[From: Makoto Nagao (ed.) *Machine translation summit*. Tokyo: Ohmsha, 1989]

2 • 6 Interlingua

— Technical Prospect of Interlingua —

Jaime G. Carbonell
Carnegie-Mellon University, U.S.A.

One of the traditional problems in being able to translate from one text in one language to one in another is that of polysemy. That is a word in the source language that has multiple meanings translating each into different words in the target language. I don't just mean homography, that is one word with multiple syntactic categories. I mean even within the syntactic category having multiple meanings. The semantic discrimination must be made and contextual analysis must lead to that discrimination in order to produce good translations.

To show an example requiring semantic discrimination for translation, consider a sentence, from the New Haven Register several years ago, "While driving down Route 72, John swerved and hit a tree." There are several problems in translating even a simple sounding sentence like that. Take that little tiny word "hit", three letters. If one is to translate "hit" into Spanish, one place to start is to look inside a dictionary, Spanish Academy Bi-Lingual Dictionary for example. There are fifty-three meanings of the word "hit". Which is the correct one? If one is to list them, one can find a word like "pegar" which means "punch" usually with intent to hurt. It is probably not John's intention to hurt the tree. Second is "chocar" which means "hit by accident", which is perfectly correct. Third is "acertar" which means to "hit" as in "hitting the bull's eye", for example when one fires an arrow and hits the target in the center. It is probably not the case that John calculated his swerve so as to hit precisely that tree, unless we knew he was suicidal from context. Other meanings of "hit" include "to smash repeatedly", which is not what John did either.

Now in order to make the correct lexical selection, one has to perform a degree of semantical analysis to know that the sentence refers to an accident, to know that if John was driving down the road he must have been driving a car. For some Slavic languages, one has to use a different verb for whether a person hit or a car did the hitting. It was not John that got off the car and punched the tree.

The next area is in concerns intersentential problems, such as resolving pronouns, other anaphors, or definite noun phrases required to translate correctly. For example, context is required to determine the gender of the pronoun "it" when translating from English into most other European languages. Also, using context from other sentences reduces ambiguity, as in the following example translating English to Japanese. Suppose we have these two fragments of dialog. In the first dialog fragment we have a customer and a repairman and the customer asks the repairman whether he has fixed his computer enable yet. The repairman can answer, "Yes I am about to take the line to your company." In the second dialog fragment, the repairman is at the subway station and the customer ask if he knows which subway line to take in order to go to the customer's building and the repairman says: "Yes, I am about to take the line to your company." In the second case, the repairman means "subway line", and in the first case he can mean "computer cable". But, in the English both sentences are identical: "Yes, I am about to take the line to your company." Yet, the previous dialog context is necessary and sufficient to figure out which meaning of "line" and which meaning of "take", to select in order to resolve the dual lexical ambiguity. The Japanese rendition of "taking the line" is different in each context.

We propose pivot-based approaches in order to retain the result of the semantic analysis in an intermediate representation. For our previous “take the line” example the representation of the source sentence is different in each dialog fragment depending upon the context, and therefore the generation in the target language would be different. In one case, it would be a “physical installation” semantic representation and in the other case, a “public transportation” one. That is, disambiguation occurs at analysis time. Once disambiguated, the pivot semantic representation can be used to generate multiple target languages — something we practice at CMU and is also practice in the Fujitsu ATLAS system. In spontaneous dialogue, in written dialog with computers and even more so in spoken translation such true simultaneous translation of speech, one needs to be able to deal with extra-grammatically constructions. That is just a fancy word to mean that people do not stick precisely to a linguistically “correct” grammar; they come close but not exactly. Also, there can be acoustical errors in recognition of speech that the translation system must be able to tolerate.

To give you an example, let us look at a natural language interface to a college register system to sign up for courses, transfer to other courses etc. developed CMU. People make spelling errors, they concoct new words, like “basketwork”, (we do not teach any course called basketwork at Carnegie Melon) and sometimes people introduce spurious phrases outside the grammar in order for the system to understand them. Of course this has precisely the opposite effect, e.g. “Please enroll Smith, if that is possible, in I think, Economics 237.” One has to be tolerant of such things. Sometimes words are typed in strange order, for example the verb at the end, as is common with Germanic speakers typing English. Some people watch television a bit too much and robots speak in strange ways in television. “Enroll Smith Economics 237” deleting articles and prepositions.

The point is simply that those sentences are all understandable, yet they do not accord perfectly to the grammar. In spontaneous dialog one has to be able to deal with extra-grammaticality for translation, as well as for natural language interfaces. In general, the kinds of problems, that have to be addressed for accurate translation include: 1) structural and syntactic ambiguity that must be resolved by semantic cues, 2) prepositional phrase attachment ambiguities, 3) style shifts, 4) pragmatics, 5) occasionally the ability to summarize. The latter is a particularly difficult problem that we are not really attacking for now. Very often the sole purpose in a fifty page document, is to understand the basic gist of it. It would be much better to have a system that instead of translating it to give different languages, produced only a precise summary. If later that document preadjust what you wanted, then the entire text could be translated and reviewed for high quality output.

In the long term we really want our machine translation systems to be multi-lingual and multi-purpose. By “multi-lingual”, I mean to be able to analyze one text and to be able to render it into a large number of different languages, so that the analysis must be deep enough to make all the distinctions required to all the different target languages. This means that the kind of pivot type approaches are more appropriate than point-to-point transfer approaches. By “multi-purpose”, I mean amortizing the big investment in time and effort required to develop detailed grammars, dictionaries, and knowledge bases. We want to use these linguistic resources for natural language data base access, natural language interfaces to expert systems, document classification, and other forms of multi-lingual text processing. These are the basic challenges on which we are focusing our research time and effort.