[From: *An encyclopaedia of translation: Chinese-Englis, English-Chinese*;
edited by Chan Sin-wai and David E.Pollard.
(Hong Kong: Chinese University Press, 1995), pp. 591-602]

MACHINE TRANSLATION

Introduction and basic concepts

The term *machine translation* (MT) refers to computerised systems responsible for the production of translations with or without human assistance. Commercial and operational MT systems are sometimes referred to as systems for *computer-aided translation*, because human aid is essential for good results. However, the field does not include the development of computer-based-translation tools to support translators, e.g. by providing access to on-line dictionaries or remote terminology databanks, the transmission and reception of texts, etc. In an MT system, the basic task of translation is undertaken by a computer program in conjunction with automated dictionaries and grammars.

Although the ideal aim of MT systems might be to produce translation as good as those from the best human translators, in practice the output has to be revised (or *'post-edited'*) for most recipients. In this respect MT does not differ from the output of most human translators which is normally revised by a second translator before dissemination. However, the types of errors produced by MT systems do differ from those of human translators (incorrect prepositions, articles, pronouns, verb tenses, etc.). Post-editing is the norm, but in certain circumstances MT output may be unedited or only lightly revised, e.g. if it is intended only for specialists familiar with the subject field of the text. Unrevised output might also serve as a rough draft for a human translator, often referred to as a 'pre-translation'.

The translation quality of MT systems may be improved either (most obviously) by developing more sophisticated methods or by imposing certain restrictions on the input. The system may be designed, for example, to deal with texts limited to the *sublanguage*' (vocabulary and grammar) of a particular subject field (e.g. biochemistry) and/or document type (e.g. patents). Alternatively, input texts may be written in a *'controlled language*', which restricts the range of vocabulary, avoids homonymy and polysemy, and eliminates complex sentence structures. A third option is to mark input texts (*'pre-edit'*) to indicate prefixes, suffixes, word divisions, phrase and clause boundaries, or to differentiate grammatical categories (e.g. to distinguish the proper name *Brown* from the adjective *brown*; or the noun *convict* from its homonymous verb *convict*.) Finally, the system itself may refer problems of ambiguity and selection to human operators (usually translators) for resolution during the processes of translation itself, i.e. in an *'interactive'* mode.

Systems are designed either for two particular languages ('*bilingual*' systems) or for more than a single pair of languages ('*multilingual*' systems). Bilingual systems may be designed to operate either in only one direction ('uni-directional'), e.g. from Japanese into English, or in both directions ('bi-directional'). Multilingual systems are usually intended to be bi-directional and to provide translations from any one language into any one or more other languages within the same configuration. In some cases, MT systems are called 'multilingual' if they combine a number of bilingual unidirectional systems based on similar principles and sharing some common dictionary and grammatical data.

In overall system design, there have been three basic types of systems. The first (and historically oldest) type is generally referred to as the *'direct translation'* approach: the MT system is designed in all details specifically for one particular pair of languages, e.g. Russian as the language of the original texts, the 'source language', and English as the language of the translated texts, the 'target language'. Translation is direct from the *source language (SL*) text to the *target language (TL)* text; the basic assumption is that the vocabulary and syntax of SL texts need not be analysed any more than strictly necessary for the resolution of ambiguities, the correct identification of TL expressions and the specification of TL word order; in other words, SL analysis is oriented specifically to one particular TL. Typically, systems consist of a large bilingual dictionary and a single monolithic program for analysing and generating texts; such 'direct translation' systems are necessarily bilingual and uni-directional.

The second basic design strategy is the *'interlingua'* approach, which assumes that it is possible to convert SL texts into representations common to more than one language. From such interlingual representations texts are generated into other languages. Translation is thus in two stages: from SL to the interlingua (IL) and from the IL to the TL. Procedures for SL analysis are intended to be SL-specific and not oriented to any particular TL; likewise programs for TL synthesis are TL-specific and not designed for input from particular SLs. Interlinguas may be based on a 'logical' artificial language, on an auxiliary language such as Esperanto, on a set of semantic primitives common to all languages, or on a 'universal' vocabulary.

