Computational Linguistics and Chinese Language Processing
Vol. 18, No. 1, March 2013, pp. 47-78

Machine Translation Approaches and Survey for Indian Languages

Antony P. J.*

Abstract

The term Machine Translation is a standard name for computerized systems responsible for the production of translations from one natural language into another with or without human assistance. It is a sub-field of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another. Many attempts are being made all over the world to develop machine translation systems for various languages using rule-based as well as statistically based approaches. Development of a full-fledged bilingual machine translation (MT) system for any two natural languages with limited electronic resources and tools is a challenging and demanding task. In order to achieve reasonable translation quality in open source tasks, corpus based machine translation approaches require large amounts of parallel corpora that are not always available, especially for less resourced language pairs. On the other hand, the rule-based machine translation process is extremely time consuming, difficult, and fails to analyze accurately a large corpus of unrestricted text. Even though there has been effort towards building English to Indian language and Indian language to Indian language translation system, unfortunately, we do not have an efficient translation system as of today. The literature shows that there have been many attempts in MT for English to Indian languages and Indian languages to Indian languages. At present, a number of government and private sector projects are working towards developing a full-fledged MT for Indian languages. This paper gives a brief description of the various approaches and major machine translation developments in India.

Keywords: Corpus, Computational Linguistics, Statistical Approach, Interlingua Approach, Dravidian Languages0.

* Professor and Head, Department of ISE, St. Joseph Engineering College, Mangalore, VTU.
E-mail: antonypjohn@gmail.com

1. Introduction

MT refers to the use of computers to automate some of the tasks or the entire task of translating between human languages. Development of a full-fledged bilingual MT system for any two natural languages with limited electronic resources and tools is a challenging and demanding task. Many attempts are being made all over the world to develop MT systems for various languages using rule-based as well as statistical-based approaches. MT systems can be designed either specifically for two particular languages, called a bilingual system, or for more than a single pair of languages, called a multilingual system. A bilingual system may be either unidirectional, from one Source Language (SL) into one Target Language (TL), or may be bidirectional. Multilingual systems are usually designed to be bidirectional, but most bilingual systems are unidirectional. MT methodologies are commonly categorized as direct, transfer, and Interlingua. The methodologies differ in the depth of analysis of the SL and the extent to which they attempt to reach a language independent representation of meaning or intent between the source and target languages. Barriers in good quality MT output can be attributed to ambiguity in natural languages. Ambiguity can be classified into two types: structural ambiguity and lexical ambiguity.

India is a linguistically rich area. It has 18 constitutional languages, which are written in 10 different scripts. Hindi is the official language of the Union. Many of the states have their own regional language, which is either Hindi or one of the other constitutional languages. In addition, English is very widely used for media, commerce, science and technology, and education only about 5% of the world's population speaks English as a first language. In such a situation, there is a large market for translation between English and the various Indian languages.

Even though MT in India started more than two decades ago, it is still an ongoing process. The third section of this paper discusses various approaches used in English to Indian languages and Indian language to Indian language MT systems. The fourth section gives a brief explanation of different MT attempts for English to Indian languages and Indian languages to Indian languages.

2. History of MT

The major changeovers in MT systems are as shown in Figure 1. The theory of MT pre-dates computers, with philosophers 'Leibniz and Descartes' ideas of using code to relate words between languages in the seventeenth century (Hutchins *et al*., 1993). The early 1930s saw the first patents for 'translating machines'. Georges Artsrouni was issued a patent in France in July 1933. He developed a device, which he called a *'cerveau mécanique'* (mechanical brain) that could translate between languages using four components: memory, a keyboard for input,

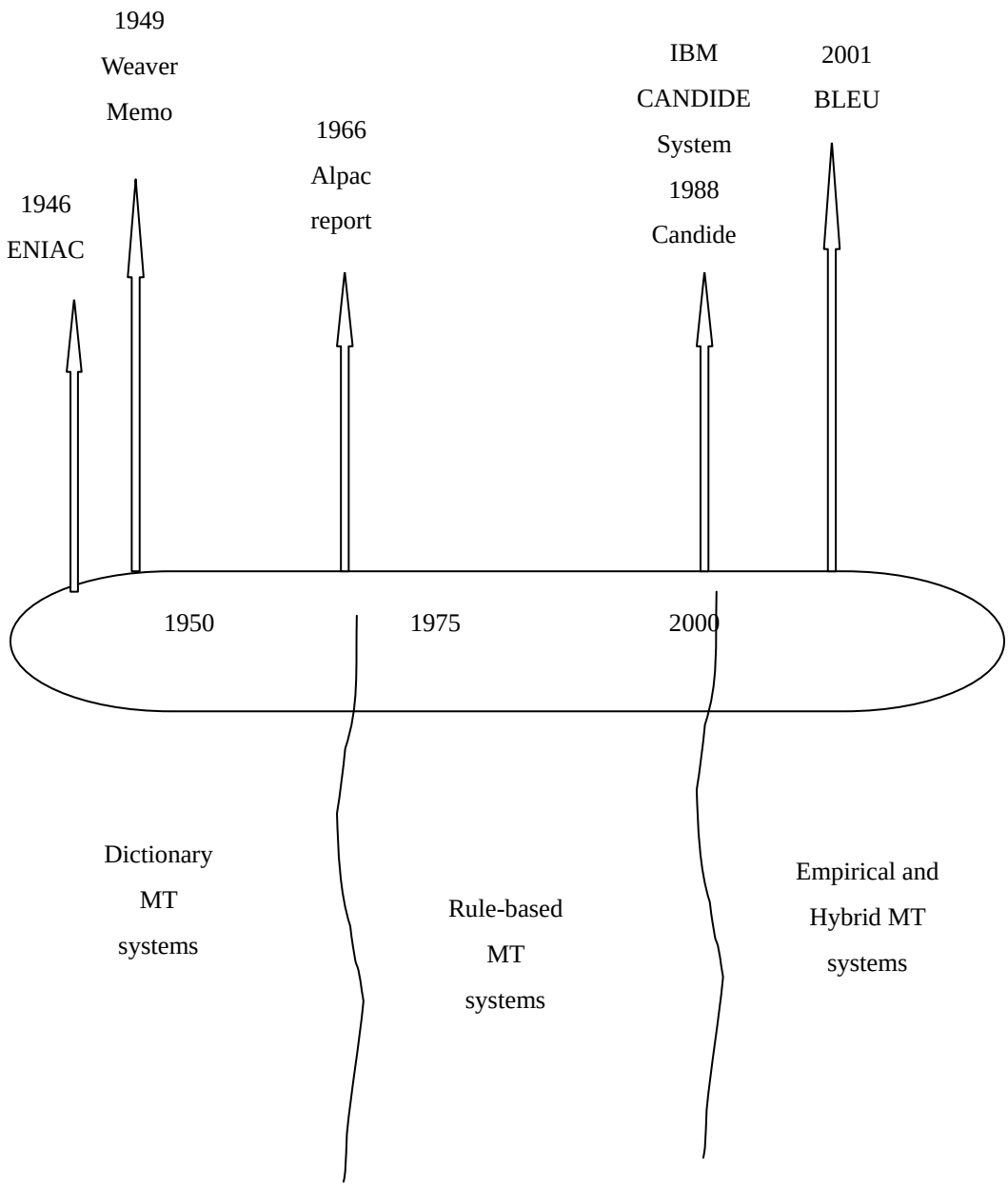

Figure 1. Major changeovers in MT Systems.

a search method, and an output mechanism. The search method was basically a dictionary look-up in the memory; therefore, Hutchins is reluctant to call it a translation system. The proposal Russian Petr Petrovich Troyanskii patented in September 1933 bears a resemblance

to the Apertium system, using a bilingual dictionary and a three-staged process, *i.e.* first a native speaking human editor of the SL (SL) pre-processed the text, then the machine performed the translation, and finally a native-speaking human editor of the TL post-edited the text (Hutchins *et al.,* 1993; Hutchins *et al.,* 2000).

After the birth of computers Electrical Numerical Integrator and Calculator (ENIAC) in 1947, research began on using computers as aids for translating natural languages (Hutchins *et al.,* 2005). The first public demonstration of MT in the Georgetown-IBM experiment, which proved deceptively promising, encouraged financing of further research in the field. In 1949, Weaver wrote a memorandum, putting forward various proposals (based on the wartime successes in code breaking) on the developments in information theory and speculation about universal principles underlying natural languages (Weaver *et al.,* 1999). In the decade of optimism, from 1954-1966, researchers encountered many predictions of imminent 'breakthroughs'. In 1966, the Automated Language Processing Advisory Committee (ALPAC) report was submitted, which said that, for 'semantic barriers', there are no straightforward solutions. The ALPAC report committee could not find any "pressing need for MT" nor "an unfulfilled need for translation (ALPAC *et al.,* 1996)".

