Engineering Outline 225 UDC 8.07 :681.14

Machine translation

Key indexing words
digital computers
natural languages
semantics
syntax

by Dr A J Szanser of the National Physical Laboratory

Definitions Machine translation (MT) is translation from one *natural language* into another by an automatic process. The only machine capable of carrying out this process is the digital computer. A natural language is one spoken by a group of people, normally identified with an ethnic unit, in contrast to an artifical language invented by man for a specific purpose (eg a programming language). The present account is restricted to written language.

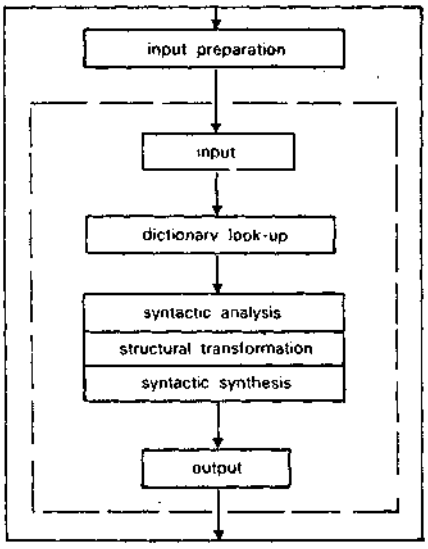
Aims Machine translation research, born out of intellectual curiosity, soon became supported by the growing demand for scientific and technological translations. Prospective users were, it seemed, ready to forgo not only style, but also grammatical correctness for the sake of speed, the latter sorely needed in available human translations. It was postulated that if a crude mechanical translation made the salient contents recognisable, a human translation could be ordered if desired, thus saving both time and money.

Development The idea of machine translation by means of an electronic computer originated in the late 1940s in England and the USA. The main activity started in the latter country, and the first authentic translation was produced in 1954 by Garvin and Dostert of Georgetown. A similar, if on a smaller scale, development followed in the USSR and other countries. In Britain, the first research group at Birkbeck College worked on French-English translation. In 1959, a research project in MT from scientific Russian into English was started at the National Physical Laboratory, Teddington (NPL). In 1961, the first international conference on MT was convened there. In 1962, there were nearly 50 MT research groups throughout the world, dealing not only with European, but also with some exotic languages.

In the meantime, the work on MT outgrew the original modest frame of word-for-word translation. It was quickly noticed that differences between languages reached far beyond word equivalence and that translation units should rather be whole phrases or even sentences. This problem brought into use the powerful tool of *syntactic analysis*.

About the middle of the last decade, MT work in the USA and elsewhere reached its peak, but at the same time its progress slowed. Early hopes based on initial successes, often exaggerated by the popular press, began to fade and soon a reaction followed. In the USA, where the research, backed by official funds, was the most active, this reaction took the strongest form and culminated in an adverse report by a special committee (ALPAC) in 1966. After this, the volume of research in that country, and to a lesser degree elsewhere, markedly decreased, although it did not cease completely. The NPL project, meant only as a pilot one, having achieved results thought to be positive, if far from perfect, was closed and submitted to an evaluation experiment earlier in the same year.

Today there still remain several MT research groups. In the USA and in the Western Europe they are mainly associated with universities and in the USSR with state research institutes.



1 The main stages of machine translation

Methods and techniques

The main stages of a MT process are shown diagrammatically in **1**.

At the start of this process, the original (*source language*) text must be introduced into a computer in machine-acceptable binary representation. An entirely automatic conversion is not yet possible. For the purpose of MT, one has to use the human faculty, by key-punching the text on paper tape or cards in a standard code. This is obviously an impediment to the translation process. Eventually it will be eliminated by preparing punched tape simultaneously with typesetting. This may become a standard

input preparation for published material, while 'electronic readers' will be reserved for typewritten, or even perhaps handwritten texts.

The input itself consists of reading the prepared tape, or cards, into a computer memory, usually on magnetic tape.

The next stage is translation of words and certain word combinations by the operation known as the *dictionary look-up*. The automatic dictionary is the collection of all required source language words, together with their equivalents in the output (*target*) language as well as the data necessary for translation, stored on an appropriate medium. The dictionary may contain either all grammatical forms of the source language words, or only their stems (invariable parts). The first type suits weakly inflected languages, such as English, the second—those strongly inflected, such as Russian. The latter type requires an additional procedure (*morphological analysis*) which recognizes the inflections and stores the grammatical information derived. The size of the dictionary is usually restricted by its intended use to a specific 'register' (eg scientific), or even a single field (eg electronics).

