Singaporean Journal Scientific Research (SJSR)
ISSN: 2231 – 0061 Vol.2, No.2 pp.110-117 year:2009

available at: : http://www.iaaet.org/sjsr

Machine Translation Systems

P.Kumaresan
Research Scholar
Dep't of Linguistics, Tamil University, Thanjavur

Dr.S.Rajendran
Professor & Head, Dep't of Linguistics
Tamil University, Thanjavur.

1. Introduction

"Translating Languages with Computer" – Machine Translation (MT) has been one of the greatest dreams in computer applications. Machine Translation systems acts as a bridge to access cross lingual information by making the documents available in one language to another language. Such systems are inexpensive, instantaneous and multiplicative when compared to human translation. Building such a system across a pair of languages is nontrivial; fully automatic high-quality translation of an arbitrary text from one language to another is far too hard to automate completely. The level of complexity in building such a system depends on the similarities and difference among the pairs of languages.

But the dream of building a deployable MT System is gradually becoming a reality. Research on MT is an intellectual challenge with worthy motive and practical objective. The challenge is to produce translations as good as those made by human translators. The motive is removal of language barriers. The practical objective is the development of economically viable systems to satisfy growing demands for translations. Contrary to general belief, there is a considerable shortage of human translators even for technical translations. To fill this vacuum there is an increasing demand, worldwide, for MT systems.

2. Computational Linguistics and Natural Language Processing

Theoretical issues in Computational Linguistics (CL) is concerned with syntax, semantics, discourse, language generation and language acquisition. Historically, it included the study of natural languages as well as artificial (Computer) languages. Applied work in computational linguistics, however, includes computer aided instructions, database interfaces, machine translation, speech understanding etc. NLP recently emerged a major area of research and the progressive developments have made it possible to provide computer aids for text processing, writing grammar and electronic dictionaries, construct efficient parsers and even to build systems for machine aided translation, speech recognition etc.

3. Machine Translation

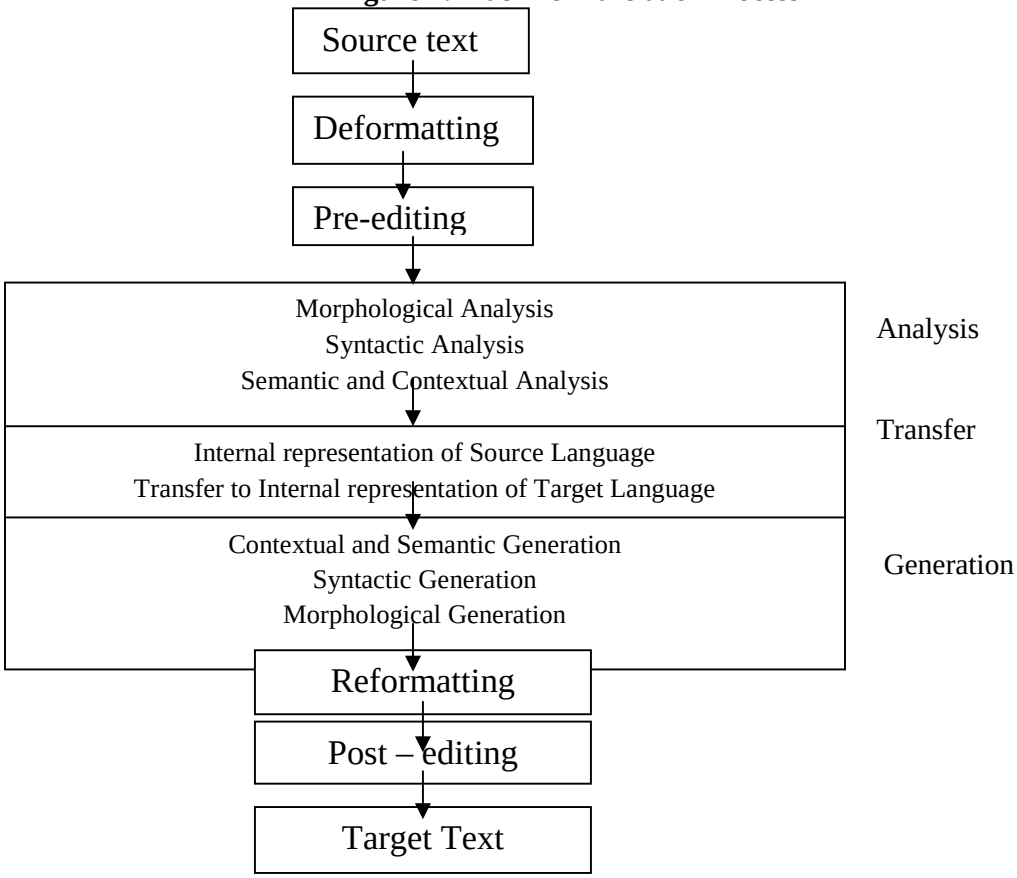
MT comes under a generic heading of Natural Language Processing (NLP). At the same time, because the technology involves many complex tasks, it is often seen as a category unto itself. This special status of MT also stems from the fact that it was the earliest kind of NLP. The theoretical and methodological bases of MT are computational linguistics theories and NLP technologies. Application of these theories and methodologies involve many issues such as dictionaries/lexicon, terminology banks, analysis of source sentences, transfer of intermediate representations, generation of target sentences, computer environments for developing and examining machine translation systems, operational environments, pre-editing of source sentences, post editing of target sentences and so on.