This report brought MT research to its knees, suspending virtually all research in the United States of America (USA) while some research continued in Canada, France, and Germany (Hutchins *et al.,* 2005). After the ALPAC report, MT almost was ignored from 1966-1980. In the year 1988, Georgetown-IBM experiment launched "IBM CANDIDE System," where over 60 Russian sentences were translated smoothly into English using 6 rules and a bilingual dictionary consisting of 250 Russian words, with rule-signs assigned to words with more than one meaning. Although Professor Leon Dostert cautioned that this experimental demonstration was only a scientific sample, or "a Kitty Hawk of electronic translation (Kitty Hawk[1])," a wide variety of MT systems emerged after 1980 from various countries and research continued on more advanced methods and techniques. Those systems mostly were comprised of indirect translations or used an 'interlingua' as an intermediary. In the 1990s, Statistical Machine Translation (SMT) and what is now known as Example-based Machine Translation (EBMT) saw the light of day (IBM, 1954). At this time the focus of MT began to shift somewhat from pure research to practical application using a hybrid approach. Moving towards the change of the millennium, MT became more readily available to individuals via online services and software for their personal computers.

[1] Kitty Hawk, North Carolina, USA was the site for the world's first successful powered human flight by the Wright brothers. "Kitty Hawk" references generally meant a break-through success in its early stages.

3. MT Approaches

Generally, MT is classified into seven broad categories: rule-based, statistical-based, hybrid-based, example-based, knowledge-based, principle-based, and online interactive based methods. The first three MT approaches are the most widely used and earliest methods. Literature shows that there have been fruitful attempts using all these approaches for the development of English to Indian languages as well as Indian languages to Indian languages. At present, most of the MT related research is based on statistical and example-based approaches. Figure 2 shows the classification of MT in Natural language Processing (NLP).

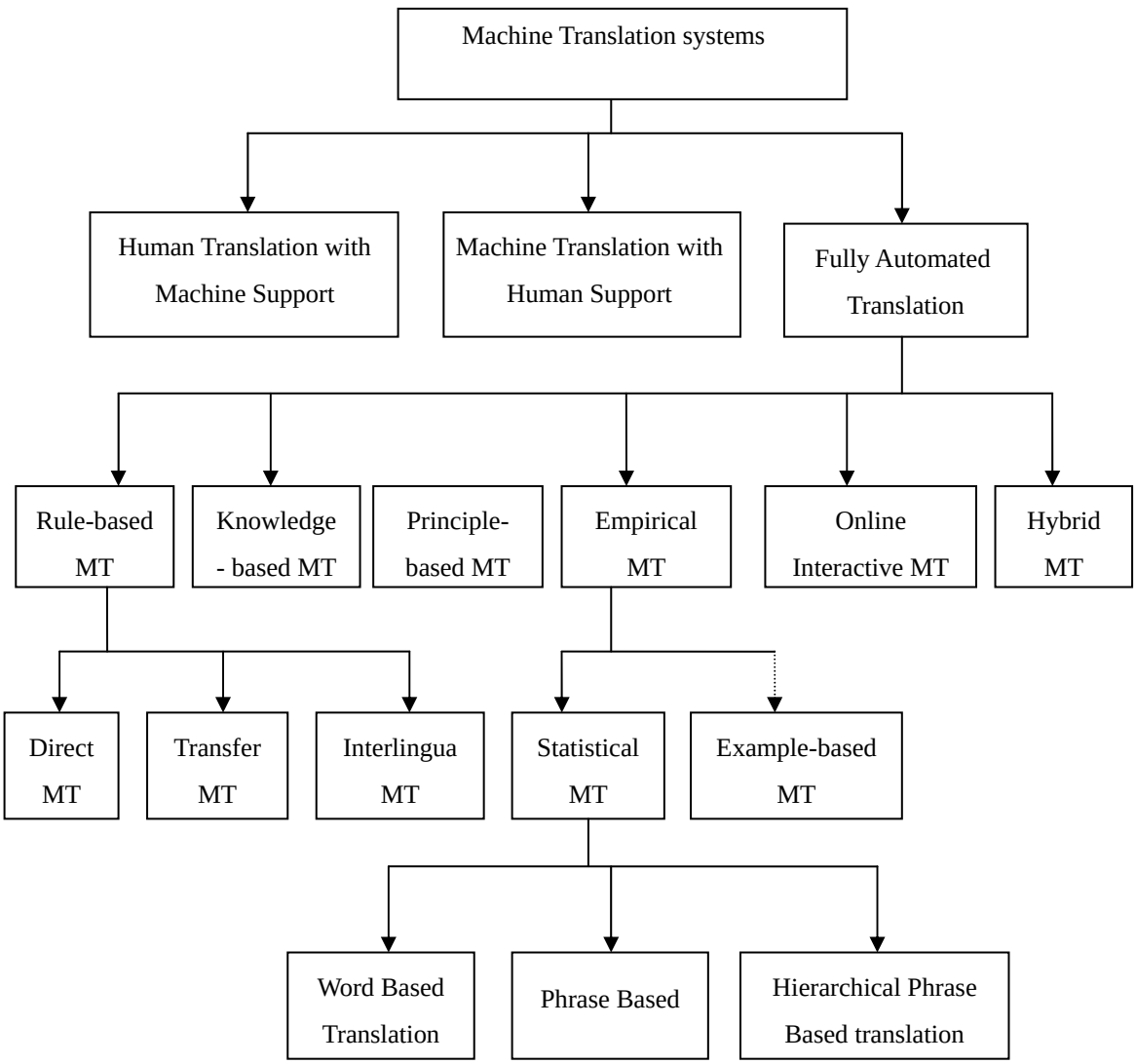

Figure 2. Classification of MT System.

3.1 Rule-based Approach

In the field of MT, the rule-based approach is the first strategy that was developed. A Rule-Based Machine Translation (RBMT) system consists of collection of rules, called grammar rules, a bilingual or multilingual lexicon, and software programs to process the rules.

Nevertheless, building RBMT systems entails a huge human effort to code all of the linguistic resources, such as source side part-of-speech taggers and syntactic parsers, bilingual dictionaries, source to target transliteration, TL morphological generator, structural transfer, and reordering rules. Nevertheless, a RBMT system always is extensible and maintainable. Rules play a major role in various stages of translation, such as syntactic processing, semantic interpretation, and contextual processing of language. Generally, rules are written with linguistic knowledge gathered from linguists. Transfer-based MT, Interlingua MT, and dictionary-based MT are the three different approaches that come under the RBMT category. In the case of English to Indian languages and Indian language to Indian language MT systems, there have been fruitful attempts with all four approaches. The main idea behind these rule-based approaches is as follows.

3.1.1 Direct Translation

In the direct translation method, the SL text is analysed structurally up to the morphological level and is designed for a specific source and target language pair (Noone *et al*., 2003; Dasgupta & Basu, 2008). The performance of a direct MT system depends on the quality and quantity of the source-target language dictionaries, morphological analysis, text processing software, and word-by-word translation with minor grammatical adjustments on word order and morphology.

3.1.2 Interlingua Based Translation

The next stage of progress in the development of MT systems is the Interlingua approach, where translation is performed by first representing the SL text into an intermediary (semantic) form called Interlingua. The advantage of this approach is that Interlingua is a language independent representation from which translations can be generated to different TLs. Thus, the translation consists of two stages, where the SL is first converted in to the Interlingua (IL) form before translation from the IL to the TL. The main advantage of this Interlingua approach is that the analyzer of the parser for the SL is independent of the generator for the TL. There are two main drawbacks in the Interlingua approach. The first disadvantage is, difficulty in defining the interlingua. The second disadvantage is Interlingua does not take the advantage of similarities between languages, such as translation between Dravidian languages. Nevertheless the advantage of Interlingua is it is economical in situations where translation among multiple languages is involved (Shachi *et al*., 2001).

Starting with the shallowest level at the bottom, direct transfer is made at the word level. Moving upward through syntactic and semantic transfer approaches, the translation occurs on representations of the source sentence structure and meaning, respectively. Finally, at the interlingual level, the notion of transfer is replaced with a single underlying representation called the 'Interlingua'. 'Interlingua' represents both the source and target texts simultaneously. Moving up the triangle reduces the amount of work required to traverse the gap between languages at the cost of increasing the required amount of analysis and synthesis.

3.1.3 Transfer Based Translation

Because of the disadvantage of the Interlingua approach, a better rule-based translation approach was discovered, called the transfer approach. Recently, many research groups have being using this third approach for their MT system, both abroad and in India. On the basis of the structural differences between the source and target language, a transfer system can be broken down into three different stages: i) Analysis, ii) Transfer and iii) Generation. In the first stage, the SL parser is used to produce the syntactic representation of a SL sentence. In the next stage, the result of the first stage is converted into equivalent TL-oriented representations. In the final step of this translation approach, a TL morphological analyzer is used to generate the final TL texts.

