[*MT evaluation: basis for future directions*. Proceedings of a workshop...2-3 November 1992, San Diego]

NEW DIRECTIONS:
It's Lucky If Concatenation Does Anything

Robert Berwick
Artificial Intelligence Laboratory
Massachusetts Institute of Technology

I'm going to show you a system that we developed over the last few months and talk about why it seems to work. And why systems that do French and Spanish can do the job that they do.

Our system is grounded on a very small set of basic principles. It claims that languages are organized out of a set of universal principles at the level of Lexical Conceptual Structure, or semantics, and that by assembling these in different ways and by slight variations, just as in atomic theory, you get different kinds of sentences bubbling to the surface. You switch it around and you get Japanese as opposed to English. So that's where we're headed, and to move forward in that direction, we're going to have to avoid some of the hazards of the field that I mentioned on the panel—namely, "simple scoring can go awry," and "it's lucky if concatenation can do anything."

1. Simple Scoring Can Go Awry

First a word about simple scoring. Here I'm going to make some remarks about the DARPA test for syntax and what they're doing, for example, on the Penn Tree Bank Project. The errors you're going to get are going to accumulate in an obvious way depending on what you want to recover. For people who aren't familiar with the Penn Tree Bank Project, they're attempting to build up a standardized bank of syntactically analyzed trees and assess the accuracy of parsers by measuring them against this.

In the case of translation, you want to recover thematic roles—who did what to whom. At least that was one of the Government's criteria. But these simple-minded scoring systems are vulnerable to the lamplight fallacy. They're easy to build, but it's all sort of normalized or boiled down to the lowest common denominator. What they do is take a tree structure and actually erase the labels (like whether something is a noun phrase or not), erase the position that something might or might not have originally been located in, and so on, and then they simply count the kinds of mistakes that occur. However, this kind of parenthesis structure isn't enough. Different surface structures can have the same parentheses, as can different structures with the same node labeling. But automatic scoring techniques actually erase information. So that's the big worry. They're going to wipe out information to get to something you can actually measure. [Indeed, after this was written this problem was recognized, and now we have an evaluation program, SEMEVAL, that tries to do justice to these issues. So in hindsight, this observation was on the mark.]

The "parenthesis counting" approach penalizes almost any sort of parser (I can name about five or six) that tries to recover thematic roles by other means. For instance, there's a valency kind of head-driven parser; there's this one that was done by Sleator at CMU; any kind of simple movement parser—in fact, even categorial grammar, as far as I understand it—would be penalized. What are the practical consequences of this? We have developed a parser that is designed to quickly recover the largest phrases it can at the fastest speed we can do it. We can now go through a year's worth of the *Wall Street Journal* in about 2 hours, analyze all the thematic roles, and actually use that as an index for a database. But this parser gets a lousy score on the DARPA sentence analysis test. We actually did the calculation on this and came up with a score of .35 while most of the other systems were around .8 or so. Yet we can recover the thematic roles perfectly easily with it. So the point about this part of the story is that you have to be very careful about the lamplight fallacy and going to the lowest common denominator, or else you'll wind up with a system that forgets what the whole thing is there for in the first place. As someone said quite accurately, we want to give the users what they need. One may well ask what the phrase structure is good for anyway if it erases the links between the thematic positions and what's actually going on.

2. Statistics and Linearity

The second point I want to make is about linear concatenation and translation by statistics—i.e., the work of Peter Brown and the rest of the team at IBM. This is actually a rather interesting thing to think about. It’s an example of a Cartesian fallacy. It’s trying to predict the occurrence of a given word based on the previous words that have been seen. How is that different from mind-reading? It’s not. It’s identical to the problem of taking a movie of people walking down the street—taking millions and millions of videotapes—and using that as a prediction of what people are going to do next. It’s perfectly obvious how this collection of videotapes fails to be a “theory” of human behavior. Of course there’s a sense in which it’s absolutely “right,” but it’s completely wrong in the sense that it doesn’t predict what people are going to do next. It’s *not* a theory of human behavior. You can’t ever predict what people are going to do next. You never know!

