MACHINE TRANSLATION

Ashok Kumar

DESIDOC, Metcalfe House, Delhi-110 054

ABSTRACT

Machine translation (MT) is an important technology which assumes significance in view of increasing globalisation and sharing of information among all the countries. The article attempts to bring the status and various features of MT to the notice of providers and users of information. MT has moved from mainframes to personal computer and is thus available to all for personal and organisational use. The paper describes various types of MT systems and the processes involved in translation. It also gives a brief idea of development of MT systems and the software available.

1. INTRODUCTION

In the present 'information age' computers are being used for processing, storage and retrieval of information. At the same time, the need for sharing of this information among all countries of the world has also been felt. Faster communication techniques are being used for easy portability of information. But the existence of information in different languages still remains one of the barriers in its use world-wide.

Machine translation (MT) is an important technology for converting information from one language into another with the help of a computer. However, the large number of languages prevalent in the world make translation a huge task, for example, just 9 languages require 72 different directions by way of combination for translation. MT attempts to provide fast translation ability on not only desktop computers but also on laptops in future.

2. HISTORICAL BACKGROUND

MT involves many complex tasks and it is the earliest kind of natural language processing. The first translation machines were designed in the early 30s and after the design of ENIAC in 1946, development of MT picked up momentum. The advent of

computers in 1940s led to increased experimentation .in this area. MT was launched as a scientific enterprise in 1949. The technical feasibility of machine was demonstrated in USA, by IBM in 1954 which encouraged government funding for research in this area. However, much of the early enthusiasm was dampened with the emergence of linguistic complexities and crude quality of translation produced by this method. High costs of inputting and delay in the development of optical character recognition technology also contributed to further disappointment.

In 1960s, research in the field of MT was slow and the area was almost ignored by the computer scientists, librarians, information scientists and linguists. It was almost forgotten by the common public. 1970s saw the revival of interest in MT and this time the researcher adopted more modest objectives. They aimed at translation that is good enough for a particular purpose and which may not be stylistically adequate but succeed in conveying the intended message. In the 80s, better and better systems were developed and now MT is available on desktop PCs.

The first generation systems for MT adopted word-to-word approach using large bilingual dictionaries. These systems were designed specifically for a specific pair of languages. In the second approach which lead to the development of second generation systems, the input text was analysed independently of the target language using syntactic and semantic representations and then the output is reconstructed into the target language. This approach facilitated development of multilingual systems. Further research lead to development of semantics-based and interactive MT systems. These systems are discussed in detail under the types of MT systems.

3. MT TODAY

Machine Translation is not a new idea, but it has found greater awareness or more favourable response now with the use of PCs. MT was earlier restricted to mainframes due to high requirements of processing power and storage capacity. Unlike interpretors, translators normally produce written output which makes translation a slow process. Computers can do this activity faster and free human translator for more creative aspects of the job.

MT analyses text in the source language and automatically generates sentences in the translated language. The computer requires machine-readable text as input and produces translated text as output. MT too involves human assistance but differs from MAT (machine-assisted translation), where a human translator performs the translation and prepares the translated text using word processors, automatic terminology managers, dictionaries text-critiquing software, repetition-processing to enhance productivity. Already most of the translators use these tools including full MT.

Machine Translation works best if the subject matter is specific or restricted. The translation quality improves further if the original text is straightforward and devoid of ambiguities, for example, maintenance manuals which are consistent in style and vocabulary.

Currently efforts in the field of MT are on to develop systems capable of handling difficult text types and language combinations with least human assistance and translating between languages having different alphabet structures. Today's PCs possess enough processing power required for such jobs which were till now mainframe-dependent. Because MT involves looking at different ways a sentence can be parsed depending on various meanings and

roles of each word, it takes a lot of time on these choices and in applying available rules. At present 386 and 486 PCs, high-end Mac and UNIX machines can handle such jobs, but it depends on the availability of software for these platforms.

The move of MT from mainframe to PC has led to other advantages. Now it is possible to use OCR, CD-ROM, modems and facsimile for inputting text, downloading databases and also to exchange computer readable information world-wide.