3.2 Statistical-based Approach

The statistical approach comes under Empirical Machine Translation (EMT) systems, which rely on large parallel aligned corpora. Statistical machine translation is a data-oriented statistical framework for translating text from one natural language to another based on the knowledge and statistical models extracted from bilingual corpora. In statistical-based MT, bilingual or multilingual textual corpora of the source and target language or languages are required. A supervised or unsupervised statistical machine learning algorithm is used to build statistical tables from the corpora, and this process is called the learning or training (Zhang *et al.*, 2006). The statistical tables consist of statistical information, such as the characteristics of well-formed sentences, and the correlation between the languages. During translation, the collected statistical information is used to find the best translation for the input sentences, and this translation step is called the decoding process. There are three different statistical approaches in MT, Word-based Translation, Phrase-based Translation, and Hierarchical phrase based model.

The idea behind SMT comes from information theory. A document is translated according to the probability distribution function indicated by $p(e|f)$, which is the Probability of translating a sentence f in the SL F (for example, English) to a sentence e in the TL E (for example, Kannada).

The problem of modeling the probability distribution *p(e|f)* has been approached in a number of ways. One intuitive approach is to apply Bayes theorem. That is, if *p(f|e)* and *p(e)* indicate translation model and language model, respectively, then the probability distribution $p(e|f) \propto p(f|e)p(e)$. The translation model *p(f|e)* is the probability that the source sentence is the translation of the target sentence or the way sentences in *E* get converted to sentences in *F*. The language model *p(e)* is the probability of seeing that TL string or the kind of sentences that are likely in the language *E*. This decomposition is attractive as it splits the problem into two sub problems. Finding the best translation $\tilde{e}$ is done by picking the one that gives the highest probability, as shown in Equation 1.

$$\tilde{e} = \arg\max_{e \in e^*} p(e \mid f) = \arg\max_{e \in e^*} p(f \mid e)\, p(e) \tag{1}$$

Even though phrase based models have emerged as the most successful method for SMT, they do not handle syntax in a natural way. Reordering of phrases during translation is typically managed by distortion models in SMT. Nevertheless, this reordering process is entirely unsatisfactory, especially for language pairs that differ a lot in terms of word-order. In the proposed project, the problem of structural differences between source and target languages is overcome successfully with a reordering task. We have also proven that, with the use of morphological information, especially for a morphologically rich language like Kannada, the training data size can be reduced considerably with an improvement in performance.

3.2.1 Word Based Translation

As the name suggests, the words in an input sentence are translated word by word individually, and these words finally are arranged in a specific way to get the target sentence. The alignment between the words in the input and output sentences normally follows certain patterns in word based translation. This approach is the very first attempt in the statistical-based MT system that is comparatively simple and efficient. The main disadvantage of this system is the oversimplified word by word translation of sentences, which may reduce the performance of the translation system.

3.2.2 Phrase Based Translation

A more accurate SMT approach, called phrase-based translation (Koehn *et al*., 2003), was introduced, where each source and target sentence is divided into separate phrases instead of words before translation. The alignment between the phrases in the input and output sentences normally follows certain patterns, which is very similar to word based translation. Even though the phrase based models result in better performance than the word based translation, they did not improve the model of sentence order patterns. The alignment model is based on flat reordering patterns, and experiments show that this reordering technique may perform

well with local phrase orders but not as well with long sentences and complex orders.

3.2.3 Hierarchical Phrase Based model

By considering the drawback of previous two methods, Chiang (2005) developed a more sophisticated SMT approach, called the hierarchical phrase based model. The advantage of this approach is that hierarchical phrases have recursive structures instead of simple phrases. This higher level of abstraction approach further improved the accuracy of the SMT system.

3.3 Hybrid-based Translation

By taking the advantage of both statistical and rule-based translation methodologies, a new approach was developed, called hybrid-based approach, which has proven to have better efficiency in the area of MT systems. At present, several governmental and private based MT sectors use this hybrid-based approach to develop translation from source to target language, which is based on both rules and statistics. The hybrid approach can be used in a number of different ways. In some cases, translations are performed in the first stage using a rule-based approach followed by adjusting or correcting the output using statistical information. In the other way, rules are used to pre-process the input data as well as post-process the statistical output of a statistical-based translation system. This technique is better than the previous and has more power, flexibility, and control in translation.

Hybrid approaches integrating more than one MT paradigm are receiving increasing attention. The METIS-II MT system is an example of hybridization around the EBMT framework; it avoids the usual need for parallel corpora by using a bilingual dictionary (similar to that found in most RBMT systems) and a monolingual corpus in the TL (Dirix *et al*., 2005). An example of hybridization around the rule-based paradigm is given by Oepen. It integrates statistical methods within an RBMT system to choose the best translation from a set of competing hypotheses (translations) generated using rule-based methods (Oepen *et al*., 2007).

In SMT, Koehn and Hoang integrate additional annotations at the word-level into the translation models in order to better learn some aspects of the translation that are best explained on a morphological, syntactic, or semantic level (Koehn *et al*., 2007). Hybridization around the statistical approach to MT is provided by Groves and Way; they combine both corpus-based methods into a single MT system by incorporating phrases (sub-sentential chunks) from both EBMT and SMT into an SMT system (Groves *et al*., 2005). A different hybridization happens when an RBMT system and an SMT system are used in a cascade; Simard proposed an approach, analogous to that by Dugast, using an SMT system as an automatic post-editor of the translations produced by an RBMT system (Simard *et al*., 2007) (Dugast *et al*., 2007).

3.4 Example-based translation

The example-based translation approach is based on analogical reasoning between two translation examples, proposed by Makoto Nagao in 1984. At run time, an example-based translation is characterized by its use of a bilingual corpus as its main knowledge base. The example-based approach comes under the EMT system, which relies on large parallel aligned corpora.

Example-based translation is essentially translation by analogy. An EBMT system is given a set of sentences in the SL (from which one is translating) and their corresponding translations in the TL, and uses those examples to translate other, similar source-language sentences into the TL. The basic premise is that, if a previously translated sentence occurs again, the same translation is likely to be correct again. EBMT systems are attractive in that they require a minimum of prior knowledge; therefore, they are quickly adaptable to many language pairs.

A restricted form of example-based translation is available commercially, known as a translation memory. In a translation memory, as the user translates text, the translations are added to a database, and when the same sentence occurs again, the previous translation is inserted into the translated document. This saves the user the effort of re-translating that sentence, and is particularly effective when translating a new revision of a previously-translated document.

More advanced translation memory systems will also return close but inexact matches on the assumption that editing the translation of the close match will take less time than generating a translation from scratch. ALEPH, wEBMT, English to Turkish, English to Japanese, English to Sanskrit, and PanEBMT are some of the example-based MT systems.

3.5 Knowledge-Based MT

Knowledge-Based Machine Translation (KBMT) is characterized by a heavy emphasis on functionally complete understanding of the source text prior to the translation into the target text. KBMT does not require total understanding, but assumes that an interpretation engine can achieve successful translation into several languages. KBMT is implemented on the Interlingua architecture; it differs from other interlingual techniques by the depth with which it analyzes the SL and its reliance on explicit knowledge of the world.

KBMT must be supported by world knowledge and by linguistic semantic knowledge about meanings of words and their combinations. Thus, a specific language is needed to represent the meaning of languages. Once the SL is analyzed, it will run through the augmenter. It is the knowledgebase that converts the source representation into an appropriate target representation before synthesizing into the target sentence.

KBMT systems provide high quality translations. Nevertheless, they are quite expensive to produce due to the large amount of knowledge needed to accurately represent sentences in different languages. The English-Vietnamese MT system is one of the examples of KBMTS.

3.6 Principle-Based MT

Principle-Based Machine Translation (PBMT) Systems employ parsing methods based on the Principles & Parameters Theory of Chomsky‘s Generative Grammar. The parser generates a detailed syntactic structure that contains lexical, phrasal, grammatical, and thematic information. It also focuses on robustness, language-neutral representations, and deep linguistic analyses.

In the PBMT, the grammar is thought of as a set of language-independent, interactive well-formed principles and a set of language-dependent parameters. Thus, for a system that uses n languages, one must have n parameter modules and a principles module. Thus, it is well-suited for use with the interlingual architecture.

PBMT parsing methods differ from the rule-based approaches. Although efficient in many circumstances, they have the drawback of language-dependence and increase exponentially in rules if one is using a multilingual translation system. They provide broad coverage of many linguistic phenomena, but lack the deep knowledge about the translation domain that KBMT and EBMT systems employ. Another drawback of current PBMT systems is the lack of the most efficient method for applying the different principles. UNITRAN is one of the examples of PBMT.