4. Machine Translation Process

Morphological analysis determines the word form including inflections, tense, number, part of

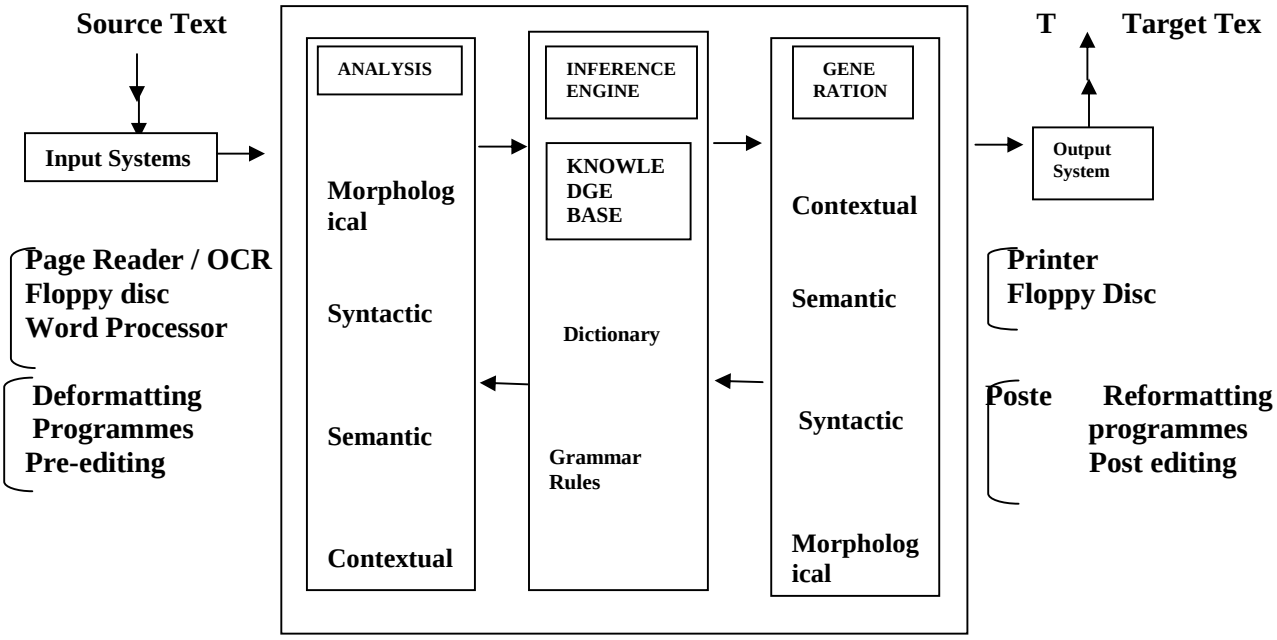
speech and so on. Syntactic analysis determines which word is the subject which one is the object, and so on. Semantic and contextual analysis determines a proper interpretation of a sentence from plural results produced by the syntactic analysis. Syntactic and semantic analysis is very often a combined operation and is executed simultaneously to produce syntactic tree structure and semantic network. The result is an internal representation of a sentence. The internal representation of the target language is often the same as that of source language, but sometimes the change of internal representation is required. The sentence generation phase is just the reverse of the analysis process.