The cost of converting the input text to machine-readable form is also debatable. But the widespread use of OCR technology and the availability of text in digital form is making it more feasible.

With database management tools for retrieving terminology and previously translated text, style checkers and DTP software, added with tools to create own dictionaries and linguistic rules, one can customise the basic architecture of an MT system.

Local Area Networks (LAN) offers cost-effective MT systems to a large group of users. The translation can be done on a less-used CPU in the network while the large lexicon and specialised glossaries can be stored at one location for use by all. Use of LAN also facilitates the control of updates at a central place for consistency in translation. Large number of dial-up services are available for MT. The files can be exchanged using a modem on networks.

Machine Translation can also be integrated with DTP process. One can retain the tags (identification name for paragraph format in DTP) of input text in translation, so that the output is also ready for DTP after translation. This can save a lot of cost and time in providing printed translated documents. Graphics and tables of the original can be retained, avoiding redrawing efforts. MT also prevents errors occurring during re-inputting.

4. THE PROCESS OF MT

Although translation is one of the most complex type of processes, it has certain common elements .

In an MT system there are three sets of data: input text, translation program and permanent resident sources. The most essential component is a dictionary of words and phrases of source language against which input text is matched during translation. Knowledge sources also include the sets of rules that are fired at various points in the translation process. In addition, there is a bank of information about the concepts invoked by the dictionary.

In large MT systems, dictionaries hold numerous words. For each word, a record holds formalised representation of information about how the word functions. These dictionaries occupy many megabytes of storage memory.

To begin with, an MT system matches the words of input text with those in dictionary through index. For each word, it retrieves a complete record which includes information about the possible functions of the word and its relationship to words that may occur with it.

5. TACKLING THE PROBLEMS OF AMBIGUITIES

Any natural language has inherent ambiguities which cannot be tackled using pure logic. It is a difficult task for the computer to translate the language accurately. Researches have indicated, that MT systems must remove at least five layers of ambiguity before it can map sentences from one language to another with some degree of accuracy.

In the first stage every word is viewed in the context of the sentence. Using grammar (stored as set of rules), the MT operates at syntactic stage. In the next stage, using more intelligent rules, it goes beyond syntax to semantic stage which is not present in most of the low end systems.

Processing at this lexical semantic stage introduces another confusion arising out of multiple meanings of words. This is generally solved using sentence level semantics in the third layer.

The fourth layer of difficulty deals with ambiguities introduced at sentence level of the semantic stage (e.g. to analyse the meaning of prepositions). The fifth layer of ambiguity concerns more technical issues, such as ellipses and amphora and calls for processing at a discourse level. Only a few systems have this ability.

6. IMPROVING QUALITY OF THE TRANSLATION

Depending upon the time availability the human translator can interact before, during, or after the automatic translation process, but for routine translation jobs, this interaction can certainly be avoided.

Pre-editing: In this method the editor (operator) intervenes before translation to eliminate lexical and structural ambiguities by either revising the text by editing software, or by customising the text for translation according to pre-established rules and vocabulary; this method is called pre-editing. Although this makes it easier for the machine, it requires editing of output too. But it is advantageous while translating from one language to many languages.

Interactive Editing: Here the editor responds to the questions posed by the computer during the process. The computer offers various alternatives, and the operator clicks on the most appropriate choices. This reduces manual editing required after or before the translation.

Post-editing: The post-editor can use word-processors and macros to speed up the process. Though post-editing is more laborious it gives control over the quality of text. This can be avoided if a rough translation is needed for a large number of readers. Depending on the text, the translation output can be doubled (compared to human output) upto 3000 to 10,000 words per day (8 hours).

Depending on the requirement, either a thorough translation or only a rough translation may be needed. The rough translation may convey the information but a thorough translation captures all the nuances of the original text. Most of the translations are still performed by people but computers have increasingly been used for this purpose.