3.7 Online Interactive Systems

In this interactive translation system, the user is allowed to suggest the correct translation to the translator online. This approach is very useful in a situation where the context of a word is unclear and there exists many possible meanings for a particular word. In such cases, the structural ambiguity can be solved with the interpretation of the user.

4. Major MT Developments in India: A Literature Survey

The first public Russian to English (Manning *et al.,* 2003) MT system was presented at Georgetown University in 1954 with a vocabulary size of around 250 words. Since then, many research projects have been devoted to MT. Nevertheless, as the complexity of the linguistic phenomena involved in the translation process together with the computational limitations of the time were made apparent, enthusiasm faded out quickly. Also, the results of two negative reports, namely ‘Bar-Hillel’ and ‘AL-PAC,’ had a dramatic impact on MT research in that decade.

During the 1970s, the focus of MT activity switched from the United States to Canada and Europe, especially due to the growing demands for translations within their multicultural societies. 'Mateo,' a fully-automatic system translating weather forecasts, enjoyed great success in Canada. Meanwhile, the European Commission installed a French-English MT system called 'Systran'. Other research projects, such as 'Eurotra,' 'Ariane,' and 'Susy,' broadened the scope of MT objectives and techniques. The rule-based approaches emerged as the correct path to successful MT quality. Throughout the 1980s, many different types of MT systems appeared with the most prevalent being those using an intermediate semantic language, such as the 'Interlingua' approach.

Lately, various researchers have shown better translation quality with the use of phrase translation. Most competitive SMT systems, such as CMU, IBM, ISI, and Google, use phrase-based systems with good results.

In the early 1990s, the progress made by the application of statistical methods to speech recognition, introduced by IBM researchers, was in purely-SMT models (Manning *et al*., 2003). The drastic increment in computational power and the increasing availability of written translated texts allowed the development of statistical and other corpus-based MT approaches. Many academic tools turned into useful commercial translation products, and several translation engines were quickly offered in the World Wide Web.

Today, there is a growing demand for high-quality automatic translation. Almost all of the research community has moved towards corpus-based techniques, which have systematically outperformed traditional knowledge-based techniques in most performance comparisons. Every year, more research groups embark on SMT experimentation, and there is regained optimism in regards to future progress within the community.

MT is an emerging research area in NLP for Indian languages, which started more than a decade ago. There have been number of attempts in MT for English to Indian languages and Indian languages to Indian languages using different approaches. The literature shows that the earliest published work was undertaken by Chakraborty in 1966 (Noone *et al*., 2003). Many government and private sector researchers, as well as individuals, are actively involved in the development of MT systems and have generated some reasonable MT systems. Some of these MT systems are in the advanced prototype or technology transfer stage, and the rest have been newly initiated. The main developments in Indian language MT systems are as follows.

4.1 ANGLABHARTI by Indian Institute of Technology, Kanpur (1991)

ANGLABHARTI is a multilingual machine aided translation project on translation from English to Indian languages, primarily Hindi, which is based on a pattern directed approach (Durgesh *et al*., 2000; Sinha *et al*., 1995; Ajai *et al*., 2009; Manning *et al*., 2003; Sudip *et al*.,

2005). The strategy in this MT system is better than the transfer approach and lies below the Interlingua approach. In the first stage, a pattern directed parsing is performed on the SL English, which generates a `pseudo-target' that is applicable to a set of Indian languages. Word sense ambiguity in the SL sentence also is resolved by a number of semantic tags. In order to transform the pseudo TL into the corresponding TL, the system uses a separate text generator module. After correcting all ill-formed target sentences, a post-editing package is used make the final corrections. Even though it is a general purpose system, it has been applied mainly in the domain of public health at present. The ANGLABHARTI system is currently implemented from English to Hindi translation called AnglaHindi which is web-enabled (http://anglahindi.iitk.ac.in) and has obtained good domain-specific results for health campaigns, successfully translating many pamphlets and medical booklets. At present, further research work is going on to extend this approach for English to Telugu/Tamil translation. The project is primarily based at IIT-Kanpur, in collaboration with ER&DCI, Noida, and has been funded by TDIL. Professor RMK Sinha, Indian Institute of Technology, Kanpur is leading this MT project.

4.2 ANGLABHARTI -II by Indian Institute of Technology, Kanpur (2004)

The disadvantages of the previous system are solved by introducing the ANGLABHARTI - II MT architecture system (Sinha *et al*., 2003). The different approach, a Generalized Example-Base (GEB) for hybridization in addition to a Raw Example-Base (REB), is used to improve the performance of the translation. Compared to the previous approach, this system first attempts a match in REB and GEB before invoking the rule-base at the time of actual usage. Automated pre-editing and paraphrasing steps are further improvements in the proposed new translation approach. The system is designed in a way that various submodules are pipelined in order to achieve more accuracy and robustness.

At present, the ANGLABHARTI technology has been transferred under the ANGLABHARTI Mission into eight different sectors across the country (Sudip *et al*., 2005). The main intention of this bifurcation is to develop Machine Aided Translation (MAT) systems for English to twelve Indian regional languages. These include MT from English to Marathi & Konkani (IIT, Mumbai): English to Asamiya and Manipuri (IIT, Guwahati): English to Bangla (CDAC, Kolkata): English to Urdu, Sindhi & Kashmiri (CDAC-GIST group, Pune): English to Malyalam (CDAC, Thiruvananthpuram): English to Punjabi (Thapar Institute of Engineering and Technology-TIET, Patiala) English to Sanskrit (Jawaharlal Nehru University - JNU, New Delhi): and English to Oriya (Utkal University, Bhuvaneshwar).

4.3 ANUBHARATI by Indian Institute of Technology, Kanpur (1995)

ANUBHARATI is a recently started MT system aimed at translating from Hindi to English (Durgesh *et al*., 2000; Sinha *et al*., 1995; Ajai *et al*., 2009; Sudip *et al*., 2005). Similar to the ANGLABHARTI MT system, ANUBHARATI is also based on machine aided translation in which a variation of the example-based approach, called a template or hybrid HEBM, is used. The literature shows that a prototype version of the MT system has been developed and the project is being extended for developing a complete system. The HEBMT approach takes advantage of pattern and example-based approaches by combining the essentials of these methods. One more added advantage of the ANUBHARATI system is that it provides a generic model for translation that is suitable for translation between any two Indian languages pair with a minor addition of modules.

4.4 ANUBHARATI-II by Indian Institute of Technology, Kanpur (2004)

ANUBHARATI-II is a revised version of the ANUBHARATI that overcomes most of the drawbacks of the earlier architecture with a varying degree of hybridization of different paradigms (Sudip *et al*., 2005). The main intention of this system is to develop Hindi to any other Indian languages, with a generalized hierarchical example-based approach. Nevertheless, while both ANGLABHARTI-I and ANUBHARTI-II did not produce the expected results, both systems have been implemented successfully with good results. Professor RMK Sinha, Indian Institute of Technology, Kanpur is leading this MT project.

4.5 Anusaaraka by Indian Institute of Technology, Kanpur and University of Hyderabad

To utilize the close similarity among Indian languages for MT, another translation system called Anusaaraka (Durgesh *et al*., 2000; Sudip *et al*., 2005), was introduced, which is based on the principles of Paninian Grammar (PG). Anusaaraka is a machine aided translation system that also is used on language access between these languages. At present, this system is applied to children's stories, and an Alpha version of the system has been developed already for language assessors from five regional languages Punjabi, Bengali, Telugu, Kannada, and Marathi into Hindi. The Anusaaraka MT approach mainly consists of two modules (Manning *et al*., 2003; Bharati *et al*., 1997). The first module is called Core Anusaaraka, which is based on language knowledge, and the second one is a domain specific module that is based on statistical knowledge, world knowledge, *etc*. That is, the idea behind Anusaaraka is different from other systems in that the total load is divided in-to parts. The machine carries out the language-based analysis of the text, and the remaining work, such as knowledge-based analysis or interpretation, is performed by the reader. The Anusaaraka project was funded by TDIL,started at IIT Kanpur, and later shifted mainly to the Centre for Applied Linguistics and

Translation Studies (CALTS), Department of Humanities and Social Sciences, University of Hyderabad. At present, the Language Technology Research Centre (LTRC) at IIIT Hyderabad is developing an English to Hindi MT system using the architecture of the Anusaaraka approach. This Anusaaraka project is being developed under the supervision of Prof. Rajeev Sangal and Prof. G U Rao.

