[from: *Concise Encyclopedia of Computer Science*; ed. Edwin D.Reilly (Chichester: Wiley, 2004), pp.477-480]

MACHINE TRANSLATION

History

Within a few years of the first appearance of the "electronic calculators" research had begun on using computers as aids for translating one natural language into another. The major stimulus was a memorandum in July 1949 by Warren Weaver, who, after mentioning tentative efforts in the United Kingdom (by Andrew Booth and Richard Richens) and in the United States (by Harry Huskey and others) put forward possible lines of research. His optimism stemmed from the war-time success in code-breaking, from developments by Claude Elwood Shannon in information theory and from speculations about universal principles underlying natural languages. Within a few years, research had begun at many US universities, and in 1954 there was the first public demonstration of the feasibility of machine translation (MT), a collaboration of IBM and Georgetown University. Although using a very restricted vocabulary and grammar, it was sufficiently impressive to stimulate massive funding of machine translation in the United States and to inspire the establishment of MT projects throughout the world.

Optimism remained at a high level for the first decade of research, but disillusion grew as researchers encountered "semantic barriers" for which they saw no straightforward solutions. There were some operational systems—the Mark II system (developed by IBM and Washington University) installed at the USAF Foreign Technology Division, and the Georgetown University system at the US Atomic Energy Authority and at Euratom in Italy—but the quality of output was disappointing. By 1964, the US government sponsors had become increasingly concerned at the lack of progress, and set up the Automatic Language Processing Advisory Committee (ALPAC). It concluded in its 1966 report that MT was slower, less accurate, and twice as expensive as human translation and that "there is no immediate or predictable prospect of useful machine translation."

The ALPAC report was widely condemned as narrow, biased and shortsighted, but the damage had been done. It brought a virtual end of MT research in the United States for over a decade and it had great impact elsewhere in the Soviet Union and in Europe. However, MT research did continue in Canada, France and Germany. In the 1960s, in the United States and the Soviet Union, MT activity had concentrated on Russian-English and English-Russian translation of scientific and technical documents for a relatively small number of potential users, most of whom were prepared to overlook mistakes of terminology, grammar and style in order to be able to read something that they would have otherwise not known about. Since the mid-1970s, the administrative and commercial demands of multilingual communities and multinational trade have stimulated the demand for translation in Europe, Canada, and Japan beyond the capacity of the traditional translation services.

The 1980s witnessed the emergence of a variety of system types from a widening number of countries. First, there were a number of mainframe systems, whose use continues to the present day. Best known is Systran, now installed worldwide and operating in many pairs of languages. Others are: Logos for German-English translation and for English-French in Canada; the internally developed systems for Spanish-English and English-Spanish translation at the Pan American Health Organization; the systems developed by the Smart Corporation for many large organizations in North America; and the Metal system from Siemens initially for German-English translation and later for other languages.

The end of the decade was a major turning point. Firstly, a group from IBM published the results of experiments on a system (Candide) based purely on statistical methods. Secondly, at the same time, certain Japanese groups began to use methods based on a corpus (collection) of translation examples, that is, using the approach now called *example-based* translation. In both approaches the distinctive feature is that no syntactic or semantic rules are used in the analysis of texts or in the selection of lexical equivalents to multiple-word phrases that occur in those texts.

A third innovation has been research on speech recognition and synthesis and translation modules, the latter mixing traditional rule-based methods and newer corpus-based approaches. Inevitably, the subject domains have been highly restricted. The major projects have been at ATR (Nara, Japan) on a system for telephone translation of conference enquiries and hotel bookings; a collaborative project (JANUS) involving ATR, Carnegie-Mellon University and the University of Karlsruhe; and in Germany the government-funded Verbmobil project for a system to aid Germans and Japanese to conduct business negotiations in English.

Linguistic Problems of MT

The basic processes of translation are the analysis of the source language (SL) text, the conversion (or *transfer*) of the *meaning* of the text into another language, and the generation (or *synthesis*) of the target language (TL) text. There are basically three overall strategies. In the direct-translation approach, systems are designed in all details specifically for one particular pair of languages. Vocabulary and syntax are not analyzed any more than strictly necessary for the resolution of ambiguities, the identification of TL equivalents, and output in correct TL word order. Analysis and synthesis are combined in single programs, sometimes of monolithic intractability (e.g. the Georgetown system).

