Input-Output Syntactic Constraints can be applied as in the original version of OSTIA(-DR).

OMEGA Learning Results

(Spanish-English experiments; similar for Spanish-German [Vilar,98])

- **Data:** A subset of Spanish-English EuTrans-I Traveler Task Data
 - Created by selecting those sentences with *at most ten words*
 - Test-Set: 588 different sentences, disjoint with training data.
- **Training:** OMEGA versus OSTIA-DR
 - *Bigram* Input and Output Syntactic Constraints. **No Categorization**.
 - Alignments obtained using the **MAR** statistical model.
- **Search:** Error Correcting parsing.

Different Training Pairs	OSTIA-DR	OMEGA-DR
1,000	27,28	16,51
2,000	19,64	11,17
4,000	11,88	8,33
8,000	8,31	5,57
16,000	5,19	4,16

- ▷ **Training data demands can be reduced by a factor of 2.**
- ▷ **Results improve using Bilingual Categorization.**

Index

- 1 Statistical Alignment Models and Finite-State Transducers ▷ 2
- 2 Alignment-controlled state merging: OMEGA ▷ 5
- 3 *Alignments and bilingual segmentation: GIATI* ▷ 12
- 4 GIATI revisited: pure statistical learning ▷ 24
- 5 Bibliography ▷ 28

Regular Grammars and finite state transducers: a morphism theorem

Theorem [Berstel 1979]:

$T \subseteq X^* \times Y^*$ is a rational translation if and only if there exist an alphabet Z , a regular language $L \subset Z^*$ and two morphisms $h_X : Z^* \rightarrow X^*$ and $h_Y : Z^* \rightarrow Y^*$ such that $T = \{(h_X(w), h_Y(w)) \mid w \in L\}$

This theorem has suggested the development of a number of transducer learning techniques, including GIATI [Casacuberta, ICGI-2000]

Explicit use of statistical alignments for FST learning: GIATI

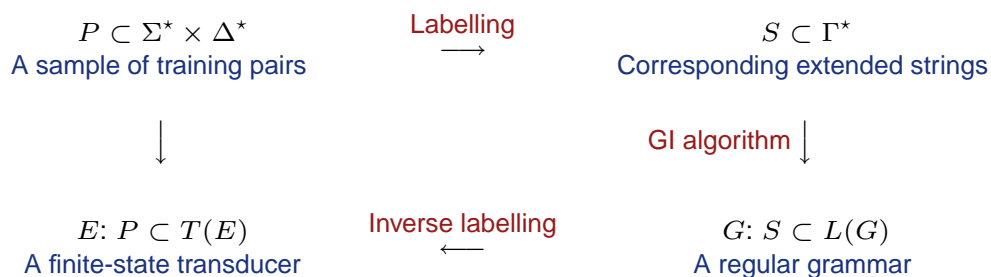
General idea in three steps:

1. Use sentence-to-sentence word alignments to convert each training *pair* (x, y) of input/output sentences from $X^* \times Y^*$ into a *single training string* z over an alphabet of “*extended symbols*” Z (composed of pairs of input/output symbols/strings)
2. Use an adequate grammar learning technique (e.g., N-Grams) to obtain a finite state “*language model*” for these strings
3. Using the adequate *morphisms*, convert back each *extended symbol* of this model into a pair of input/output symbols/strings. This effectively transforms the *language model* into a *finite state transducer*

This general method is referred to as

Grammatical Inference and Alignments for Transducer Inference (GIATI)

GIATI: general training procedure



LEARNING APPROACH:

1. Build a labelled corpus (extended symbols) using statistical alignments.
2. Infer a (stochastic) regular grammars using the labelled corpus.
3. Transform the extended symbols of transitions into input/output symbols.

GIATI: First step (Example)

USING STATISTICAL ALIGNMENTS TO CONVERT TRAINING PAIRS INTO TRAINING STRINGS

Training pairs:

una camera doppia	→	a double room
una camera	→	a room
la camera singola	→	the single room
la camera	→	the room

Aligned sentences:

una camera doppia a (1) double (3) room (2)	una camera a (1) room (2)	la camera singola the (1) single (3) room (2)	la camera the (1) room (2)
una camera doppia a double room	una camera a room	la camera singola the single room	la camera the room

GIATI: First step: the labelling procedure

Let x, y and a be an input string, an output string and an alignment function, respectively, z is the labelled string with $|z| = |x|$ and:

For $1 \leq i \leq |z|$

$$z_i = \begin{cases} x_i + y_j + y_{j+1} + \dots + y_{j+l} & \text{if } \exists j : a(j) = i \text{ and } \neg \exists j' < j : a(j') > a(j) \\ & \text{and for } j'' : j \leq j'' \leq j+l, a(j'') \leq a(j) \\ x_i & \text{otherwise} \end{cases}$$

Aligned training pairs:

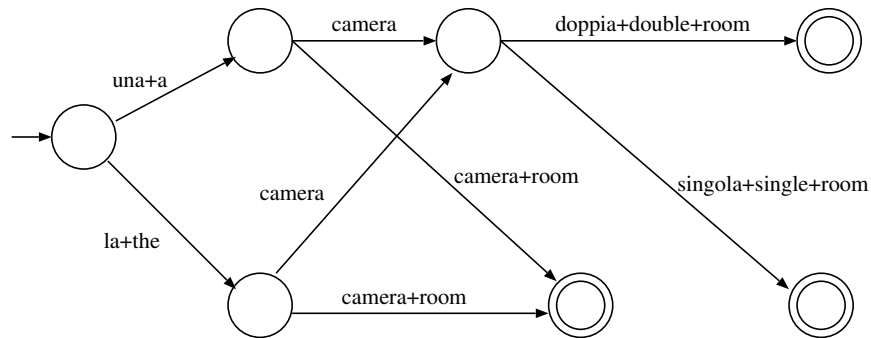
una camera doppia	a (1) double (3) room (2)	⇒	una+a camera doppia+double+room
una camera	a (1) room (2)	⇒	una+a camera+room
la camera singola	the (1) single (3) room (2)	⇒	la+the camera singola+single+room
la camera	the (1) room (2)	⇒	la+the camera+room

Training strings:

GIATI: Second step

FROM TRAINING STRINGS TO GRAMMARS: N-GRAMS

$$\Pr(z) \approx \prod_{i=1}^{|z|} \Pr(z_i | z_{i-n+1}, \dots, z_{i-1})$$



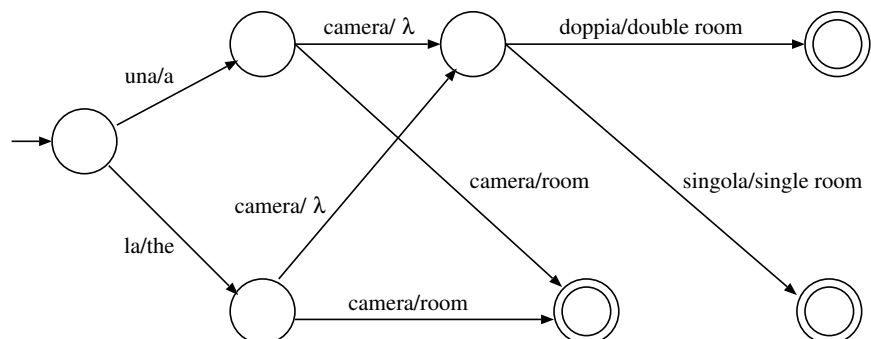
PROBLEM: Non-seen events in the training set.

COMMON SOLUTION: Smoothing.

GIATI: Third step

FROM GRAMMARS TO TRANSDUCERS: INVERSE LABELLING

GRAMMAR	TRANSDUCER
$(q, a + b_1 + b_2 + \dots + b_k, q')$	$(q, a, b_1 b_2 \dots b_k, q')$



GIATI results

With IBM Model 5 alignments and back-off smoothed n -grams, for the standard corpus EUTRANS-0
(171,481 different training pairs, Vocabularies: 689/514 words)

n	states	transitions	WER (%)	SER (%)
2	4,056	67,235	8.8	50.1
3	33,619	173,500	4.7	27.2
4	110,321	364,373	4.1	23.2
5	147,790	492,840	3.8	20.5
6	201,319	663,447	3.6	19.0
7	264,868	857,275	3.4	18.0
8	331,598	1,050,949	3.3	17.4
9	391,812	1,218,367	3.3	17.2
10	438,802	1,345,278	3.2	16.8
11	471,733	1,432,027	3.1	16.4
12	492,620	1,485,370	3.1	16.4

Comparative experiments: benchmark corpora

EUTRANS-I CORPUS [VIDAL 1997]

	Spanish	English
Train: Sentences	10,000	
Words	97,131	99,292
Vocabulary	686	513
Test: Sentences	2,996	
Words	35,023	35,590
Bigram Perplexity	8.6	5.2

Semiautomatically generated Spanish-English sentences, human-to-human communication at a reception desk of a hotel.

EUTRANS-II CORPUS (ITI 2000)

	Italian	English
Train: Sentences	3,038	
Words	55,302	64,176
Vocabulary	2,459	1,712
Test: Sentences	300	
Words	6,121	7,243
Bigram Perplexity	31	25

Transcriptions of Italian-English spontaneous sentences, person-to-person communication in the hotel framework.

OSTIA / OMEGA / GIATI comparative results

[EUTRANS Final Report, 2000], [EUTRANS Deliv.D2.1a , 2000], [Casacuberta, 2002]

Corpus	Method	Assited by	n-grams	WER
EUTRANS-I	OSTIA	ECP	2	8.3
EUTRANS-I	OMEGA	ECP, IBM2'	2	6.6
EUTRANS-I	GIATI	BOS, IBM5	5	6.6
EUTRANS-II	OMEGA	ECP, IBM2	2	41.7
EUTRANS-II	OMEGA	ECP, IBM2, ABS	2	36.5
EUTRANS-II	GIATI	BOS, IBM5	2	28.1
EUTRANS-II	GIATI	BOS, IBM5, ABS	2	24.9

ECP = Error-Correcting Parsing

BOS = Back-Off Smoothing

ABS = Automatic Bilingual Segmentation

IBM_k = IBM Model *k* statistical alignments

IBM2' = Symetrized IBM2

Summary of Stochastic Finite-State MT results

Translation Word Error Rate (TWER %)

Task	MLA	EUTRANS-0	EUTRANS-I	EUTRANS-II	TT2-XRCE	AMETRA	TT2-UE
Languages	Sp-En	Sp-En	Sp-En	It-En	En-Sp	Sp-Ba	En-Sp
Vocabularies	30	689/514	689/514	2.5K/1.7K	26K/30K	719/1.3K	84K/97K
Training (words)	110K	4.5M	100K	50K	600K	90K	6M
Year	1993	1996	1998	1999	2004	2003	2004
OSTIA	3	≈1	-	-	-	-	-
OSTIA-DR	1	<1	10	>80	-	-	-
OMEGA	-	<1	4	37	-	-	-
GIATI	-	3	7	25	32	40	56
Best result	-	-	4	25	28	36	47
Non FS system	-	-	AT	AT	PB	PB	PB

Languages: **E**nglish, **S**panish, **I**talian, **B**asc

PB = Phrase-based alignment models

AT = Alignment Templates

Index

- 1 Statistical Alignment Models and Finite-State Transducers ▷ 2
- 2 Alignment-controlled state merging: OMEGA ▷ 5
- 3 Alignments and bilingual segmentation: GIATI ▷ 12
- 4 *GIATI revisited: pure statistical learning* ▷ 24
- 5 Bibliography ▷ 28

Pure statistical approach: GIATI revisited

Let $\Pr(x, y)$ be the joint probability of a pair of sentences (x, y)

- Let J and I be the *given* lengths of x and y , respectively.
- Assume that y is **segmented into J segments**,

$$\mu : \{1, \dots, J\} \rightarrow \{1, \dots, I\} \quad \text{with} \quad \mu_{j+1} > \mu_j \quad \text{for} \quad 1 \leq j < J \quad \text{and} \quad \mu_J = I$$

Further assumptions:

- The distributions that rule I , J and μ are uniform.
- The correspondence among source symbols and target segments is **monotone**.
- By using a n -grams approximation with an special “end” symbol \$.

$$\Pr(x, y) \propto \sum_K \sum_{\mu_1^K} \prod_{k=1}^J \Pr(x_k, y_{\mu_{k-1}+1}^{\mu_k} | x_{k-n+1}^{k-1}, y_{\mu_{k-n+1}}^{\mu_{k-1}}) \cdot \Pr(\$, \$ | x_{J-n+2}^J, y_{\mu_{J-n+2}}^{\mu_J})$$

Pure statistical approach: GIATI revisited

Features:

- Main feature: All possible segmentations of the training set are considered.
- Parameter estimation: E-M algorithm.
- A SFST implementation:
 - The states are all possible $(x_{k-n+1}^{k-1}, y_{\mu_{k-n+1}}^{\mu_{k-1}})$ in the training set;
 - The probability of a transition between two states $(x_{k-n+2}^k, y_{\mu_{k-n+2}}^{\mu_k})$ and $(x_{k-n+1}^{k-1}, y_{\mu_{k-n+1}}^{\mu_{k-1}})$ is $\Pr(x_k, y_{\mu_{k-1}+1}^{\mu_k} | x_{k-n+1}^{k-1}, y_{\mu_{k-n+1}}^{\mu_{k-1}})$ with x_k as source symbol and $y_{\mu_{k-n+1}}^{\mu_{k-1}}$ as the target string;
 - The probability that $(x_{k-n+1}^{k-1}, y_{\mu_{k-n+1}}^{\mu_{k-1}})$ of a final state is $\Pr(\$ | x_{k-n+1}^{k-1}, y_{\mu_{k-n+1}}^{\mu_{k-1}})$.
- Generalization to arbitrary segmentations of the source sentence.

Conclusions

- We have thoroughly explored the learning of FST and its applications in MT
- Other contributions in this area: [Knight & Al-Onaizan, 98], [Mäkinen, 99], [Bangalore, Ricardi et al., 01]
- As task complexity and/or data scarceness increases, it becomes more and more important to make use of methods borrowed from statistical language processing.

Particularly relevant: *statistical alignments* and *smoothing* techniques

- Making explicit use of these techniques, GIATI is among the most promising approaches for FST MT
- A new pure statistically based development of GIATI is under way

Index

- 1 Statistical Alignment Models and Finite-State Transducers ▷ 2
- 2 Alignment-controlled state merging: OMEGA ▷ 5
- 3 Alignments and bilingual segmentation: GIATI ▷ 12
- 4 GIATI revisited: pure statistical learning ▷ 24
- 5 *Bibliography* ▷ 28

Bibliography

- J.M.Vilar: Improve the learning of subsequential transducers by using alignments and dictionaries. In “Grammatical Inference: Algorithms and Applications”, vol.1891 of *Lecture Notes in Artificial Intelligence*, pp.298–312. Springer-Verlag, 2000.
- F. Casacuberta: Inference of finite-state transducers by using regular grammars and morphisms. In “Grammatical Inference: Algorithms and Applications”, vol.1891 of *Lecture Notes in Artificial Intelligence*, pages 1–14. Springer-Verlag, 2000.
- F.Casacuberta and E.Vidal. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(2):205-225, 2004.
- F.Casacuberta, E.Vidal, and D.Picó. Inference of finite-state transducers from regular languages. *Pattern Recognition*, In press, 2005.

Pattern Recognition Approaches to Machine Translation

E. Vidal and F. Casacuberta

Pattern Recognition and Human Language Technology Group

Departament de Sistemes Informàtics i Computació

Institut Tecnològic d'Informàtica

Universitat Politècnica de València

8: Recursive Alignment Models

Francisco Casacuberta Nolla

`fcn@iti.upv.es`

24-28 January 2005

F. Casacuberta – DSIC-ITI-UPV

[Pattern Recognition approaches to Machine Translation](#)

[Recursive Alignment Models](#)

Index

- 1 Introduction ▷ [2](#)
- 2 A recursive alignment model: MAR ▷ [11](#)
- 3 Stochastic inversion transduction grammar ▷ [30](#)
- 4 Bilingual Recursive Alignments ▷ [34](#)
- 5 Bibliography ▷ [47](#)

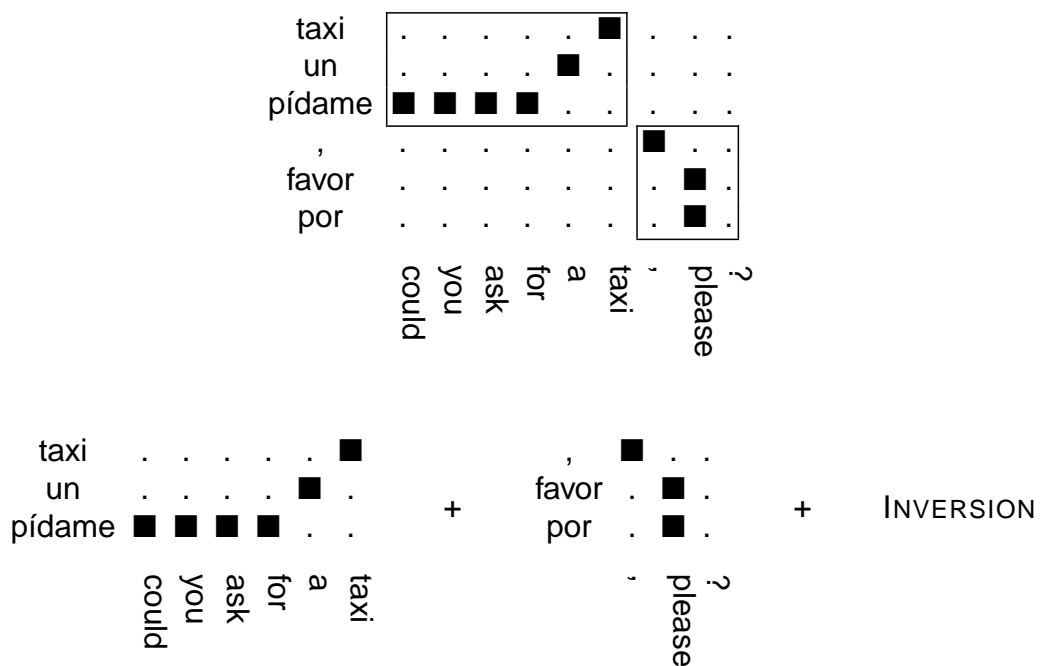
Index

- 1 *Introduction* ▷ 2
- 2 A recursive alignment model: MAR ▷ 11
- 3 Stochastic inversion transduction grammar ▷ 30
- 4 Bilingual Recursive Alignments ▷ 34
- 5 Bibliography ▷ 47

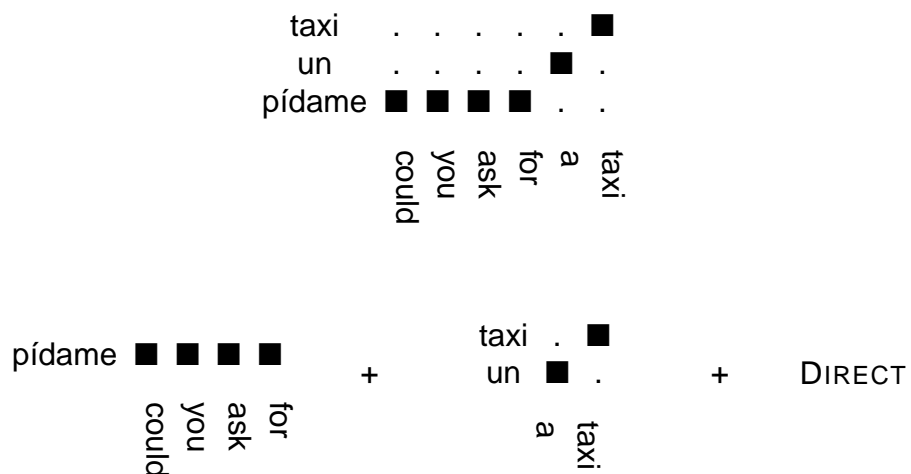
Exemple of word alignments

taxi	■	.	.	.
un	■
pídame	■	■	■	■
,	■	.	.
favor	■	.
por	■	.
	could	you	ask	for	a	taxi	,	please	?

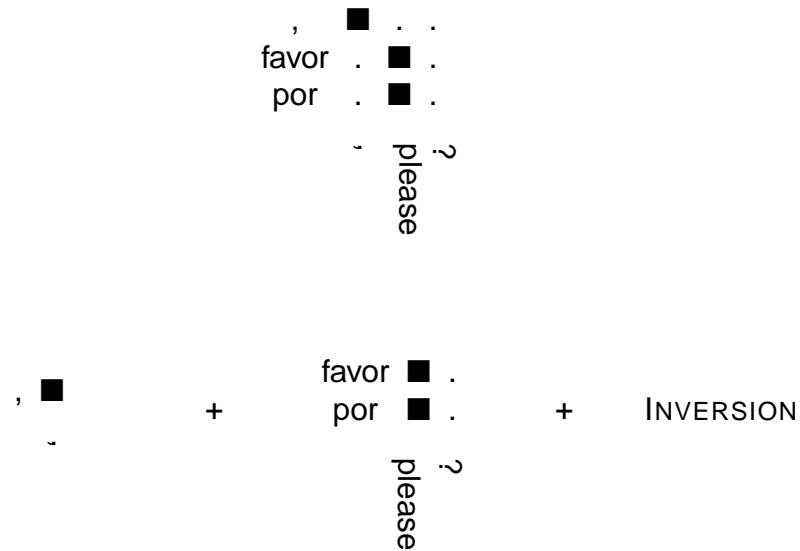
Example of word alignments



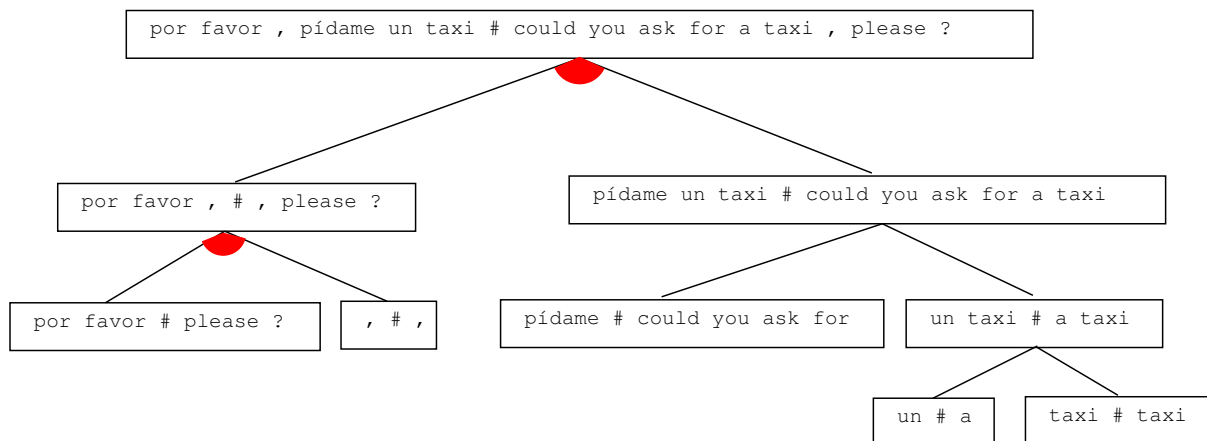
Example of word alignments



Exemple of word alignments

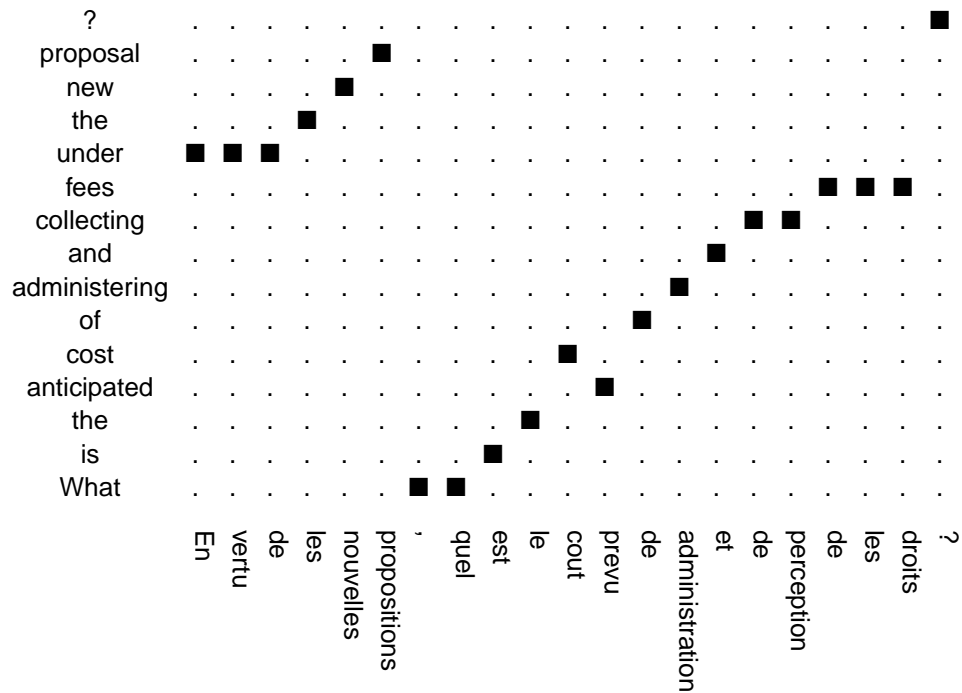


Exemple of word alignments



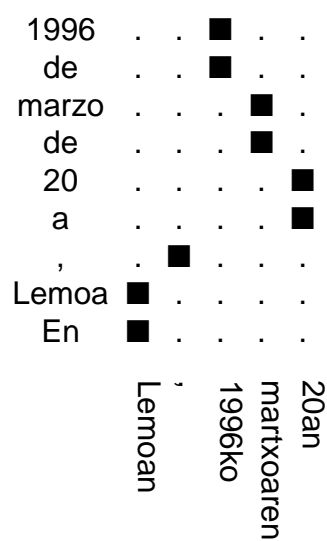
Exemple of word alignments

H. Ney, *Statistical Natural Language Processing*, 2003: Canadian Hansards



Example of word alignments

AMETRA corpus



Exemple of word alignments

METEO corpus

sud	■
meitat	■	.
seva	■	.	.
la
en	■	.	.	.
Llevant	.	.	.	■
de	.	.	■
des	.	.	■
sobretot	■	■
	sobre	todo	desde	Levante	en	su	mitad	sur

Index

- 1 Introduction ▷ 2
- 2 *A recursive alignment model: MAR* ▷ 11
- 3 Stochastic inversion transduction grammar ▷ 30
- 4 Bilingual Recursive Alignments ▷ 34
- 5 Bibliography ▷ 47

A Recursive Alignment Model: MAR

$$MAR = (\text{Recursive Alignment Model})^{-1}$$

J.M. Vilar: *Aprendizaje de transductores subsecuenciales para su empleo en tareas de dominio restringido*. PhD thesis, UPV. 1998.

*The slides on MAR are modified versions of some material supplied by J.M. Vilar.

A Recursive Alignment Model: MAR

- Accounts for differences in word order between languages.
- Assumes hierarchical structured alignments.
- The alignments obtained are particularly adequate to be used in combination with finite-state techniques:

Allow to use automatically obtained short phrases (rather than words) for:

- Bilingual clustering.
- Reordering of source-target pairs.