4.6 Anusaaraka System from English to Hindi

The Anusaaraka system from English to Hindi preserves the basic principles of information preservation and load distribution of original Anusaaraka (Manning *et al*., 2003; Bharati *et al*., 1997). To analyze the source text, it uses a modified version of the XTAG based super tagger and light dependency analyzer that was developed at the University of Pennsylvania. The advantage of this system is that, after the completion of the source text analysis, the user may read the output and can always move to a simpler output if the system produces the wrong output or fails to produce output.

4.7 MaTra (2004)

MaTra is an English to Indian languages (at present Hindi) Human-Assisted translation system based on a transfer approach using a frame-like structured representation that resolves the ambiguities using rule-based and heuristics approaches (Durgesh *et al*., 2000; Sudip *et al*., 2005; Manning *et al*., 2003). MaTra is an innovative system, which provides an intuitive GUI, where the user visually can inspect the analysis of the system and can provide disambiguation information to produce a single correct translation. Even though the MaTra system is intended to be a general purpose system, it has been applied mainly in the domains of news, annual reports, and technical phrases. MaTra is an ongoing project and the system currently is able to translate domain-specific simple sentences. Current development is towards covering other types of sentences. The Natural Language group of the Knowledge Based Computer Systems (KBCS) division at the National Centre for Software Technology (NCST), Mumbai (currently CDAC, Mumbai) has undertaken the task developing the MaTra system and is funded by TDIL.

4.8 MANTRA by Centre for Development of Advanced Computing, Bangalore (1999)

The Mantra MT system is intended to perform translation for the domains of gazette notifications pertaining to government appointments and parliamentary proceeding summaries between English and Indian languages as well as from Indian languages to English, where source and TL grammars are represented using Lexicalized Tree Adjoining Grammar (LTAG) formalism (Durgesh *et al*., 2000; Sudip *et al*., 2005). The added advantage of this system is

that the system can also preserve the formatting of input Word documents across the translation. After the successful development of MANTRA-Rajyasabha, language pairs like Hindi-English and Hindi-Bengali translation already have started using the Mantra approach. The Mantra project is being developed under the supervision of Dr. Hemant Darbari and is funded by TDIL and the Department of Official Languages, Ministry of Home Affairs, Government of India.

4.9 UCSG-based English-Kannada MT by University of Hyderabad

Using the Universal Clause Structure Grammar (UCSG) formalism, the Computer and Information Sciences Department at the University of Hyderabad, under the supervision of Prof. K. Narayana Murthy, developed a domain-specific English-Kannada MT system (Durgesh *et al*., 2000; Sudip *et al*., 2005; Manning *et al*., 2003). This UCSG-based system is based on a transfer-based approach and has been applied to the translation of government circulars. The system work is done at the sentence level and requires post-editing. At its first step of translation, the source (English) sentence is analysed and parsed using UCSG parser (developed by Dr. K. Narayana Murthy). Then, using translation rules, an English-Kannada bilingual dictionary, and network based Kannada Morphological Generator (developed by Dr. K. Narayana Murthy), the system translates in-to the Kannada language. This project has been funded by government of Karnataka and work is going to improve the performance of the system. Later, the same approach was applied for English-Telugu translation.

4.10 UNL-based MT between English, Hindi and Marathi by Indian Institute of Technology, Mumbai

Universal Networking Language (UNL) MT between English, Hindi, and Marathi is based on the Interlingua approach (Durgesh *et al*., 2000; Sudip *et al*., 2005; Manning *et al*., 2003). Under the supervision of Prof. Pushpak Bhattacharya, IIT Bombay is the Indian participant in UNL, which is an international project of the United Nations University, aimed at developing an Interlingua for all major human languages in the world. In the UNL based MT, the knowledge of the SL is captured or converted into UNL form and reconverted from UNL to the TL, like Hindi and Marathi. The SL information is represented sentence by sentence which is later converted into a hypergraph having concepts as nodes and relations as directed arcs (Shachi *et al*., 2002). The document knowledge is expressed in three dimensions as word knowledge, conceptual knowledge, and attritute labels.

4.11 Tamil-Hindi Anusaaraka MT

The KB Chandrasekhar Research Centre of Anna University at Chennai is active in the area of Tamil NLP. A Tamil-Hindi language assessor has been built using the Anusaaraka formalism (Durgesh *et al*., 2000; Sudip *et al*., 2005; Manning *et al*., 2003). The group has developed a Tamil-Hindi machine aided translation system under the supervision of Prof. CN Krishnan, with a performance of 75%.

4.12 English-Tamil machine Aided Translation system

Recently, the NLP group also developed a prototype of English-Tamil Human Aided MT System (Manning *et al*., 2003; Dwivedi *et al*., 2010). The system mainly consists of three major components: an English morphological analyzer, a mapping unit, and the Tamil language morphological generator.

4.13 SHIVA MT System for English to Hindi

This project was developed jointly by the Indian Institute of Science, Bangalore, and International Institute of Information Technology, Hyderabad, in collaboration with Carnegie Mellon University based on an example-based approach (Sudip *et al*., 2005; Dwivedi *et al*., 2010). An experimental system has been released for experiments, trials, and user feedback and is publicly available.

4.14 SHAKTI MT System for English to Hindi, Marathi and Telugu

This is a recently started project that also was developed jointly by Indian Institute of Science, Bangalore, and International Institute of Information Technology, Hyderabad, in collaboration with Carnegie Mellon University (Sudip *et al*., 2005; Dwivedi *et al*., 2010). The system follows a hybrid approach by combining both rule and statistical-based approaches. An experimental system for English to Hindi, Marathi, and Telugu is publicly available for experiments, trials, and user feedback.

4.15 Anuvadak English-Hindi MT

Anuvadak 5.0 English to Hindi software is a general-purpose tool developed by the private sector company Super Infosoft Pvt Ltd., Delhi, under the supervision of Mrs. Anjali Rowchoudhury (Durgesh *et al*., 2000; Sudip *et al*., 2005; Manning *et al*., 2003; Dwivedi *et al*., 2010). The system has inbuilt dictionaries in specific domains and supports post-editing. If the corresponding target word is not present in the lexicon, the system has a facility to translate that source word into the target. The system can run in Windows and a demonstration version of the system is publicly available.

4.16 English-Hindi Statistical MT

A statistical-based English to Indian languages, mainly Hindi, MT system was started by IBM India Research Lab at New Delhi, using the same approach as its existing work on other languages (Durgesh *et al*., 2000; Manning *et al*., 2003).

4.17 English-Hindi MAT for news sentences

A rule-based English to Hindi Machine Aided Translation system was developed by Jadavpur University, Kolkata, under the supervision of Prof. Sivaji Bandyopadhyay (Durgesh *et al*., 2000). The system uses the transfer based approach and is currently working on domain specific MT system for news sentences.

4.18 A hybrid MT system for English to Bengali

Under the supervision of Prof. Sivaji Bandyopadhyay, a hybrid-based MT system for English to Bengali was developed at Jadavpur University, Kolkata, in 2004 (Dwivedi *et al*., 2010). The current version of the system works at the sentence level.

4.19 Hinglish MT system

In 2004, Prof. Sinha and Prof. Thakur developed a standard Hindi-English MT system called Hinglish by incorporating an additional level in the existing ANGLABHARTI-II and ANUBHARTI-II systems (Dwivedi *et al*., 2010). The system produced satisfactory results in more than 90% of the cases, except the case with polysemous verbs.

4.20 English to (Hindi, Kannada, Tamil) and Kannada to Tamil language-pair EBMT system (2006)

An example-based English to Hindi, Kannada, and Tamil, as well as Kannada to Tamil (Dwivedi *et al*., 2010), MT system was developed by Balajapally *et al*. (2006). A set of bilingual dictionaries comprised of a sentence dictionary, phrase-dictionary, word-dictionary, and phonetic-dictionary of parallel corpora of sentences, phrases, words, and phonetic mappings of words is used for the MT. A corpus size of 75,000 most commonly used English-{Hindi, Kannada and Tamil} sentence pairs are used for MT.

4.21 Punjabi to Hindi MT system (2007)

A direct word-to-word translation approach, a Punjabi to Hindi MT system, was developed by Josan and Lehal at Punjabi University, Patiala, and reported 92.8% accuracy (Dwivedi *et al*., 2010). In addition to the Punjabi-Hindi lexicon and morphological analysis, the system also consists of modules that support word sense disambiguation, transliteration, and post-processing.

4.22 MT System among Indian language - Sampark (2009)

Consortiums of institutions (including IIIT Hyderabad, University of Hyderabad, CDAC (Noida, Pune), Anna University, KBC, Chennai, IIT Kharagpur, IIT Kanpur, IISc Bangalore, IIIT Alahabad, Tamil University, Jadavpur University) started to develop MT systems among Indian languages, called Sampark and have already released experimental systems for {Punjabi, Urdu, Tamil, Marathi} to Hindi and Tamil-Hindi in 2009 (Dwivedi *et al*., 2010).