Figure 1: Machine Translation Process



Machine Translation System

Figure 2. Machine Translation System



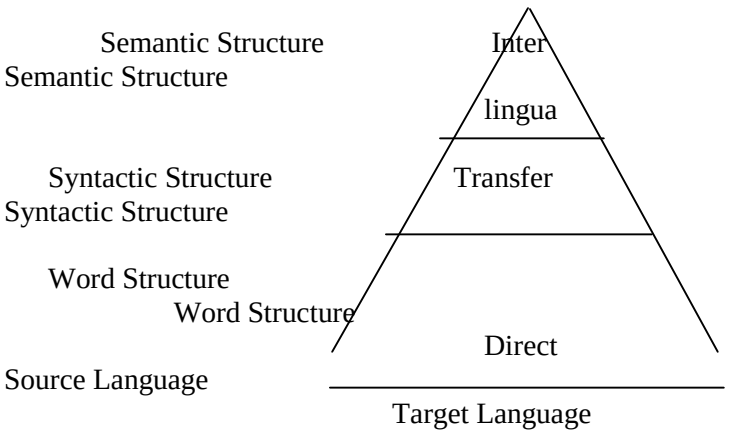
Central portion of the figure, viz. inference engine is nothing but a knowledge base which stores various rules of source language as well as target language, dictionaries and grammar supplemented by common sense rules for translation.

Thus machine translation is basically a knowledge information processing system and is a large scale Artificial Intelligence problem. FGCS project in 1982 by MITI in Japan did not directly include machine translation. However one of the focal research themes of the project was identified as NLP and it was recognized that high level machine translation systems are likely to be the future application of FGCS technology.

5. Evolution of Various Approaches

Research in the area of machine translation started much earlier to that in the area of Natural Language Processing. Various approaches for machine translation have been suggested and tried out. With the advances in computational linguistics and availability of powerful computers further refinements in these approaches were noticed, besides the emergence of newer approaches. These are reviewed in the following sections.

Figure 3: Approaches to Machine Translation



6. Direct Approach: First Generation

The very first proposals of the "first generation" of MT for using computers to translate till about 60s were called "direct" approach. These were essentially dictionary driven methods with no or low-level syntactic analysis and use of semantic features. These efforts came to an end with the publication of ALPAC report in 1966. There was a very little activity for about a decade thereafter.

7. Transfer Approach: Second Generation

The interests were revived in 1976 with the installation of SYSTRAN, a weather forecast translation system in Canada. Increasing activities followed throughout the world and rule based, syntax oriented abstract representation approaches became the "Second generation" of MT. Examples are Ariane, of the GETA Project in France, EUROTRA of the commission of European Communities and Mu of the Kyoto University, Japan. At the same time, considerable work was also going on rule based Inter-lingua models. In machine translation systems, an intermediate representation is necessary to express the results of sentence analysis. This represents syntactic and semantic structures of an input sentence given as a character string. Syntactic structure is shown as tree and semantic structure as a network. This intermediate representation of inter-lingua approach is called inter-lingua. Notable amongst the inter-lingua models was the CMU project at Carnegi Mellon University in USA and multilingual machine translation project at Centre of International co-operation for computerization in Japan.

Around this time commercial systems began appearing on the marked scene, most of them being for translations from English to Japanese and vice-versa. They also migrated from main frames to PCs and became popular.

8. Corpus Based approach: Third Generation

By about 1989, a new era and the "third generation" began in MT research with the emergence of wide range of collectively called "corpus based" approaches. This represented a new departure in MT research. The corpus-based approaches include two categories, the first one uses direct information derived from corpora for analysis, transfer and generation of translations. This category includes statistics based, example based and connectionist approaches. Second category makes an indirect use of corpora as a source of information for deriving or compiling lexical, grammatical and knowledge databases. This second category consists of a range of activities such as database compilation, lexical and knowledge acquisition and statistical information to aid or complement rule based methods.