The use of MT depends on the need of the users. In the case of time-bound jobs, the quality need not be perfect because the style and subject matter of sources varies. Such jobs can be handled by the general purpose systems. These systems have large and detailed dictionaries, knowledge sources and grammar rule bases. However, the output is inferior to the output produced by specific domain systems.

Translation quality improves as the systems acquire more rules and larger detailed lexicons through use. As the dictionaries and rules bases grow larger, less and less human intervention is required. The other factors which affect quality are structural proximity of two languages, the domain, and the type of text. The quality required of the translation also depends on its intended use. Human intervention is needed to make the raw machine translation acceptable by further polishing.

General purpose MT systems can speed up the work of in-house translators who produce publication quality copy in various

subject areas. The most widespread use of MT is in translating text in limited domains (manuals etc). This reduces ambiguity and keeps terminology consistent even when the quantity of job is very large.

7. TYPES OF MT SYSTEMS

Languages are complicated, and this makes translation a formidable challenge. Each language is unique, with elements that make it colourful and special. But it also makes a language difficult to translate for a human or machine.

Automated translation is feasible because some applications like technical documents are straightforward. As MT systems evolve, they will take on the challenge of general translation.

MT systems can be categorised as direct, transfer or still complex inter-lingual systems. They can also be categorised as those whose knowledge bases are built by humans and those that collect their rules statistically.

7.1 Direct-translation Systems

These systems include software that translates languages by replacing source language words with target-language words. These are suitable for applications where text having limited vocabulary and definite style is translated. The systems contain correspondence lexicons and clarity of replacement patterns. Such a system can handle similar languages like English and German but have problems in handling non-similar languages like English and Chinese or Japanese because of difference in grammar and other unique features of these languages.

Direct MT systems need massive lexicons of specific words and phrases. These lexicons become more cumbersome as they grow and contain many redundant entries. The latest systems include analysis of wordform and structure and remove commonalties by creating a root and rules for deviation.

7.2 Syntactic-transfer Systems

These systems have evolved from direct systems and use software to analyse the input sentence and then apply linguistic and lexical rules, called transfer rules, to map grammatical information from one language to another. Using 'parser' software, these systems identify the structure of the input sentence to an acceptable level. After creating a parsing tree, the system uses its transfer rules to rewrite the tree according to the syntax of target language. Then a sentence generator program builds sentences and takes care of proper tense and number.

For deeper analysis and better translation, shallow semantic transfer systems analyse sentences for their meaning. Such systems have been developed using advances in computational linguistics and AI. These capture the meaning of sentences and then apply transfer rules to rewrite the result in target language.

Programs called 'analysers' have been developed, which often contain a parser and identify the appropriate representation term for each word or phrase and assemble the terms into a coherent structure. The systems generator finds appropriate target language words for the semantic frames it encounters.

7.3. Inter-lingual Systems

These systems use software to translate text by using a central data representation notation called inter-lingual. Only experimental prototypes have been developed which run on workstations. Inter-lingual systems can usually produce several paraphrases of the input sentence using semantic analysers and generation rules. More and more powerful generators

are being developed to handle the complexity of languages.

Inter-lingual systems require more detailed analysis than is necessary for any language combination, resulting in overloading of the generator and failure if it exceeds the flexibility of the interlingua.

These systems also use less number of rules and thus perform better on multi-lingual MT systems instead of having transfer rules for each language pair. They create one representation midway between all the languages which means, less human effort is required to construct transfer rules. Adding a new language to an inter-lingual system is therefore easier and the system is language neutral.

7.4 Statistical Systems

The inter-lingual and transfer systems require large lexicons and rule sets to make it robust when handling new text. In direct MT, the information is encoded in dictionaries, phrases and words; in transfer system, the data is placed in grammar, lexicons and transfer rules. Inter-lingual systems contain the information in representation databases and lexicons.

The work required for creating such systems is enormous and therefore attempts to acquire the requisite information automatically have gained momentum recently, leading to the statistical systems. In statistical systems, the computer creates a correspondence lexicon from an existing body of parallel text in two languages, using programs to line up each word in a sentence with its foreign counterpart. Over 90 per cent accuracy has been achieved in automatically developing correspondence lexicon of words and phrases. Translation is then done by direct substitution followed by the recording of words using statistically derived rules.