4.23 English to Bengali (ANUBAAD) and English to Hindi MT System by Jadavpur University

Using a phrasal example-based approach, Jadavpur University developed a domain- specific translation of English news to Bengali called ANUBAAD, with current system work at the sentence level (Sudip *et al*., 2005). Also, the university started to develop a translation system for English news headlines to Bengali using a semantics-example-based approach. Using the same architecture, the university also developed a MT system for English-Hindi, and the system works currently at the simple sentence level. Recently the university also started to develop an Indian languages (Bengali, Manipuri) to English MT system. These translation systems are developing under the supervision of Prof. Sivaji Bandyopadhyay. The university uses these translation systems for guiding students and researchers who work in the MT area.

4.24 Oriya MT System (OMTrans) by Utkal University, Vanivihar

Utkal University, Bhuvaneshwar is working on an English-Oriya MT system OMTrans under the supervision of Prof. Sanghamitra Mohanty (Sudip *et al*., 2005; Manning *et al*., 2003). In addition to the parser and Oriya Morphological Analyser (OMA), the system also consists of an N-gram based word sense disambiguation module.

4.25 English-Hindi EBMT system by IIT Delhi

The Department of Mathematics, IIT Delhi, under the supervision of Professor Niladri Chatterjee developed an example-based English-Hindi MT system (Sudip *et al*., 2005). They have developed divergence algorithms for identifying the divergence for English to Hindi example-based system and a systematic scheme for retrieval from the English-Hindi example base.

4.26 Machine Aided Translation by Centre for Development of Advanced Computing (CDAC), Noida

Using the Machine Aided Translation system approach, a domain-specific translation system for translating public health related sentences from English to Hindi was developed (Manning *et al*., 2003). The system supports the advantage of post-editing and reportes 60%

performance.

4.27 Hindi to Punjabi MT system (2009)

Goyal and Lehal of Punjabi University, Patiala, developed a Hindi to Punjabi MT system based on a direct word-to-word translation approach (Goyal *et al*., 2009; Dwivedi *et al*., 2010). The system consists of the following modules: pre-processing, a word-to-word Hindi-Punjabi lexicon, morphological analysis, word sense disambiguation, transliteration, and post-processing. They also have developed an evaluation approach for a Hindi to English translation system and have reported 95% accuracy. Still, work is being carried out to achieve a better system.

4.28 A Statistical MT Approach to Sinhala-Tamil Language (2011)

Ruvan Weerasinghe developed an SMT Approach to Sinhala-Tamil Language Translation (Weerasinghe *et al*., 2011). This work reports on SMT based translation performed between language pairs, such as the Sinhala-Tamil and English-Sinhala pairs. The experiments results show that current models perform significantly better for the Sinhala-Tamil pair than the English-Sinhala pair and prove that the SMT system works better for languages that are not too distantly related to each other.

4.29 An Interactive Approach for English-Tamil MT System on the Web (2002)

Dr. Vasu Renganathan, University of Pennsylvania, developed an interactive approach for an English-Tamil MT System on the Web (Samir *et al*., 2010). The system is set on a rule-based approach, containing around five thousand words in the lexicon and a number of transfer rules used for mapping English structures to Tamil structures. This is an interactive system in that users can update this system by adding more words into the lexicon and rules into the rule-base.

4.30 Translation system using pictorial knowledge representation (2010)

Samir Kr. Borgohain and Shivashankar B. Nair introduced a new MT approach for Pictorially Grounded Language (PGL) based on their pictorial knowledge (Samir *et al*., 2010). In this approach, symbols of both the source and the TLs are grounded on a common set of images and animations. PGL is a graphic language and acts as a conventional intermediate language representation. While preserving the inherent meanings of the SL, the translation mechanism can also be scalable into a larger set of languages. The translation system is implemented in such a way that images and objects are tagged with both the source and target language equivalents, which makes the reverse translation much easier.

4.31 Rule-based Reordering and Morphological Processing For English-Malayalam SMT (2009)

This is an attempt to develop a statistical-based MT for English to Malayalam language by a set of MTech students under the guidance of Dr. K P Soman (Rahul *et al*., 2009). In this approach, they showed that a SMT based system can be improved by incorporating the rule-based reordering and morphological information of source and target languages.

4.32 SMT using Joshua (2011)

A piloted SMT based English to Telugu MT (MT) System called "enTel" was developed by Anitha Nalluri and Vijayanand Kommaluri, based on Johns Hopkins University Open Source Architecture (JOSHUA) (Anitha *et al*., 2011). A Telugu parallel corpus from the Enabling Minority Language Engineering (EMILLE) developed by CIIL Mysore and English to Telugu Dictionary, developed by Charles Philip Brown, is considered for training the translation system.

4.33 Multilingual Book Reader

The NLP team, including Prashanth Balajapally, Phanindra Bandaru, Ganapathiraju, N. Balakrishnan and Raj Reddy, introduced a multilingual book reader interface for DLI that supports transliteration and good enough translation (Prashanth) based on transliteration, word to word translation and full-text translation for Indian language. This is a simple, inexpensive tool that exploits the similarity between Indian languages. This tool can be useful for beginners who can understand their mother tongue or other Indian languages, but cannot read the script, and for an average reader who has the domain expertise. This tool can be also be used for translating either the documents or the queries in a multilingual search purpose.

4.34 A Hybrid Approach to EBMT for English to Indian Languages (2007)

Vamshi Ambati and U Rohini proposed a hybrid approach to EBMT (EBMT) for English to Indian languages that makes use of SMT methods and minimal linguistic resources (Ambati *et al*., 2007). Currently work is going on to develop English to Hindi as well as other Indian language translation systems based on manual and a statistical dictionary built from an SMT tool using an example database consisting of source and target parallel sentences.

4.35 SMT by Incorporating Syntactic and Morphological Processing

Ananthakrishnan Ramanathan, Pushpak Bhattacharyya, Jayprasad Hegde, Ritesh M. Shah, and M. Sasikumar proposed a new idea to improve the performance of the SMT based MT by incorporating syntactic and morphological processing (Ananthakrishnan). In this contest, they proved that performance of a baseline phrase-based system can be substantially improved by i)

reordering the source (English) sentence as per target (Hindi) syntax, and (ii) using the suffixes of target (Hindi) words.

4.36 Prototype MT System from Text-To-Indian Sign Language (ISL)

This is a very different approach to MT that is intended for dissemination of information to the deaf people in India and was proposed by Tirthankar Dasgupta, Sandipan Dandpat, and Anupam Basu (Dasgupta *et al*. 2008; Harshawardhan *et al*., 2011). At present, a prototype version of English to Indian Sign Language has been developed and the ISL syntax is represented based on Lexical Functional Grammar (LFG) formalism.

4.37 An Adaptable Frame based system for Dravidian language Processing (1999)

In the proposed work, a different approach that makes use of the karaka relations for sentence comprehension is used in the frame-based translation system for Dravidian languages (Idicula *et al*., 1999). Two pattern-directed application-oriented experiments are conducted, and the same meaning representation technique is used in both cases. In the first experiment, translation is done from a free word order language to fixed word order one, where both the source and destination are natural languages. In the second experiment, however, the TL is an artificial language with a rigid syntax. Even though there is a difference in the generation of the target sentence, the results obtained in both experiments are encouraging.

4.38 English-Telugu T2T MT and Telugu-Tamil MT System (2004)

CALTS in collaboration with IIIT, Hyderabad; Telugu University, Hyderabad; and Osmania University, Hyderabad developed an English-Telugu and Telugu-Tamil MT system under the supervision of Prof. Rajeev Sangal (CALTS). The English-Telugu system uses an English-Telugu machine aided translation lexicon of size 42000 words and a wordform synthesizer for Telugu. The Telugu-Tamil MT system was developed based on the available resources at CALTS: Telugu Morphological analyzer, Tamil generator, verb sense disambiguator, and Telugu-Tamil machine aided translation dictionary. The performance of the systems is encouraging, and it handles source sentences of varying complexity.

4.39 Developing English-Urdu MT Via Hindi (2009)

R. Mahesh K. Sinha proposed a different strategy for deriving English to Urdu translation using an English to Hindi MT system (R. Mahesh *et al*., 2009). In the proposed method, an English-Hindi lexical database is used to collect all possible Hindi words and phrases. These words and phrases are further augmented by including their morphological variations and attaching all possible postpositions. Urdu is structurally very close to Hindi and this

augmented list is used to provide mapping from Hindi to Urdu. The advantage of this translation system is that the grammatical analysis of English provides all the necessary information needed for Hindi to Urdu mapping and no part of speech tagging, chunking, or parsing of Hindi has been used for translation.