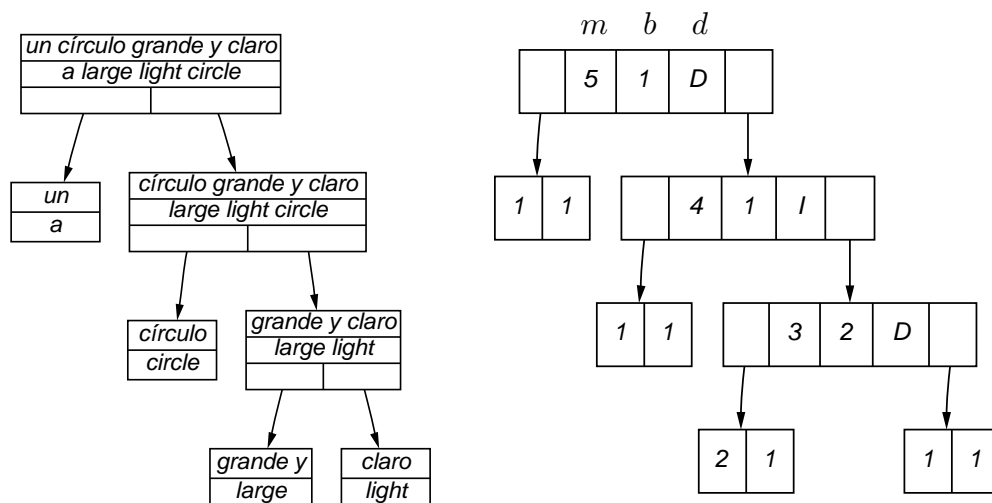
MAR's generative process

The translation of a sentence segment can be decomposed in:

1. Decide whether MAR or IBM has to be used. If IBM is chosen, the segment is translated by it and the process ends.
2. If MAR is used, the sentence is divided in two segments.
3. Each segment is recursively translated
(hence the name: (*Recursive Alignment Model*)⁻¹).
4. The resulting translations are concatenated in the original or in the inverse order.

A simple example

Leaves are labelled by $J = |x|$, $I = |y|$, while internal nodes are labelled by J, b, d : $1 \leq b \leq J = |x|$, $d \in \{D, I\}$, where b is the cut-point of x and D, I indicate that the source-target segments are “Directly” or “Inversely” aligned, respectively.



Formal derivation

First, approximate the translation probability by

$$\begin{aligned}\Pr(\mathbf{y}|\mathbf{x}) &\approx P_M(\mathbf{y}|\mathbf{x}) \\ &= \Pr(M = \text{IBM}|\mathbf{x}) \cdot P_{M1}(\mathbf{y}|\mathbf{x}) \\ &\quad + \Pr(M = \text{MAR}|\mathbf{x}) \cdot P_{MAR}(\mathbf{y}|\mathbf{x})\end{aligned}$$

$P_{M1}(\mathbf{y}|\mathbf{x})$ is given by IBM-1; $P_{MAR}(\mathbf{y}|\mathbf{x})$ can be written as:

$$\begin{aligned}P_{MAR}(\mathbf{y}|\mathbf{x}) &= \sum_{b=1}^{J-1} \Pr(b|\mathbf{x}) \\ &\quad \cdot \sum_{d \in \{D, R\}} \Pr(d|b, \mathbf{x}) \\ &\quad \cdot \sum_{c=1}^{I-1} \Pr(\mathbf{y}_1^c|b, d, \mathbf{x}) \Pr(\mathbf{y}_{c+1}^I|b, d, \mathbf{x}, \mathbf{y}_1^c)\end{aligned}$$

Simplifications

1. The choice of the model depends only on the length of the source sentence:

$$\Pr(M = \text{IBM}|\mathbf{x}) \approx \mathcal{M}_1(J) \quad \Pr(M = \text{MAR}|\mathbf{x}) \approx \mathcal{M}_M(J)$$

2. The place for the boundary depends only of the two words adjacent to it:

$$\Pr(b|\mathbf{x}) \approx \frac{\mathcal{B}(\mathbf{x}_b, \mathbf{x}_{b+1})}{\sum_{i=1}^{J-1} \mathcal{B}(\mathbf{x}_i, \mathbf{x}_{i+1})}$$

3. The direction of the concatenation depends on these two words:

$$\Pr(d = D|b, \mathbf{x}) \approx \mathcal{D}_D(\mathbf{x}_b, \mathbf{x}_{b+1}) \quad \Pr(d = R|b, \mathbf{x}) \approx \mathcal{D}_R(\mathbf{x}_b, \mathbf{x}_{b+1})$$

4. The translations of the two halves are independent:

$$\Pr(\mathbf{y}_1^c|b, d, \mathbf{x}) \approx \begin{cases} P_M(\mathbf{y}_1^c|\mathbf{x}_1^b) & \text{if } d = D \\ P_M(\mathbf{y}_1^c|\mathbf{x}_{b+1}^J) & \text{if } d = R \end{cases}$$

$$\Pr(\mathbf{y}_{c+1}^I|b, d, \mathbf{x}, \mathbf{y}_1^c) \approx \begin{cases} P_M(\mathbf{y}_{c+1}^I|\mathbf{x}_{b+1}^J) & \text{if } d = D \\ P_M(\mathbf{y}_{c+1}^I|\mathbf{x}_1^b) & \text{if } d = R \end{cases}$$

Final form of MAR

J.M.Vilar *Aprendizaje de transductores subsecuenciales*. PhD thesis, UPV. 1998.

$$\begin{aligned}
 \Pr(y | x) &\approx P_M(y|x) \\
 &= \mathcal{M}_1(J) \cdot P_{M1}(y|x) \\
 &\quad + \mathcal{M}_M(J) \sum_{b=1}^{J-1} \frac{\mathcal{B}(x_b, x_{b+1})}{\sum_{i=1}^{J-1} \mathcal{B}(x_i, x_{i+1})} \\
 &\quad \cdot \left(\mathcal{D}_D(x_b, x_{b+1}) \sum_{c=1}^{I-1} p_T(y_1^c | x_1^b) \cdot P_M(y_{c+1}^I | x_{b+1}^J) \right. \\
 &\quad \left. + \mathcal{D}_I(x_b, x_{b+1}) \sum_{c=1}^{I-1} p_T(y_{c+1}^I | x_1^b) \cdot P_M(y_1^c | x_{b+1}^J) \right)
 \end{aligned}$$

where $P_{M1}(y|x) = \frac{n(J | I)}{J^I} \prod_{j=1}^I \sum_{i=1}^J l(y_j | x_i)$ corresponds to IBM-1 model.

Parameter estimation

Maximum Likelihood criterion: Given a sample of example pairs, A , find model parameter values such that the likelihood of A is maximum. That is, find the maximum of:

$$\mathcal{L}_A = \prod_{(x,y) \in A} P_M(y|x)$$

This can be (locally optimally) solved through *Expectation Maximization*. Baum Eagon's inequality is used to estimate all the parameters, except for the β s which are reestimated using Gopalakrishnan's.

Parameter estimation: “polynomial” and “rational parameters”

Let p be a parameter such that \mathcal{L} is polynomial with p and let $\mathcal{F}(p)$ be all the other parameters related with p (i.e.: $\sum_{q \in \mathcal{F}(p)} q = 1$). A reestimated value of p , $T(p)$ ([Baum & Eagon, 1968]):

$$T(p) = \frac{p \frac{\partial \mathcal{L}}{\partial p}}{\sum_{q \in \mathcal{F}(p)} q \frac{\partial \mathcal{L}}{\partial q}} = \frac{p \sum_{(\mathbf{x}, \mathbf{y}) \in A} (P_M(\mathbf{y} | \mathbf{x}))^{-1} \frac{\partial P_M(\mathbf{y} | \mathbf{x})}{\partial p}}{\sum_{q \in \mathcal{F}(p)} q \sum_{(\mathbf{x}, \mathbf{y}) \in A} (P_M(\mathbf{y} | \mathbf{x}))^{-1} \frac{\partial P_M(\mathbf{y} | \mathbf{x})}{\partial q}}$$

Let p be a parameter such that \mathcal{L} is rational with p and let $\mathcal{F}(p)$ be all the other parameters related with p (i.e.: $\sum_{q \in \mathcal{F}(p)} q = 1$). A reestimated value of p , $T(p)$ ([Gopalakrishnan et al, 1991]):

$$T(p) = \frac{p \frac{\partial \mathcal{L}}{\partial p} + C}{\sum_{q \in \mathcal{F}(p)} q \frac{\partial \mathcal{L}}{\partial q} + C} = \frac{p \sum_{(\mathbf{x}, \mathbf{y}) \in A} (P_M(\mathbf{y} | \mathbf{x}))^{-1} \frac{\partial P_M(\mathbf{y} | \mathbf{x})}{\partial p} + C}{\sum_{q \in \mathcal{F}(p)} q \sum_{(\mathbf{x}, \mathbf{y}) \in A} (P_M(\mathbf{y} | \mathbf{x}))^{-1} \frac{\partial P_M(\mathbf{y} | \mathbf{x})}{\partial q} + \sum_{q \in \mathcal{F}(p)} C}$$

Derivative of MAR probabilities

- Let m be a fixed source sentence length. $\mathcal{M}_I(m)$ provides the probability of choosing the IBM-1 model, given m : $\partial P_M(\mathbf{y} | \mathbf{x}) / \partial \mathcal{M}_I(m)$
- Let m be a fixed source sentence length. $\mathcal{M}_M(m)$ provides the probability of choosing the MAR model, given m : $\partial P_M(\mathbf{y} | \mathbf{x}) / \partial \mathcal{M}_M(m)$
- Let x, x' be two fixed source words. $\mathcal{B}(x, x')$ accounts for the probability of placing a boundary point between x and x' : $\partial P_M(\mathbf{y} | \mathbf{x}) / \partial \mathcal{B}(x, x')$
- Let x, x' be two fixed source words. $\mathcal{D}_D(x, x')$ provides the probability of choosing a *Direct* alignment: $\partial p_T(\mathbf{y} | \mathbf{x}) / \partial \mathcal{D}_D(x, x')$
- Let x, x' be two fixed source words. $\mathcal{D}_R(x, x')$ provides the probability of choosing an *Inverse* alignment: $\partial p_T(\mathbf{y} | \mathbf{x}) / \partial \mathcal{D}_R(x, x')$
- Let m, n be fixed lengths of source/target sentences. $n(m | n)$ accounts for the length distribution of IBM-1 model: $\partial P_M(\mathbf{y} | \mathbf{x}) / \partial n(m | n)$
- Let x, y be fixed source/target words. $l(y | x)$ determines the probability that y be a *translation* of x : $\partial P_M(\mathbf{y} | \mathbf{x}) / \partial l(y | x)$

Parameter estimation: Details and simplifications

- 10 Expectation Maximization iterations with neutral initialization.
- The value of $\mathcal{M}_I(l)$ set to 0 for $l > 4$.
- The value of $n(l, m)$ is not estimated for l or m greater than four.
- The values of $l(x|y)$ are not estimated for pairs with $l(x|y) < 10^{-5}$.

The resulting estimation algorithm has polinomial time complexity, though it still is very computationally intensive.

Results: Training perplexity evolution

Training set perplexity computed as:

$$PP = \sqrt[m]{\prod_{(x,y) \in \mathcal{S}} (P_M(y|x))^{-1}}, \quad m = \sum_{(x,y) \in \mathcal{S}} I.$$

Iteration	English		German	
	1,000	32,000	1,000	30,000
0	506.05	625.79	558.24	703.76
1	23.10	24.52	23.61	25.88
2	8.78	8.33	10.61	10.31
3	5.07	4.90	6.26	6.19
4	3.98	3.89	4.80	4.85
5	3.45	3.40	4.12	4.16
6	3.19	3.22	3.74	3.82
7	3.08	3.09	3.51	3.58
8	3.01	3.06	3.38	3.45

Convergence is achieved after a moderate number of iterations.

Spanish-English MAR alignment

deseo reservar dos habitaciones para tres días.	I want to book two rooms for three days.
deseo reservar dos habitaciones	I want to book two rooms
deseo reservar	I want to book
deseo	I want
reservar	to book
dos habitaciones	two rooms
dos	two
habitaciones	rooms
para tres días.	for three days.
para tres	for three
para	for
tres	three
días.	days.
días	days
.	.
.	.

Spanish-English MAR alignment

deseo una habitación con televisión para esta noche.	I want a room with a tv for tonight.
deseo una habitación con	I want a room with
deseo una	I want a
deseo	I want
una	a
habitación con	room with
habitación	room
con	with
televisión para esta noche.	a tv for tonight.
televisión para	a tv for
televisión	a tv
para	for
esta noche.	tonight.
esta noche	tonight
.	.
.	.

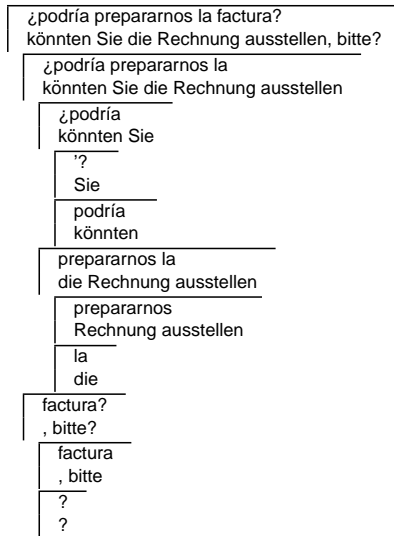
Spanish-English MAR alignment

¿podríamos pagar el recibo con cheques de viaje?	could we pay the bill by traveler check?
¿podríamos pagar el	could we pay the
¿podríamos	could we
'?	could
podríamos	we
pagar el	pay the
pagar	pay
el	the
recibo con cheques de viaje?	bill by traveler check?
recibo con cheques	bill by traveler check
recibo con	bill by
recibo	bill
con	by
cheques	traveler check
de viaje?	?

Spanish-German MAR alignment

¿nos llama a nuestro taxi, por favor?	würden Sie unser Taxi bestellen, bitte?
¿nos llama a nuestro taxi	würden Sie unser Taxi bestellen
¿nos	würden Sie
'?	Sie
nos	würden
llama a nuestro taxi	unser Taxi bestellen
llama a	bestellen
nuestro taxi	unser Taxi
nuestro	unser
taxi	Taxi
, por favor?	, bitte?
, bitte?	, por
, por	, favor?
favor?	bitte?
bitte?	favor
favor	bitte
?	?
?	?

Spanish-German MAR alignment

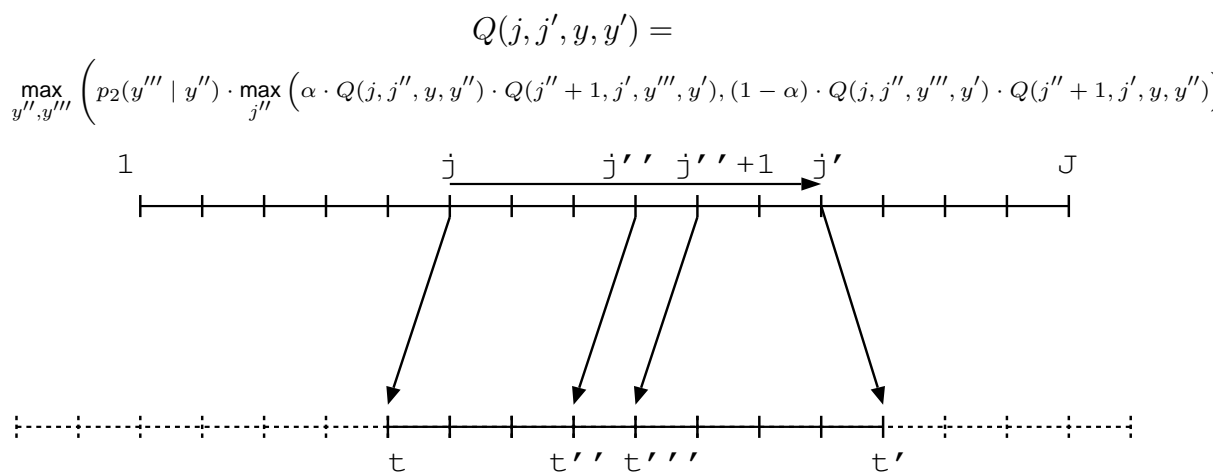


A simplified recursive alignment model

H. Ney, *Statistical Natural Language Processing*, STC Doctorate Program, UPC. 2003

$$\Pr(\mathbf{y}, \mathbf{x}) \approx \sum_{i,j} \left(\alpha \cdot \Pr(y_1^i, \mathbf{x}_1^j) \cdot \Pr(y_{i+1}^I, \mathbf{x}_{j+1}^J) + (1 - \alpha) \cdot \Pr(y_1^i, \mathbf{x}_{j+1}^J) \cdot \Pr(y_{i+1}^I, \mathbf{x}_1^j) \right)$$

Searching using a bigram target language model and a maximum approach:



Index

- 1 Introduction ▷ 2
- 2 A recursive alignment model: MAR ▷ 11
- 3 *Stochastic inversion transduction grammar* ▷ 30
- 4 Bilingual Recursive Alignments ▷ 34
- 5 Bibliography ▷ 47

Stochastic inversion transduction grammars

D. Wu: *Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora*. Comp. Ling. 1997.

A context-free based approach to bilingual segmentation

For a non-terminal symbol A, B and C and for any source word s and any target word t ,

$$A \rightarrow \langle B, C \rangle$$

$$A \rightarrow [B, C]$$

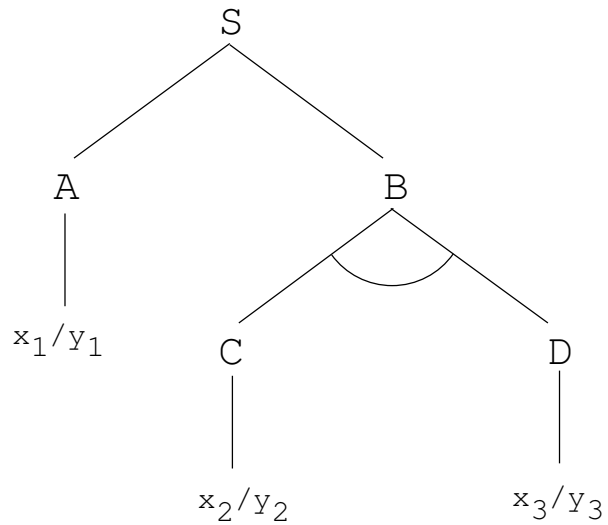
$$A \rightarrow x/y$$

$$A \rightarrow x/\lambda$$

$$A \rightarrow \lambda/y$$

An example

$S \rightarrow [A, B]$
 $A \rightarrow x_1/y_1$
 $B \rightarrow \langle C, D \rangle$
 $C \rightarrow x_2/y_2$
 $D \rightarrow x_3/y_3$



$x_1 x_2 x_3 \Rightarrow y_1 y_3 y_2$

Stochastic inversion transduction grammars

- Learning:
 - Adapted context-free grammatical inference
 - Inside-outside estimation
- Translation:
 - Adapted Cooker-Younger-Kasami parser algorithm
 - Inside or outside algorithms

Index

- 1 Introduction ▷ 2
- 2 A recursive alignment model: MAR ▷ 11
- 3 Stochastic inversion transduction grammar ▷ 30
- 4 *Bilingual Recursive Alignments* ▷ 34
- 5 Bibliography ▷ 47

Recursive Bilingual Alignments

An alignment between phrases of a source sentence and phrases of a target sentence.

- It represents the translation relations between two sentences.
- It also includes information about the possible reorderings needed in order to generate the target sentence from the source sentence.
- Representation: binary tree.
 - The inner nodes store the reordering directions.
 - The leaf nodes store the translation relations.

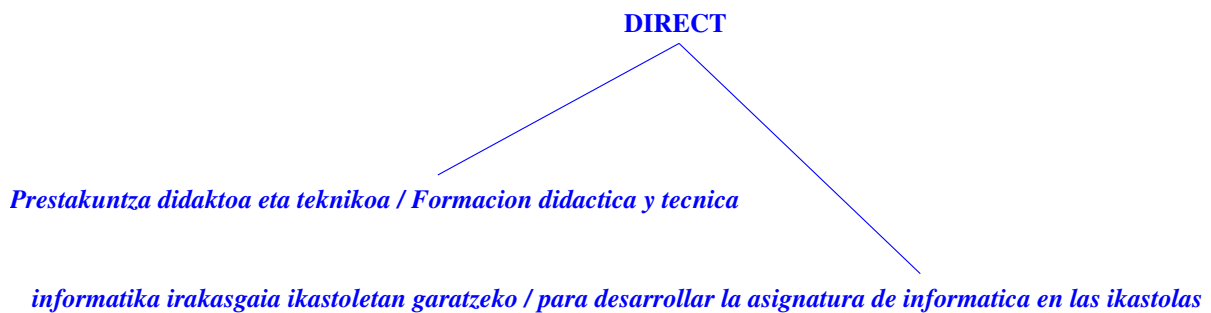
*The slides on RBA are modified versions of some material supplied by F. Nevado.

An example

[illegible]

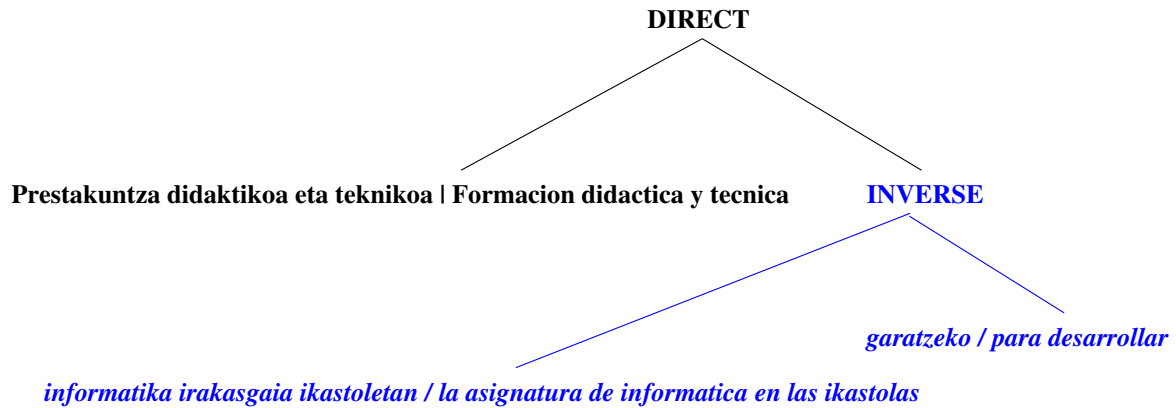
An example

Prestakuntza didaktoa eta teknikoa informatika irakasgaia ikastoletan garatzeko
Formación didáctica y técnica para desarrollar la asignatura de informática en las ikastolas



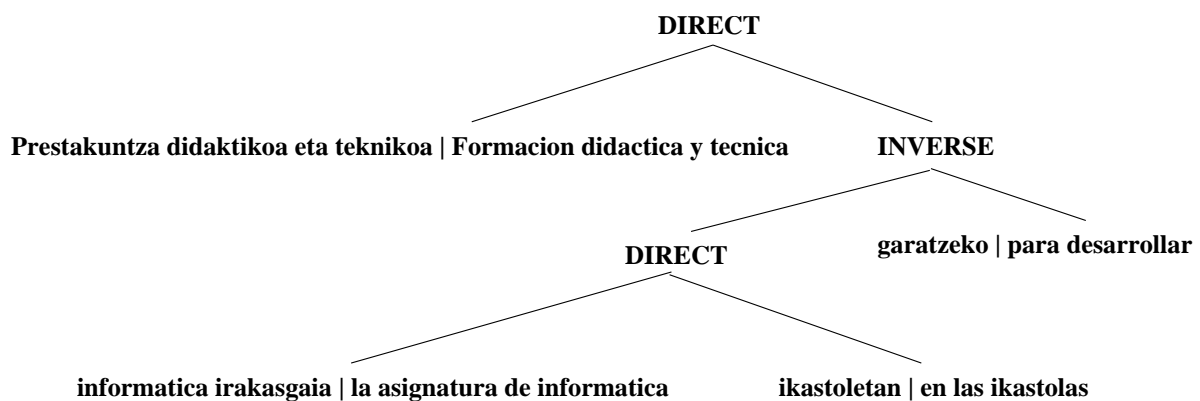
An example

Prestakuntza didaktoa eta teknikoa informatika irakasgaia ikastoletan garatzeko
Formación didáctica y técnica para desarrollar la asignatura de informática en las ikastolas



An example

Prestakuntza didaktoa eta teknikoa informatika irakasgaia ikastoletan garatzeko
Formación didáctica y técnica para desarrollar la asignatura de informática en las ikastolas



Greedy bilingual recursive alignment

$$\Pr(\mathbf{y}, \mathbf{x}) \approx \max_{i,j} \left(\alpha \cdot \Pr(\mathbf{y}_1^i, \mathbf{x}_1^j) \cdot \Pr(\mathbf{y}_{i+1}^I, \mathbf{x}_{j+1}^J) + (1 - \alpha) \cdot \Pr(\mathbf{y}_1^i, \mathbf{x}_{j+1}^J) \cdot \Pr(\mathbf{y}_{i+1}^I, \mathbf{x}_1^j) \right)$$

- $\alpha = 0.5$
- $\Pr(\mathbf{y}_i^{i'}, \mathbf{x}_j^{j'}) \approx P_{M1}(\mathbf{y}_i^{i'}, \mathbf{x}_j^{j'})$

$$(\hat{i}, \hat{j}) = \operatorname{argmax}_{i,j} \left\{ \max \left(P_{M1}(\mathbf{y}_1^i, \mathbf{x}_1^j) \cdot P_{M1}(\mathbf{y}_{i+1}^I, \mathbf{x}_{j+1}^J), P_{M1}(\mathbf{y}_1^i, \mathbf{x}_{j+1}^J) \cdot P_{M1}(\mathbf{y}_{i+1}^I, \mathbf{x}_1^j) \right) \right\}$$

Recalign algorithm

A greedy algorithm to compute recursive alignments from a bilingual corpus.

Probability of translating a source phrase into a target phrase \longrightarrow Model 1.

Algorithm:

1. Given \mathbf{x} and \mathbf{y} , it computes the most probable breakpoint in each sentence using Model 1.
2. If the translation probability for \mathbf{x} and \mathbf{y} is higher than the translation probability of dividing them:

- It creates a leaf node where the output sequence is considered to be the translation of the input sequence.

Else:

- It creates a new inner node of the tree.
- Apply recursively the algorithm to the left and the right children.

Recalign: variants

- To control the medium length of the generated segments

⇒ Combine the translation probabilities with a distribution over the sequences length: **LEN** modification.

- Model 1 can obtain imprecise divisions

⇒ Only allow divisions that are compatible with a word alignment: **ALI** restriction.

- Source-to-target (Target-to-source).
- Symmetrization: union, intersection, refined.