But statistically-based systems require large mainframe systems to build and store their correspondence lexicons. They also need millions of sentences of parallel text containing the words to be translated.

7.5 Hybrid Systems

Recent developments show that it is possible to create hybrid statistically-based transfer systems, where the correspondence lexicon contains not only words and phrases but syntactic terms that represent line, number and part of speech. Such systems can thus minimise the human effort required in lexicon, grammar, and rule construction.

8. DEVELOPMENTS IN MT SYSTEMS

Japan has been one of the most active countries in the development of MT systems. ATLAS-1 is a syntax-based system—world's first commercially available English-Japanese translation system. The prototype was completed in 1982. ATLAS-11 is a semantic-based system which translates from Japanese to English although its ultimate aim is to translate into three or more languages. Both these systems require per- as well as post- editing work. These systems also have benefited from research in the area of artificial intelligence.

In USA, the first commercial computer-based system—MicroCAT appeared in 1983. At present the Sun Microsystem Sparc station and other midrange Unix workstations are host to many commercial MT systems. Engspan is a recent system designed on 386 machines by American Health Organisation. This system translates from English to Spanish and is derived from the earlier system developed on mainframe computers.

East Asian Consortium CICC (Centre of the International Co-operation for

Computerisation) has been working on building such system working with groups in East Asian countries. Such systems in US include KBMT (Carnegie Mellon University), Ultra (New Mexico State University) and Panglose (jointly developed by the above two universities and University of Southern California) with many multiple chained workstations. PC versions may be available by the year 2000 AD.

Japan is leading in R&D work for speech translation. The ATR (Advanced Telecom Research) Institute at Kyoto, Japan has developed a prototype translating telephone, enquiries about configuration registration which aims at translating anyone's speech without customisation to specific voice patterns using a vocabulary of 1500 words.

Speech-recognition systems use microphone with software which match incoming sounds against stored patterns to produce a string of text with 90 per cent accuracy. Another system from NEC can recognise 5000 isolated words or a 1500-word vocabulary of fluid speech.

In Germany, a project called Verbmobile, to build a portable translator device is under progress. This system uses English as a kind of interlingua for translating business discussions.

The multi-lingual information retrieving systems are expected to be commercially available to provide information on selected topics by giving a collection of keywords or a few relevant paragraphs. A multilingual information retrieval system developed by Fujitsu Labs (Japan) can translate query words from German or English into Japanese to retrieve relevant articles, translate their titles and a request to translate even full article.

An MT machine developed at Carnegie Mellon University, using optical scanning and OCR technology, translates the text placed on its face plate. This technology may

Deciding for MT

- Have an application for which MT is appropriate.
- Make the input for translation available in machine-readable form, otherwise the cost increases significantly.
- There should be a large volume of material for translation (approx. 100,000 words per month).
- Initially have only one subject (domain); it can be expanded as experience is gained in using MT.
- The text to be translated should contain no ambiguities.
- Good word processing support and skilled operators should be available.
- Have positive attitude about using MT, especially in the beginning stage which involves customizing the dictionaries and gaining proficiency in post-editing.
- Buy software according to Identified criteria specific for your needs, also look for customer support from vendor.
- Test the systems on randomly selected texts, making sure that outputs to be compared are under the same conditions.
- The software should be easy to use and the system should be well maintained.
- Know the language combinations that have been developed for the system, the size of dictionaries or knowledge bases.
- Know the ease with which dictionaries can be added and potential expandability to include the domain of you interest.

be available commercially in general domain in the next 10 years.

A hand-held word-translation device, called Passport, uses holography to input text. The user can photograph a piece of text which is shown on its display. You select the words by circling them and the machine provides a translation on an LCD using bilingual dictionary chip for each language.

Another machine called Signspeak can translate American sign language for hearing-impaired and spoken English using a glove with sensors to track arm movements.