4.40 Bengali-Assamese automatic MT system-VAASAANUBAADA (2002)

Kommaluri Vijayanand, S. Choudhury and Pranab Ratna proposed an automatic bilingual MT for Bengali to Assamese using an example-based approach (Kommaluri *et al.*, 2002). They used a manually created aligned bilingual corpus by feeding real examples using pseudo code. The quality of the translation was improved by preprocessing the longer input sentences and also via the backtracking techniques. Since the grammatical structure of Bengali and Assamese is very similar, lexical word groups are required.

4.41 Phrase based English-Tamil Translation System by Concept Labeling using Translation Memory (2011)

The Computational Engineering and Networking research centre of Amrita School of Engineering, Coimbatore, proposed an English-Tamil translation system. The system is set on a phrase-based approach by incorporating concept labeling using translation memory of parallel corpora (Harshawardhan *et al.*, 2011). The translation system consists of 50,000 English-Tamil parallel sentences, 5000 proverbs, and 1000 idioms and phrases, with a dictionary containing more than 2,00,000 technical words and 100,000 general words. The system has an accuracy of 70%.

4.42 Rule-based Sentence Simplification for English to Tamil MT System (2011)

This work is aimed at improving the translation quality of an MT system by simplifying the complex input sentences for an English to Tamil MT system (Poornima *et al.*, 2011). In order to simplify the complex sentences based on connectives, like relative pronouns or coordinating and subordinating conjunctions, a rule-based technique is proposed. In this approach, a complex sentence is expressed as a list of sub-sentences while the meaning remains unaltered. The simplification task can be used as a preprocessing tool for MT where the initial splitting is based on delimiters and the simplification is based on connectives.

4.43 Manipuri-English Bidirectional SMT Systems (2010)

Using morphology and dependency relations, a Manipuri to English bidirectional SMT system was developed by Thoudam Doren Singh and Sivaji Bandyopadhyay (Doren Singh *et al.*, 2010). The system uses a domain-specific parallel corpus of 10350 sentences from news for

training purposes and the system is tested with 500 sentences.

4.44 English to Kannada SMT System (2010)

P.J. Antony, P. Unnikrishnan and Dr. K.P Soman proposed an SMT system for English to Kannada by incorporating syntactic and morphological information (Unnikrishnan *et al*., 2010). In order to increase the performance of the translation system, we have introduced a new approach in creating the parallel corpus. The main ideas that we have implemented and proven effective in the English to Kannada SMT system are: (i) reordering the English source sentence according to Dravidian syntax, (ii) using the root suffix separation on both English and Dravidian words, and iii) use of morphological information that substantially reduces the corpus size required for training the system. The results show that significant improvements are possible by incorporating syntactic and morphological information into the corpus. From the experiments we have found that the proposed translation system successfully works for almost all simple sentences in their twelve tense forms and their negatives forms.

4.45 Anuvadaksh

This system is an effort of the English to Indian Language MT (EILMT) consortium. Anuvadaksh is a system that allows translating the text from English to six other Indian languages, *i.e*. Hindi, Urdu, Oriya, Bangla, Marathi, and Tamil. Anuvadaksh being a consortium based project has a hybrid approach that is designed to work with platform and technology independent modules.

This system has been developed to facilitate the multi-lingual community, initially in the domain-specific expressions of tourism, and it would subsequently foray into various other domains in a phase-wise manner. It integrates four MT Technologies:

Tree-Adjoining-Grammar (TAG) based MT.
SMT.
Analyze and Generate rules (Anlagen) based MT.
Example-based MT (EBMT).

4.46 Google Translate

Google Translate is a free translation service that provides instant translations between 57 different languages. Google Translate generates a translation by looking for patterns in hundreds of millions of documents to help decide on the best translation. By detecting patterns in documents that have already been translated by human translators, Google Translate makes guesses as to what an appropriate translation should be. This process of seeking patterns in large amounts of text is called "SMT".

4.47 English to Assamese MT System

An English to Assamese MT system is in progress (Sudhir *et al*., 2007). The following activities are in progress in this direction.

• The graphical user interface of the MT system has been re-designed. It now allows the display of Assamese text. Modifications have been made in the Java modules.

• The existing Susha encoding scheme has been used. In addition, a new Assamese font set has been created according to that of Susha font set. The system is now able to display properly consonants, vowels, and matras of Assamese characters properly.

• The mapping of the Assamese keyboard with that of Roman has been worked out.

• The process of entering Assamese words (equivalent of English words) in the lexical database (nouns and verbs) is in progress.

The system developed basically a rule-based approach and relies on a bilingual English to Assamese dictionary. The dictionary-supported generation of Assamese text from English text is a major stage in this MT. Each entry in the dictionary is supplied with inflectional information about the English lexeme and all of its Assamese equivalents. The dictionary is annotated for morphological, syntactic, and partially semantic information. It currently can handle translation of simple sentences from English to Assamese. The dictionary contains around 5000 root words. The system simply translates source language texts to the corresponding target language texts phrase to phrase by means of the bilingual dictionary lookup.

4.48 Tamil University MT System

Tamil University, Tanjore, initiated a machine oriented translation from Russian-Tamil during 1983-1984 under the leadership of Vice-Chancellor Dr. V.I Subramaniam (Sudhir *et al*., 2007). It was taken up as an experimental project to study and compare Tamil with Russian in order to translate Russian scientific text into Tamil. A team consisting of a linguist, a Russian language scholar, and a computer scientist were entrusted to work on this project. During the preliminary survey, both Russian SL and Tamil were compared thoroughly for their style, syntax, morphological level, *etc*.

4.49 Tamil-Malayalam MT System

Bharathidasan University, Tamilnadu, is working on translation between languages belonging to the same family, such as Tamil-Malayalam translation (Sudhir *et al*., 2007). The MT consists of the following modules that are in progress.

Lexical database- This will be a bilingual dictionary of root words. All the noun roots and verb roots are collected.

Suffix database- Inflectional suffixes, derivative suffixes, plural markers, tense markers, sariyai, case suffixes, relative participle markers, verbal participle markers, *etc* will be compiled.

Morphological Analyzer- This is designed to analyze the constituents of the words. It will help to segment the words into stems and inflectional markers.

Syntactic Analyzer- The syntactic analyzer will find the syntactic category, like Verbal Phrase, Noun Phrase, and Participle Phrase. This will analyze the sentences in the source text.

Table 1 below provides a summary of all 49 MT systems.

Table 1. Comparison of MT systems in India

Sr. No	MT System (Year)	Source-Target Language	Developer	Approach	Domain
1	ANGLABHARTI (1991)	English to Indian languages (primarily Hindi)	IIT, Kanpur	Pseudo-interlingua	General
2	ANGLABHARTI - II (2004)	English to Indian languages	IIT, Kanpur	Pseudo-interlingua	General
3	ANUBHARATI (1995)	Hindi to English	IIT, Kanpur	GEBMT	General
4	ANUBHARATI-II (2004)	Hindi to any other Indian languages	IIT, Kanpur	GEBMT	General
5	Anusaaraka (1995)	Punjabi, Bengali, Telugu, Kannada, and Marathi to Hindi.	IIT, Kanpur and University of Hyderabad	PG	General
6	Anusaaraka (1995)	from English to Hindi	IIT, Kanpur and University of Hyderabad	PG	General
7	MaTra (2004)	English to Indian languages (at present Hindi)	CDAC, Mumbai	Transfer based	General
8	MANTRA (1999)	English to Indian languages and Reverse	CDAC, Pune	TAG	Administration, office orders
9	UCSG-based MT	English-Kannada	University of Hyderabad	transfer based	government circulars
10	UNL-based (2003)	Between English, Hindi, and Marathi	IIT, Mumbai	Interlingua	General
11	Tamil-Hindi Anusaaraka MT	Tamil-Hindi	KBC Research Centre, Anna University,	PG	General
12	English-Tamil HAMT	English-Tamil	NLP group	HAMT	General