Essence of statistics based method is the alignment of sentence in the two languages and the calculation of the probabilities that any one word in a sentence of one language corresponds to two, one or zero words in the translated sentence in target language. The IBM-candide research project based on a large corpus of the Canadian Hansrad, records of parliamentary debates in English and French is an example.

Example based method was first proposed in 1980s but was implemented only a decade later. The basic philosophy in this method is that translation is often a matter of finding analogues examples. The method essentially relies on a bilingual database examples derived from a large corpus of texts and their translations. Example based machine translation approach has been extensively used in ATR project of Japan for spoken language translation. Another example is knowledge base machine translation (KBMT) at Carnegi Mellon University in USA.

Connectionist approach is a result of research in parallel computation, neural network or connectionism. Being the latest, these developments have attracted the MT researchers also. Connectionist method computes the distance between input text segments and bilingual text data in example based MT model. It offers the prospect of systems "learning" from past successes and failures. Previously, learning has meant that systems suggest changes on the basis of statistics about corrections made by users at the post-editing stage.

9. Hybrid Approaches on the Horizons

The hybrid approaches are also on the horizons. In some such approaches, corpus information is used for tuning analysis and transfer grammar. In others a standard transfer based MT approach is followed using traditional analysis and generation technique but having transfer component based on aligned bilingual corpora. In yet others, statistical information is used as the source of preference assignment during text disambiguation.

10. Components in Machine Translation System: Electronic Dictionaries and Lexical Databases

In figure 2, a typical machine translation system; Dictionary is one component in the central portion. This implies monolingual / bilingual / multilingual, machine-readable dictionaries for a particular pair of pairs of languages which the machine is required to translate. The dictionaries provide definitions of words, their syntactic categories and at times usage by way of examples. They are, however, a poor source of much needed information for sophisticated semantic processing, i e. how a word is used in relation to other in a sentence.

The electronic dictionary is not simply a machine-readable dictionary; it is a dictionary containing all the information necessary for computers to understand natural language. Thus an electronic dictionary must contain meanings of words, i.e., concepts expressed by words, their grammatical characteristics when they express concepts, and knowledge necessary for understanding concepts. Large-scale electronic dictionaries are being developed for machine translations in each of the MT projects currently under way at several places. Small-scale electronic dictionaries are used for question answering and speech recognition systems.

10.1 Study of Corpus

It has now been well recognized that a large text corpus is very useful and is much informative in the construction of electronic dictionaries, lexical databases and allows testing of grammar formalism etc. Many NLP and MT projects have acknowledged the need of a large corpus. Japan's EDR project and multilingual machine translation projects have built their corpora fist and thereafter started development of electronic dictionaries etc. Founding of Linguistic Data Consortium (LDC) in 1992 by the US Govt .for large scale development and widespread sharing of resource for research in linguists technology is yet another acknowledgement of importance of study of corpus. Goal of LDC is to collect, create and disseminate corpora of texts as well as speech to the researchers in NLP, MT and speech recognition. Yet another use of corpus is found in evaluation of machine translation system itself.

10.2 **Morphological Analyser and Generator**

Computational morphology deals with the recognition, analysis and generation of words. Most regular and productive morphological process a cross languages is inflection, while other aspects such as derivation, affixes and combining forms etc. are also included. In many languages nouns and adjectives vary according to number, gender, and case. Inflection alters the form of the word in number, gender, mood, tense, aspect, person and case. Similarly the verbs also take different forms depending upon person, tense etc.

A morphological analyser or generator supplies information concerning morphosyntactic properties of the words it analyses or constructs. In principle, there are two ways to deal with morphologically related forms. One is to store all the word forms with associated relevant properties, for example, walk (verb present plural), walks (Verb present singular), walking (verb present progressive) and walked (verb past). The other way is to store one base form walk (verb) with rules to relate variants. These options will have to be chosen depending upon how expensive the storage will be. Further, languages are creative and hence new words enter the language. So storing all variants is likely to become an intractable proposition.