What I want to point out is that simple concatenation—and we have done some experiments with this—can’t possibly really work for languages like Japanese. We’ve tried it. In fact, we did it with a language that’s even less predictable than Japanese, namely Warlpiri. If I say, “Take the boomerang from the child,” that can essentially appear in any permutation. When you calculate the conditional probabilities, there’s no prediction of what word follows any other word. The same is largely true of Japanese. So if you use a bi-gram or a tri-gram model, that’s disastrous. But then, why do statistics work at all in French or Romance languages? The answer, as someone commented on the panel, is that English and French are very much alike. If you have binary trees, which everybody agrees on, and you have Romance languages, then with a few twists the same branching structure is locally linear. You can get the concatenation windows you want. But in other languages it doesn’t work like that. As soon as you move to languages like Japanese, there are real problems. What are you going to get when you try to line up Japanese and English? It's not going to match up at all. I guess the conclusion, however, is that this business of linear alignment with *modest* deletions actually does make a very strong prediction about what human languages look like—a prediction, it would appear, that is incorrect.

Let me give you a simple example of why linear statistics are too powerful and need grammars. The reason is that conditional probabilities obey transitivity and natural languages do not always obey this. Consider a string of three words, W1, W2, W3. Now, given PR(W2|W1)—that is, W3 can be selected by W1, given the right numbers. The catch is that if the three “words” are S(ubject), V(erb), and 0(bject), the inexorable transitivity of the probability calculation makes it possible for subjects to select verbs—correct—but also that subjects can select objects, which is not seen in natural languages. Of course, what blocks this is the “verb phrase”—the notion of government and barriers. But this is lost in the reduction to numbers and the absence of grammar.

3. A Different Approach

Now I would like to talk about a different way of looking at things. What we have developed is a small, axiom set of about 25 or so principles that interact to give many thousands of different construction types on the surface. This set may well be augmented by other specific construction types that vary from language to language—it doesn’t exclude that. The power of this, of course, is that the additive sum of these principles is much smaller than their multiplicative interaction. What do I mean by “principle”? One of the basic ones that shows up in English and Japanese, for example, is that English has a function-argument structure (“green with envy”) whereas in Japanese it’s the mirror image, i.e., argument-function structure (“ice cream eat”).

Under this kind of model, what we do to analyze a sentence is to superimpose each one of these boxes—like the notion that something is head-first or head-final, the notion that something has to have a thematic role, etc. We can superimpose these linearly, so it is a constraint-based system in the sense that we can put these boxes one on top of another. We can plunk a sentence in and get some kind of logical form out—some canonical form that at least standardizes some of the relations about things from language to language.

To get this so that we can output Japanese on the other wide, we change exactly four binary parameters: we say that Japanese is head-final, that you can drop elements freely, that there's none of this adjacency between verb and its case, and that it doesn't have what we call WH- in syntax (you don't form questions by saying "What did John eat?"; you leave the WH- word there—although that's a complicated story). That interacts together to give a lot of effects that you see in Japanese: the scrambling around of things, the use of case, free word order.

I like to call what we're doing "deformation of character": if it's wrong to say that languages are linear, then what you really want to do is say is that there's an abstraction that does linearize things. We get a relatively clean analysis in most cases. Once you've got that linear structure, then it can be mapped from one language to another. You get an exact translation back out. It's dead simple because it's straight compositional. I'm not implying that that's all you need, but I'm saying that if you did this transformation, then you could probably use the statistical analysis to get you someplace. But if you're going to assume that languages are linear like that, you won't. Things never work out as simply in nature as you think they might.

So the overall picture looks like this. There's a set of graph deformations: one of them does morphological graph deformations at the level of morphology; there's syntactic graph deformation; there's one that does thematic grid deformation using Lexical Conceptual Structure and then maps it back out again. This is an interlingual approach, of course, as it must be, but I think that's the right approach.

Once you do this kind of analysis you can use something like a bi-gram or tri-gram statistical method to get you rarer constructions if you want to. Then you can use a statistical approach, correctly, to extract what is linear in natural languages. But before you do this kind of deformation you can't, at least not completely. There are two advantages to this approach. One is that it's almost like automatic programming; you just flip these four switches and you can actually get the kind of graph deformation you need. And the second is that it's actually rather robust. If the system breaks down at some point and doesn't meet one of the constraints, it will tell you where it broke down and it will do the analysis up to that point. So it actually handles so-called ill-formed sentences quite naturally. In fact, under this view there is no such thing as an ill-formed sentence: these are just strings that meet more or fewer of the constraints.