The third basic strategy is the less ambitious *'transfer'* approach. Rather than operating in two stages through a single interlingual representation, there are three stages involving underlying (abstract) representations for both SL and TL texts. The first stage converts SL texts into abstract SL-oriented representations; the second stage converts these into equivalent TL-oriented representations; and the third generates the final TL texts. Whereas the interlingua approach necessarily requires complete resolution of all ambiguities in the SL text so that translation into any other language is possible, in the transfer approach only those ambiguities inherent in the language in question are tackled; problems of lexical differences between languages are dealt with in the second stage (transfer proper).

Within the stages of analysis and synthesis (or generation), most MT systems exhibit clearly separated components involving different levels of linguistic description: morphology, syntax, semantics. Hence, *analysis* may be divided into morphological analysis (identification of word endings, word compounds), syntactic analysis (identification of phrase structures, dependency, subordination, etc.), semantic analysis (resolution of lexical and structural ambiguities); *synthesis* may likewise pass through semantic synthesis (selection of appropriate compatible lexical and structural forms), syntactic synthesis (generation of required phrase and sentence structures), and morphological synthesis (generation of correct word forms). In transfer systems, the *transfer* component may also have separate programs dealing with lexical transfer (selection of vocabulary equivalents) and with structural transfer (transformation into TL-appropriate structures).

In many older systems, particularly those of the 'direct translation' type the components of analysis, transfer and synthesis were not always clearly separated. Some of them also mixed data (dictionary and grammar) and processing rules and routines. Later systems have exhibited various degrees of *modularity*, so that system components, data and programs can be adapted and changed without damage to overall system efficiency. A further stage in some recent systems is the *reversibility* of analysis and synthesis components, i.e. the data and rules used in the analysis of a particular language are applied in reverse when generating texts in that language.

Problems and methods

The main linguistic problems encountered in MT systems may be treated under four main headings: 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. Many examples could be given, some English ones are: homonyms (*cry* as ‘weep' or 'shout', *bank* as 'edge of river' or 'financial institution') require different translations (e.g. in French *pleurer*: *crier*; *rive*: *banque*); nouns can function as verbs (*control*, *plant*, *face*) and are hence 'ambiguous', since the TL may well have different forms (e.g. French *contrôle*: □*diriger*, *plante*: *planter*, *face*: *affronter*). A polysemous word such as English *field* has often many possible translations, e.g. in Japanese: *hatake* (field for crops), *nohara* (open space), *kyougiba* (sports field), *bun'ya* (sphere of activity), etc. In many cases, target languages make distinctions which are quite absent in the source, e.g. *wear* is unambiguous (non-polysemous) for English speakers, but Chinese distinguishes between 'wearing clothes, shoes, etc.' (*chuan*), 'wearing jewelry, spectacles, etc.' (*dai*^{4}) and 'wearing a tie' (*da*^{2}).

In many cases, differences between the vocabulary of the source and target languages are also accompanied by structural differences. A familiar example involves the translation of the English verb *know* into French or German, where there are two verbs which express 'knowledge of a fact' (*connaître* and *kennen*) and 'knowledge of how to do something' (*savoir* and *wissen*):

(1) I know the man - Je connais l'homme; Ich kenne den Mann.

(2) I know what he is called - Je sais ce qu'il s'appelle; Ich weiss wie er heisst.

Translation into unrelated languages typically involves considerable structural change. For example, the English sentence (3) must be completely reformulated in Japanese (4):

(3) The earthquake destroyed the buildings

(4) Jishin de kenbutsu ga kowareta
Earthquake-by buildings collapsed
i.e. 'The buildings collapsed due to the earthquake'

Various aspects of syntactic relations can be analysed. There is the need (a) to identify valid sequences of grammatical categories, (b) to identify functional relations: subjects and objects of verbs, dependencies of adjectives on 'head' nouns, etc., (c) to identify the constituents of sentences: noun phrases, verb groups, prepositional phrases, subordinate clauses, etc. Each aspect has given rise to different types of parsers: the predictive syntactic analyser concentrated on sequences of categories (developed subsequently by Woods as the Augmented Transition Network parser); parsers based on dependency grammar (of Tesnière, Hays, etc.) look for functional relationships; and phrase structure parsers identify the kinds of constituency structures familiar in Chomskyan generative grammars. Each have their strengths and weaknesses, and modern MT systems often adopt an eclectic mixture of parsing techniques, now often within the framework of a 'unification grammar' formalism.