10.3 Grammars and its Characteristics

(i) Phrase structure Grammar (PSG): There are quite a few variants in this Category, such as, context free PSG, context sensitive PSG, Augmented Transition network Grammar (ATN), Definite clause (DC) Grammar, Categorical Grammar, Lexical Functional Grammar (LFG), Generalized PSG, Head driven PSG, Tree Adjoining Grammar (TAG).
(ii) Dependency Grammar
(iii) Case Grammar
(iv) Systematic Grammar
(v) Montague Grammar

Which grammar would suit a particular language depends on many factors specific to that language. For example, PSG has a serious problem in the analysis of sentences in Japanese. On the other hand Dependency Grammar has been very popular for Japanese but its draw back comes out clearly in disambiguation. Case grammar offers certain advantage as sentence representation is done by case frames. Advantage is that, a sentence in different languages, which express same contents, may have the same case frames. Due to this advantage many MT systems seen to have adopted it for sentence

analysis and also form sentence generation. However, the difficulty arises in a situation where different usages of same verb of noun have to be distinguished. And if in language such usage is large, the number of semantic markers also becomes large. A sentence includes lot of information such as syntactic, semantic, textual, inter personal and so on. Grammar formalisms are just a framework to explain basic structure of a language and any one-grammar formalism may not apply across the languages.

11. Major Obstacles in MT

11.1. Text Input

The first module in any MT systems is Text Input of source language as shown in Figure 2. If machine-readable text in source language is available the process can be started by inserting a floppy disc. But if printed pages are to be input, either manual typing or optical character reader (OCR) is to be used. In manual typing speed and the cost of typists are to be accounted for. In case of OCR proof reading and error correction by human are to be managed.

11.2. De-formatting and Reformatting

An operative machine translation system is expected to do much more than simply translate individual sentences. Most of the text, which needs translation quickly, e.g. technical documentation, is heavily formatted. In some texts major portion on a page may be non-translatable material in the form of figures, flow charts of tables. Therefore, specific modules have to be built into the MT system, which will identify text portion to be translated and generate a template of that page. The individual translation units, usually sentences, but in case of headlines or table entries, single word or phrases, are automatically recognized and numbered consecutively. These are written into a text file and transferred for translation. After translation, the file containing the target language text units is returned to the user. This text has now to be reformatted after making appropriate correction (Post editing) etc. This reformatting step takes care of ensuring that the target language text is available with same layout as original including figures, flow charts, tables etc.

11.3. Pre-Editing and Post Editing

At times, segmentation of long sentences into two or more short sentences is also required which is often done manually at the pre-editing stage. Pre-editing and Post editing has a certain correlation. When a heavy of elaborate pre-editing is performed, very simple post editing may be sufficient and vice-versa. Post editing is generally unavoidable and hence many machine translation users do only post editing. It is essential to avoid ambiguities and also to improve the quality and style of translation. There may be words unknown to the system and difficult to analyse. In such cases post editing provides a facility to update the system lexicon. Post editing the machine output is not the same as revising a “human” translation. While the machine will make severe errors in syntax, human translator will make fewer but random errors, which are less predictable.

It is reported after a mail in poll conducted by Word Perfect magazine in June 1993 of the MT software now available on PCs, that the Pre editing is basically division of long sentences into shorter ones, fixing up punctuation marks and blocking material that does not require translation. Hence not much time is spent on it. Some users of Japanese to English system however, reported that pre-editing takes about 40% of total translation time.

On the other hand the poll reports that post editing, generally, accounts for larger share of production time and also the cost. Some language combinations are reported to give better results and consequently require lesser post editing. There is also another important factor, which require consideration at the time of post editing and that is the quality of expected output. A few may require very high quality output e.g. insurance contracts. Some others may require editing for accuracy but not for style. Yet others may require “information only” for “understanding only” and much time may not be spent on post editing.

11.4. Introduction of Machine Translation Systems in an Organization

The state of the art in computational linguistics does not permit the perfect translation of random texts. Therefore, if a text is translated with an aim of publication, post editing by human translation will remain a necessity, even if a system is tuned for specific subject area. The quality of translation does not hinge only on MT system but is equally dependent on the quality of source text. One also has to consider intended purpose of text,

expectations of readers and even stylistic preferences of post editor.