The second strategy is the *interlingua* approach that assumes the possibility of converting SL texts into (semantic) representations common to a number of languages, from which texts can be generated in one or more TLs. In interlingua systems SL analysis and TL synthesis are monolingual processes independent of any other languages, and the interlingua is designed to be language-independent or "universal".

The third strategy is the *transfer* approach, which operates in three stages: from the SL text into an abstract "intermediary" representation that is not language-independent but oriented to the characteristics of the SL (analysis); from such an SL-oriented representation to an equivalent TL-oriented representation (transfer); and from the latter to the final TL text (synthesis). Major examples of the transfer approach are the GETA, SUSY, Mu and Eurotra systems.

The main linguistic problems encountered in MT systems are fourfold: lexical, structural, contextual, and pragmatic or situational. In each case, the problems are primarily caused by the inherent ambiguities of natural languages and by the lack of direct equivalences of vocabulary and structure between one language and another. Some English examples are:

Lexical: homonyms (*fly* as "insect" or "move through air", *bank* as "edge of river" or "financial institution") require different translations (*mouche*: *voler*; *rive*: *banque*).

Structural: nouns can function as verbs (*control*, *plant*, *face*) and are hence "ambiguous", since the TL may well have different forms (*contrôle*: *diriger*, *plante*: *planter*, *face*: *affronter*)

Contextual: other languages make distinctions which are absent in English: *river* can be French *rivière* or *fleuve*, German *Fluß* or *Strom*; *blue* can be Russian *sinii* or *goluboi*.

Often all combine, as illustrated by a simple but common example, the word *light*. This can be a noun meaning "luminescence", an adjective meaning "not dark", an adjective meaning "not heavy", or a verb meaning "to start burning". In French, the meanings are conveyed by four different words *lumière*, *léger*, *clair*, and *allumer*.

Various aspects of syntactic relations can be analyzed. There is the need to (1) identify valid sequences of grammatical categories; (2) identify functional relations: subjects and objects of verbs, dependencies of adjectives on "head" nouns, and so on; and (3) identify the constituents of sentences: noun phrases, verb groups, prepositional phrases, subordinate clauses, and so on. Each aspect has given rise to different types of parsers: the *predictive syntactic analyzer* of the 1960s concentrated on sequences of categories; the *dependency grammar* has concentrated on functional relationships; and the *phrase structure grammars* have been the models for parsers of constituency structure. All have their strengths and weaknesses, and modern MT systems often adopt an eclectic mixture of parsing techniques within the framework of a "unification grammar" formalism.

MT in Practice

Many researchers have been persuaded that for the foreseeable future, it is unrealistic to attempt to build fully automatic systems capable of the translation quality achieved by human translators. The most obvious recourse, which has been adopted since the first MT systems, is to employ human translators to revise and improve the crude and inaccurate texts produced by MT systems. Initially "post-editing" was undertaken manually; later systems incorporate online revision and in some cases special facilities for dealing with the most common types of error (e.g. transposition of words, insertion of articles). Revision for MT differs from the revision of traditionally produced translations; the computer program is regular and consistent with terminology, unlike the human translator, but typically it contains grammatical and stylistic errors which no human translator would commit. The development of powerful microcomputer text editing facilities has led to the introduction of interactive MT systems. During the translation process, a human operator (normally a translator) may be asked to help the computer resolve ambiguities of vocabulary or structure.

Another possibility is to constrain the variety of language in the input texts. There are two approaches: either the system is designed to deal with one particular subject matter or the input texts are written in a vocabulary and style which it is known that the MT system can deal with. The former approach is illustrated by the METEO system, introduced in 1976, which translates weather forecasts from English into French for public broadcasts in Canada. The latter approach has been taken by the Xerox Corporation in its use of the Systran system; manuals are written in a controlled English (unambiguous vocabulary and restricted syntactic patterns), which can be translated with minimal revision into five languages. Other examples are the Smart systems installed at a number of large US and Canadian institutions that combine online editing to ensure clear documentation in English and "restricted language" MT to produce translations for subsequent editing.

Bibliography

1992. Hutchins, W.J. and Somers, H.L. *An Introduction to Machine Translation*. London: Academic Press.
1992. Newton, J. (ed.) *Computers in Translation: a Practical Appraisal*. London, New York: Routledge.
1995. Mason, J. and Rinsche, A. *Translation Technology Products*. London: Ovum Ltd.

W.John Hutchins