The most serious weakness of all syntactic parsers is precisely their limitation to structural features. An English prepositional phrase can in theory modify any preceding noun in the sentence as well as a preceding verb:

(5) The car was driven by the teacher with great skill

(6) The car was driven by the teacher with defective tyres

(7) The car was driven by the teacher with red hair

In (5) the phrase *with great skill* modifies the verb phrase *was driven*; in (6) *with defective tyres* is attached to *the car*; and in (7) *with red hair* is an attribute of *the teacher*. However, these attachments are based on semantic or pragmatic information (e.g. knowledge that cars do not have red hair). But syntactic analysis can go no further than offer each possibility, and the specific relationship has to be identified by later semantic analysis involving lexical and situational context.

To overcome some of these problems, many parsers now include the identification of case relations. Consider, for example:

(8) The house was built by a doctor for his son last year.

In this sentence, the Agent of the action ('building') is *a doctor*, the Object of the action is *the house*, the Recipient (or Beneficiary) is *his son* and the Time of the action is *last year*. Many languages express these relations explicitly, suffixes of Latin, German, Russian nouns *(-ibus*, *-en*, *-ami*), prepositions of English and French (*to*, *à*), particles of Japanese (*ga*, *wa*); but they are often implicit (as in English direct objects). There are rarely any direct correspondences between languages and most markers of cases are multiply ambiguous in all languages, cf. *with* expressing Manner (5) or Attribute (6 and 7). Nevertheless, there is a sufficient regularity and universality in such 'case relations' to have encouraged their widespread adoption in many MT systems, particularly in those with Japanese as source or target language.

There is also some agreement about the use of semantic features, i.e. the attachment of such categories as 'human', 'animate', 'liquid' to lexical items and their application in the resolution
of ambiguities. For example, in:

(9) He was beaten with a club

In this sentence, the 'social' sense (meeting place) of *club* is excluded by the verb-type which requires an 'inanimate' Instrument, i.e. it must have the 'hammer' or 'weapon' meaning. Similarly:

(10) The sailor went on board

(11) The sailor was examined by the board

The 'physical' sense of *board* (ship) in (10) is confirmed by the verb-type (motion, *go*) and the preposition of Location (*on*), and the 'social' (committee) sense in (11) is confirmed by the verb *examine* which requires an 'animate' Agent.

Few operational MT systems involve any deeper levels of semantic or pragmatic analysis, yet the resolution of many linguistic problems clearly transcends sentence boundaries. A common and persistently difficult problem involves the use of pronouns. Consider the following:

(12) The soldiers shot the women. They were buried next day.

We know that the pronoun *they* does not refer to *soldiers* and must refer to *women* because we know that 'shooting' implies 'killing' and 'injury' or 'death' and that 'death' is followed (normally) by 'burial'. This identification is crucial when translating into languages where the pronoun must indicate whether the referents are male or female (e.g. French *elles* and *ils*.) Such examples demonstrate that the disambiguation and correct selection of TL equivalents would often seem to be impossible without reference to knowledge of the (non-linguistic) features and properties of the actual objects and events described. Recent advances in Artificial Intelligence (AI) have encouraged the investigation of knowledge-based MT systems, at least for systems restricted to specific domains. For example, in a system designed for translating texts in computer science and data processing the English word *tape* could mean either magnetic tape or adhesive tape. In the following sentence (13), reference to the knowledge base should establish that only the 'adhesive tape' interpretation is possible since diskettes do not contain 'magnetic tapes' which can be removed.

(13) Remove the tape from the diskette

In view of the linguistic limitations of MT systems it should be clear that the most suitable texts are either those of a technical or scientific nature, where there is often a high degree of direct terminological equivalence and where problems of homonymy and polysemy can be reduced by the restriction of dictionaries to specific subject domains, or administrative texts with a high degree of repetition, where stylistic considerations are unimportant (e.g. the minutes of meetings, internal reports, etc.) Obviously unsuitable are literary and philosophical texts, where nuances of vocabulary and cultural and stylistic factors play an important role; and equally unsuitable are texts with particularly complex sentence structures, e.g. patents and legal documents. The suitability of texts intended for publication depends on various economic factors, such as whether input texts are in machine-readable form, whether long documents change little between editions (e.g. operational manuals for equipment), whether a great deal of terminology work has to be done, and so forth. In certain circumstances, the control of input text (e.g. restrictions on vocabulary and syntax) can be cost-effective if output is to be in more than one target language. There is now a substantial body of practical experience with MT systems, which can be called upon to assist potential users.