Corpus description

EUTRANS-I English-Spanish

Training:

	English	Spanish
Sentences	12,960	
Words	134,435	131,707
Vocabulary	514	688

Test:

	English	Spanish
Sentences	40	
Words	487	491
Vocabulary	126	149
Trigram Perplexity	3.6	4.6

DFB Basque-Spanish

Training:

	Basque	Spanish
Sentences	284,842	
Words	4,203,117	5,661,564
Vocabulary	144,670	62,412

Test:

	Basque	Spanish
Sentences	20	
Words	481	609
Vocabulary	342	311
Trigram Perplexity	776.3	135.4

Assessment

Given a segmentation S produced by a system and given a reference segmentation S_r produced by an expert,

- *Recall*: Number of bilingual segments that are correct with respect to the number of references:

$$Recall = \frac{S \cup S_r}{S_r}$$

- *Precision*: Number of bilingual segments that are correct with respect to the number of bilingual segments supplied by the system:

$$Precision = \frac{S \cup S_r}{S}$$

- *F-measure*:

$$F - measure = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$$

Results: Eutrans-I (Spanish-to-English)

Bilingual segmentation	Recall	Precision	F-measure
GIATI-labelling	39.22	87.96	54.25
<i>Realign</i>	37.23	87.10	52.16
<i>Realign</i> + LEN	76.86	72.00	74.35
<i>Realign</i> + ALI(S-E)	40.01	87.12	54.84
<i>Realign</i> + ALI(S-E) + LEN	77.38	71.78	74.48
<i>Realign</i> + ALI(\cup)	52.21	82.41	63.92
<i>Realign</i> + ALI(\cup) + LEN	81.63	67.52	73.91
<i>Realign</i> + ALI(\cap)	38.20	86.57	53.01
<i>Realign</i> + ALI(\cap) + LEN	76.86	72.00	74.35
<i>Realign</i> + ALI(ref)	49.14	84.28	62.08
<i>Realign</i> + ALI(ref) + LEN	81.17	68.37	74.22

Results: DFB (Spanish-to-Basque)

Bilingual segmentation	Recall	Precision	F-measure
GIATI-labelling	63.16	39.13	48.32
<i>Realign</i>	75.00	24.21	36.61
<i>Realign</i> + LEN	65.03	36.46	46.73
<i>Realign</i> + ALI(S-B)	78.26	24.08	36.83
<i>Realign</i> + ALI(S-B) + LEN	79.07	24.66	37.60
<i>Realign</i> + ALI(\cup)	92.16	14.69	25.34
<i>Realign</i> + ALI(\cup) + LEN	93.37	14.88	25.66
<i>Realign</i> + ALI(\cap)	76.77	24.29	36.91
<i>Realign</i> + ALI(\cap) + LEN	74.09	34.01	46.62
<i>Realign</i> + ALI(ref)	84.54	21.21	33.91
<i>Realign</i> + ALI(ref) + LEN	83.49	30.58	44.76

Index

- 1 Introduction ▷ 2
- 2 A recursive alignment model: MAR ▷ 11
- 3 Stochastic inversion transduction grammar ▷ 30
- 4 Bilingual Recursive Alignments ▷ 34
- 5 *Bibliography* ▷ 47

Bibliography

1. H. Ney: *Statistical Natural Language Processing*. STC Doctorate Program, UPC. 2003
2. F. Nevado, F. Casacuberta and E. Vidal: *Parallel corpora segmentation by using anchor words*. EACL 2003 workshop on EAMT, Budapest, Hungary, 2003.
3. F. Nevado, F. Casacuberta and J. Landa: *Translation memories enrichment by statistical bilingual segmentation*. IV International Conference on Language Resources and Evaluation - LREC2004, 335-338, Lisbon, 2004.
4. F. Nevado and F. Casacuberta: *Bilingual corpora segmentation using bilingual recursive alignments*. III Jornadas en Tecnologías del Habla, 3JTH, Valencia, 2004.
5. J.M. Vilar: *Aprendizaje de transductores subsecuenciales para su empleo en tareas de dominio restringido*. PhD thesis, UPV. 1998.
6. D. Wu: *A polynomial-time algorithm for statistical machine translation*. Annual Meeting of the ACL archive Proceedings of the 34th conference on Association for Computational Linguistics. Santa Cruz, California. 1996.
7. D. Wu: *Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora*. Computational Linguistics. 23(3): 377-403. 1997.
8. D. Wu and H. Wong: *Machine Translation with a Stochastic Grammatical Channel*. Proc. of the 17th international conference on Computational linguistics. Montreal, Quebec, Canada, 1998.

Pattern Recognition Approaches to Machine Translation

E. Vidal and F. Casacuberta

Pattern Recognition and Human Language Technology Group

Departament de Sistemes Informàtics i Computació

Institut Tecnològic d'Informàtica

Universitat Politècnica de València

9: Speech-to-Speech Translation

Francisco Casacuberta Nolla

`fcn@iti.upv.es`

24-28 January 2005

F. Casacuberta – DSIC-ITI-UPV

[Pattern Recognition approaches to Machine Translation](#)

[Speech-to-speech translation](#)

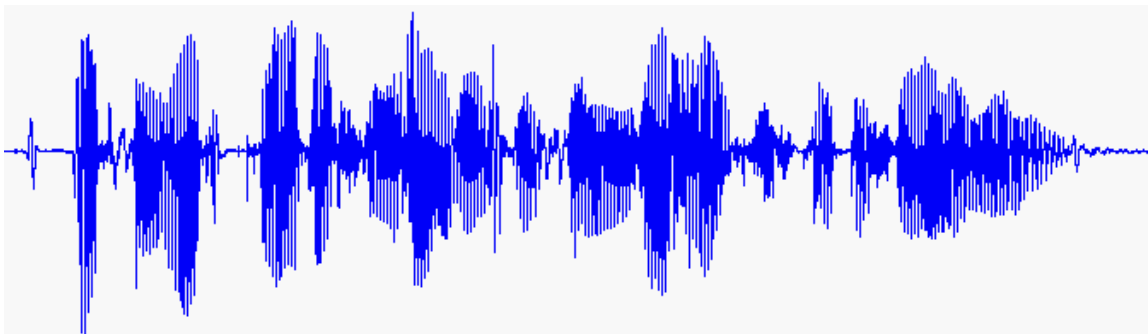
Index

- 1 Speech processing ▷ [2](#)
- 2 Automatic speech recognition ▷ [7](#)
- 3 Speech to speech translation ▷ [19](#)
- 4 Bibliography ▷ [33](#)

Index

- 1 *Speech processing* ▷ 2
- 2 Automatic speech recognition ▷ 7
- 3 Speech to speech translation ▷ 19
- 4 Bibliography ▷ 33

An utterance



/por favor, quiero reservar una habitación doble hasta pasado mañana/

Speech technologies

- Speech synthesis: From text to speech
- Speaker recognition/verification: From speech to the speaker identity.
- Dictation: From speech to text.
- Speech summarization: From speech to text.
- Speech categorization: From speech to simple semantic classes.
- Speech understanding: From speech to “semantic” information.
- Dialog processing: From speech to “semantic” information through complex interactions.
- Speech translation: From speech to speech.

Speech recognition, understanding and translation

SPEECH RECOGNITION :

por favor , quiero reservar una habitación doble hasta pasado mañana .

SPEECH UNDERSTANDING:

(ACTION=RESERVATION) (ROOM_TYPE=DOUBLE)
(DATE_OF_ENTRANCE=TODAY) (DATE_OF_LEAVING=TODAY+2)

SPEECH TRANSLATION:

I want to book a double room until the day after tomorrow, please.

Some speech characteristics

- There are not clear separation between two adjacent words
- The words can be uttered in different ways (also by the same speaker)
- Noise and distortion.
- A speech sentence can not be well formed (gramatically)

Index

- 1 Speech processing ▷ 2
- 2 *Automatic speech recognition* ▷ 7
- 3 Speech to speech translation ▷ 19
- 4 Bibliography ▷ 33

Statistical framework for speech recognition

Given an acoustic sequence \mathbf{v} , search for the sentence $\hat{\mathbf{x}}$:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmax}} \Pr(\mathbf{x} \mid \mathbf{v})$$

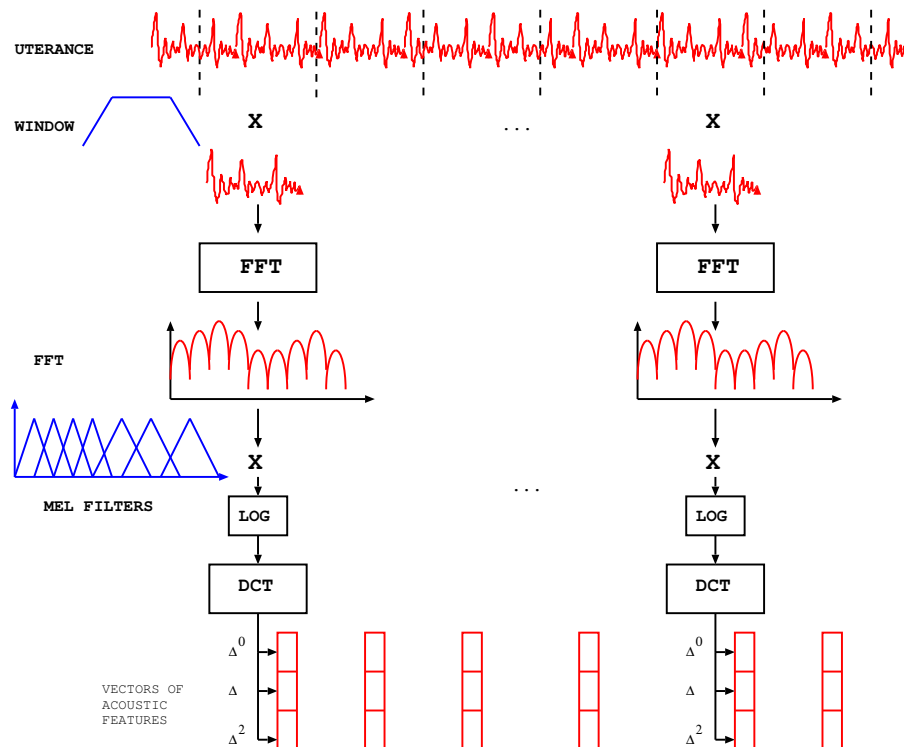
Using the Bayes' rule

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmax}} \Pr(\mathbf{x}) \cdot \Pr(\mathbf{v} \mid \mathbf{x})$$

STATISTICAL MODELS FOR SPEECH RECOGNITION

- $\Pr(\mathbf{v} \mid \mathbf{x})$: **Acoustic models** (HIDDEN MARKOV MODELS)
- $\Pr(\mathbf{x})$: **Language model** (N-GRAMS or STOCHASTIC GRAMMARS)

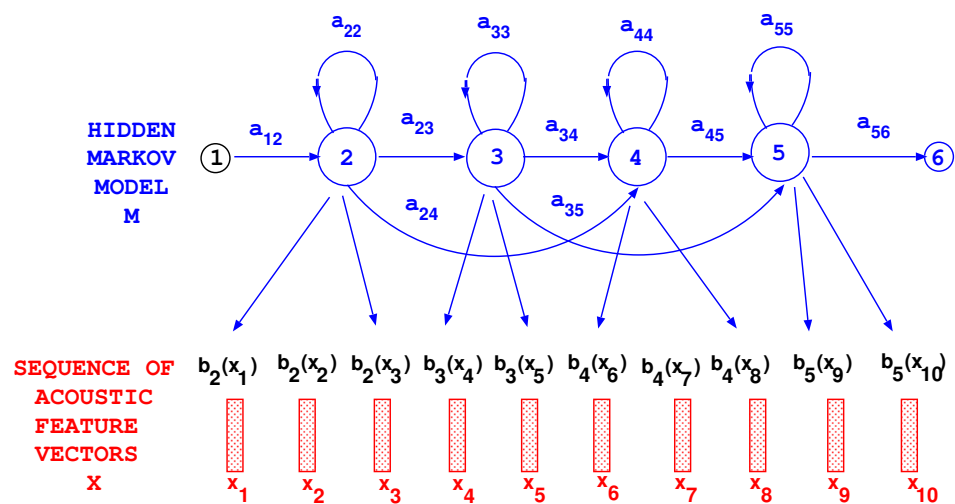
Speech preprocessing



Acoustic units

- **Words:**
 - Include contextual information (coarticulation).
 - Too many units \Rightarrow difficult training.
- **Phoneme:**
 - Context dependent (allophons).
 - Few units \Rightarrow easy training.
- **Compromise:**
 - Adequate number of units.
 - With some coarticulation information.
 - Proposals: syllables, semi-syllables, diphones, contextual phones, ...

Hidden Markov models (HMM)

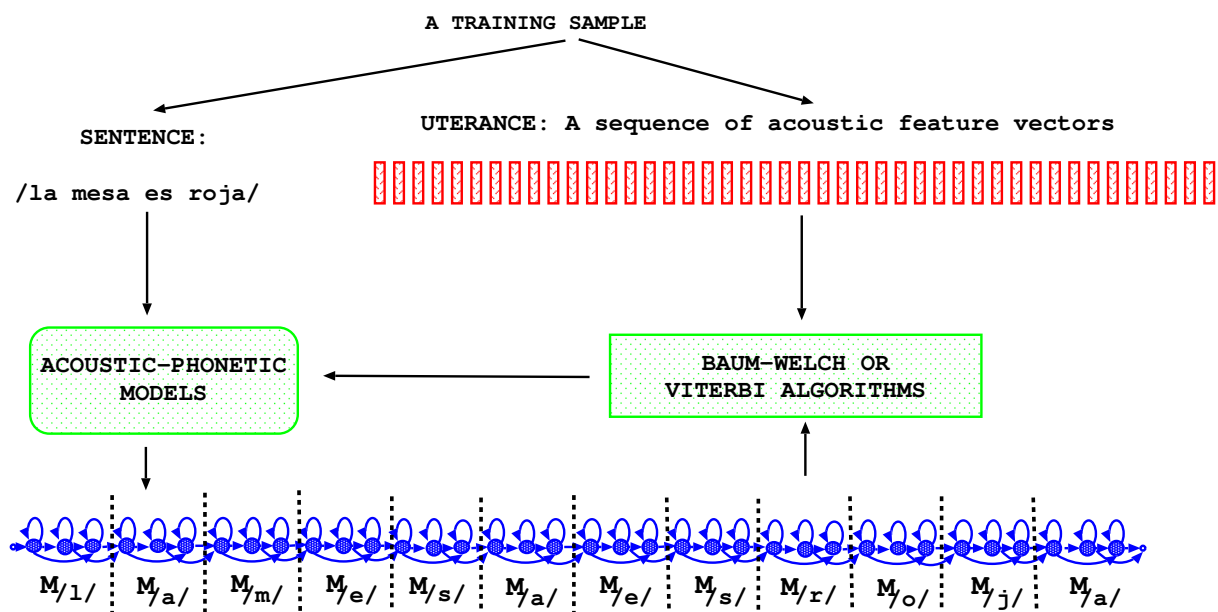


$$a_{1,2} \cdot b_2(x_1) \cdot a_{2,2} \cdot b_2(x_2) \cdot a_{2,2} \cdot b_2(x_3) \cdot a_{2,3} \cdot b_3(x_4) \cdot a_{3,3} \cdot b_3(x_5) \cdot a_{3,4} \cdot b_4(x_6) \cdot a_{4,4} \cdot b_4(x_7) \cdot a_{4,4} \cdot b_4(x_8) \cdot a_{4,5} \cdot b_5(x_9) \cdot a_{5,5} \cdot b_5(x_{10}) \cdot a_{5,6}$$

Hidden Markov models (HMM)

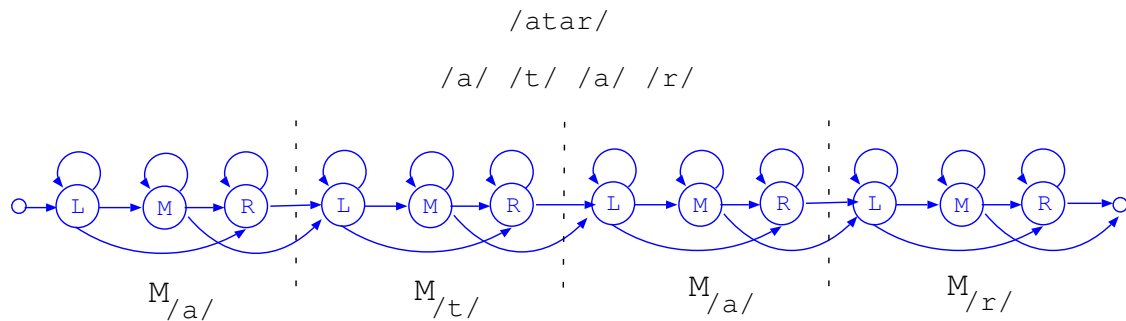
- **Components of a HMM** $\mathcal{M} = \langle Q, E, a, \pi, b \rangle$
 - **Topology:** Q : set of states. $E (= \mathbb{R}^d)$: space of acoustic features.
 - **Probabilistic distributions:**
 - * between states ($a : Q \times Q \rightarrow [0, 1]$),
 - * initial state ($\pi : Q \rightarrow [0, 1]$)
 - * emission (density) ($b : Q \times E \rightarrow [0, 1]$).
- **Decoding algorithms:** Forward and Backward.
- **An approximation:** Viterbi (+ Beam Search + Histogram Pruning).
- **Training algorithms:**
 - Maximum likelihood Baum-Welch, Viterbi.
 - Other criteria: Maximum mutual information, minimum discriminative information, discriminative.

Training hidden Markov models



Word acoustic models

Concatenation of phone units.



Language models

$$\Pr(\mathbf{y}) = \prod_{i=1}^I \Pr(y_i \mid \mathbf{y}_1^{i-1})$$

- **Stochastic grammars** $G = (N, \Sigma, R, S, p)$.

$$\Pr(\mathbf{y}) \approx P_G(\mathbf{y}) = \sum_{d(\mathbf{y})} P_G(d(\mathbf{y})) \approx \max_{d(\mathbf{y})} P_G(d(\mathbf{y}))$$

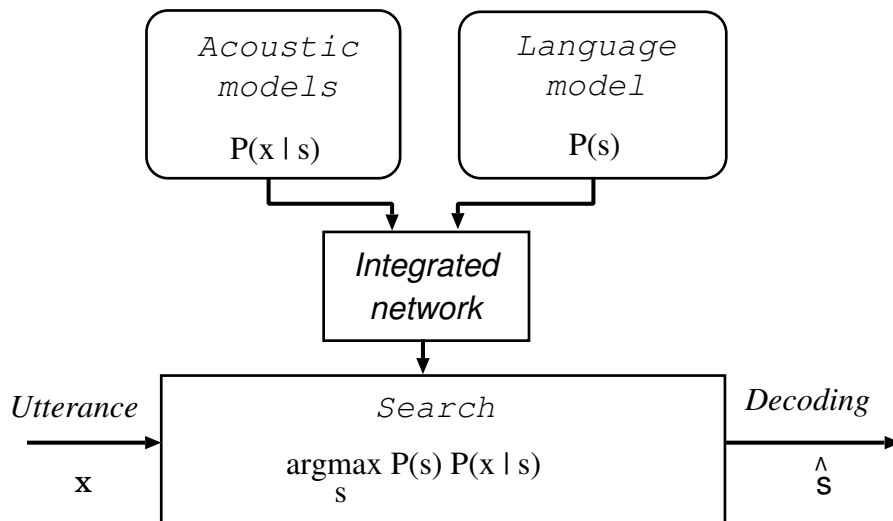
- **N-grams**

$$\Pr(\mathbf{y}) \approx \prod_{i=1}^I p_n(y_i \mid \mathbf{y}_{i-n+1}^{i-1})$$

- **Learning:**

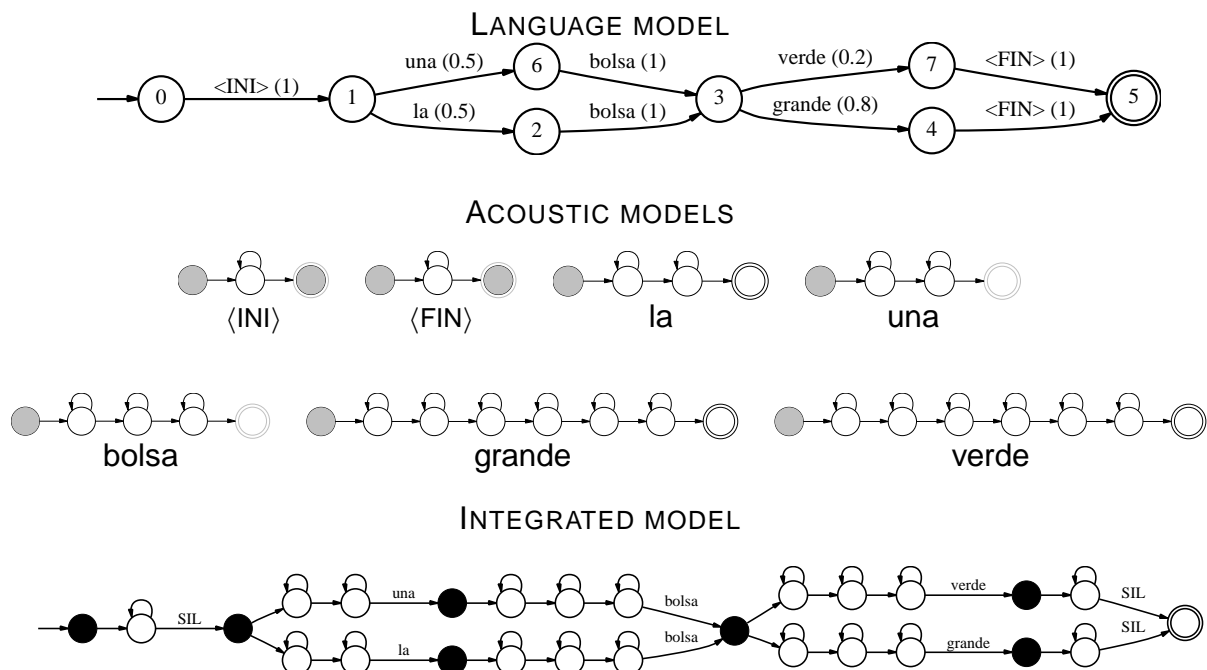
- Grammatical inference techniques.
- Maximum likelihood, maximum entropy.
- Smoothing.
- Extensions: categories, cache, triggers, etc.

Integrated architecture for speech recognition

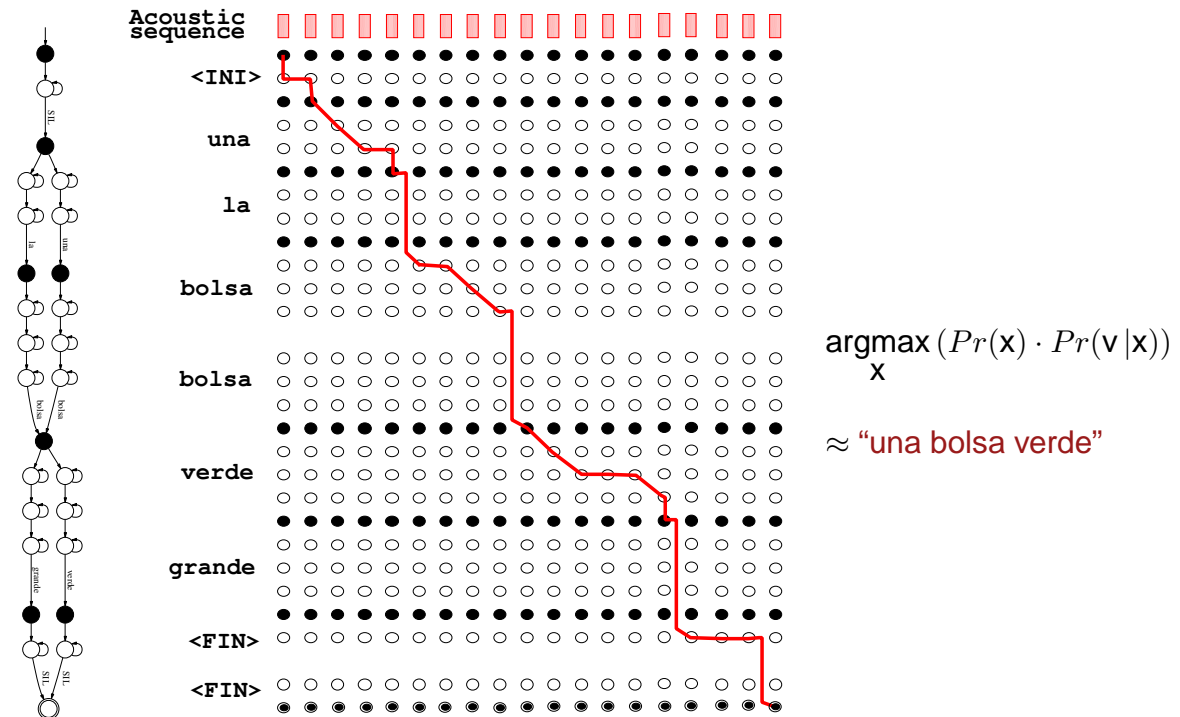


Search engine:
THE VITERBI ALGORITHM (+ beam search + ...)

Integrated architecture for speech decoding



An example of speech decoding



Index

- 1 Speech processing ▷ 2
- 2 Automatic speech recognition ▷ 7
- 3 *Speech to speech translation* ▷ 19
- 4 Bibliography ▷ 33

General statistical framework for speech translation

Given an acoustic sequence v , search for the target sentence \hat{y} :

$$\hat{y} = \underset{y}{\operatorname{argmax}} \Pr(y | v)$$

The translation can be viewed as:

$$v \longrightarrow x \longrightarrow y$$

where x is a possible decoding of v , and y is the translation of x .

$$\underset{y}{\operatorname{argmax}} \sum_x \Pr(y, x | v) \approx \underset{y}{\operatorname{argmax}} \max_x \Pr(y, x | v)$$

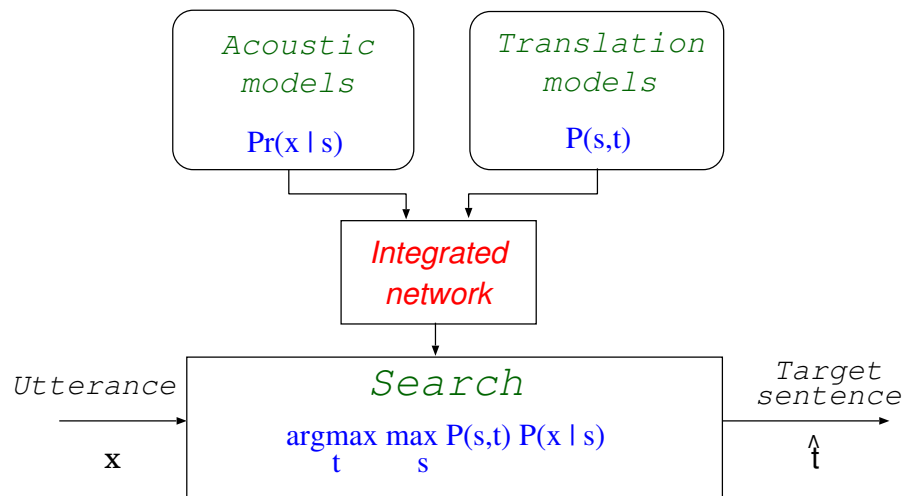
Statistical framework for speech translation

$$\underset{y}{\operatorname{argmax}} \max_x \Pr(y, x | v) = \underset{y}{\operatorname{argmax}} \max_x (\Pr(x, y) \cdot \Pr(v | x))$$

- $\Pr(v|x)$: **Acoustic models**
 - **HIDDEN MARKOV MODELS**
- $\Pr(x,y)$: **Translation models**
 - **STOCHASTIC FINITE-STATE TRANSDUCERS**

INTEGRATED ARCHITECTURE TO SPEECH TRANSLATION.