C-Star system jointly developed by Camping & University of Karbuls (Germany) trained networks recognises English text and generate Chinese output. Another project deals with English and Russian.

9. MT SOFTWARE FOR DESKTOP

Today's PCs and workstations are powerful enough to handle MT software. Therefore there is a migration of application software from mainframe to desktop.

Machine Translation software for desktop computers translate language sentence by sentence using AI or linguistic rules to deal with syntax and grammar. The documents are processed in either batch mode, where the sentence allows post-editing of each translated sentence or in interactive mode where a number of optionals are displayed for the operator to choose. The price of the package range from $79 to $1200 and are available in Arabic, Danish, Dutch, Finnish, French, German, Greek, Italian, Japanese, Korean, Russian, Spanish and Swedish. For Portuguese, Chinese and Norwegian, packages have also been developed. Software is generally for one way or two way language pairs. e.g. Language Assistant Series 5.0 from Microtac and GTS-Basic 1.0 and GTS- Professional from Globalink respectively. Professional Translation System 2.0 from Toltran uses a modular language translation concept for selling language as source or target language module.

Details on some MT software packages available on personal computers has been provided in the Appendix.

10. FUTURE OF MT

All over the world, there is move towards networking and to have more electronic and paper-less communication. Thus there is a greater need for inexpensive technology for translation assistance of e-mail messages.

In the near future, MT is going to become a way of life to help in breaking the language barrier. Translating telephones, multilingual e-mail, and small computers that would scan and translate. This would enable business across the world, travel and bargain to foreign countries, without learning language.

Breakthrough in speech recognition technology will make it possible to develop speaker-dependent applications from the written-text MT systems of today. Research in developing knowledge sources and internal representations for MT will enable faster and less human dependent systems. This will enable better ways to capture, store, represent knowledge development in general purpose systems; parsed information achieved will be a boon for building large knowledge bases.

Many technologies would be involved in doing this. These include automated speech recognition, speech generation, OCR and machine translation. Intermediate tools are already available. Already systems can recognise limited amount of speech without customisation. Further training on voice patterns may give unlimited capacity. When the specific vocabulary in a limited environment of subject (included to a few theorems and words) is used, the error rate is negligible.

The spread of e-mail and computer networks have created need for an inexpensive technology for browsing type translation assistance of e-mail messages.

At present, fax machines have quietly changed our lives and business. In coming years, automated MT systems will make speech and writing in a multitude of languages, easier to find, read and understand.

BIBLIOGRAPHY

1. Fransman, Martin. The market and beyond: Information technology in Japan. Cambridge University Press, USA, 1993. pp. 219-22.
2. Hovy, Edward. How MT works. *Byte*, 1993, **18**(1), 167-76.
3. Hovy, Edward. MT at your service. *Byte*, 1993, , 160.
4. Hutchins, W J. Why computers do not translate better. *Aslib Proceedings*, 1992, **44**(10), 351-59.
5. Hutchins, W J. Machine translation and machine-aided translation. *Journal of Documentation*, 1978, **38**(2), 119-59.
6. Ishida, Richard. Future translation work-benches some essential requirements. *Aslib Proceedings*, 1994, **46**(6), 163-70.
7. Keynes, Multon, *et al.* ELINOR: The Electronic Library Project at De Montfort University. *Aslib Proceedings*, 1993, **45**(1), 3-6.
8. Miller, L.C. Babelware for the Desktop. *Byte*, 1993, **18**(1) , 177-83.
9. Poulter, Alan. Building a browsable virtual reality Library. *Aslib Proceedings*, 1994, **46**(6), 151-55.
10. Riley, Christopher. The foreign language barrier in information transfer at the University of Tasmania. *Aslib Proceedings*, **44**(10), 1992, 331-40.
11. Scott, B.E. Five layers of ambiguity. *Byte*; 1993, **18**(1), 174.
12. Wilks, Yorick. Developments in MT research in the US. *Aslib Proceedings*, 1994, **46**(4), 111-16.
13. Vasconcellos, M. Machine translation. *Byte*, 1993, **18**(1), 153-64.