History and future

Research on using computers as aids for translating natural languages began within a few years of the first appearance of the newly invented 'electronic calculators'. The major stimulus was a memorandum in July 1949 by a director of the Rockefeller Foundation in New York, Warren Weaver, who described tentative efforts in Great Britain (by Booth and Richens) and in the United States (by Huskey and others in Los Angeles) and proposed various approaches. Within a few years research began at many US universities, and in 1954 the first public demonstration of the feasibility of machine translation (MT) was given (a collaboration of IBM and Georgetown University). Although using a very restricted vocabulary and grammar it was sufficiently impressive to encourage massive funding of MT in the United States and to inspire the establishment of MT projects throughout the world.

The earliest systems consisted primarily of large bilingual dictionaries where entries for words of the source language (SL) gave one or more equivalents in the target language (TL) and some rules for producing the correct word order in the output. It was soon recognised that this 'word-for-word' approach was inadequate and ad hoc; the need for more systematic methods of syntactic analysis became evident, and a number of projects were inspired by contemporary developments in linguistics, particularly Zellig Harris' and Noam Chomsky's ideas on syntactic transformations, but also other models such as dependency grammar and stratificational grammar.

Optimism remained at a high level for the first decade of MT research, with many predictions of imminent "breakthroughs", 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 (although satisfying many recipients' needs for information). In 1964, the US government sponsors set up the

Automatic Language Processing Advisory Committee (ALPAC), which concluded in its famous 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 to 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, in France, in Germany, and in Hong Kong. Within a few years Peter Toma, one of the members of the Georgetown University project, had developed Systran for operational use by the USAF (1970) and by NASA (in 1974/5), and shortly afterwards Systran was installed by the Commission of the European Communities for translating from English into French (1976) and later between other Community languages. At the same time, in Canada the METEO system developed at Montreal University was successfully put into operation for the daily translation of weather reports into French - in continuous use from 1976 to the present day - and in Hong Kong the CULT system (Chinese University Language Translator) began to be used for the regular production of translations of a Chinese mathematics journal into English. METEO is an -example of a fully automatic system in a highly restricted 'sublanguage'; CULT is a system demanding substantial pre- and post-editing.

In the 1960s in the US 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 which they would have otherwise not known about. Since the mid-1970s the demand for MT has come from quite different sources with different needs and different languages. 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 demand is now for cost-effective computer-based translation systems which can deal with commercial and technical documentation in the principal languages of international commerce.

The 1980s have witnessed the emergence of a variety of system types and from a widening number of countries. There are a number of mainframe systems. 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 systems developed at the Pan American Health Organization for Spanish-English and English-Spanish translation; the systems developed by the Smart Corporation for many large organisations in North America; and the recently marketed METAL system from Siemens for German-English translation. Major systems for English-Japanese and Japanese-English translation have come from Japanese computer companies, Fujitsu, Hitachi and Toshiba. The wide availability of microcomputers and of text-processing software has led to the commercial market for cheaper MT systems, exploited in North America and Europe by companies such as ALPNET, Weidner, Linguistic Products, Tovna and Globalink, and by numerous Japanese companies, e.g. Sharp, NEC, Oki, Mitsubishi, Sanyo. Other microcomputer-based systems have appeared from China (TRANSTAR), Korea (e.g. NARA), Bolivia (ATAMIRI), the Soviet Union, etc.

Throughout the 1980s research on more advanced methods and techniques has continued. The dominant strategy is now that of 'indirect' translation via intermediate representations, sometimes interlingual in nature, involving semantic as well as morphological and syntactic analysis and sometimes non-linguistic 'knowledge bases'. There is increasing emphasis on devising systems for particular subject areas and particular specific purposes, for monolingual users as well as bilingual users (translators), and for interactive operation rather than batch processing. The most notable projects have been the GETA-Ariane system at Grenoble, SUSY and ASCOF at Saarbrücken, Mu at Kyoto, DLT at Utrecht, Rosetta at Eindhoven, the knowledge-based MT project at Carnegie-Mellon University (Pittsburgh), the ArchTran project in Taiwan, and two ambitious international multilingual projects: Eurotra, supported by the European Communities, involving teams in each member country; and the Japanese CICC project with participants in China, Indonesia and Thailand.