The choice of the medium on which to store the dictionary depends on the hardware (computer and its peripheral equipment) available. The huge bulk of an automatic dictionary precludes storage in the inner core of the computer. It may be stored on magnetic tape, in which case there is no direct access to its various parts and the look-up proceeds by running it simultaneously with the text tape, the text words being pre-arranged in the dictionary order. Each time a match is found, the target language equivalents and the grammatical data are transferred to the text word, and finally the now 'augmented' text is re-sorted into its original order. This method is called *serial access* and takes, obviously, rather a long time. Alternatively, a shorter look-up time, *random access*, can be obtained by using other media, such as a magnetic disk, or a special 'photostore', introduced by the IBM.

The dictionary look-up is usually combined with a search for idioms. If one component of a recognized idiom (usually the less common word) is found, cross-references point to other words, and if these are detected, all these entries are replaced by the common, idiom entry.

The next three stages, namely, syntactic analysis, structural transformation and syntactic synthesis, amount to translation of phrases and sentences. The respective procedures are carried out entirely within the operational store of the computer.

Syntactic analysis is necessary for two main reasons. Many words have multiple equivalents in another language and the choice can be guided by their syntactic role. What is even more important, languages differ widely in their syntactic structure and, in order to preserve the invariant 'meaning', one must detect the original structure in the first place. Examples of the NPL syntactic analysis procedures are shown, in a simplified form, in **2**.

The kind of analysis used in a MT system depends on the accepted *grammar*, ie rules of interdependence of the constituents (words, to start with) within higher-ranking linguistic units. There are various sets of such rules (usually termed *linguistic models*) and the method of analysis depends on the model used.

Having analysed the linguistic unit, normally a sentence, the next thing to do is to convert its syntactic structure into that proper to the target language. This process is called *structural transformation*. The more complete the model, the less transformation is required.

The last step in the syntactic translation is synthesis in the target language. The synthesis procedure follows the structure delineated at the preceding stage and adds necessary grammatical finish, such as inflections proper to particular words, thus preparing the output.

In early MT, the three stages described above were often fused in an ad hoc manner in programs dealing with specific problems, eg translation of compound predicates. In more developed systems (such as the NPL one), these stages are separated for economy and consistence. A sentence is first completely analysed, allowing for variants caused by ambiguities, and the results of the analysis are expressed in a determined form (for example, a 'list structure' representing a dependency tree). These are, in turn, modified according to structural transformation rules and finally the target language sentence is synthesized. In addition to these basic procedures others, destined to improve the quality of the output (such as approximate recovery of the meaning of words not found in the dictionary), can be added.

A sample of the NPL MT output is shown in **3**.

The output itself is carried out either directly (on the computer line printer) or indirectly (off-line) via punched paper tape or cards. The output is usually controlled by another program (*format control*) regulating its final appearance.

Verb government program

A. Prevention of incorrect insertion of prepositions

RUSSIAN TEXT:	Практика	этой теории	не подтверждает
TRANSLATION:	Practice	of this theory	does not confirm.
OPERATION OF THE PROGRAM:	The verb in negative form requires genitive (or accusative) noun complement. The only noun block in one of these cases is "this theory," which should not be linked by means of "of" to the preceding noun, but should be transferred after the governing verb		
FINAL TRANSLATION:	Practice does not confirm this theory.		

B. Resolution of ambiguities

RUSSIAN TEXT:	Работникам	завода	премии	выдаются.
TRANSLATION	To workers	of factory	of prize prizes	are handed out
OPERATION OF THE PROGRAM:	The predicate in the form of short participle requires a plural subject, which can only be the ambiguous word. This rules out the upper equivalent. The noun block in dative is governed by the verb of the participle and generally comes after it.			
FINAL TRANSLATION:	Prizes are handed out to workers of factory.			

Personal pronoun resolution program

This is the only program so far that requires some information to be carried from one sentence to another. It helps to solve various ambiguities: personal or impersonal use (she/it), possessive or personal (his/him), attributive or predicative (their/theirs). In each case the same Russian word corresponds to both forms.

PREVIOUS SENTENCE:	Автор не соглашается с этой теорией.	
TRANSLATION:	Author does not agree with this theory.	
CURRENT SENTENCE	Процитируем его: . . .	Приведем его слова: . . .
TRANSLATION:	We shall cite his/him/its/it: . . .	We shall quote his/him/its/it words: . . .
OPERATION OF THE PROGRAM:	Since the previous sentence contains only one masculine noun and it is personal, therefore the ambiguous word retains only "his/him" equivalents	
	The ambiguous word is not followed by a noun, and it agrees in case with the predicate verb as its complement, therefore it is a personal pronoun	The ambiguous word is immediately followed by a noun which can be a noun complement of the predicate verb, it is, therefore, a possessive pronoun
FINAL TRANSLATION:	We shall cite him: . . .	We shall quote his words

2 Examples of the National Physical Laboratory's machine translation syntactic analysis procedures (simplified). The kind of analysis used in machine translation depends on the accepted grammar

In practice to reach accurate correspondence of model and object is impossible,
also

therefore arises problem of optimal or rational selection of model.