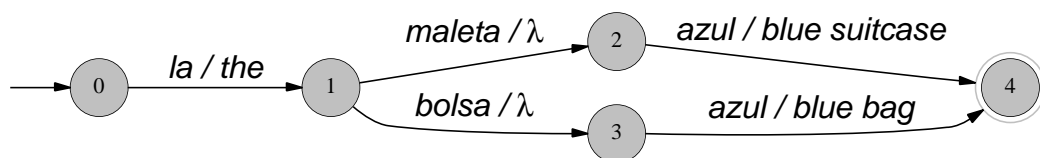
Integrated architecture for speech translation



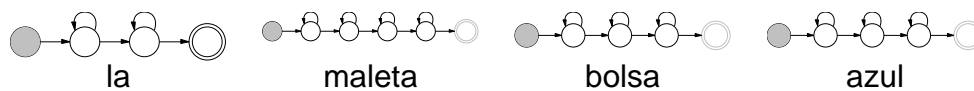
Search engine:
THE VITERBI ALGORITHM (+ beam search + ...)

Integrated architecture for speech translation

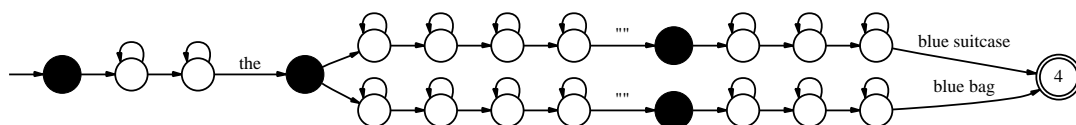
ORIGINAL FINITE-STATE TRANSDUCER



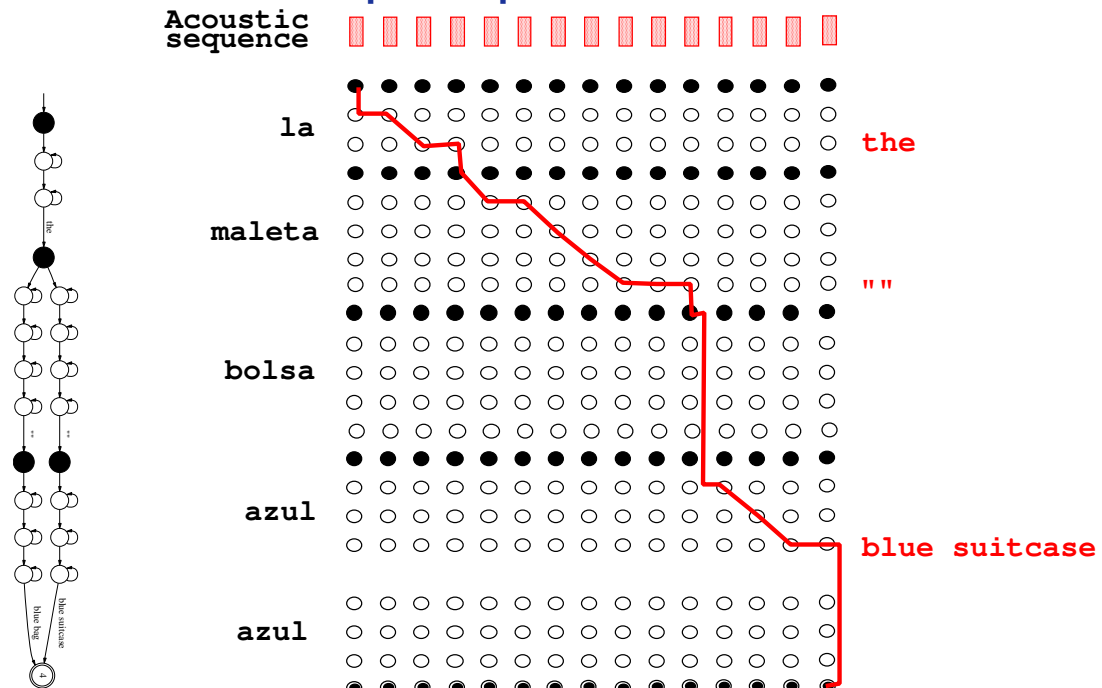
ACOUSTIC MODELS



PHONETIC EXPANSION



An example of speech translation



$$\operatorname{argmax}_{y,x} \Pr(v | x) \cdot \Pr(y, x) \approx \text{"the blue suitcase / la maleta azul"}$$

Statistical framework for speech translation

$$\operatorname{argmax}_y \max_x \Pr(y, x | v) = \operatorname{argmax}_y \max_x (\Pr(y | x) \cdot \Pr(x) \cdot \Pr(v | x))$$

- $\Pr(v|x)$: **Acoustic models**
 - **HIDDEN MARKOV MODELS**
- $\Pr(x)$: **Source language models**
 - **N-GRAMS**
- $\Pr(y | x)$: **Translation models**
 - **STOCHASTIC FINITE-STATE TRANSDUCERS**
 - **STATISTICAL ALIGNMENT MODELS + STOCHASTIC DICTIONARIES**

SERIAL ARCHITECTURE TO SPEECH TRANSLATION.

Serial architecture for speech translation

$$\operatorname{argmax}_y \max_x \{Pr(y|x) \cdot Pr(x) \cdot Pr(v|x)\}$$

1. *Word decoding of v.*

$$\hat{x} = \operatorname{argmax}_x \{Pr(x) \cdot Pr(v|x)\}$$

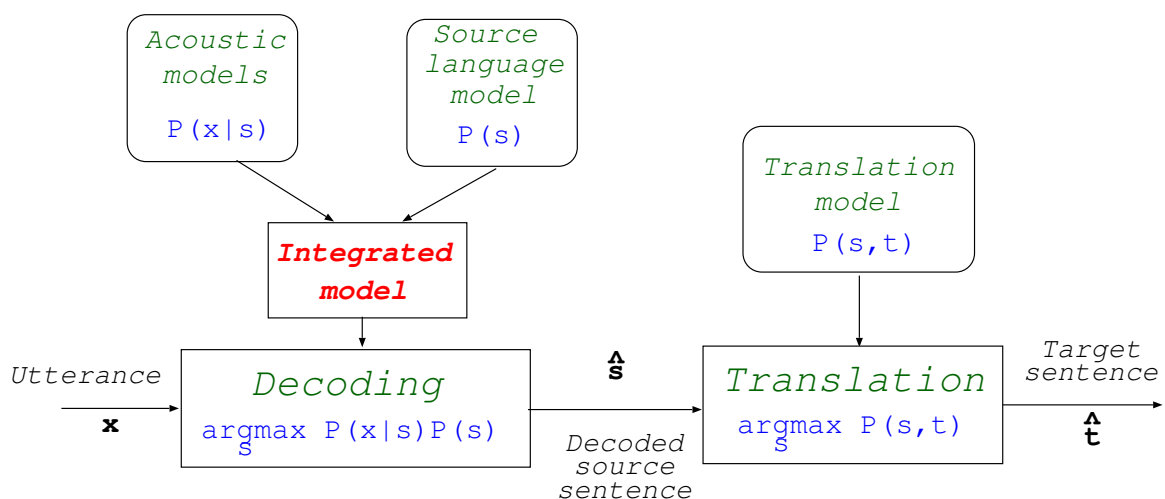
$Pr(x)$: source language model; $Pr(v|x)$: acoustic models.

2. *Translation of \hat{x} .*

$$\hat{y} = \operatorname{argmax}_y Pr(y|\hat{x}) = \operatorname{argmax}_y Pr(y, \hat{x}) = \operatorname{argmax}_y Pr(\hat{x}|y) \cdot Pr(y)$$

$Pr(y, \hat{x})$ or $Pr(\hat{x}|y)$: translation model; $Pr(y)$: target language model.

Serial architecture for speech translation



Search engine for decoding and text translation:
THE VITERBI ALGORITHM (+ beam search + ...)

Experimental results with EUTRANS-0 (Spanish to English)

- **Vocabulary:** 686 Spanish words and 513 English words.
- **Text training:** 490,000 pairs (4,655,000/4,802,000 running words)
- **Speech training:** 11,000 running words for 25 CDHMM of monophones.
- **Speech test:** 336 sentences (3,000 running words) (PP=6.8)
- **Source language models for the serial architecture:** trigrams.

Models	Architecture	Source Language Model	WER(%)	TWER(%)
OMEGA	Integrated	OMEGA	8.4	7.6
OMEGA	Serial	Trigrams	8.6	9.4
GIATI	Integrated	GIATI	7.5	10.7
GIATI	Serial	Trigrams	8.6	11.6
ALTEMP	Serial	Trigrams	8.6	9.9

Experimental results with EUTRANS-II (Italian to English)

- **Vocabulary:** 2,459 Italian words and 1,701 English words.
- **Text training:** 3,038 pairs (61,232/72,446 running words)
- **Speech training:** 52,511 running words for 2,700 CDHMM of triphones.
- **Speech test:** 278 sentences (5,381 running words) (PP=6.8)
- **Source language models for the serial architecture:** trigrams.

Models	Architecture	Source Language Model	WER(%)	TWER(%)
GIATI	Serial	Trigrams	22.1	37.9
GIATI	Integrated	GIATI	32.0	44.8
OMEGA	Serial	Trigrams	22.1	49.4
OMEGA	Integrated	OMEGA	52.5	57.0
ALTEMP	Serial	Trigrams	22.1	37.8

Iterative search (1)

$$\operatorname{argmax}_y \Pr(y | v) \approx \operatorname{argmax}_y \max_x \Pr(y) \cdot \Pr(x | y) \cdot \Pr(v | x)$$

a) INITIALIZATION

1. *Decoding v:* $\hat{x} \approx \operatorname{argmax}_x \{\Pr(x) \cdot \Pr(v | x)\}$

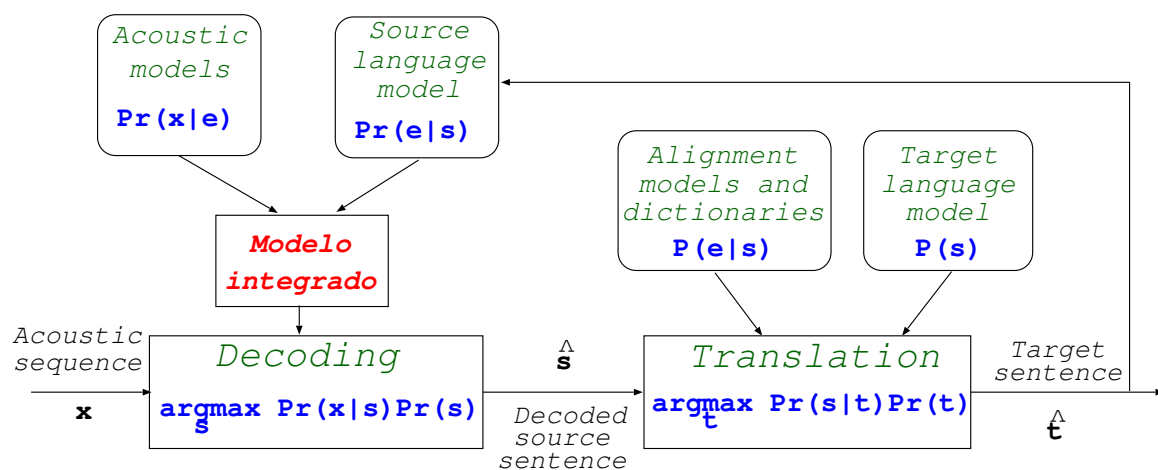
2. *Translating \hat{x} :* $\hat{y} \approx \operatorname{argmax}_y \Pr(\hat{x} | y) \cdot \Pr(y)$

b) GENERAL ITERATION

1. *Decoding v using \hat{y} :* $\hat{x} \approx \operatorname{argmax}_x \{\Pr(x | \hat{y}) \cdot \Pr(v | x)\}$

2. *Translating \hat{x} :* $\hat{y} \approx \operatorname{argmax}_y \Pr(\hat{x} | y) \cdot \Pr(y)$

Iterative search (2)



EUTRANS demos

EuTrans

What is this for? What is EuTrans?

Eutrans' translation tool Dial +34 96 387 72 34

[Index](#) [Italian](#) [Spanish](#) [Catalan to Spanish](#) [Catalan to English](#)

Operation Instructions	
Phone Key	Action
1	Listen to translated sentence
2	Perform a new recognition

[More commands](#)

Results of last call

Recognized Sentence
buongiorno , vorrei prenotare una stanza singola con bagno .

Translated Sentence
Good morning , I would like to reserve a single room with bathroom .

On-line demos

<http://prhltdemos.iti.es/demo/>

Index

- 1 Speech processing ▷ 2
- 2 Automatic speech recognition ▷ 7
- 3 Speech to speech translation ▷ 19
- 4 *Bibliography* ▷ 33

Bibliography

1. Jelinek, *Statistical Methods for Speech Recognition*. The MIT Press, 1998.
2. Rabiner. *Fundamentals of Speech Recognition*, Prentice Hall, 1993.
3. Ney, *Speech translation: Coupling of recognition and translation*, Proceedings of International Conference on Acoustic, Speech and Signal Processing (ICASSP99), 1999.
4. Casacuberta and de la Higuera. *Linguistic decoding is a difficult computational problem*. Pattern Recognition Letters, 20:813–821, 1999.
5. Amengual, Benedí, Casacuberta, Castaño, Castellanos, Jiménez, Llorens, Marzal, Pastor, Prat, Vidal, and Vilar, *The EuTrans-I Speech Translation System*. Machine Translation, 15:75–103, 2000.
6. Ney, Niessen, Och, Sawaf, Tillmann, and Vogel, *Algorithms for statistical translation of spoken language*. IEEE Transactions on Speech and Audio Processing, 8(1):24–36, 2000.
7. Casacuberta, Ney, Och, Vidal, Vilar, Barrachina, García-Varea, Llorens, Martínez, Molau, Nevado, Pastor, Picó, Sanchis, and Tillmann, *Some approaches to statistical and finite-state speech-to-speech translation*. Computer Speech and Language, 18:25–47, 2004.
8. Casacuberta, Vidal, Sanchis, and Vilar. *Pattern recognition approaches for speech-to-speech translation*. Cybernetic and Systems: an International Journal, 35(1):3–17, 2004.

Pattern Recognition approaches to Machine Translation

F. Casacuberta and E. Vidal

Pattern Recognition and Human Language Technology Group
Instituto Tecnológico de Informática
Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de Valencia, Spain

Computer Assisted Translation

Enrique Vidal

`evidal@iti.upv.es`

January 2005

E. Vidal – ITI-UPV-DSIC

[Pattern Recognition Machine Translation](#)

[Computer Assisted Translation](#)

Index

- 1 Computer Assited Translation (CAT) ▷ [2](#)
- 2 Statistical Framework for (text-input) CAT ▷ [7](#)
- 3 Interactive Search ▷ [9](#)
- 4 Using Speech in the CAT Framework ▷ [22](#)
- 5 Bibliography ▷ [31](#)

Index

- 1 *Computer Assited Translation (CAT)* ▷ 2
- 2 Statistical Framework for (text-input) CAT ▷ 7
- 3 Interactive Search ▷ 9
- 4 Using Speech in the CAT Framework ▷ 22
- 5 Bibliography ▷ 31

Introduction to Computer Assited Translation (CAT)

- MT systems are not perfect: they often produce erroneous (portions) of target-language text
- To correct these errors, human post-processing is generally needed
- CAT aims to increase the overall (MT + human) productivity by incorporating human correction activities within the translation process itself

Main idea:

Iterative process where human activity is embedded in the loop

- Use a MT system to produce target text segments that can be accepted or amended by a human translator; these correct(ed) segments are then used by the MT system as additional information to achieve further, hopefully improved suggestions

CAT Human-Machine (keyboard) interactive process

- In each iteration, a correct prefix (y_p) of the target sentence is available and the CAT system computes its best (or N -best) translation suffix hypothesis (\hat{y}_s) to complete this prefix.
- Given $y_p\hat{y}_s$, the CAT cycle proceeds by letting the user establish a new, longer acceptable prefix.

This prefix is typically formed by y_p , followed by an initial part of \hat{y}_s *accepted* by the user (a), followed by text obtained by means of additional user keystrokes (k) generally aimed to amend remaining incorrect parts of \hat{y}_s .

This prefix becomes a new y_p , thereby starting a new CAT prediction cycle

- Ergonomics and user preferences dictate exactly when the system can start its new cycle, but typically, it is started after each user-entered word or even after each new user keystroke.
- These ideas were studied in [Foster02] and have been thoroughly explored in the TT2 project

CAT human-machine (keyboard) interactive process: example

Translating the source sentence “Click OK to close the print dialog” into Spanish:

ITER-0	(y_p)	()
ITER-1	(\hat{y}_s)	(Haga clic para cerrar el diálogo de impresión)
	(a)	(Haga clic)
	(k)	(en)
	(y_p)	(Haga clic en)
ITER-2	(\hat{y}_s)	(<i>ACEPTAR para cerrar el diálogo de impresión</i>)
	(a)	(<i>ACEPTAR para cerrar el</i>)
	(k)	(cuadro)
	(y_p)	(Haga clic en <i>ACEPTAR para cerrar el cuadro</i>)
FINAL	(\hat{y}_s)	(<i>de diálogo de impresión</i>)
	(a)	(<i>de diálogo de impresión</i>)
	(k)	(#)
	($y_p \equiv y$)	(Haga clic <u>en</u> <i>ACEPTAR</i> para cerrar el <u>cuadro</u> de diálogo de impresión)

System suggestions are printed in cursive and user input in boldface typewriter font.

In the final translation, y , text that have been typed by the user is underlined

Evaluating MT and CAT systems

THREE MEASURES

- TRANSLATION WORD ERROR RATE (TWER):
Minimum number of *word* insertions, deletions and substitutions needed to edit the system output into a (single) target reference
- TRANSLATION CHARACTER ERROR RATE (TWER):
Minimum number of *character* insertions, deletions and substitutions needed to edit the system output into a (single) target reference
- KEY-STROKE RATIO (KSR):
Number of key-strokes that are necessary to achieve a (single) target reference divided by the number of running characters.

Index

- 1 Computer Assited Translation (CAT) ▷ 2
- 2 *Statistical Framework for (text-input) CAT* ▷ 7
- 3 Interactive Search ▷ 9
- 4 Using Speech in the CAT Framework ▷ 22
- 5 Bibliography ▷ 31

Text prediction for Computer-Assisted Translation (CAT)

Given a source text x and a “correct” *prefix* y_p of the target text, search for a *suffix* \hat{y}_s , that maximizes the posterior probability over all possible suffixes:

$$\hat{y}_s = \underset{y_s}{\operatorname{argmax}} \Pr(y_s \mid x, y_p)$$

Taking into account that $\Pr(y_p \mid x)$ does not depend on y_s , we can write:

$$\begin{aligned} \hat{y}_s &= \underset{y_s}{\operatorname{argmax}} \Pr(y_p y_s \mid x) \\ &= \underset{y_s}{\operatorname{argmax}} \Pr(x \mid y_p y_s) \cdot \Pr(y_p y_s) \end{aligned} \quad (1)$$

$$= \underset{y_s}{\operatorname{argmax}} \Pr(x, y_p y_s) \quad (2)$$

- (1): Statistical Alignment and Language models
- (2) Stochastic Finite State Transducers
- Text-input MT is a particular case, where $y_p = \lambda$
- Main difference of CAT vs. MT: **search over the set of suffixes**

Index

- 1 Computer Assisted Translation (CAT) ▷ [2](#)
- 2 Statistical Framework for (text-input) CAT ▷ [7](#)
- 3 *Interactive Search* ▷ [9](#)
- 4 Using Speech in the CAT Framework ▷ [22](#)
- 5 Bibliography ▷ [31](#)

CAT Interactive Search

High speed is needed because typically a new system hypothesis must be produced in real time after each user keystroke

WORD-GRAPH BASED APPROACH:

- For each source sentence, *a graph representing all its possible translations according to the translation model is generated*
- *In each CAT iteration, the Word-Graph is searched for a best path compatible with the prefix given in this iteration*
- *Error-Correcting smoothing (edit distance) is used to allow for user-given prefixes that may not exist in the Word-Graph*
- *Computation is carried out in an incremental manner: in each iteration the results from the previous iteration are updated*

Example of CAT human-machine (keyboard) interaction

S: Load your originals into the Document Feeder

H: [Cargue los originales en la](#)

Example of CAT human-machine (keyboard) interaction

S: Load your originals into the Document Feeder

H: Cargue los originales en la

P: Cargue los originales en e

Example of CAT human-machine (keyboard) interaction

S: Load your originals into the Document Feeder

H: Cargue los originales en la

P: Cargue los originales en e

H: Cargue los originales en el alimentador de originales

Example of CAT human-machine (keyboard) interaction

S: Load your originals into the Document Feeder

H: Cargue los originales en la

P: Cargue los originales en e

H: Cargue los originales en el alimentador de originales

T: Cargue los originales en el alimentador de originales

S: Source sentence (x)

P: Current human-validated Prefix (y_p)

H: System Hypothesis (\hat{y}_s)

T: Final Translation

More examples of CAT human-machine (keyboard) interaction

S: It also contains a section to help users of previous software versions adapt more quickly to the new software

H: Se se para ayudar a los usuarios de versiones anteriores del software a que se a dapten más rápidamente a este nuevo software

P: T

H: También se ofrece una sección para ayudar a los usuarios de versiones anteriores del software a que se adapten más rápidamente a este nuevo software

P: También c

H: También contiene una sección para ayudar a los usuarios de versiones anteriores del software a que se adapten más rápidamente a este nuevo software

T: También contiene una sección para ayudar a los usuarios de versiones anteriores del software a que se adapten más rápidamente a este nuevo software

More examples of CAT human-machine (keyboard) interaction

S: Dirección de la alimentación para tamaños de papel estándar 1-9

H: Feed direction for standard stock names 1-9

P: Feed direction for standard p

H: Feed direction for standard paper sizes 1-9

T: Feed direction for standard paper sizes 1-9

More examples of CAT human-machine (keyboard) interaction

S: Edición de la lista de impresoras

H: Editing printers

P: Editing t

H: Editing the printers

P: Editing the printer l

H: Editing the printer list

T: Editing the printer list

Benchmark Xerox printer manuals corpus

Data		English	Spanish	English	German	English	French
Train	Sent. pairs	56K		53K		49K	
	Run. words	572K	657K	543K	583K	507K	441K
	Vocabulary	26K	30K	25K	27K	25K	37K
Test	Sentences	1 125		984		996	
	Run. words	7.6K	9.4K	9.6K	10.0K	10.8K	9.8K
	Out of Voc.	341	362	219	552	252	255
	Run. chars.	46K	58K	55K	63K	61K	71K
	Perplexity	107	60	93	169	193	135

Benchmark EU bulletin corpus

Data		English	Spanish	English	German	English	French
Train	Sent. pairs	214K		223K		215K	
	Run. words	5.9M	6.6M	6.5M	6.1M	6.0M	6.6M
	Vocabulary	84K	97K	87K	152K	85K	91K
Test	Sentences	800		800		800	
	Run. words	20K	25K	22K	21K	22K	24K
	Out of Voc.	108	140	107	227	113	119
	Perplexity	96	72	95	153	97	71

CAT results with the Xerox corpus

DATA: XRCE2	GIATI 3-gram (1-best)			GIATI 3-gram (5-best)		
	KSR	CER	TWER	KSR	CER	TWER
En-Es	17.6	30.3	43.1	15.6	25.0	37.8
Es-En	21.5	35.5	51.4	18.9	28.1	45.2
En-Fr	37.1	54.3	73.8	34.3	48.5	69.6
Fr-En	39.4	55.3	71.9	36.7	49.5	67.7
En-De	38.8	62.8	81.3	35.4	56.7	77.2
De-En	36.4	61.5	78.5	32.9	55.1	73.3

CAT results with the EU corpus

DATA: EU	GIATI 5-gram (1-best)			GIATI 5-gram (5-best)		
	KSR	CER	TWER	KSR	CER	TWER
En-Es	27.5	37.6	55.8	24.6	34.8	51.7
Es-En	25.4	38.0	52.5	22.7	35.1	48.0
En-Fr	26.2	36.0	53.9	23.5	33.4	50.1
Fr-En	23.1	36.1	49.2	20.6	32.8	44.4
En-De	29.4	41.2	65.5	26.8	38.1	60.3
De-En	31.0	44.4	66.6	28.0	41.4	61.2

Index

- 1 Computer Assisted Translation (CAT) ▷ 2
- 2 Statistical Framework for (text-input) CAT ▷ 7
- 3 Interactive Search ▷ 9
- 4 *Using Speech in the CAT Framework* ▷ 22
- 5 Bibliography ▷ 31

Using Speech Recognition in CAT

- Early idea: a human translator dictates aloud the translation in the *target language*. As the source text is known by the system, this knowledge can be used to reduce recognition errors.
- Alternative idea within the CAT framework: the human translator determines acceptable prefixes of the suggestions made by the system by reading (with possible modifications) parts of these suggestions.
 - A much lower degree of freedom is possible and the correspondingly lower perplexity allows for sufficiently high recognition accuracy.
 - As this is fully integrated within the CAT paradigm, the user can make use of the conventional means (keyboard and/or mouse) to guarantee that the produced text exhibits an adequate level of quality.

Target language dictation in CAT

A *human* translator *dictates* the translation of a source text, x , producing a *target language* acoustic sequence v .

Given v and x , the system should search for a most likely decoding of v :

$$\hat{y} = \underset{y}{\operatorname{argmax}} \Pr(y \mid x, v)$$

By the assumption that $\Pr(v \mid x, y)$ does not depend on x ,

$$\hat{y} = \underset{y}{\operatorname{argmax}} \Pr(v \mid y) \cdot \Pr(x \mid y) \cdot \Pr(y)$$

- $\Pr(v \mid y) \approx$ (TARGET LANGUAGE) ACOUSTIC MODELS
- $\Pr(x \mid y) \approx$ TRANSLATION MODEL
- $\Pr(y) \approx$ TARGET LANGUAGE MODEL

Similar to plain speech decoding, where: $\hat{y} = \underset{y}{\operatorname{argmax}} \Pr(v \mid y) \cdot \Pr(y)$

Further use of speech recognition in CAT

Let x be the source text and y_p a “correct” prefix of the target sentence.

As in pure text CAT the system suggests an optimal suffix:

$$\hat{y}_s = \underset{y_s}{\operatorname{argmax}} \Pr(y_s \mid x, y_p) \quad (3)$$


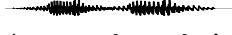
The user is now allowed to *utter some words*, v , generally aimed at amending parts of \hat{y}_s and the system has then to obtain a most probable decoding of v :

$$\hat{d} = \underset{d}{\operatorname{argmax}} \Pr(d \mid x, y_p, \hat{y}_s, v) \quad (4)$$

Finally, the user can enter additional amendment keystrokes k , to produce a new consolidated prefix, y_p , based on the previous y_p , \hat{d} , k and parts of \hat{y}_s .