In the immediate future, there will clearly be continued expansion and improvement of systems for the business and administrative communities. As at present, the MT market will include both microcomputer and mainframe systems. The cheaper microcomputer systems will produce relatively poor output needing substantial revision but which can be applied cost-effectively in commercial services. More expensive mainframe (or minicomputer) systems will be developed on transfer and interlingua approaches with some use of AI techniques. These will be producing higher quality output, which, although still requiring revision for many purposes (e.g. publication), will be satisfying basic information needs without revision.

It is probable that other types of systems will appear. Nearly all systems at present require users to know both source and target languages, generally to the level expected of regular translators. There is clearly a need for systems which can be used by individual non-translators ignorant of the source language in order to get translations giving at least the essential content of documents. At a further stage, these systems should be integrated with other documentation systems (information retrieval, abstracting, paraphrasing). There is an equally clear need for systems for those ignorant of target languages, e.g. businessmen (and others) wanting to convey simple messages to make travel arrangements, to book hotel accommodation, to arrange meetings, etc. Already some research has been done on 'interactive analysis' systems: writers would be asked (in computer-initiated dialogues conducted in their own language) for information to resolve ambiguities and to enable the generation of appropriate texts.

However, probably the most obvious area of future development will be speech translation. Research is already in progress (particularly in Japan) on systems for international telephone communication (initially restricted to standard business messages) which combine voice interpretation and voice production with machine translation. Given the problems of speech recognition in addition to the peculiarities of conversational language, operational prototypes are regarded very much as long-term objectives.

Nearly all developments depend on improvements in the automation of the basic translation processes. Many researchers still aspire to the ultimate ideal of fully automatic high quality translation but it seems increasingly unrealistic. MT output suffers still from what appear to be low-level problems: wrong pronouns, incorrect prepositions and tenses, erroneous translations of common vocabulary. Progress is slow, but developments in -artificial intelligence, in linguistic theory, in computational linguistics and in computer technology promise future improvements in overall quality.

At a more basic level much progress depends on the continued efforts to standardise terminology both within and across languages, which is of benefit to translators and technical writers generally. More specifically, the wasteful duplication involved in the creation of large MT dictionaries calls for inter-project cooperation, a process which has already started in Japan with the Electronic Dictionary Research project.

The translation of natural languages by machine, first dreamt of in the seventeenth century, has become a reality in the late twentieth. Computer programs are producing translations - not perfect translations, for that is an ideal to which no human translator can aspire; nor translations of literary texts, for the subtleties and nuances of poetry are beyond computational analysis; but translations of technical manuals, scientific documents, commercial prospectuses, administrative memoranda, medical reports. Machine translation is not primarily an area of abstract intellectual inquiry but the application of computer and language sciences to the development of systems answering practical needs. MT is no longer seen as a threatening replacement of translators but as an aid to multilingual communication. The future development of MT rests on fruitful interaction between the researchers of experimental systems investigating new methods and theories, the developers of commercial systems exploiting well-tested methods in cost-effective practical systems, and the perception of the real needs of translators and other potential users of translation systems.

Further reading

For general introduction to MT see Hutchins & Somers (1992); for the general history of MT see Hutchins (1986); for descriptions of current systems and developments see Hutchins (1988), Slocum (1988) and Vasconcellos (1988).

References

ALPAC. (1966) *Language and machines: computers in translation and linguistics*. A report by the Automatic Language Processing Advisory Committee. Washington, D.C.: National Academy of Sciences.

Hutchins, W.J. (1986) *Machine translation: past, present, future*. Chichester: Ellis Horwood. New York: Halsted Press.

Hutchins, W.J. (1988) 'Recent developments in machine translation: a review of the last five years.' In: *New directions in machine translation*, ed. D.Maxwell et al. (Dordrecht: Foris), 7-62

Hutchins, W.J. & Somers, H.L. (1992) *An introduction to machine translation*. London: Academic Press.

Slocum, J. (1988), ed. *Machine translation systems*. Cambridge: University Press.

Vasconcellos, M. (1988), ed. *Technology as translation strategy*. Binghamton, NY: State University of New York.