

[from website of Linguistics Association of America.]

Machine Translation

by Martin Kay of Xerox-PARC, Palo Alto, CA, and Stanford University

History of Machine Translation

At the end of the 1950s, researchers in the United States, Russia, and Western Europe were confident that high-quality machine translation (MT) of scientific and technical documents would be possible within a very few years. After the promise had remained unrealized for a decade, the National Academy of Sciences of the United States published the much cited but little read report of its Automatic Language Processing Advisory Committee. The ALPAC Report recommended that the resources that were being expended on MT as a solution to immediate practical problems should be redirected towards more fundamental questions of language processing that would have to be answered before any translation machine could be built. The number of laboratories working in the field was sharply reduced all over the world, and few of them were able to obtain funding for more long-range research programs in what then came to be known as computational linguistics.

There was a resurgence of interest in machine translation in the 1980s and, although the approaches adopted differed little from those of the 1960s, many of the efforts, notably in Japan, were rapidly deemed successful. This seems to have had less to do with advances in linguistics and software technology or with the greater size and speed of computers than with a better appreciation of special situations where ingenuity might make a limited success of rudimentary MT. The most conspicuous example was the METEO system, developed at the University of Montreal, which has long provided the French translations of the weather reports used by airlines, shipping companies, and others. Some manufacturers of machinery have found it possible to translate maintenance manuals used within their organizations (not by their customers) largely automatically by having the technical writers use only certain words and only in carefully prescribed ways.

Why Machine Translation Is Hard

Many factors contribute to the difficulty of machine translation, including words with multiple meanings, sentences with multiple grammatical structures, uncertainty about what a pronoun refers to, and other problems of grammar. But two common misunderstandings make translation seem altogether simpler

than it is. First, translation is not primarily a linguistic operation, and second, translation is not an operation that preserves meaning.

There is a famous old example that makes the first point well. Consider the sentence:

The police refused the students a permit because they feared violence.

Suppose that it is to be translated into a language like French in which the word for 'police' is feminine. Presumably the pronoun that translates 'they' will also have to be feminine. Now replace the word 'feared' with 'advocated'. Now, suddenly, it seems that 'they' refers to the students and not to the police and, if the word for students is masculine, it will therefore require a different translation. The knowledge required to reach these conclusions has nothing linguistic about it. It has to do with everyday facts about students, police, violence, and the kinds of relationships we have seen these things enter into.

The second point is, of course, closely related. Consider the following question, stated in French: *Où voulez-vous que je me mette?* It means literally, "Where do you want me to put myself?" but it is a very natural translation for a whole family of English questions of the form "Where do you want me to sit/stand/sign my name/park/tie up my boat?" In most situations, the English "Where do you want me?" would be acceptable, but it is natural and routine to add or delete information in order to produce a fluent translation. Sometimes it cannot be avoided because there are languages like French in which pronouns must show number and gender, Japanese where pronouns are often omitted altogether, Russian where there are no articles, Chinese where nouns do not differentiate singular and plural nor verbs present and past, and German where flexibility of the word order can leave uncertainties about what is the subject and what is the object.

The Structure of Machine Translation Systems

While there have been many variants, most MT systems, and certainly those that have found practical application, have parts that can be named for the chapters in a linguistic text book. They have lexical, morphological, syntactic, and possibly semantic components, one for each of the two languages, for treating basic words, complex words, sentences and meanings. Each feeds into the next until a very abstract representation of the sentence is produced by the last one in the chain.

There is also a 'transfer' component, the only one that is specialized for a particular pair of languages, which converts the most abstract source representation that can be achieved into a

corresponding abstract target representation. The target sentence is produced from this essentially by reversing the analysis process. Some systems make use of a so-called 'interlingua' or intermediate language, in which case the transfer stage is divided into two steps, one translating a source sentence into the interlingua and the other translating the result of this into an abstract representation in the target language.