Example of speech-enabled CAT human-machine interaction

Translating the source sentence “Click OK to close the print dialog” into Spanish:

ITER-0	(y_p)	$()$
ITER-1	(\hat{y}_s)	<i>(Haga clic para cerrar el diálogo de impresión)</i>
	(v)	
	(\hat{d})	(Haga clic a)
	(k)	(en ACEPTAR)
	(y_p)	<i>(Haga clic en ACEPTAR)</i>
ITER-2	(\hat{y}_s)	<i>(para cerrar el diálogo de impresión)</i>
	(v)	
	(\hat{d})	(cerrar el cuadro)
	(k)	()
	(y_p)	<i>(Haga clic en ACEPTAR para cerrar el cuadro)</i>
FINAL	(\hat{y}_s)	<i>(de diálogo de impresión)</i>
	(k)	(#)
	$(y_p \equiv y)$	(Haga clic en ACEPTAR para cerrar el cuadro de diálogo de impresión)

System suggestions are printed in cursive, text decoded from user speech in boldface and typed text in boldface typewriter font. In the final translation, y , text obtained from speech decoding is marked in boldface, while typed text is underlined.

Models for speech recognition in CAT

From Eq. (4):

$$\hat{d} = \underset{d}{\operatorname{argmax}} \Pr(d \mid x, y_p, \hat{y}_s, v) = \underset{d}{\operatorname{argmax}} \Pr(d \mid x, y_p, \hat{y}_s) \cdot \Pr(v \mid x, y_p, \hat{y}_s, d)$$

and, by making the assumption that $\Pr(v \mid x, y_p, \hat{y}_s, d)$ only depends on d :

$$\hat{d} = \underset{d}{\operatorname{argmax}} \Pr(d \mid x, y_p, \hat{y}_s) \cdot \Pr(v \mid d)$$

- $\Pr(v \mid d) \approx$ **(TARGET LANGUAGE) ACOUSTIC MODELS**
- $\Pr(d \mid x, y_p, \hat{y}_s) \approx$ **TARGET LANGUAGE MODEL CONSTRAINED BY THE SOURCE SENTENCE, THE PREFIX AND THE SUFFIX**

Less and more restricted scenarios, depending on the latter model:

- CAT-PREF: Ignore the dependency on the system suggestion \hat{y}_s
- CAT-SEL: Restrict d to be just a prefix of \hat{y}_s

Speech recognition in CAT: CAT-PREF

Starting from:

$$\hat{d} = \underset{d}{\operatorname{argmax}} \Pr(d \mid x, y_p, \hat{y}_s) \cdot \Pr(v \mid d)$$

a *less restricted* scenario arises if only the prefix y_p is available; that is, the previous system prediction \hat{y}_s is ignored and the user is assumed to produce free target speech, only constrained to be a translation of the source text and a continuation of the given prefix:

$$\hat{d} = \underset{d}{\operatorname{argmax}} \Pr(d \mid x, y_p) \cdot \Pr(v \mid d)$$

As compared with the dictated-translation framework, this adds the constraint provided by the target text prefix, y_p , thereby allowing for higher speech decoding accuracy.

Most restricted speech recognition in CAT: CAT-SEL

Starting from:

$$\hat{d} = \underset{d}{\operatorname{argmax}} \Pr(d \mid x, y_p, \hat{y}_s) \cdot \Pr(v \mid d)$$

a *most restricted* scenario appears if the decoding of v is constrained to be *exactly* a prefix of the suffix suggested by the system, \hat{y}_s .

The uttered prefix would help the user determine an accepted part of the system suggestion.

In this case, $\Pr(d \mid x, y_p, \hat{y}_s) = \Pr(d \mid \hat{y}_s)$ and the above equation can be written as:

$$\hat{d} = \underset{d}{\operatorname{argmax}} \Pr(d \mid \hat{y}_s) \cdot \Pr(v \mid d)$$

As compared with all the previous scenarios involving speech, here $\Pr(d \mid \hat{y}_s)$ can be modeled by a very low perplexity language model, which allows for much higher speech decoding accuracy.

CAT speech recognition results

- SPEECH DATA: Utterances of fragments of target language sentences from the test XEROX CORPUS (485 fragments, 10 speakers, 5,796 utterances)
- MODELS: derived from both source and target sentences of the training XEROX corpus
- DEC and DEC-PREF used for comparison:
 - DEC: Conventional speech recognition of target language utterances (source text ignored)
 - DEC-PREF: Target speech recognition constrained by the known prefix (source text ignored)

	DEC	DEC-PREF	CAT-PREF	CAT-SEL
Word Error Rate (%)	18.6	16.1	10.6	1.6
Sentence Error rate (%)	50.2	44.4	30.0	3.6

Using knowledge about the source sentence is more important than using only user-validated prefixes

Index

- 1 Computer Assisted Translation (CAT) ▷ [2](#)
- 2 Statistical Framework for (text-input) CAT ▷ [7](#)
- 3 Interactive Search ▷ [9](#)
- 4 Using Speech in the CAT Framework ▷ [22](#)
- [5 Bibliography](#) ▷ [31](#)

Bibliography

- G. Foster, P. Langlais, G. Lapalme. User-Friendly Text Prediction for Translators. Conference on EMNLP. 2002.
- F.Casacuberta and E.Vidal. Machine translation with inferred stochastic finite-state transducers. Computational Linguistics, 30(2):205-225, 2004.
- J. Civera, J. Vilar, E. Cubel, A. Lagarda, F. Casacuberta, E. Vidal, D. Picó, and J. González, A syntactic pattern recognition approach to computer assisted translation. Advances in Statistical, Structural and Syntactical Pattern Recognition – S+SSPR 2004 IAPR workshop. A. Fred, T. Caelli, A. Campilho, R. P. Duin, and D. de Ridder, Eds. LNCS, Springer-Verlag, Lisbon, 2004.
- E.Vidal, F.Casacuberta, L.Rodríguez, J.Civera and C.Martínez Computer-Assisted Translation Using Speech Recognition. To be published, 2005.

PingPongPlus: Augmentation and Transformation of Athletic Interpersonal Interaction

Craig Wisneski, Julian Orbanes, Hiroshi Ishii

MIT Media Laboratory

20 Ames St.

Cambridge, MA 02139, U.S.A.

{wiz, joules, ishii}@media.mit.edu

ABSTRACT

PingPongPlus (PP+) is a digitally enhanced version of the classic ping-pong game. We have designed a digital layer of audio/visual augmentation on top of a conventional ping-pong table using a newly developed ball tracking system and video projection. The "reactive table" displays patterns of light and shadow as a game is played, and the rhythm and style of play drives accompanying sound. In the process, this project explores new ways to couple athletic recreation and social interaction with engaging digital enhancements. This paper describes the basic idea, research agenda, several applications, technical implementation, and initial experiences.

Keywords

augmented reality, reactive surface, athletic / kinesthetic interaction, computer-supported collaborative play, interactive media art.

INTRODUCTION

Computer-Supported Collaborative Play can take many forms. It runs the technical gamut from highly sophisticated networked video games to electronic board games. Most of the work in today's digital multiplayer games has lost the element of the physical presence of people and their kinesthetic interactions. We are interested in designing systems for collaborative play that push the physical world back into the forefront of design, without relying on simple GUI controllers (such as a mouse, keyboard, or joystick) [1]. Rather, in our model of collaboration, more emphasis is placed on the physicality of the people involved. We believe that a person's physical prowess, and sense of kinesthesia, can be leveraged to strengthen the quality of collaborative play. To do this, we have investigated new ways to interact with a surface and to sense activity. We seek to examine some of the ways digital augmentations can change traditional, physically-based game play and allow new interfaces with the digital world.

Permission to make digital/hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, the copyright notice, the title of the publication and its date appear, and notice is given that copyright is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires specific permission and/or a fee.

CHI '97, Atlanta GA USA

Copyright 1997 ACM 0-89791-802-9/97/03 ...\$3.50

PINGPONGPLUS

PingPongPlus is a digitally enhanced ping pong game using a "reactive table" that incorporates sensing, sound, and projection technologies. The table displays patterns of light and shadow as a game is played, and the rhythm and style of play drives accompanying sound. For example, in one mode, a bouncing ball leaves images of

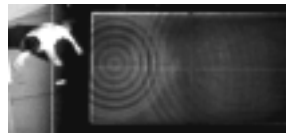


Fig. 1. Water Ripple

rippling water (Fig. 1).

Technical Overview

PingPongPlus consists of two main elements: a ball-tracking system, and a graphics projection system (Fig. 2).

The ball position sensing is done solely through sound. When a ball hits, the sound travels through the table. Eight microphones mounted on the underside of the table pick up the sound. When a microphone detects a hit, a time value is assigned to that microphone, and sent to a computer through a custom made electronic circuit. The time values are evaluated on a 200 MHz PC by an algorithm that determines the location of the hit. The algorithm we have developed can pinpoint the ball's position within a few inches in a matter of milliseconds, which is good enough for our application.

The graphics are created in accordance with the ball tracking information. They are written in Visual C++ with a custom-made graphics package. A projector suspended 20 ft. above the table displays the graphics on to its surface.

APPLICATIONS

Over 12 different applications have been developed and tested on the table. Five of the applications are discussed here. Through laboratory sponsor meetings, demonstrations, and exhibitions, hundreds of people have played with PP+, and their feedback was reflected in our iterative design.

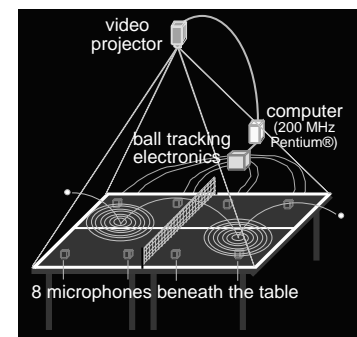


Fig.2. System configuration of PingPongPlus

Water Ripple

The *Water Ripple* is a simple, causal augmentation. When a ball hits the table, an image of a water ripple flows out from the spot the ball landed (Fig. 1). Players found this to be one of the less distracting applications from the normal game of ping-pong, allowing them to concentrate on the game at hand, yet augmenting the game in a non-traditional sense. People often played with curiosity, rather than competitiveness, trying to examine what kinds of interference wave patterns they could create and view on the table. One child even climbed up on the table and created water ripples with his foot, rather than a ball.

Thunderstorm

The Thunderstorm application incorporates game logic into its structure. By keeping the ball in play, rallying back and forth, the players “build up a thunderstorm.” At the beginning of a point, only calm, flowing waves appear on the table (Fig. 3). As the rally duration increases, the sound of a heartbeat gets faster, wind whips around the sound space, and the waves speed up. If the ball is kept in play for a long time, lightning bolts shoot from one side of the table to the other, connecting the ball’s last two locations. In this mode, we found that the style of game, the way people play, is changed due to the additional effects. When the wind picks up and the heartbeat gets faster, players tend to hit the ball faster and harder.

Black-Out

With the Black-Out mode, we experimented with how augmentation can change strategies employed in a game. This mode is intended to be played in a completely dark room, where the only light comes from the bright white projection on the table. In this mode, a large black spot appears wherever the ball hits, effectively “taking light away” from the other person’s side of the table (Fig. 4). By concentrating hits in a single area, all the opponent’s light can be taken away in that space. The removal of light can be used strategically.

Painting

This application explores the collaborative possibilities of the project. One side of the table is a blank canvas, and the other is a collection of two colors of “ink”. When a ball hits the black area of the “ink,” it leaves a black spot on the canvas (Fig. 5). Accordingly, when it hits the white “ink,” it leaves a white spot on the canvas side of the table. Through collaboration on color choices and placements by expert players, an interactive artwork can be made on the canvas. There is a shift here, from normal ping-pong to a different kind of collaborative game. The object is not to win a game, but it is to collaboratively create an image. This shows how augmentation can not only change the nature of game play, but it can change the object of the game itself.

Comets

The Comets application continues to change the object of the game. In this mode, when a ball hits the table, it

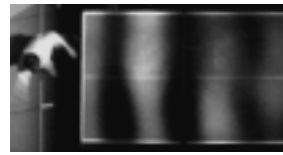


Fig. 3. Thunderstorm



Fig. 4. Black-Out

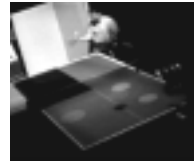


Fig. 5. Painting

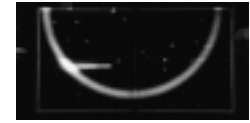


Fig. 6. Comets

“releases a comet” which travels up towards the net (Fig 6). When the comet hits the net, it creates a sound that is mapped to the place on the table the comet originated from. Experts using this mode could potentially use PP+ to creating music, or at least an interesting sound sculpture.

DISCUSSION

We have been exploring a design space along the axis of competition-collaboration and augmentation-transformation. The more subtle augmentation of the *Water Ripple* mode does not change the basic nature of ping-pong play very much. In contrast, *Black-Out* provides players with new strategies to win a game. The *Painting* mode gives a new collaborative goal where players have tried to coordinate their play to paint on a “canvas” table.

Fig. 7 illustrates our design axis of augmentation-transformation and sample applications.

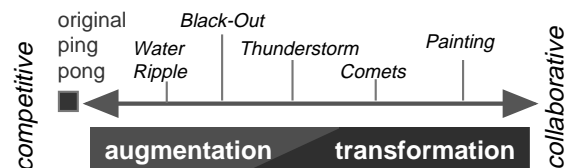


Fig. 7. Design space of PingPongPlus project

CONCLUSION

We expect PingPongPlus will suggest new directions to couple athletic recreation and social interaction with engaging digital enhancements. By augmentation and transformation of physically-based games, new, engaging games can be developed in the physical/digital world.

ACKNOWLEDGMENTS

We thank the colleagues of the Digital Life and TTT consortium at the MIT Media Laboratory for their support and collaboration. Special thanks to Dr. Joe Paradiso for his advice in designing the tracking system.

REFERENCES

1. Ishii, H., and Ullmer, B. (1997). Tangible Bits: Towards Seamless Interfaces between People, Bits, and Atoms. In Proc. of CHI '97, ACM, March 1997, pp. 234-241.

PingPongPlus: Design of an Athletic-Tangible Interface for Computer-Supported Cooperative Play

Hiroshi Ishii, Craig Wisneski, Julian Orbanes, Ben Chun, and Joe Paradiso*

Tangible Media Group

*Physics and Media Group

MIT Media Laboratory

20 Ames St., Cambridge, MA 02139, U.S.A.

{ishii, wiz, joules, benchun, joep}@media.mit.edu

ABSTRACT

This paper introduces a novel interface for digitally-augmented cooperative play. We present the concept of the "athletic-tangible interface," a new class of interaction which uses tangible objects and full-body motion in physical spaces with digital augmentation. We detail the implementation of PingPongPlus, a "reactive ping-pong table", which features a novel sound-based ball tracking technology. The game is augmented and transformed with dynamic graphics and sound, determined by the position of impact, and the rhythm and style of play. A variety of different modes of play and initial experiences with PingPongPlus are also described.

Keywords

tangible interface, enhanced reality, augmented reality, interactive surface, athletic interaction, kinesthetic interaction, computer-supported cooperative play.

INTRODUCTION

When an expert plays ping-pong, a well-used paddle becomes *transparent*, and allows a player to concentrate on the task – playing ping-pong. The good fit of grasp is vital to making a paddle transparent [10]. To achieve a "good fit," a user has to choose a paddle of the right size, right form, and right weight for his or her hand and style of play. To achieve a "better fit," the user has to *customize* the tool by scraping the edge of the paddle with a knife and sandpaper. The "best fit" is, however, achieved by using a paddle over a long period of time.

Figure 1 shows the author's paddle and the traces of the body left on it [4]. After twenty years of use, the grip of the paddle has captured the traces of his right hand (marks of the thumb and index finger in front and marks of the middle finger on back). The right-bottom picture shows the dent made on the back of the paddle by a strong grasp with the tip of the middle finger.

The ping-pong paddle, which can co-evolve with a user by changing its physical form and being united with the human hand, suggests an important direction for HCI – transparent physical extensions of our body and mind into both physical and digital worlds.

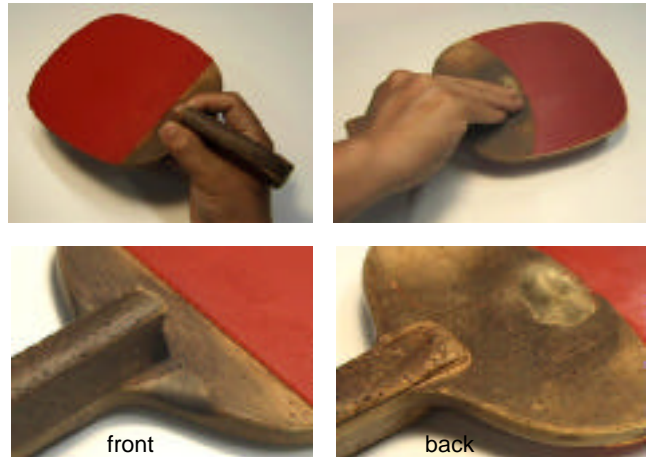


Figure 1 Traces of grasping hand left on the well-used ping-pong paddle

Moreover, the full-body motion, speed, and rhythm of a ping-pong game make the interaction very engaging and entertaining. Kinesthesia is one of the keys of what makes ping-pong enjoyable.

Modern graphical user interface (GUI) technologies provide very limited, generic physical forms (e.g. mouse, keyboard, and monitor) and allow limited physical motions (only clicking and typing). Thus, the GUI is difficult to adapt to human bodies and to take advantage of kinesthesia.

Goals of the PingPongPlus Project

We have designed PingPongPlus on top of the classic game of ping-pong [21]. Its goals are:

1. to demonstrate an instance of an *athletic-tangible interface*, developed on top of existing skills and protocols of familiar competitive/cooperative play.
2. to develop an underlying technology for an "interactive architectural surface" which can track the activities happening on the surface.
3. to study the impact of digital augmentation on the competitive/cooperative nature of play.

COMPUTER-SUPPORTED COOPERATIVE PLAY

Sport is an activity governed by a set of rules or customs that involves skill and physical exertion. It is often

undertaken competitively against opponents, while it is played cooperatively within a team. By playing sports, people can not only learn athletic skills and develop physical strength, but they can also develop social communication and coordination skills.

Computer support is gradually embedding itself in, and transforming the way we play sports and games. Traditional computer games are now extending their reach out from the sole domain of the keyboard, mouse, joystick, and twitch-controllers [8]. Children can create and teach robots, interact with their dolls, and experience complex skiing and motorcycle simulators. With the rise of networks, in the home and in the arcade, play can occur cooperatively more than ever before.

We may give a generic label “CSCP” (Computer-Supported Cooperative Play) to uses of computer technology that enhance physical exertion, social interaction, and entertainment in sport and play. Our research interests in CSCP encompass both the *augmentation* and *transformation* of sports and games. We expect that CSCP research will guide us to design a new form of HCI that we call the “athletic-tangible interface.” This refers to a new class of interaction that uses tangible objects and full-body motion in physical spaces with digital augmentation. We believe that a person’s physical prowess and sense of kinesthesia can be leveraged to strengthen the quality of a collaborative play experience in physical/digital domain.

Our athletic-tangible interface research looks at augmentation and transformation of *real* sports and games, rather than partial simulations of them. Arcade simulation games, while moving in very promising physically-based directions, can only imitate portions of real experience. Immersive virtual environments, such as VIDEOPLACE [7] and ALIVE [9], allow users to use unencumbered full body motion. Although these systems are engaging, they are designed to provide only a simulated experience and the interaction is limited to simple gesturing.

We see the opportunity to explore the design of new games and play experiences where physical interaction is of central importance. We have begun to explore this by adding digital layers of graphics and sound on top of existing skills and protocols of classic games.

DESIGN OF PINGPONGPLUS

We have chosen ping-pong as a target sport of our athletic-tangible interface research, and have designed a computer-augmented version called “PingPongPlus.” PingPongPlus is a digitally enhanced ping-pong game using a “reactive table” that incorporates sensing, sound, and projection technologies. The table displays graphics patterns as a game is played, and the rhythm and style of play drives accompanying sound.

Figure 2 shows a snapshot of PingPongPlus in the water ripples mode, and Figure 3 shows the system architecture of PingPongPlus. In the water ripples mode, a bouncing ball leaves images and the sound of rippling water.

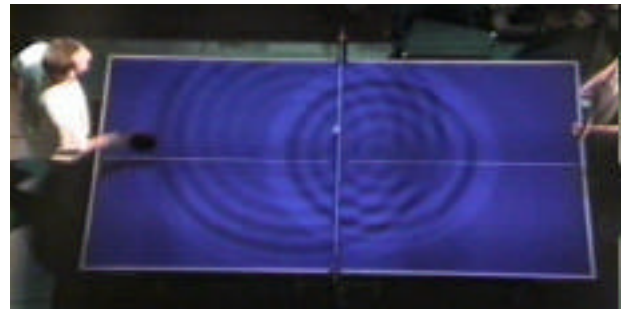


Figure 2 PingPongPlus in water ripples mode

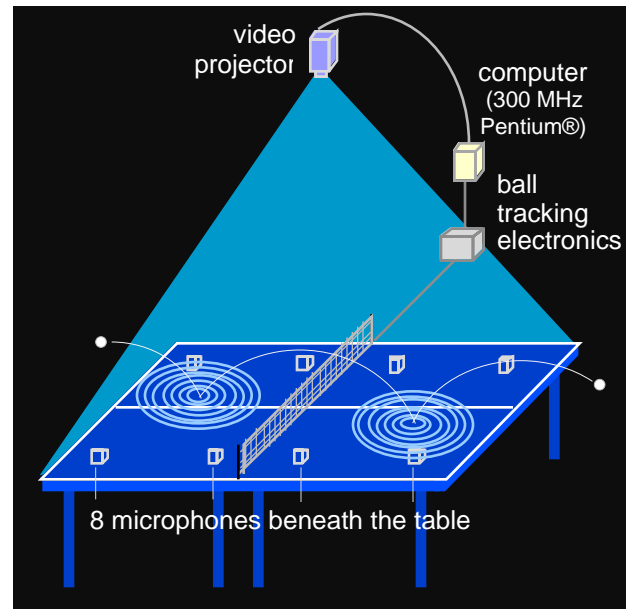


Figure 3 System architecture of PingPongPlus

A series of “tangible interfaces” have been created which give physical form to online digital information [3, 5, 16]. In these projects, users can directly *grasp* and *manipulate* digital information by coupling graspable objects and online digital information. We have also demonstrated the concept of an *interactive surface* that can sense and track the graspable objects on it and project digital shadows [15, 17].

In PingPongPlus, we are extending this notion of tangible interfaces by integrating the kinesthesia of athletic interaction. With PingPongPlus, users experience dynamic and athletic interactions using the full-body in motion, a paddle in hand, a flying ball, and a reactive table. PingPongPlus requires sophisticated realtime coordination among the body, paddle, ball, and digital effects of graphics and sound.

IMPLEMENTATION TECHNOLOGY

The PingPongPlus system consists of ball-tracking hardware, software algorithms for ball-hit location detection, and a graphics projection system. The technology behind creating “interactive surfaces” is of

utmost importance to this system, and is further described here.

Ball Tracking System

We have developed a sound-based ball tracking system. When a ball hits, the sound travels through the table at roughly twice its speed in air. Eight microphones mounted on the underside of the table pick up the sound. When a microphone detects a hit, a time value is assigned to that microphone, and it is sent to a computer through a custom-made electronic circuit. The time values are evaluated on a 300 MHz PC by an algorithm that determines the location of the hit. The algorithm we have developed can pinpoint the ball's position within a few inches in a matter of milliseconds, which is good enough for our application.

Figure 4 shows a schematic diagram of a ball hit. The four microphones ($m1$, $m2$, $m3$, and $m4$) on the underside of each table top pick up the ball hit sound at different times ($t1$, $t2$, $t3$, and $t4$). Given this information, there are a few different algorithms that can determine the original location of a ball hit. We implemented two different methods along with the necessary hardware.

Hardware Implementation

A custom-built hardware circuit connects the ping-pong table to the computer via the serial port (Fig. 5). This circuit only outputs a microphone number ($m1$, $m2$, $m3$, or $m4$) along with its associated time value ($t1$, $t2$, $t3$, $t4$). Software running on a host PC does the rest of the work.

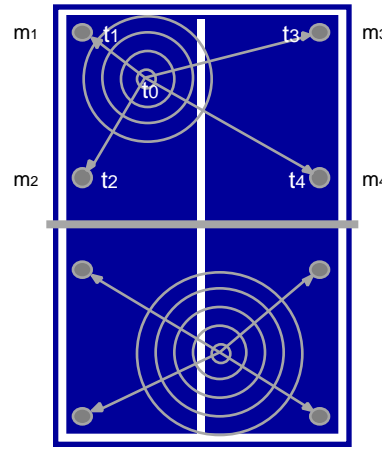
The hardware is realized by doing peak thresholding on signals from the microphones. The microphones themselves are electret pickups, which output a voltage around 0.25 volts for a typical hit. First, their signal is passed through an op-amp which increases their gain by a factor of 20, such that there is a signal between 0 and 5 volts, quiescently at 2.5 volts. This signal is sent through two comparators and an or-gate that compare the signal's absolute value (relative to the 2.5 volt center) against a threshold voltage (both high and low). The comparator/or-gate pair returns true to a PIC chip if there is an impact. This PIC chip is running at 20 MHz, and polls its input about 100,000 times a second. If there is a hit, the PIC chip assigns a time value to that microphone input, and sends this information out a serial connection. Fig. 6 shows a photo and a block diagram of the electronic circuit.

Including the microphones, the total cost for this hardware is nominal. A future improvement to this system is to implement peak detection and to match the various incoming waveforms (as opposed to simple thresholding) to more accurately determine the time differences, and perhaps enable us to extract impact characteristics. It is expected that this will produce significant gains in accuracy and reliability.

Software Algorithms for Location Detection

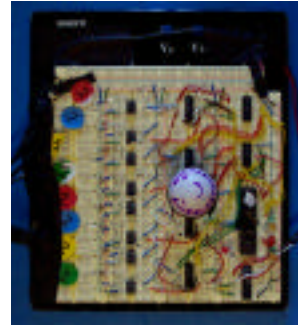
Given the hit timing information from the hardware, the software can calculate a ball-hit coordinate in a number of different ways.

The first algorithm we implemented is by a direct inspection of the time differences. If the ball lands directly

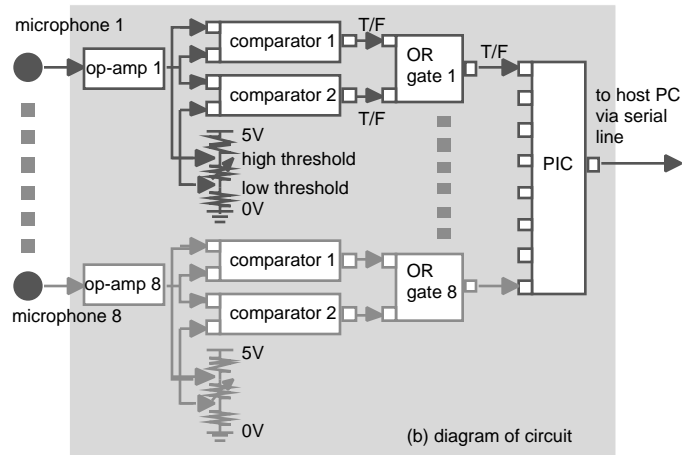


m#: microphones

Figure 4 Ball tracking algorithms



(a) photo of circuit



(b) diagram of circuit

Figure 5 Ball tracking electronic circuit

at a midpoint between two microphones, the time differences between the two points will be the same ($t1 = t2$, for instance), and you can infer that the ball landed on a straight line equidistant from those points. If the ball lands closer to one microphone than another, it can be inferred that the ball landed on a hyperbolic shaped curve between the two points.