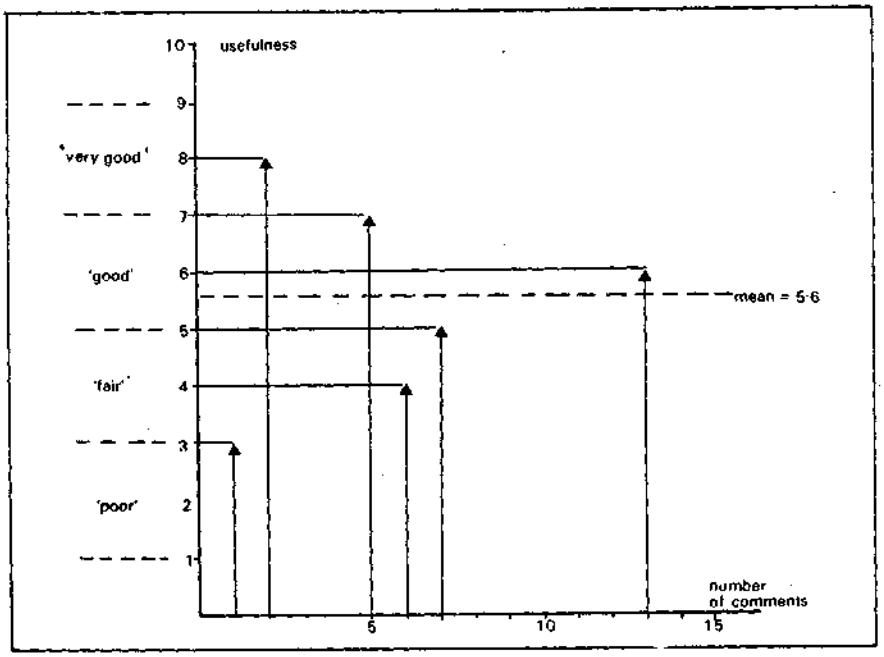
Apparently, successful decision of this problem can be implemented applicably to
solution

specific object on basis of study of its characteristics.
concrete in

Is here offered most simple and evident approach.

Suppose object of control is described by system of equations

3 Sample of the National Physical Laboratory's machine translation output. The output itself is carried out either directly (on the computer line printer) or indirectly (off-line) via punched tape or cards



4 Assessment of usefulness of the National Physical Laboratory's machine translation. The results show that the system was viable

Evaluation

Evaluation of MT quality is not an easy task. In various experiments of this kind such incongruous qualities as intelligibility, grammatical correctness and fidelity, often contradicting each other, were taken into account, with widely differing results. In the NPL experiment, the criterion accepted was 'usefulness'. The readers were assumed not to have any knowledge of Russian but to be experts in their own fields. A number of invited people from universities and industry agreed to comment on translations of the articles selected by themselves. Their comments were then reduced to a uniform scale and the results are shown in **4**. Too much must not be built on this experiment, but it does show that the system was viable.

Apart from the primary aim of MT research, it has produced valuable side profits, resulting from its strong interdependence with the modern science of linguistics. No real progress in MT would be possible without modern linguistic theory, but also the latter owes much to MT.

Prospects

Syntactic translation has been well developed by the leading MT groups, and also theoretically at university centres. It is, however, generally recognized that this is not enough. Many differences between languages arise out of the 'meaning' of words and, being impervious to syntactic treatment, create what has been named the *semantic barrier*. Today, the work on semantic analysis has already begun with encouraging results at several research centres. It should not be assumed, however, that a purely automatic translation can attain the ideal quality, equal to that of the human mind, for many reasons which go beyond the scope of this outline. This end can be achieved only by the interaction of man and computer and to find the optimum balance should be the purpose of further research. On the other hand, practical or 'useful' MT is undoubtedly possible.

There is also an economic aspect to consider. With the standards so far achieved, commercial demand is obviously wanting. As regards the future, however, things not only may, but are likely to change. Foreign languages will remain, and they show no tendency towards any unifying process. The cost of human labour will in all probability continue to rise, and computer technology will continue to progress. Sooner or later MT is certain to re-enter the limelight.

Some centres of research or further information

ERIC Clearinghouse for Linguistics, Center for Applied Linguistics, Washington, DC
'Machine Translation', Graduate Library School, University of Chicago
Mechanical Translation Project, University of Montreal