Therefore, the introduction of MT system into an existing organization, whether it is a large company or a translation bureau requires several steps. First, purpose of translation usage must be clearly established. Then kinds of documents must be specified and the expected quality, speed and cost must be clarified. In-appropriate use is likely to lead to frustration. For productive use of MT system, an initial training period of about two weeks will be necessary. At least two specialists, one for pre editing and other for post editing will have to be assigned. They are required to have some background in linguistics or languages, which the machine will translate. They must be trained for operations such as pre editing, post editing, dictionary changes and enhancements. Sentential styles of input and output documents must be carefully studied and translation equivalents must be determined. The documents must be translated by trial basis for sufficient volume and the dictionary and grammar must be tuned to the environment. Second training for about a week after a few months of trial translations may be beneficial as it can answer questions which had arisen during actual application.

12. Objectives and Activities IAMT

The specific objectives of IAMT are collection of information, exchange and dissemination of the same and standardization. To meet these objectives IAMT under-takes the following activities:

(i) To convene biennial General Assembly
(ii) Sponsor workshops, symposia and conferences on MT and related technologies and applications.
(iii) Organize tutorials and training courses.
(iv) Establishment of technical committees, special interest groups and study teams.

13. Future of Machine Translation

With the increasing interest in NLP and Machine Translation all over the world, there will be continuous improvements in the existing systems with a view to make them more robust and easily adaptable to the needs of the users. Certain areas have been already identified. These are enumerated below:

(i) **Development of dictionaries and lexical databases using Corpora:** Methods for producing dictionaries even from untagged texts have been presented in seminars and conferences. Lexical databases are being prepared by organizations, such as Microsoft corporation, USA, Institute of Machine Translation at the University of Stuttgart etc.
(ii) **MT for specialized applications:** Large corporations with multinational operations are devoting considerable efforts to develop MT systems for in house use in specific domains. Examples are “Simplified English” developed and being used by Boeing Corporation for translation of maintenance manuals of Boeing aircrafts. General Motors are also working in developing similar systems.
(iii) **MT with human intervention:** This is a paradigm shift with realistic attitude, which is gaining ground. It was a dominant theme at the conference “Language Technology 2000” held in Germany in 1993.
(iv) **MT on Network:** Availability of MT software on a network appears to be a distinct reality in near future.
(v) **Information retrieval:** Some researchers have suggested that MT should be introduced in information retrieval systems and databases systems so that users all over the world can have access to any on information source.

14. Conclusion

We have seen the state of art of developing MT Systems in and outside India. The development of MT systems outside India, especially in European and America, is remarkable. India is also making attempts to develop MT systems for Indian Language to Indian Language transfer as well as English to Indian languages transfer. Ministry of Communication and Information Technology and Ministry of Human resources give financial support to these programs. Of course we have to travel a lot to achieve this goal. Though such attempts are expensive, at least for the sake of Research and Development and from the point of view of experiment we have to try to build such systems. The present thesis explores one such possibility.

References

[1] ALPAC. 1966. Language and Machines: Computers in Translation and Linguistics. Washington D.C.: National Academy of Sciences.

[2] Buchmann, B. 1987. Early History of Machine Translation. In: King, M. (ed.) 1987. Machine Translation: The State of the Art. Edinburgh: Edinburgh University Press.

[3] Geetha, K. 1985. Subsystems of Principles: A Study in Universal Based on Tamil Syntax. Ph.D. Thesis. Kanpur: IIT.

[4] Isabelle, P. 1993. Machine-Aided Human Translation and the Paradigm Shift (manuscript). Japan: MT Smmit IV.

[5] King, M. (ed.) 1987. Machine Translation: The State of the Art. Edinburgh: Edinburgh University Press.

[6] Newmark, P. 1988. A Text Book of Translation. New York: Prentice Hall.

[7] Nida, E.A. 1974. Theory and Practice of Translation. Leiden: United Bible Societies.

[8] ------1975. Language Structure and Translation. Stanford, California: Stanford University Press.

[9] Nirenberg, S. (ed.) 1987. Machine Trasnlsation. Cambridge: Cambridge University Press.

[10] Rohrer, C. 1993. The Future of MT Technology (manuscript).Japan: MT Summit IV.