The time differences between many microphones can be compared, which results in a system of hyperbolas that

intersect at different points. This system of equations can be solved to yield these points. By throwing out intersection points that do not occur on the table, and looking at which points have the largest number of intersecting hyperbolas, a very good approximation of the original hit location can be made. This method is efficient, as no calibration whatsoever is required.

This algorithm, however, has drawbacks. First, it requires solving equations for two variables that go out to infinity. This is computationally expensive. Second, this method is sometimes not accurate. Sometimes there might be multiple intersection points, or possibly, no points at all. In these cases, a best guess must be made based on the data, and the system is fairly prone to error.

While hyperbolic locator algorithms have been further refined in the literature (e.g. [2]), we have developed a much simpler algorithm to calculate the ball hit position that is better suited to this application. This method is based on a comparison of the time-difference data to a set of model parameters that are acquired by a linear least-squares fit of calibration/training data. The model for this method is:

$$AX = Y$$

Where:

Y = the ball landing coordinate vector (x,y)

X = sensor data vector (time differences information)

A = model parameters (matrix obtained by linear least-squares fit)

When an impact occurs, the sensor values, X , are multiplied by the model parameters, A , which returns a ball landing coordinate, Y . Matrix A , the model parameters, is set through a calibration routine. This calibration routine, however, only needs to be performed once in the life of the table, unless the microphone placement is changed.

Training data is acquired by dropping a ping-pong ball on certain known spots on the table a number of times. In our case, we chose to calibrate the table with 18 distinct points; the A matrix was then calculated through a least-squares fit to this data [14].

Although it involves a linear approximation to hyperbolic relation, this method works well here for a variety of reasons. Since it is a simple matrix multiplication, it is very fast. Also, the linear least-squares fit error metric in the creation of the model parameters makes the system somewhat adaptive to imperfect tables. Performance does not degrade as drastically around edges as compared to the first algorithm (This is important, as most hard surfaces have different kinds of edge effects.). Using this method makes the sensing system more portable to other kinds of tables and surfaces. Although the linear approximation introduces some distortion, it provided accuracy on the order of a few inches, while being fast enough to appear perceptually instant.

At the early stage of this PingPongPlus project, we evaluated the use of computer vision technology for ball

tracking, but we concluded that it was slower, more complicated, and computationally more expensive than sound-based tracking technology. Computer-vision, however, is attractive because the system can capture not only the ball but also the motion of players with paddles. Computer vision could be a reasonable and more interesting alternative technology when the computation speed becomes fast enough and the price drops.

Creation and Projection of Graphics

The graphics are created in accordance with the ball tracking information. They are written in Visual C++ with a custom-made graphics package. In the following APPLICATION section, we describe several patterns of graphics we have developed.

A projector suspended 20 ft. above the table displays the graphics on to its surface. We used a Mitsubishi LCD projector LVP-G1A for the experiments, but the brightness of this projector was not enough. To see the graphics on the surface of ping-pong table, we had to darken the room, making it difficult for human eyes to track the ball. We expect the next generation of brighter video projection technology and, potentially, "e-ink" technology [6] to resolve this problem.

In order to make the graphics less "pixelated," we out-focused the video projector slightly so that the image became softer and naturally merged into a wooden table surface.

APPLICATIONS

We have designed and implemented over a dozen different application modes on the PingPongPlus table. The goal of our application design was to explore the design space characterized by the two axes: 1) augmentation vs. transformation, and 2) competition vs. collaboration.

We had two phases of application development.

Phase 1: 1997 Summer-Fall

Artistic and collaborative play modes: water ripples, thunderstorm, spots, painting, comets, etc.

Phase 2: 1998 Spring-Summer

An enhanced artistic mode (school of fish) and a new competitive game mode (Pac-Man®).

PingPongPlus was demonstrated from October 1997 until July 1998 at the MIT Media Lab to the faculty, students, and sponsors. In July 1998, PingPongPlus was exhibited at SIGGRAPH '98 Enhanced Realities in Orlando [20].

Although we have not yet conducted formal experiments to evaluate those applications, informal feedback from casual users was reflected in the iterative design of these applications. In this section, we illustrate and discuss seven examples of those applications.

Water Ripples mode

The *Water Ripple* mode is a simple, causal augmentation. When a ball hits the table, an image of a water ripple flows out from the spot the ball landed (Fig. 2). Players found this to be one of the less distracting applications from the normal game of ping-pong, allowing them to concentrate

computer
screen
shot

Figure 6 Spots mode

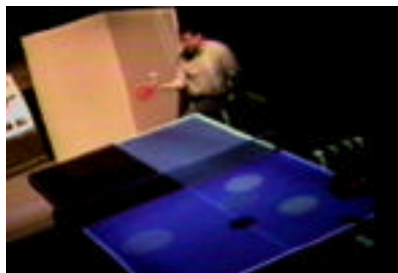
computer
screen
shot

Figure 7 Painting mode

on the game at hand, yet augmenting the game in a non-traditional sense. People often played with curiosity, rather than competitiveness, trying to examine what kinds of interference wave patterns they could create on the table. Once a child even climbed up on the table and created water ripple with his foot. When a player makes an error by hitting a ball into the net, it is usually disappointing. However, in water ripples mode, it turns into an opportunity to enjoy a sequence of small water ripples making a beautiful pattern of interference and sound.

Spots mode

The *Spots* mode was originally intended to be played in a completely dark room where the only light source is the bright white projection on the table. In this mode, a large black spot appears wherever the ball hits, effectively “taking light away” from the other person’s side of the table (Fig. 6). The removal of light can be used strategically, changing the strategies employed in a game.

Painting mode

The *Painting* mode was derived from spots mode. The *Painting* mode was designed to explore the collaborative aspects of PingPongPlus. In *Painting* mode, one side of the table is a blank canvas, and the other is a black and white “ink” pallet. When a ball hits the black area of the “ink,” it leaves a black spot on the canvas (Fig. 7). Accordingly, when it hits the white “ink,” it leaves a white spot on the canvas side of the table. Through collaboration on color choices and placements by expert players, an interactive artwork can be made on the canvas. There is a shift here away from normal ping-pong to a collaborative painting game. The object is not to win a game, but to create an image. This suggests digital augmentation can not only change the nature of the game, but also change the object of the game itself.

In practice, however, the precise control of the ball is too difficult for most users. They could not succeed in painting what they intended. Rather than coordinating the ball movement to create images, they simply enjoy painting visual effects. This motivated us to design the *Comets* mode.

Comets mode

In the *Comets* mode, when a ball hits the table, it “releases a comet” which travels up towards the net (Fig 8). When the comet hits the net, it creates a sound that is mapped to the place on the table from which the comet originated. Experts using this mode could potentially use PingPongPlus to create/play music. We are planning to further explore the integration of playing music and ping-pong by using the speed of play as a metronome that controls the tempo of music being played.

Thunderstorm mode

The *Thunderstorm* mode was designed to encourage collaboration by continuing to rally rather than scoring points. By keeping the ball in play, rallying back and forth, players “build up a thunderstorm.” At the beginning of a point, calm, flowing waves appear on the table (Fig. 9 top). As the rally duration increases, a sound of a heartbeat in the background gets faster, wind whips around the sound space, and waves speed up. If the ball is kept in play for a long time, lightning bolts shoot from one side of the table to the other, connecting the ball’s last two locations (Fig. 9 bottom).

In this mode, we found that the way people play is changed due to the additional effects of the thunderstorm. When the wind picks up and the heartbeat gets faster, players tend to be more nervous and hit the ball faster and harder. Players try to rally until they see the lightning. The lightning at the end of a long rally encourages players to cooperate.

Pac-Man® mode

In *Pac-Man®* mode, the Namco classic video game is reinterpreted for the PingPongPlus environment (Fig. 10). The ball serves the same functions as Pac-Man® did in the video game; it is controllable by the players and results in the scoring of points, which is the goal of the game. Points are awarded for accuracy in hitting the various fruit targets, and points are taken away for hitting the ghosts.

computer
screen
shot

Figure 8 Comets mode

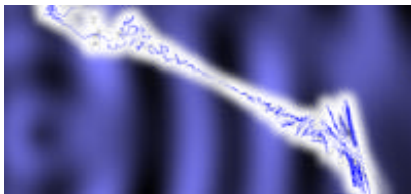
computer
screen
shot

Figure 9 Thunderstorm mode

computer
screen
shot

Figure 10 Pac-Man® mode

We designed this *Pac-Man* mode to see if we could transform ping-pong into a very different engaging, competitive game. However, it was found that it was difficult to divide visual attention between tracking a ball and watching the Pac-Man screen on a table. This indicates that highly detailed display elements on the table do not work as well as simple visual patterns. The best results seem to occur when a simple visual pattern is combined with some level of complexity to keep the game interesting. The School of Fish mode is a good example of this concept.

School of Fish mode

The school of fish with water ripples seemed to be the most popular mode for players. In this mode, a school of fish swims on the table (Fig. 11) following a behavior pattern set forth from the algorithms that Craig Reynolds developed for flock behavior [12]. (Top three pictures are the images from a computer screen, and the bottom picture is a picture from the installation.) The ball causes a splash and a ripple in the “water” where it hits, scaring the fish. In time, the fish, following their individual behavior models, school back together. The simplicity of the visual display, combined with the complexity of the emergent activity from a behavior model made this mode continually compelling, even after days of play.

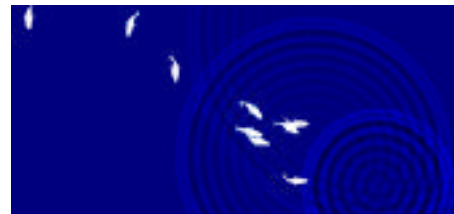
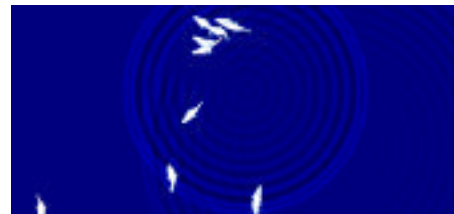
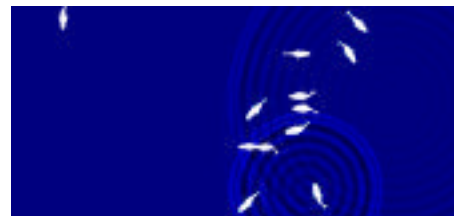
computer
screen
shots

Figure 11 School of fish and water ripples mode

DISCUSSION AND FUTURE WORK

Through the PingPongPlus project, we intended to explore a design space that can be characterized by two axes: augmentation vs. transformation, and competition vs. collaboration. Figure 13 illustrates the seven applications plotted in this 2D design space based on our intention and experiences.

Originally ping-pong is a competitive game, and modes such as water ripples and spots did not change the basic nature of ping-pong play very much; it was still primarily a competitive game.

In contrast, *Comets*, *Painting* and *Thunderstorm* modes added new collaborative goals. For example, *Thunderstorm* and *Comets* encouraged players to keep playing to see the lightning effects or to hear the music of the comets. The *Painting* mode was intended to encourage coordination to paint on a "canvas" table.

Pac-Man was intended to test the transformation of the game into another competitive game. Originally, we expected that the experienced players could place the ball accurately to score points. This assumption proved to be false, showing that careful design is needed in "target" games.



Figure 12 PingPongPlus in use (SIGGRAPH '98, Enhanced Reality)

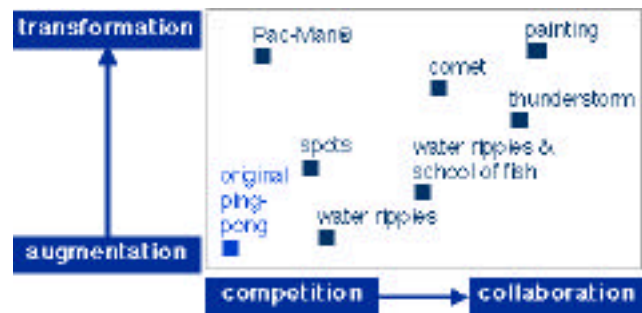


Figure 13 Design Space of PingPongPlus Applications

School of Fish mode was most successful in keeping the attention of both players, and those watching around the table. Even when no one was playing, people enjoyed watching fish swim in a virtual pond.

Although we have focused on the *transformation* of interaction in this paper, we see promising applications in the augmentation of players' performance. We plan to design a ping-pong expert training system using PingPongPlus.

Besides new modes tailored for the PingPongPlus table, there exist a number of extensions that can be made in the realm of *interactive surfaces*. Interactive surfaces absorb information from the physical world, move it into a digital world, process it, and then radiate the results back to a physical world. This is one of the key concepts of the Tangible Bits vision [5]. We plan to use the PingPongPlus sensing system in conjunction with various new wireless sensor technologies to extend the application domain of interactive surfaces.

RELATED WORK

Research in Augmented Reality [1, 19] and Ubiquitous Computing [18] stimulated this work. VIDEOPLACE [7], ALIVE [9], and many other computer-vision based interactive systems have been developed that allow people to use human body motion as a means of interacting with the digital world using vision tracking techniques.

There are also a variety of *virtual reality* (VR) systems [7, 13] which enable people to interact with computational 3D space using a HMD (head-mounted display) and a data glove. AR² Hockey (Augmented Reality AiR Hockey) [11] is a good example of a *mixed-reality* (MR) system for digitally augmented competitive multi-user games. The players of AR² Hockey use physical mallets to hit a virtual puck with a see-through head-mounted display.

CONCLUSION

We have presented the concept of the athletic-tangible interface through the example of PingPongPlus, an augmented ping-pong table. We developed new sound-based ball tracking technology that is robust and inexpensive. Through experiments with various application modes, we explored the design space of interactions with special focus on two axes: augmentation vs. transformation and competition vs. collaboration.

We expect PingPongPlus to suggest new directions to integrate athletic recreation and social interaction with engaging digital enhancements. By the augmentation and transformation of physical games, new, engaging interactions can be developed in the physical/digital world.

ACKNOWLEDGMENTS

We thank the members of Tangible Media Group and our colleagues in the Digital Life and Things That Think Consortia at the MIT Media Laboratory for their support and collaboration. Thanks are also due to Rich Gold at Xerox PARC and Michael Naimark at Interval Research for their valuable comments on an early prototype of PingPongPlus.

REFERENCES

1. Azuma, R., A Survey of Augmented Reality, Presence, Vol. 6, No. 4, August 1997, pp. 355-385.
2. Chan, Y.T., A Simple and Efficient Estimator for Hyperbolic Location, IEEE Transactions on Signal Processing, Vol. 42, No. 8, August 1994, pp. 1905-1915.
3. Gorbet, M., Orth, M. and Ishii, H., Triangles: Tangible Interface for Manipulation and Exploration of Digital Information Topography, in *Proceedings of Conference on Human Factors in Computing Systems (CHI '98)*, (Los Angeles, April 1998), ACM Press, pp. 49-56.
4. Ishii, H. "The Last Farewell": Traces of Physical Presence. interactions 5, 4 (July + August 1998), ACM, pp. 55-56.
5. Ishii, H. and Ullmer, B., Tangible Bits: Towards Seamless Interfaces between People, Bits and Atoms, in *Proceedings of Conference on Human Factors in Computing Systems (CHI '97)*, (Atlanta, March 1997), ACM Press, pp. 234-241.
6. Jacobson, J., et al., The Last Book. IBM Systems Journal, Vol. 36, No. 3, 1997, pp. 457-463.
7. Krueger, M., Artificial Reality II, Addison-Wesley, 1990.
8. Levy, Steven., Hackers: Heroes of the Computer Revolution, Delta Books; February 1994.
9. Maes, P., Darrell, T., Blumberg, B., and Pentland, A., The ALIVE System: Wireless, Full-Body Interaction with Autonomous Agents, *ACM Multimedia Systems*, Special Issue on Multimedia and Multisensory Virtual Worlds, ACM Press, Spring 1996.
10. MacKenzie, C. and Iberall, T., The Grasping Hand, North-Holland, 1994.
11. Ohshima, T., Satoh, K., Yamamoto, H., and Tamura, H., AR² Hockey, in *Conference Abstracts and Applications, SIGGRAPH '98*, ACM, July 1998, pp. 110.
12. Reynolds, C. Flocks, Herds, and Schools: A Distributed Behavior Model, in *Proceedings of SIGGRAPH '87*, ACM Press, pp. 25-34.
13. Rheingold, H., Virtual Reality, Summit Books, 1988.
14. Strang, G., Introduction to Applied Mathematics, Wellesley-Cambridge Press, 1986
15. Ullmer, B. and Ishii, H., The metaDESK: Models and Prototypes for Tangible User Interfaces, in *Proceedings of Symposium on User Interface Software and Technology (UIST '97)*, (Banff, Alberta, Canada, October, 1997), ACM Press, pp. 223-232.
16. Ullmer, B., Ishii, H. and Glas, D., mediaBlocks: Physical Containers, Transports, and Controls for Online Media, in *Proceedings of SIGGRAPH '98*, (Orlando, Florida USA, July 1998), ACM Press, pp. 379-386.
17. Underkoffler, J. and Ishii, H., Illuminating Light: An Optical Design Tool with a Luminous-Tangible Interface, in *Proceedings of Conference on Human Factors in Computing Systems (CHI '98)*, (Los Angeles, April 1998), ACM Press, pp. 542-549.
18. Weiser, M. The Computer for the 21st Century. Scientific American, 1991, 265 (3), pp. 94-104.
19. Wellner, P., Mackay, W., and Gold, R. Computer Augmented Environments: Back to the Real World. *Commun. ACM*, Vol. 36, No. 7, July 1993
20. Wisneski, C., Orbanes, J. and Ishii, H., PingPongPlus, in *Conference Abstracts and Applications, SIGGRAPH '98*, ACM, July 1998, pp. 111.
21. Wisneski, C., Orbanes, J. and Ishii, H., PingPongPlus: Augmentation and Transformation of Athletic Interpersonal Interaction (short paper), in *Summary of Conference on Human Factors in Computing Systems (CHI '98)*, (Los Angeles, April 1998), ACM Press, pp. 327-328.

Vision between action and perception

Giuseppe Boccignone

Dipartimento di Ingegneria dell'Informazione e Ingegneria Elettrica

Università di Salerno and INFN

via Ponte Melillo 1, 84084 Fisciano (SA), Italy

boccig@unisa.it

1 Introduction

Since the time of Aristotle, via Descartes [1], the mind has been regarded as intrinsically sensory in nature: a passive black box that is impressed with input from the environment. In this perspective, vision, conceived as the passive reception of information, became the paradigm exemplar of mental processing. For the common sense, vision is synonymous with sight. Marr himself appears to endorse this conception: in computational vision, classical, reconstructive approaches conceive the task of vision as the construction of a detailed representation of the physical world, so to transform two dimensional data into a description of the three dimensional spatiotemporal world [2]. This "paradigm" is summarized at a glance in Figure 1. However, it has been argued

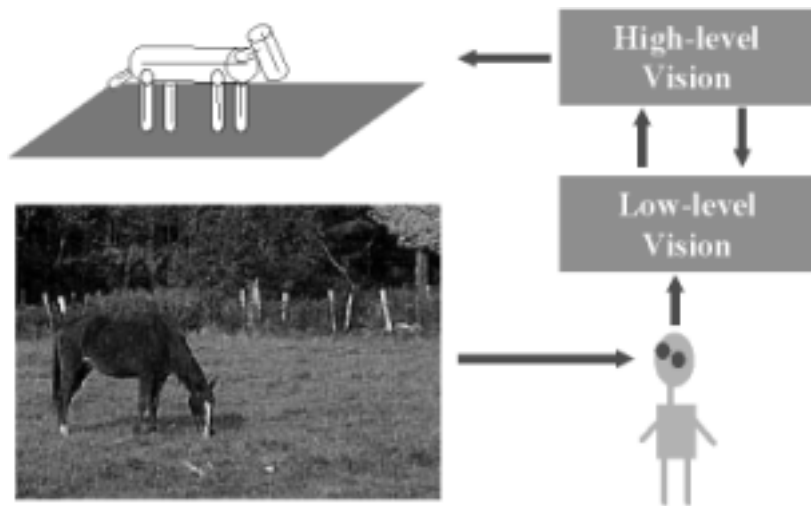


Figure 1: The reconstructive paradigm

that when an agent interacts with its environment, the visual system is used to subserve problem-solving behaviors and such behaviors often do not require an accurate model of the world in the traditional sense. Vision used in behaviors associated with intelligence, has been termed, since Ballard's seminal paper [3], *animate vision*, and has its roots in the theories of robot behavior [4], [5], in the study of lower animal vision [6], and in psychology [7], [8], [9]. Animate vision research investigates ways in which fast,

fluent, adaptive responses can be supported by less computationally intensive routines, which intertwine sensing with acting and moving in the world. An important example, is provided by the use of fast and repeated eye movements to survey a visual scene and to extract detailed information only at selected foveated locations. An average of three eye fixations per second generally occurs during active looking; these eye fixations are intercalated by rapid eye jumps, called saccades, during which vision is suppressed. An example is provided in Figure 2. On the one hand, frequent saccades enable agents to



Figure 2: Left: original image. Center: with each fixation the eye jumps (saccades) from region of interest to region of interest; each region (Focus of attention) is detected at high resolution. Right: composite view, as achieved by the eye. The eye/brain collects both a low-resolution overview using the entire retina and small high-resolution views in the fovea (central region of the retina).

circumvent the need to build enduring and detailed models of the visual surroundings [10]. On the other hand, gaze-shifting capabilities are related to the problem of visual attention. It is well known that the amount of incoming information to the primate visual system is much greater than that which can be fully processed. Only part of this information is processed in full detail while the remainder is left relatively unprocessed [3], [9]. For example, from the high-resolution foveal representation in the retina where most processing resources are allocated to the central 5 degrees of the visual field, to late stages of visual cortical processing where receptive fields invariably grow to encompass the fovea, the neural architecture disproportionately represents the central visual field. Additionally, dynamic mechanisms of selective attention focus the processing resources of the visual system by functioning as an information gating mechanism. Together, attentional mechanisms and neural architecture determine what visual information is or is not processed.

From the general standpoint of attention research, over the last 40 years the major themes have been defined by a small set of metaphors [11]. Two are particularly important in the context of this paper, since providing the rationale (either implicit or explicit) behind many works discussed in the sequel.

The Attention as Filter Metaphor (Broadbent, [12]. According to this position, attention acts as the gateway/filter to memory. Thus, stimuli that pass through the filter compete for perceptual capacity, while an unattended stimulus cannot be stored in memory.

The Attention as Spotlight Metaphor. Given a field with objects in it, the spotlight sheds light over the field. When the spotlight focuses the target object (Focus of Attention, FOA), this is immediately perceived by the subject.

Both metaphors give rise to relevant research questions. For example, considering the filtering metaphor: What are the properties of the filter? What is the locus of such filter:

early (prior to object recognition, but after some initial feature analysis) or late? How much is filtered? Does the filter work as an inhibitory device, with respect to irrelevant information, or as excitatory device? Or, in the case of the spotlight assumption: what are the properties of the FOA? Can a FOA be tuned in size and shape? Can the FOA be split in multiple foci of attention?

Such research questions have animated visual attention research, and related computational models (space-based vs object-based, bottom-up vs. top-down). For an in-depth review of the field, the reader can refer to [13]. All such models in many ways investigate few or more components of the generic architecture outlined in Figure 3. For

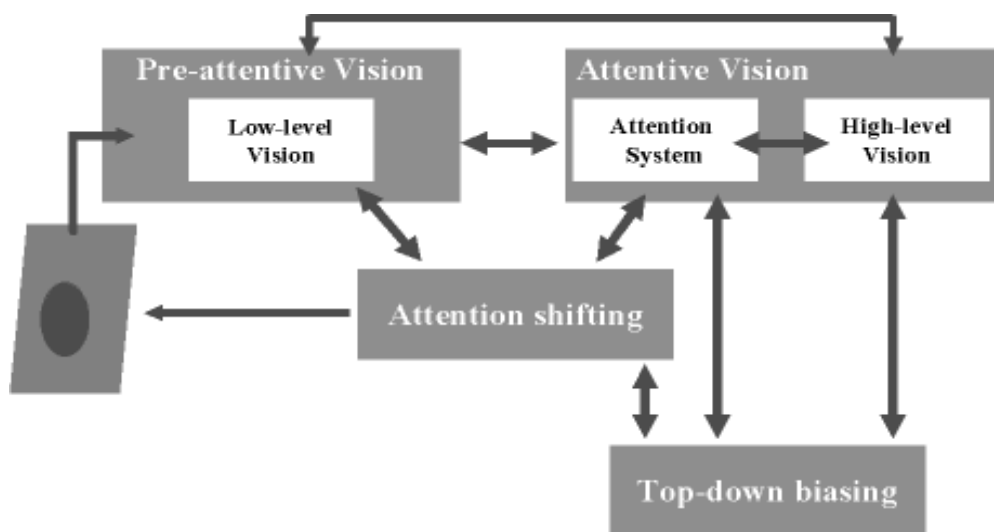


Figure 3: Functional model of an animate vision system

example, in the bottom-up direction, early visual features are computed pre-attentively, across the entire visual field, and the attention component may provide feature integration through a saliency map [14], [15], [16] or associative memory/tables [17], [18]; the attention shifting mechanism can be implemented through a winner-take-all (WTA) strategy [14], [19], or pattern field dynamics [18]. Top-down models might instead use simple top-down biasing in the formation of the saliency map [18],[15], or explicitly use prior knowledge to direct FOA choice [20], and to recognize objects [21], [22], [23]; different forms of more complex behaviors [24], [16] can also be adopted for top-down biasing. It is worth noting, that few works aim at realizing the whole architecture (Fig. 3) [16], [24]. Even few, for instance, consider the problem of attention related to dynamic scenes [15], [18], [16]

Coming back to the initial point, eye movements are an essential part of the whole game, because they must carry the fovea and, consequently, the visual attention to each part of an image to be fixated upon and processed with high resolution. But how is the selection of one particular spatial location accomplished? How is this choice related to the information embedded within an image? Does it involve primarily bottom- up, sensory-driven cues or does expectation of the targets characteristics play a decisive role?

2 Background on eye movements

Beyond the very early studies of eye scanning, by Guy Buswell [25], who first noted that eye fixations in general concentrated on particular areas of the picture, Alfred Yarbus pioneered studies on basic aspects of eye movement control, and recorded the scanning of observers looking at pictures [26] (Figure 4). Yarbus, although at times suggesting that

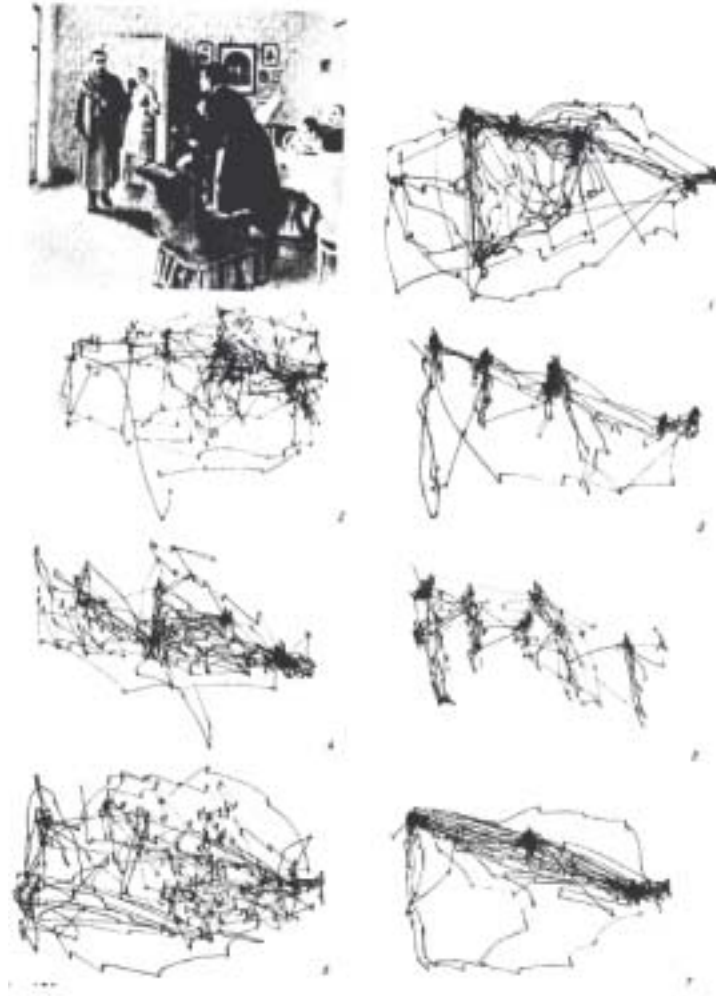


Figure 4: Examples of eye scanning records obtained by Yarbus (1967). Each trace shows a three minute record of eye scanning. Observers were given different instructions: 1) Free examination, 2) estimate the material circumstances of the family, 3) give the ages of the people, 4) surmise what the family were doing before the visitors arrival, 5) remember the clothes worn by the people in the picture, 6) remember the position of the people and objects in the room, 7) estimate how long the "unexpected visitor" had been away.

a scan is first used to form a general impression, used his data to draw the conclusion that additional time spent on perception is not used to examine the secondary elements, but to re-examine the most important elements ([26]). This is quite surprising given the records often comprised an abnormally long viewing period (many seconds) during which there is opportunity for the observers mental set to modify.

Buswell, Yarbus and a number of researchers have accumulated much data on average parameters of the eye movements made during picture scanning. Fixation durations

when viewing pictures and scenes show a skewed distribution, with a mode at 230 ms and a mean of 330 ms [27]. The finding that the mean fixation duration increases as viewing continues was noted by Buswell [25] and has been replicated several times [28]. For example, Antes [29] reported that the mean duration of early fixation was 215 ms and that this increased progressively to 310 ms after a few seconds inspection.

A variety of low level and high level, cognitive factors affect fixation durations during picture viewing. In this perspective, an extreme point of view is the scanpath theory by Noton and Stark [30] who claimed that, i) when a particular visual pattern is viewed, a particular sequence of eye movements is executed and ii) that this sequence is important in accessing the visual memory for the pattern. They obtained some experimental support for their first postulate by showing reproducible scanpaths. The second postulate is more controversial. Walker-Smith et al. [31] recorded scanning patterns when observers carried out tasks involving facial recognition. When an individual test face was examined to determine a match to a memorized face, some reproducible scanning sequences were observed. However, when observers were asked to make a direct comparison between two simultaneously presented faces, there was no suggestion that a scanpath occurred on one face at a time. Rather a repeated scanning between the two faces, resembling a feature by feature comparison, occurred.

Closely related to the Spotlight metaphor, is how visual attention may be object-driven. Indeed, the visual world presents a scene, which in general contains a number of well defined and often well located objects. Individual objects viewed foveally can be recognized rapidly and generally within a single fixation. The process by which this occurs has been, and is still a matter of debate but all contenders accept that foveal recognition of single objects is an area where the passive vision approach, i.e. massive parallel processing of retinal information, is legitimate. Nevertheless, object recognition deteriorates quite rapidly in the parafovea and periphery. Nelson and Loftus [32] allowed subjects a limited amount of time to scan a picture and then performed a recognition test, viewing a version of the picture in which one object had been changed. Their task was to detect the changed object. Detection rate for small objects (1 degree) was above 80% for objects that had been directly fixated and higher if they received two fixations. However it fell to 70% for objects where the closest original fixation had been at a distance from the object in the range 0.5 degrees - 2 degrees, and to a chance level for objects viewed more peripherally. This suggests that the gaze needs to be directed to within 2 degrees of an object in a scene for it to be reliably encoded.

It is worth noting that scene perception involves more global properties of the visual world than simply the collection of objects. Object perception can be studied in isolation from scene perception but it is less clear that the converse is true. It is not in general possible to envisage a scene devoid of objects, which creates a practical problem in disentangling scene perception from object perception. The scene provides context for the objects and one question that has received extensive investigation is whether objects are more readily perceived in an appropriate scene context.

Biederman [33] has argued that scene information is captured within a single glimpse. However, as in the case of visual search, only a limited number of objects can be processed in parallel. On the other hand, Intraub [34], using a set of 250 pictures cut from magazines, compared recognition memory after a subset had been shown. With 6 seconds viewing time, recognition rates were 94%; with tachistoscopic exposure durations, the rate fell to about 80% (and the false positive rate rose from 8% to 11%). This con-

firms that the memory ability for material shown in a single glimpse is impressive, but the opportunity to inspect a picture with eye movements is beneficial. Further, Henderson et al. [27] experiments give evidence that peripheral information can be usefully extracted from pictorial material viewed in isolation, although only at the destination point of a planned saccade.

The overall conclusions from this work are that objects viewed foveally or in the close parafovea can be identified and categorized very rapidly. Evidence suggests that objects not fixated closer than 2 or 3 degrees are not recognized and thus eye movements are in general necessary for the identification of objects within scenes. A certain amount of scene information can be extracted from a single glimpse of a scene and this allows the scene schema to be evoked with very little delay, but it is likely that eye movements over a scene will provide additional information, as well as information about specific objects.

A rather alternative view, that our subjective impression of an immediate pictorial reality is illusory, has received increasing support from recent studies. In the experiments reported by Grimes [35], an observer views a picture and at some point whilst a saccade is in progress, aspects of the picture were changed. Participants were told to expect a memory test and were also forewarned that something in the scene might occasionally change and were asked to report any such changes. Surprisingly large changes were often undetected. Thus, in a picture of two men wearing hats, swapping the hats was never detected. Movement of a child in playground scene, involving an image size change of 30% showed a detection rate of under 20%. Failure to detect changes in Grimes experiment suggests strongly that, contrary to expectation, the relevant information for detecting changes is just not available in any visual representation of the scene. If this were so, making the change simultaneous with a saccade would not be important. Of course if changes are made to a scene during a fixation, visual transients are introduced and these transients have the effect of directing attention to the area of the change. This striking phenomenon has been termed *change blindness*. One effective approach has been to mask a scene change by making several simultaneous conspicuous changes at different locations in the scene (*mudsplash* technique). Investigations using this technique [36] have shown that surprisingly large changes can fail to be detected. Notice that change blindness is clearly incompatible with the picture in the head account which we have termed passive vision.

An important series of studies witnessing how eye movements support every day action has been carried out by Michael Land. He devised and built a light head mounted video based eye tracking system [37], which enabled a record to be built up of the fixation positions adopted by an observer during a variety of active tasks. Tasks studied have ranged through driving [37], table tennis [38], piano playing [38] and tea-making [39]. In certain cases, the eye records have revealed, in an unexpected way, how vision is used. For instance, while playing table tennis, the eyes are very active. Their activity takes roughly the same path as the ball. Contrary to popular belief, they do not follow the ball, but the eye works in an anticipatory way: as the opponent makes the return shot, the player fixates the top of the net, using the clearance at this point to judge the balls trajectory.

3 Theoretical models based on information theory and statistical physics

It is clear from the above discussion, that, from a general point of view, what meets the fovea is only a part of what meets the (minds) eye. Even more fundamental is the fact that informativeness depends on the cognitive task being undertaken. Thus information-based eye scanning should be properly matched to an articulated theory of cognitive activity for the task in question. Yet, it is important to model bottom-up processes, since they might describe how attention is deployed the first few hundreds of milliseconds after the presentation of a new scene [40]. In that case, the attention mechanism can be articulated in two steps: i) compute some measure of saliency over the images; ii) generate attentional shifts, to move the FOA on points of interest.

3.1 Eye movements and information

Buswell and Yarbus both reported that eye fixations tended to fall on the important details of pictures. Mackworth and Morandi [41] generated an informativeness measure for different regions of a picture, by cutting it up into 64 separate segments. A separate group of 20 observers were shown these segments individually and asked to rate how informative they appeared to be using a nine-point scale. The informativeness rating obtained for a region was highly predictive of the probability that the region would be fixated when an observer viewed the picture in a preference task.

It is tempting to believe that the idea of informativeness might be the key that allows cognitive activity to be revealed from patterns of eye scanning. A computational example of such an approach has been given by Ferraro et al. [42]. In order to compute the "context-free" information embedded in a picture, the model assumes that the message source is represented by a dynamical system, namely the pattern together with a given transformation. Images are considered as isolated thermodynamical systems, by identifying the image intensity with some dynamical variable, e.g. temperature or concentration of particles, evolving in time. According to thermodynamics, information in physical systems is related to the rate of variation of the entropy H , which is usually written as $dH/dt = dH_e/dt + dH_i/dt$, where dH_e/dt is the variation due to external sources, while the entropy production term $dH_i/dt \geq 0$ is generated by changes inside the system; for the case of isolated systems, $dH/dt = dH_i/dt$. In general, dH_i/dt can be defined by spatial integration of entropy production density σ , which is a function of both position and time, along a fine to coarse transformation realized through a diffusion process (scale space). In the case of grey level images, entropy variation along scales is used to characterize basic, low-level information and to identify perceptual components of the image, such as shape and texture. For dealing with natural scenes, an extension of the approach to color images has been proposed [43]. For each color channel, the transition from fine to coarse scales is obtained through the generalized diffusion equation $\frac{\partial f_i}{\partial t} = -\text{div} \vec{J}_i$, f_i being the intensity of the i -th color channel. For each i , the flow density is given by $\vec{J}_i = \sum_{j=1}^n L_{ij} \vec{X}_j$. Interactions among color components are modelled by setting $\vec{X}_i = \nabla \left(\frac{1}{f_i(x,y,t)} \right)$ and $L_{ij} = \kappa_{ij} f_i f_j$, where κ_{ij} , are coefficients weighting the strength of the interactions between channels i and j . Then the fine to coarse transformation is ruled by the system of coupled evolution equations, $\frac{\partial f_i}{\partial t} =$

$-\text{div} \left(\sum_j L_{ij} \vec{X}_j \right) = \sum_j \nabla \cdot \left(\kappa_{ij} f_i f_j \frac{\nabla f_j}{f_j^2} \right)$, which can be developed as [43]:

$$\begin{aligned} \frac{\partial f_i}{\partial t} &= \nabla^2 f_i + \sum_{j \neq i} \kappa_{ij} \\ &\times \left\{ \frac{f_i}{f_j} \nabla^2 f_j + \frac{1}{f_j^2} \left[f_j \left(\frac{\partial f_i}{\partial x} \frac{\partial f_j}{\partial x} + \frac{\partial f_i}{\partial y} \frac{\partial f_j}{\partial y} \right) - f_i \left(\frac{\partial f_j}{\partial x} \frac{\partial f_j}{\partial x} + \frac{\partial f_j}{\partial y} \frac{\partial f_j}{\partial y} \right) \right] \right\}. \end{aligned} \quad (1)$$

Hence, local spatio-chromatic entropy production $\Sigma(x, y, t) = \sum_{i,j} L_{ij} \vec{X}_i \cdot \vec{X}_j$ is given by:

$$\Sigma = \sum_i \frac{\nabla f_i \cdot \nabla f_i}{f_i^2} + \sum_{i \neq j} \kappa_{ij} \frac{\nabla f_i \cdot \nabla f_j}{f_i f_j}, \quad (2)$$

the first terms $\frac{\nabla f_i \cdot \nabla f_i}{f_i^2}$ accounting for the density of entropy production for every channel i , considered in isolation, and the cross terms $\kappa_{ij} \frac{\nabla f_i \cdot \nabla f_j}{f_i f_j}$ holding for the dependence of entropy production on interactions among channels.

It has been shown that entropy Σ can be used to form a conspicuity map which represents the relevance of image points. An activity function is defined as $a_\Sigma(x, y) = \int_0^\infty \Sigma(x, y, t) dt$ and the map $a_\Sigma : (x, y) \rightarrow a_\Sigma(x, y)$ represents the activity map. Next, a transformation $\mathcal{C} : a_\Sigma \rightarrow \mathcal{C}(a_\Sigma)$ is applied to a_Σ , which approximates a lateral inhibition mechanism [14]. Eventually $\mathcal{C}(a_\Sigma)$ is the *information-based conspicuity map*. In [42], \mathcal{C} has been fed into a dynamical neural network, a 2D layer of leaky integrate-and-fire neurons representing a dynamical saliency map (DSM) as proposed by Itti et al. [14]. Figure 5 depicts the scanpath computed by the visual attention module on the "Horse" image and on a region of interest, which has been previously foveated in the same image.

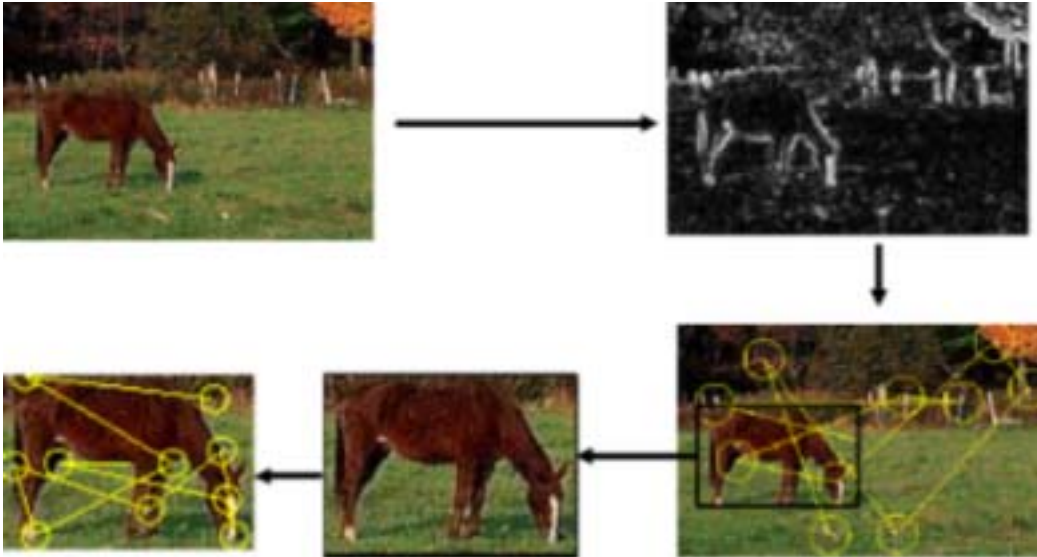


Figure 5: The original picture (top left), the entropy-based conspicuity map (top right), the scanpath on the original image (bottom right), and a refined scanpath on a region of interest (the horse) extracted from the same image. Circles represent the foveation areas.

3.2 Physics-based models for gaze-shifts

Arising in part from interest in scanpaths, a number of workers have developed techniques to capture statistical regularities in the pattern of eye scanning. One approach has considered statistical dependencies in parameters such as saccade direction or fixation location [44]. The simplest form of such a sequential dependence is the Markov process in which the properties of the immediately preceding saccade constrain the probabilities of the one currently programmed. Markov analysis has been proposed as a way of controlling robot vision by Rimey and Brown [45]. Several papers have explored the relationship between scanning statistical properties and image statistics [46], [47].

Although probabilistic approaches have been applied to various problems in eye movement research, the specific functional form of saccadic magnitude distributions has attracted surprisingly little attention. An exception is represented by the work of Brockmann and Geisel who proposed a phenomenological model for the generation of human visual scanpaths [48]. Successions of saccadic eye movements are treated as realizations of a stochastic jump process in a random quenched salience field. Based on the assumption that the visual system minimizes the typical time needed to process a visual scene, the theory predicts that scanpaths are geometrically similar to a prominent class of random walks known as Levy flights. Interestingly enough, Levy flights play a role in systems ranging from global climate fluctuations to animal foraging behavior [49]

The dynamical quantity of interest in random walk models is the probability of fixating a location x at discrete times t in a visual field of size L . The transition probability density of shifting the gaze from a point y to x , $\rho(y \rightarrow x)$ is defined as the product $\rho(y \rightarrow x) \propto \mathcal{S}(x)f(|x - y|)$ of a random quenched salience field $\mathcal{S}(x) > 0$, which quantifies the salience at the target location x , that is the probability of attracting the gaze, and a term $f(|x - y|)$, the probability of generating a saccade of magnitude $|x - y|$, averaged over the salience field ensemble. In particular, the Cauchy-Levy distribution is experimented,

$$f(|x - y|) = \frac{D_C}{\pi(D_C^2 + |x - y|^2)} \quad (3)$$

The authors provide interesting results on simulated random salience fields.

This approach is certainly a promising research direction provided that effective algorithms and actual salience field are taken into account. On the one hand, it grounds on a solid theoretical foundation. But, most important, it has the advantage of accounting for both the structural information embedded in $\mathcal{S}(x)$, and the random component which is implicit in eye movements (see, e.g. [50]).

To push further the approach we have experimented with the following algorithmic dynamical procedure, which in the same vein combines the use of a saliency field in the form of a conspicuity map $\mathcal{S} \equiv \mathcal{C}$ and a flight length probability drawn from different densities (purely random, gaussian, Cauchy's).

Set $y \leftarrow$ image center

$n_{FOA} \leftarrow 0$

repeat

Set y as the current fixation point (FOA)

Generate randomly a jump length L with probability $f(|x - y|)$

```

    Choose  $x$  at length  $L$  from  $y$ , which maximizes the saliency field  $\mathcal{C}(x)$ 
    Store  $x$  and set  $y \leftarrow x$ 
     $n_{FOA} \leftarrow n_{FOA} + 1$ 
until  $n_{foa} \leq N$ 

```

where n_{foa} is the number of FOA computed and N the maximum number of FOA. Some preliminary results are shown in Figure 6. An interesting question is how to measure

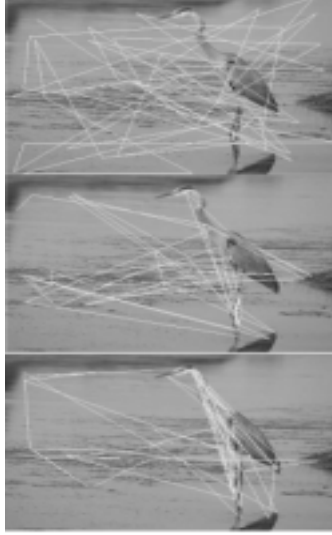


Figure 6: Top image: the scanpath obtained using unconstrained random walk. Center: the scanpath obtained through the proposed algorithm (Gaussian jump probability). Bottom: another scanpath obtained through the proposed algorithm (Levy-Cauchy jump probability). 50 fixations have been considered

the information provided by the process. A viable method is to consider the symbol dynamics provided by the system. For instance, it is possible to code the sequence of fixation in a finite string s [50]. Then, given s the problem turns in determining the quantity of information $\mathcal{I}(s)$. In principle, one such measure is the algorithmic information content, or Kolmogorov's complexity [51], that is the length of the smallest binary program p , which given in input to a partial recursive function (universal computer) \mathcal{U} , generates s :

$$\mathcal{I}(s) = \min\{|p| : \mathcal{U}(p) = s\} \quad (4)$$

Unfortunately, $\mathcal{I}(s)$ is not computable, so suitable approximations have to be investigated [52], which is matter of on-going research.

4 An example: animate queries in Content Based Image Retrieval

Retrieval is performed by using the visual content of multimedia data, usually considering some low/medium level features, such as colors, texture, spatial position, objects shape. Content Based Retrieval Systems (*CBRS*) are in general characterized by means of the capability of supporting image retrieval relying upon a specific image similarity model. The problem of assessing the similarity between two images can be reformulated

as a task of visual search: given a target image I^q and a test image I^i , is there an instance of the target in the test image? To this end an appealing solution could be that of exploiting the natural parameters of visual computation.

In a recent work [53], it has been argued that digital image retrieval performances can gain advantage from mechanisms through which human beings perceive visual similarity between images. The animate framework may play a twofold role: the selection of certain aspects of the input stimulus while causing the effects of other aspects of the stimulus to be minimized (filtering metaphor); the introduction of the dimension of time: features and relationships are not established as static structures, but are incrementally set-up along the visual inspection task. In other terms, attention "linearizes" the 2D structure, naturally reducing visual matching complexity.

In order to compare a target image I^q with a test image I^i , two scanpaths (FOA sequences) are extracted, respectively, from I^q and I^i . This step is performed by using the Itti and Koch model [13]. To provide a similarity measure of the two images, each scanpath undergoes further processing/measurement, so that relevant information/features are derived: the "flow" of such features, is denoted Information-Path (IP) (see Fig. 7). The chromatic similarity of two $FOAs$, respectively $C^q(p)$ and $C^i(p)$, is

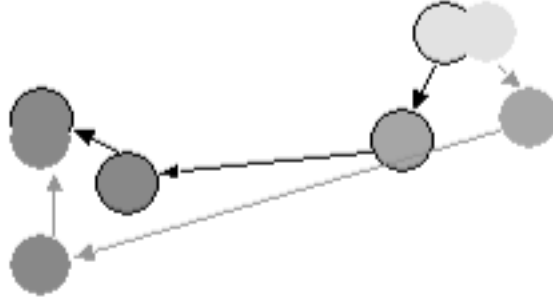


Figure 7: Information Paths of two similar images, indicated using light and dark lines, respectively

evaluated through color histogram intersection [54]. Given two color histograms of a target and a test FOA , respectively $h(C^q(p))$ and $h(C^i(p))$, on the same number of bins $b = [0, \dots, B]$, it is possible to define a similarity measure

$$\mathcal{M}_{col} = 1 - \frac{\sum_b (\min(h_b(C^q(p)), h_b(C^i(p'))))}{\sum_b h_b(C^q(p))}. \quad (5)$$

To evaluate FOA 's similarity in terms of shape and texture, the following distance is used:

$$\mathcal{M}_{tex} = 1 - \sum_{j \in \mathcal{N}(p, p')} \frac{\text{abs}(Cov_{WT}(C^q(p)) - Cov_{WT}(C^i(p')))}{\min(\text{abs}(Cov_{WT}(C^q(p)), \text{abs}(Cov_{WT}(C^i(p'))))}, \quad (6)$$

$Cov_{WT}(C^q(p))$ and $Cov_{WT}(C^i(p'))$ being the wavelet transform's covariance signatures of target and test $FOAs$ respectively. FOA 's content similarity is computed through the weighed mean of these two terms, $\mathcal{M}_{content} = \mu_1 \mathcal{M}_{tot} + \mu_2 \mathcal{M}_{tex}$

To evaluate the IP similarity in terms of spatial position, the Euclidean distance between homologous $FOAs$'s centers, $d_{p, p'}$, p and p' being FOA center coordinates. This distance is "penalized" if, for the two images, the movement between the current

FOA and the next one is not in the same direction; thus, $\hat{d}_{p,p'} = d_{p,p'} \cdot e^{-\Delta}$, Δ being the difference of direction between two $FOAs$, namely $\Delta = \Omega \cdot \text{sign}[(x_{I_q}^j - x_{I_q}^{j-1}) \cdot (x_{I_i}^j - x_{I_i}^{j-1})] \cdot \text{sign}[(y_{I_q}^j - y_{I_q}^{j-1}) \cdot (y_{I_i}^j - y_{I_i}^{j-1})]$. Ω represents the penalization constant and $(x_{I_q}^j, y_{I_q}^j), (x_{I_i}^j, y_{I_i}^j)$ centers coordinates relative to j -th $FOAs$ of I_q and I_i images. Thus, after $\hat{d}_{p,p'}$, normalization, $\mathcal{M}_{spatial} = 1 - \hat{d}_{p,p'}$.

Eventually, FOA similarity is given by the *weighted mean* of $FOAs$ content similarity and $FOAs$ spatial similarity, $\mathcal{M}_{FOA} = \alpha \mathcal{M}_{content} + \beta \mathcal{M}_{spatial}$, where $\alpha, \beta \in [0, 1]$. Clearly, due to variations of lighting conditions, pose, different background, one should expect a certain degree of variability in the feature range on similar FOA . Thus, a fuzzyness degree is associated to the matching result between homologous $FOAs$.

The similarity or matching \mathcal{M}_{image} is obtained through the following matching algorithm. At each step the algorithm compare homologous $FOAs$ of two images, evaluating image similarity in terms of nodes's content and spatial position. In a first step the algorithm compare the first k target image $FOAs$, $k < K$ where K is the number of extracted $FOAs$, with the homologous $FOAs$ of the various images in database. To start, a FOA clusters considered. At the end of this first match a fuzzy IP similarity measure is obtained. For all images that have a fuzzy similarity value which falls in a confidence region CR the algorithm stops. The algorithm can be stopped also in the case of very dissimilar $FOAs$. In the other cases, the matching goes on FOA by FOA . The algorithm stops if the matching value falls in a CR or if all nodes are examined, returning a *similarity* or *not similarity* answer according to a threshold T .

In order to take into account the difference of time that a human eye spend on two $FOAs$ pertaining to two different images (evaluated in terms of *firing times* of the WTA net used to determine gaze-shifts [14]), the image will be decomposed in a data flow of activating tokens represented by $FOAs$. The indexing mechanism exploits the sequenced arrival of tokens to rapidly recognize the query image as similar to the image present in the DB. In turn, the indexing mechanism can modify the latency of tokens by delaying those whose effects are incompatible with the current interpretation state. Thus, after normalization, the time similarity is introduced, $\mathcal{M}_{time} = 1 - \text{abs}(t_{C^q} - t_{C^i})$, t_{C^q} and t_{C^i} being the *WTA firing times* relative to homologous $FOAs$ of images I^q and I^i . Eventually, FOA similarity becomes:

$$\mathcal{M}_{FOA} = \alpha \mathcal{M}_{content} + \beta \mathcal{M}_{spatial} + \gamma \mathcal{M}_{time} \quad (7)$$

where $\alpha, \beta, \gamma \in [0, 1]$. Eventually, the matching algorithm is the following.

Time-based IP Matching Algorithm.

Compute the Information Path of the Target Image

$j \leftarrow 0$

while ($\neg stop \wedge j < K$)

do Compute content similarity $\mathcal{M}_{content}$ between $C_j^q(p)$ and $C_j^i(p')$

 Compute spatial position similarity $\mathcal{M}_{spatial}$ between $C_j^q(p)$ and $C_j^i(p')$

 Compute time similarity \mathcal{M}_{time} between $C_j^q(p)$ and $C_j^i(p')$

 Compute FOA similarity \mathcal{M}_{FOA} between $C_j^q(p)$ and $C_j^i(p')$

$j \leftarrow j + 1$

if ($j=k$) **then**

 Compute the *arithmetic mean* of the first three $FOAs$ similarity

```


$$\mathcal{M}_{image} = \frac{1}{k} \sum_{f=1}^k \mathcal{M}_{FOA_f}$$

if ( $\mathcal{M}_{image} < \text{thresholdMIN}$ )  $\vee$  ( $\mathcal{M}_{image} > \text{thresholdMAX}$ ) then
    stop
if ( $j > k$ ) then
    Compute the arithmetic mean of  $j - 1$  FOAs similarity
    
$$\mathcal{M}_{image} = \sum_{z=1}^j \mathcal{M}_{FOA_z} / j$$

    if ( $\mathcal{M}_{image} < \text{thresholdMIN}$ )  $\vee$  ( $\mathcal{M}_{image} > \text{thresholdMAX}$ ) then
        stop
if ( $\mathcal{M}_{image} > T$ ) then
    target image is similar to test image

```

A retrieval example is shown in Fig. 8. Another example of bottom-up process, but



Figure 8: Animate query results. Left: query image (bear). Right: a collection of retrieved images

applied to video analysis is reported in a note presented at this Workshop [55]. A recent application to foveated video compression has been proposed by Bovik [56].

5 The role of learning and behavior

Actually, limiting the discussion on gaze-shift mechanisms from the standpoint of static scene analysis, though important, would result in a rather narrow view of the problem. A crucial function of the brain is to achieve appropriate responses to complex situations. Examining a scene involves, as discussed, sequential FOA selection at the rate of about three per second. Thus, vision in humans involves continual sequential interactions with the world, while saccades are related to observer time-dependent problem solving. As remarked by Ballard [57], “*At 300 ms timescale cognition appears as sequential programs that depend on moment-by-moment interactions with the environment*”. Sequentiality costs in time, but leads to enormously compact information encoding, namely in the sense of Kolmogorov [51].

This question brings back to how such programs may be learned. At the 300-ms timescale, from a biological perspective, learning models can be conceived as a set of actions that are available for each of a set of discrete states; such actions are probabilistic

and the value of taking an action is increased/decreased using a model of (chemical) reward. Also, rewards may be delayed. To fulfil these constraints, different approaches are viable. A useful way, which handles uncertainty by allowing probabilistic transitions between states, is provided by hidden Markov models [45] or, analogously, in the optimal control formulation, by Kalman filtering [23], [58].

Unfortunately, Markov models cannot cope with agent's actions that change the world. This task requires an extension to Markov decision processes (MDP), that can incorporate/learn the different actions available to the agent. One such tool is reinforcement learning (RL).

5.1 Reinforcement learning

RL is a kind of MDP ([59]), in which, unlike HMM, actions are performed under control of the agent. Define a finite set of states S and a set of actions A . At each discrete time, the agent observes state $s_t \in S$ and chooses action $a_t \in A$, then receives immediate reward r_t and his state changes to s_{t+1} . The Markov assumption holds, $s_{t+1} = \delta(s_t, a_t)$ and $r_t = r(s_t, a_t)$, i.e., r_t and s_{t+1} depend only on *current* state and action. Note that functions δ and r may be nondeterministic and not necessarily known to agent.

An agent's learning task can be defined as follows: execute actions in environment, observe results, and learn the action policy $\pi : S \rightarrow A$ that maximizes $E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots]$ from any starting state in S . Here $0 \leq \gamma < 1$ is the discount factor for future rewards. In the deterministic case, for each possible policy π the agent might adopt, we can define an evaluation function over states

$$V^\pi(s) \equiv r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots \equiv \sum_{i=0}^{\infty} \gamma^i r_{t+i} \quad (8)$$

where r_t, r_{t+1}, \dots are generated by following policy π starting at state s . Restated, the task is to learn the optimal policy π^* , where $\pi^* \equiv \arg \max_{\pi} V^\pi(s), (\forall s)$.

We might try to have the agent learn the evaluation function V^{π^*} (which we write as V^*). The agent could then perform a lookahead search to choose best action from any state s since $\pi^*(s) = \arg \max_a [r(s, a) + \gamma V^*(\delta(s, a))]$. One problem here is that this works well if agent knows $\delta : S \times A \rightarrow S$, and $r : S \times A \rightarrow \mathbb{R}$. But when it doesn't, it can't choose actions this way. Then, one can define a new function very similar to V^*

$$Q(s, a) \equiv r(s, a) + \gamma V^*(\delta(s, a)) \quad (9)$$

If agent learns the evaluation function Q , it can choose optimal action even without knowing δ . In fact, $\pi^*(s) = \arg \max_a [r(s, a) + \gamma V^*(\delta(s, a))] = \arg \max_a Q(s, a)$. Note that Q and V^* are closely related, $V^*(s) = \max_{a'} Q(s, a')$, which allows us to write Q recursively as $Q(s_t, a_t) = r(s_t, a_t) + \gamma \max_{a'} Q(s_{t+1}, a')$. Let \hat{Q} denote learner's current approximation to Q . Consider the training rule

$$\hat{Q}(s, a) \leftarrow r + \gamma \max_{a'} \hat{Q}(s', a') \quad (10)$$

where s' is the state resulting from applying action a in state s . Thus, the learning algorithm is the following:

```

for each  $s, a$  initialize table entry  $\hat{Q}(s, a) \leftarrow 0$ 
Observe current state  $s$ 
do forever
  Select an action  $a$  and execute it
  Receive immediate reward  $r$ 
  Observe the new state  $s'$ 
  Update the table entry for  $\hat{Q}(s, a)$  using eq. 10
   $s \leftarrow s'$ 

```

A central idea of RL is called temporal difference (TD) learning. TD methods are general learning algorithms to make long-term predictions about dynamical systems. They are based on reducing discrepancy between successive Q estimates. For instance, consider, one step time difference, $Q^{(1)}(s_t, a_t) \equiv r_t + \gamma \max_a \hat{Q}(s_{t+1}, a)$, then two step difference $Q^{(2)}(s_t, a_t) \equiv r_t + \gamma r_{t+1} + \gamma^2 \max_a \hat{Q}(s_{t+2}, a)$. Thus, $TD(\lambda)$ learning can be written in the form of the following training rule:

$$Q^\lambda(s_t, a_t) = r_t + \gamma[(1 - \lambda) \max_a \hat{Q}(s_t, a) + \lambda Q^\lambda(s_{t+1}, a_{t+1})] \quad (11)$$

$TD(\lambda)$ algorithm sometimes converges faster than Q learning; also, it converges for learning V^* for any $0 \leq \lambda \leq 1$.

An application in an active vision context is reported by Minut and Mahadevan [60]. States are defined as clusters of images representing the same region in the environment. Each image is represented by color histograms [54] computed in the within-fixation processing. Agent's actions a_1, a_2, \dots, a_s represent saccades to the most salient points in one of the eight 45 central angles in the image. The agent receives positive reward for a saccade that brings the target object in the field-of-view, and a small negative reward otherwise

The model consists of two interacting modules. A top module, which uses RL, trains on a set of clusters (states) consisting of images with similar color histograms. This module learns a policy for saccading from one region to another towards the region(s) most likely to contain the target object, by selecting the gaze direction according to its utility (Q-value). The second module consists of low-level visual routines and performs within fixation processing by computing either two saliency maps (color and symmetry maps) providing the agent with a set of locations of interest in the image, and identifies the target object. An example of finding a target in a cluttered environment is shown in Figs. 9 and 10.



Figure 9: Finding an object in a cluttered environment (adapted from [60])



Figure 10: Learned sequence to reach the object (adapted from [60])

5.2 Vision and action

The various situations discussed up to this point have had in common the fact the observer was simply taking in information from the visual environment. Although such observing behavior occurs frequently in everyday life, a situation that is probably even more common is one in which the viewer is also engaged in carrying out some action. Under these circumstances, the pattern of eye scanning must be integrated into an overall action sequence.

For example, Lands analysis of tea making previously discussed [39] proposes the concept of object-related action units (in the tea-making case, find the kettle and transport to sink are examples of such units). These units, with very rare exceptions, are carried out sequentially and involve engagement of all sensorimotor activity on the relevant object or objects. The eyes move to the object, or the point at which the objects activity is directed, before the manipulation starts. In general the eyes anticipate the action by about 0.6 sec. A highly important feature of these analysis is the potential for a long overdue integration of work in eye scanning with emerging theories of sequential action. Some recent studies have explored this integration .

Ballard et al [10] devised an artificial manipulative task (block assembly) carried out by mouse-controlled manipulations of a computer screen display. The set up allowed a detailed record to be kept of both the manipulative actions and the eye scanning of the individual carrying out the action. The subject operates on a computer display and has the task of assembling a copy of the model in the workspace. To do this it is necessary to operate with the mouse, obtaining blocks from the workspace using a click and drag procedure.

Here the eyes are used *deictically*. The term deictic refers to the intrinsic ability of a certain action to serve as a pointer to information. From general considerations of computational theory, Ballard et al. argue that the use of pointers is essential for a cognitive system to operate. Eye pointing is one such way.

Fixation on an object allows the brains internal representations to be implicitly referred to an external point . The deictic strategy employs this pointer as part of a general *do it where I am looking* strategy to select objects for action. This is the crucial reason for directing the eyes to an object and of course is supplemented by the enhanced visual resolution that occurs through foveal vision. Further additional pointers relate to purely memory-based activity. However, a consistent finding that emerges in part from the experimental investigations presented in the paper is that use of memory pointers is avoided if an alternative is available.

The data collected from the block assembly task supported this characterization of the underlying cognitive operations. Blocks were invariably fixated before they were operated on. Furthermore, there was clear evidence that the preferred strategy involved making minimal demands on any internalized memory. In the block assembly task, as has been found in other tasks, many more saccades were made than appear necessary on a logical basis. The most common sequence observed in the block assembly task was eye-to-model, eye-to-resource, pick-from-resource, eye-to model, eye-to-construction, drop-at-construction.

The authors conclude that memory minimization is a significant feature of activity in the situation. Cognitive representations (in this case the position of the block in the model) are computed as late as possible before the necessary action. This *just-in-time* strategy, it is argued, minimizes both memory and computational loads.

These results are in complete contrast with the traditional passive vision account involving intensive computation of an internalized representation. Representations are computed only as they are needed.

6 Concluding remarks: an evolutionary perspective

If vision has to play the role of paradigm exemplar of mental processing, then the animate approach suggests a limited use of representations (e.g., recall Ballard's *just-in-time* policy). This result is an explicit challenge to classical, symbolic AI. For instance, the models proposed in "situated" robotics (e.g., Brooks' subsumption architecture, [61]) appear as a radical alternative to the classical AI robotics approach, since interacting with the environment without being supervised by a centralized control and action planning, as is the case in AI robotics. In this sense, Brooks' extreme claim of intelligence without reasoning [4] was close to Gibson's conception [7] that the agent has direct perception of the world, without the mediation of some representation. This sort of active and fierce debate, which has deeply touched classical AI [62], might come to a point, if the whole matter is considered from an evolutionary perspective. Indeed, the ability to orientate rapidly towards salient objects in a cluttered visual scene has evolutionary significance. For instance, it may allow the observer to detect quickly possible prey, predators, or mates. In this sense, the effectiveness in directing gaze rapidly towards object of interest, is likely to be the most important function of selective visual attention. The assumption that the main function of the visual system is the construction of some sort of internal model or percept of the external world is a quite narrow view of the whole story. The visual control of much more complex behaviors, such as reaching out and grasping an object, also appear to depend on mechanisms that are functionally and neurally separate from those mediating our perception of that object. Indeed, the origins of vision may be related more to its contribution to the control of

action than to its role in conscious perception, a function which appears to be a relative newcomer on the evolutionary scene ([63],[57]).

As Goodale and Humphrey argued [63], this fact becomes apparent from an evolutionary standpoint: vision evolved in animals, not to enable them to see the world, but to guide their movements through it. Indeed, the visual system of most animals, rather than being a general-purpose network dedicated to reconstructing the rather limited world in which they live, consists instead of a set of relatively independent input-output lines, or visuomotor modules, each of which is responsible for the visual control of a particular class of motor outputs. Some of the most compelling demonstrations have been provided by experiments with so-called *rewired* frogs (the retinotectal projections projecting to the optic tectum on the same side of the frogs brain instead of to the optic tectum on the opposite side, as is the case in the normal animal). Rewired frogs showed mirror-image predator avoidance and jumped towards rather than away from the looming visual stimuli, while showing quite normal visually-guided barrier avoidance as they locomoted from one place to another, even when the edge of the barrier was placed in the visual field where mirror-image feeding and predator avoidance could be elicited. Thus, it would appear that there are at least two independent visuomotor systems in the frog: a tectal system, which mediates visually elicited prey-catching and predator-avoidance, and a pretectal system which mediates visually guided locomotion around barriers [63]. These results do not fit well with the common view of a visual system dedicated to the construction of a general-purpose representation of the external world.

Visuomotor modularity of the kind found in the frog also exists in the mammalian brain [63]. However, the very complexity of day-to-day living in many mammals, particularly in higher primates, demands much more flexible organization of the circuitry. In monkeys (and thus presumably in humans as well), there is evidence that many of the phylogenetically ancient visuomotor circuits that were present in more primitive vertebrates are now modulated by more recently evolved control systems in the cerebral cortex. The idea of separate visuomotor channels is consistent with animate vision [3], [5].

Although the need for more flexible visuomotor control was one of the demands on the evolving primate brain, another was related to the need to identify the objects, to understand their significance and causal relations, to plan a course of action, and to communicate with other members of the species. In short, the emergence of cognitive systems and complex social behavior created a whole new set of demands on vision and the organization of the visual system. Direct sensory control of action was not enough, as interactions with the world become more complicated and subtle. In fact, humans and other primates, behave as though their actions are driven by some sort of internal model of the world in which they live. The representational systems that use vision to generate such percepts of the world must carry out very different transformations on visual input than the transformations carried out by the visuomotor modules described earlier. Moreover, these systems are not linked directly to specific motor outputs, but instead to cognitive systems involving memory, semantics, spatial reasoning, planning, and communication. But even though such higher-order representational systems permit the formation of goals and the decision to engage in a specific act without reference to particular motor outputs, the actual execution of an action may nevertheless be mediated by dedicated visuomotor modules that are not dissimilar in principle from

those found in frogs and toads. Summing up, biological vision has two *distinct but interactive* functions: 1) the perception of objects and their relations, which provides a foundation for the organisms cognitive life; 2) the control of actions directed at those objects, in which specific sets of motor outputs are programmed

This twofold approach to high-level vision suggests that Marrian or reconstructive approaches [2] and Gibsonian or "animate-behaviorist" approaches [7], [3], [4], [5], [6], [8], [9], need not be mutually exclusive and may be actually complementary. Here, the long-term computational challenge lies in the integration of bottom-up and top-down cues, in order to provide coherent gaze-shifting, and in the interplay among action, attention, scene or object recognition. Figure 11 attempts at sketching an overall functional model (cfr. Figs. 1 and 3) accounting for all such issues. This gives evidence

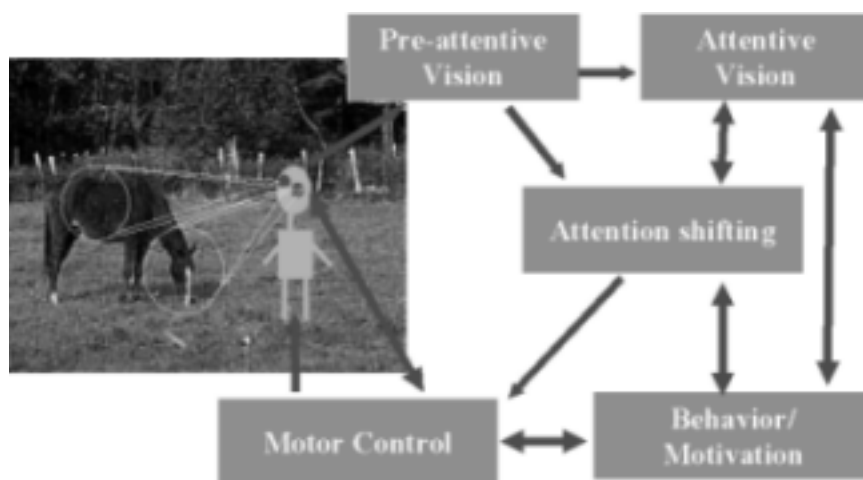


Figure 11: An active agent

that the agent's interaction with the world cannot be explored without the mediation of representations (in a broad sense), of different degree of complexity. As argued by Cordeschi, "*future developments in experimental research in AI and situated models, as well as in neuroscience and in cognitive ethology, will be able to suggest new hints for approaching this issue*", [62].

However, in pursuing this goal, one should not be further seduced by Descartes' vision of mind as a realm distinct from body and world [1], [61].

Acknowledgement

The author would like to thank M. Ferraro, R. Cordeschi, A. Marcelli and A. Picariello for enlightening discussions. This research was funded by MURST ex 60% and INFM.

References

- [1] A.Damasio, Descartes' Error. The Grosset Putnam, New York, NY (1994).
- [2] D. Marr, Vision. Freeman, S. Francisco, CA (1982).
- [3] D.Ballard, "Animate vision," Art. Intell., no. 48 (1991) pp.57-86.

- [4] R. Brooks, "Intelligence without representation," *Art. Intell.*, no. 47 (1991) pp. 139–159.
- [5] R. Bajcsy, "Active perception," *Proceedings IEEE*, vol. 76(1988), pp.996–1005.
- [6] M.Arbib, "Perceptual structures and distributed motor control," in *Handbook of Physiology: The Nervous System II. Motor Control* (V. Brooks, ed.), (Bethesda, MD), American Physiological Society (1981), pp. 1449–1480.
- [7] J. Gibson, *The ecological approach to visual perception*. Lawrence Erlbaum Associates, Hillsdale, NJ (1987).
- [8] U.Neisser, *Cognition and Reality. Principles and Implications of Cognitive Psychology*. W.H. Freeman, S. Francisco, CA (1976).
- [9] K. O'Regan, "Solving the 'real' mysteries of visual perception: The world as an outside memory," *Can. J. of Psychology*, vol. 46, no. 3 (1992), pp. 461–488.
- [10] Hayhoe, M. M., Ballard, D. H. and Bensinger D.G., "Task constraints in visual working memory", *Vis. Res.*, vol. 38, n. 1 (1998), pp. 125–137.
- [11] D. Fernandez-Duque and M.L. Johnson, "Attention metaphors: how metaphors guide the cognitive psychology of attention", *Cog. Science*, vol 23, (1999), pp. 83–116.
- [12] D. Broadbent, *Perception and Communication*. Pergamon, New York, NY (1958).
- [13] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews*, vol. 2 (2001) pp. 1–11.
- [14] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on PAMI*, vol. 20 (1998) pp. 1254–1259.
- [15] R. Milanese, S. Gil and T. Pun, "Attentive mechanisms for dynamic and static scene analysis", *Opt. Eng.*, vol. 34 (1995), pp. 2428–2434 .
- [16] B. Adams, C. Breazeal, R.A. Brooks, B. Scassellati "Humanoid Robots: A New Kind of Tool," *IEEE Int. Systems* (2000), pp. 25–31.
- [17] I. A. Rybak, V. I. Gusakova, A. V. Golovan, L.N. Podladchikova, and N.A. Shevtsova, "A model of attention-guided visual perception and recognition", *Vis. Res.*, vol. 38 (1998), pp.2387–2400
- [18] G.J. Gieffing, H.Janssen, H. Mallot, "Saccadic Object Recognition with an Active Vision System," *Proc. 10th Eur. Conf. Art. Intell.*, (1992), pp.803–805.
- [19] J. K. Tsotsos, et al. "Modeling visual-attention via selective tuning", *Art. Intell.*, vol.78 (1995), pp. 507–545
- [20] D. A. Chernyak and L. W. Stark, "TopDown Guided Eye Movements", *IEEE Trans. on SMC - Part B*, vol. 31, n. 4 (2001), pp. 514–522
- [21] K. Schill, E. Umkehrer, S. Beinlich, G. Krieger, and C. Zetzsche, "Scene analysis with saccadic eye movements: top-down and bottom-up modeling", *J. Electronic Imaging* (in press).
- [22] J. Denzler and C.M. Brown, "Information theoretic sensor data selection for active object recognition and state estimation", *IEEE Trans. on PAMI*, vol. 24, n. 2 (2002), pp.145–157
- [23] R. P. N. Rao and D. H. Ballard, "Dynamic model of visual recognition predicts neural response properties in the visual cortex", *Neur. Comp.*, vol 9 (1997), pp. 721–763.
- [24] G. Backer, B. Mertshing and M. Bollmann, "Data and Model-Driven Gaze Control for an Active-Vision System", *IEEE Trans. on PAMI*, vol. 23, n. 12 (2001), pp.1415–1429.
- [25] G. T. Buswell, *How people look at pictures*. University of Chicago Press, Chicago (1935).
- [26] A.L. Yarbus, *Eye movements and vision*. Plenum Press, New York, NY (1967).
- [27] J. M. Henderson, and A. Hollingworth, "High-level scene perception", *Ann. Rev. of Psych.*, vol. 50 (1999), pp.243–271.

- [28] P. Viviani, "Eye movements in visual search. Cognitive, perceptual and motor control aspects. In: Eye movements and their role in visual and cognitive processes, E. Kowler ed., Elsevier, Amsterdam (1990).
- [29] J. R. Antes, "The time course of picture viewing", *J. of Exp. Psych.*, vol. 103, (1974), pp. 62–70.
- [30] D. Noton and L. Stark, "Scanpaths in saccadic eye movements while viewing and recognising patterns", *Vis. Res.*, vol. 11 (1971), pp. 929–942.
- [31] G.J. Walker-Smith, A.G. Gale, and J.M. Findlay, "Eye movement strategies involved in face perception", *Perception*, vol. 6 (1977), pp. 313–326.
- [32] W. W. Nelson and G. R. Loftus, "The functional visual field during picture viewing", *J. of Exp. Psych., Human Learning and Memory*, vol. 6 (1980), pp. 391–399.
- [33] I. Biederman, R. J. Mezzanotte and J. C. Rabinowitz, "Scene perception: detecting and judging objects undergoing relational violations", *Cog. Psych.*, vol. 14 (1982) pp. 143–177.
- [34] H. Intraub, "Presentation rate and the representation of briefly glimpsed pictures in memory", *J. of Exp. Psych., Human Learning and Memory*, vol. 6 (1980), pp. 1–12.
- [35] J. Grimes, "On the failure to detect changes in scenes across saccades". In *Perception*, vol. 5, Vancouver Studies in Cognitive Science, (ed. K. Akins), Oxford University Press, New York (1996), pp. 89–110.
- [36] R.A. Rensink, J.K. O'Regan, J.J. Clark, "To see or not to see: the need for attention to perceive changes in scenes", *Psych. Sc.*, vol. 8 (1997), pp. 368–373.
- [37] M. F. Land and D. Lee, "Where we look when we steer", *Nature*, vol. 369 (1994), pp. 742–743.
- [38] M. F. Land and S. Furneaux, "The knowledge base of the oculomotor system", *Phil. Trans. R. Soc. Series B*, 352B (1997), pp. 1231–1239.
- [39] M. F. Land, N. Mennie, and J. Rusted, "The roles of vision and eye movements in the control of activities of everyday living", *Perception*, vol. 28 (1999), pp. 1311–1328.
- [40] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention", *Vis. Res.*, vol. 42 (2002) pp. 1071–1123.
- [41] N. H. Mackworth, and A. J. Morandi, "The gaze selects informative detail within pictures", *Perception and Psychophysics*, vol. 2 (1967), 547–552.
- [42] M. Ferraro, G. Boccignone and T. Caelli, "Entropy-based representation of image information", *Patt. Rec. Lett.*, vol. 23 (2002), pp. 1391–1398.
- [43] G. Boccignone, M. Ferraro, and T. Caelli, "Generalized Spatio-chromatic Diffusion," *IEEE Trans. on PAMI*, vol 24, no. 10 (2002).
- [44] S. R. Ellis and L. Stark, "Statistical dependency in visual scanning", *Human Factors*, vol. 28 (1986), pp. 421–438.
- [45] R. D. Rimey and C. M. Brown, "Controlling Eye Movements with Hidden Markov Models", *Int. J. of Comp. Vis.*, vol. 7 (1991) pp. 47–65.
- [46] S. K. Mannan, K. H. Ruddock, and D. S. Wooding, "Fixation sequences during visual examination of briefly presented 2D images", *Spat. Vis.*, vol. 11 (1997), 157–178.
- [47] G. Krieger, I. Rentschler, G. Hauske, K. Schill, and C. Zetsche, "Object and scene analysis by saccadic eye-movements: an investigation with higher order statistics", *Spat. Vis.*, vol. 13(2000), pp. 201–214.
- [48] D. Brockmann and T. Geisel, "The ecology of gaze shifts", *Neurocomputing*, vol. 32–33 (2000), pp. 643–650.
- [49] G.M. Viswanathana, V. Afanasyev, S. V. Buldyrev, S. Havlin, M.G.E. da Luz, E.P. Raposo, H. Eugene Stanley, "Levy flights in random searches", *Physica A*, vol. 282 (2000), pp. 1–12.

- [50] C. M. Privitera and L. W. Stark, "Algorithms for Defining Visual Regions-of-Interest: Comparison with Eye Fixations", IEEE Trans. on PAMI, vol. 22, no. 9 (2000), pp. 970–982.
- [51] T.M. Cover and J. A. Thomas, Elements of Information Theory. Wiley and Sons, New York (1991).
- [52] C. Bonanno, S. Galatolo and G. Menconi, "Information in sequences and applications", Physica A, vol 305 (2002) pp. 196–199.
- [53] G. Boccignone, A. Picariello, V. Moscato and M. Albanese, "Image Similarity based on Animate Vision: Information-Path Matching", Proc. Multimedia Information Systems '02 (2002).
- [54] M.J. Swain and D.H. Ballard, "Color indexing", Int. Journal of Computer Vision, vol. 7, n. 1 (1991) pp. 11–32.
- [55] G. Boccignone, A. Marcelli, G. Somma, "Analysis of dynamic scenes based on visual attention", Proc. Workshop Percezione e Visione nelle Macchine, Siena, Italy (2002).
- [56] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated Video Compression with Optimal Rate Control", IEEE Trans. on IP, vol. 10, no. 7 (2001), pp. 977–992.
- [57] D.H. Ballard. An Introduction to Natural Computation. The MIT Press, Cambridge, MA (1997).
- [58] R.P.N. Rao, "An optimal estimation approach to visual perception and learning", Vis. Res., vol. 39 (1999), pp. 1963–1989.
- [59] T. Mitchell, Machine Learning, McGraw-Hill (1997).
- [60] S. Minut and S. Mahadevan, "A reinforcement learning model of selective visual attention", Proceedings of the Fifth Int. Conf. on Autonomous Agents, ACM Press, Montreal, Canada, eds. J. P. Müller, E. Andre. S. Sen and C. Frasson (2001), pp. 457–464.
- [61] A. Clarke, Being There. Putting Brain, Body, and World Together Again. The MIT Press, Cambridge, MA. (1997).
- [62] R. Cordeschi, The Discovery of the Artificial - Behavior, Mind and Machines Before and Beyond Cybernetics. Kluwer Academic Publishers, Dordrecht (2002).
- [63] M. A. Goodale, G. K. Humphrey, "The objects of action and perception," Cognition, vol. 67 (1998), pp. 181–207.