Sr. No	MT System (Year)	Source-Target Language	Developer	Approach	Domain
13	SHIVA (2004)	English to Hindi	IISc- Bangalore, IIIT Hyderabad, and Carnegie Mellon University	EBMT	General
14	SHAKTI (2004)	English to Hindi, Marathi and Telugu	IISc- Bangalore, IIIT Hyderabad, and Carnegie Mellon University	RBM	General
15	Anuvaadak	English-Hindi	Super Infosoft Pvt Ltd., Delhi	Not-Available	Not-Available
16	English-Hindi Statistical MT	English to Indian languages	IBM India Research Lab, New Delhi	EBMT & SMT	Not-Available
17	English-Hindi MAT	English to Hindi	Jadavpur University, Kolkata	transfer based	newssentences
18	Hybrid MT system	English to Bengali	Jadavpur University Kolkata	Hybrid	Sentence level
19	Hinglish MT system (2004)	Hindi - English	IIT-Kanpur	Pseudo interlingua	General
20	English to Indian and Kannada to Tamil language-pair EBMT system (2006)	i) English to Hindi, Kannada, and Tamil ii) Kannada to Tamil	Balajapally	Example-based	Most Commonly used sentences
21	Punjabi to Hindi MT system (2007)	Punjabi to Hindi	Punjabi University, Patiala	Direct word to word	General
22	Sampark (2007)9	Among Indian languages	Consortiums of institutions	CPG	Not-Available
23	ANUBAAD (2004)	English to Bengali and English to Hindi	Jadavpur University	RBMT & SMT	News Sentences
24	OMTrans	English-Oriya	Utkal University, Bhuvaneshwar	Not-Available	SchoolbookSentences
25	English-Hindi EBMT system	English-Hindi	IIT Delhi	Example-based, Divergence algorithms	Not-Available
26	Machine Aided Translation	English to Hindi	CDAC, Noida	Machine Aided Translation	Public health related sentences
27	Hindi to Punjabi MT system (2009)	Hindi to Punjabi	Punjabi University, Patiala	direct word-to-word	General

Sr. No	MT System (Year)	Source-Target Language	Developer	Approach	Domain
28	Sinhala-Tamil MT (2011)	Sinhala to Tamil	RuvanWeerasinghe	SMT based	General
29	English-Tamil MT on Web (2002)	English to Tamil	University of Pennsylvania	rule-based	General
30	Pictorial knowledge Based MT (2010)	English to Assamese	Samir Kr. Borgohain and Shivashankar B. Nair	pictorial knowledge	People not well versed in each other‟s languages
31	English-Malayalam Statistical MT (2009)	English to Malayalam	AMRITA University, Coimbatore	SMT based	General
32	enTel (2011)	English to Telugu	AnithaNalluri and VijayanandKommaluri	SMT based	Not-Available
33	Multilingual book reader interface for DLI	Translation for Indian languages	PrashanthBalajapally and Team	Word-to-Word Translation	documents or the queries
34	English to Indian Languages MT (2007)	English to Indian Languages	VamshiAmbati and Rohini U proposed ,	Example-based	Not-Available
35	Incorporating Syntactic and Morphological based MT	English-Hindi	AnanthakrishnanRamanathan and Team	Stasticalphrase-based	Not-Available
36	Text-To-Indian Sign Language (ISL) MT	English to Indian Language	TirthankarDasgupta, SandipanDandpat, and AnupamBasu	Lexical Functional Grammar (LFG) formalism	Deaf people in India
37	Dravidian language Processing System (199)	Dravidian language	SumamUMAM MARY IDICULA	Adaptable Frame based	Not-Available
38	English-Telugu T2T MT and Telugu-Tamil MT (2004)	English-Teluguand Telugu-Tamil	CALTS; IIIT-Hyderabad; Telugu University- Hyderabad, Osmania University-Hyderabad,	Not-Available	Not-Available
39	English-Urdu MT via Hindi (2009)	English-Urdu	R. Mahesh K. Sinha	Not-Available	Not-Available
40	VAASAANUBAADA (2002)	Bengali- Assamese	KommaluriVijayanand S Choudhury and PranabRatna	EBMT	News

Sr. No	MT System (Year)	Source-Target Language	Developer	Approach	Domain
41	Phrase based English - Tamil MT (2011)	English - Tamil	CEN, AMRITA University, Coimbatore	Phrase based	General
42	Sentence Simplification System for English to Tamil (2011)	English - Tamil	Not-Available	Rule-based	Not-Available
43	Manipuri-English Bidirectional MT (2010)	Manipuri-English and -English-Manipuri	ThoudamDoren Singh and SivajiBandyopadhyay	Statistical	news
44	English to Dravidian Language MT (2010)	English to Malayalam	CEN, AMRITA University, Coimbatore	SMT Based	Simple sentences
45	Anuvadaksh	English to six other Indian languages *i.e.* Hindi, Urdu, Oriya, Bangla, Marathi, Tamil	EILMT consortium	hybrid approach	Tourism
46	Google Translate	Translations between 57 different languages	Google	SMT	General
47	English to Assamese MT	English to Assamese	Not-Available	Rule-based	Not-Available
48	Russian-Tamil MT (1983-1984)	Russian-Tamil	Tamil University, Tanjore	Not-Available	scientific text
49	Tamil - Malayalam MT	Tamil - Malayalam	Bharathidasan University, Tamil Nadu	Not-Available	Not-Available

5. Conclusion

This survey described machine translation (MT) techniques in a longitudinal and latitudinal way with an emphasis on the MT development for Indian languages. Additionally, we tried to describe briefly the different existing approaches that have been used to develop MT systems. From the survey, we found that almost all existing Indian language MT projects are based on a statistical and hybrid approach. We also identified the following two reasons that most of the developed MT systems for Indian languages have followed the statistical and hybrid approach. The first reason is, since Indian languages are morphologically rich in features and agglutinative in nature, rule-based approaches have failed in many situations for developing full-fledged MT systems. Second the general benefits of statistical and hybrid approaches have encouraged researchers to choose these approaches to develop MT systems for Indian languages.

Reference

ALPAC. (1966). *Language and Machines: Computers in Translation and Linguistics*. A report by the Automatic Language Processing Advisory Committee (Tech. Rep. No. Publication 1416), 2101 Constitution Avenue, Washington D.C., 20418 USA: National Academy of Sciences, National Research Council.

Ambati, V., & Rohini, U. (2007). A Hybrid Approach to EBMT for Indian Languages. *ICON 2007*.

Badodekar, S. (2003). *Translation Resources, Services and Tools for Indian Languages*. Computer Science and Engineering Department, Indian Institute of Technology, Mumbai, 400019, India.

Balajapally, P., Bandaru, P., Ganapathiraju, M., Balakrishnan, N., & Reddy, R. (2006). Multilingual Book Reader: Transliteration, Word-to-Word Translation and Full-text Translation. In *VAVA 2006*.

Bharati, A., Chaitanya, V., Kulkarni, A. P., & Sangal, R. (1997). ANUSAARAKA: Machine Translation in Stages. *A Quarterly in Artificial Intelligence*, 10(3), 22-25.

Borgohain, S. K., & Nair, S. B. (2010). Towards a Pictorially Grounded Language for Machine-Aided Translation. *International Journal on Asian Language Processing*, 20 (3), 87-109.

CALTS in collaboration with, IIIT Hyderabad. English-Telugu T2T Machine Translation and Telugu-Tamil Machine translation System. Indo-German Workshop on Language technologies, AU-KBC Research Centre, Chennai, 2004 . www.au-kbc.org/dfki/igws/Machine_Translation.ppt.

Dasgupta, T., & Basu, A. (2008). An English to Indian Sign Language Machine Translation System, www.cse.iitd.ac.in/embedded/assistech/Proceedings/P17.pdf.

Dasgupta, T., Dandpat, S., & Basu, A. (2008). Prototype Machine Translation System From Text-To-Indian Sign Language. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, 19-26.

Dave, S., Parikh, J., & Bhattacharya, P. (2001). Interlingua-based English-Hindi Machine Translation and Language Divergence. *Journal of Machine Translation*, 16(4), 251-304.

Dirix, P., Schuurman, I., & Vandeghinste V. (2005). Metis II: Example-based machine translation using monolingual corpora - system description. In P*roceedings of the 2nd Workshop on Example-Based Machine Translation*, 43-50.

Dugast, L., Senellart, J., & Koehn, P. (2007). Statistical post-editing on SYSTRAN's rule-based translation system. In *Proceedings of the Second Workshop on SMT*, 220-223.

Dwivedi, S. K., & Sukhadeve, P. P. (2010). Machine Translation System in Indian Perspectives. *Journal of Computer Science*, 6(10), 1111-1116.

Goyal, V., & Lehal, G. S. (2009). Evaluation of Hindi to Punjabi Machine Translation System. *IJCSI International Journal of Computer Science*, 4(1), 36-39.

Groves, D. & Way, A. (2005). Hybrid example-based SMT: the best of both worlds. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, 183-190.

Harshawardhan, R., Augustine, M. S., & Soman, K. P. (2011). Phrase based English - Tamil Translation System by Concept Labeling using Translation Memory. *International Journal of Computer Applications* (0975 - 8887), 20(3), 1-6.

Hutchins, J. (1993). The first MT patents. *MT News International*, 14-15.

Hutchins, J. (2005). The history of machine translation in a nutshell. http://www.hutchinsweb.me.uk/Nutshell-2005.pdf.

Hutchins, W. J., & Lovtskii, E. (2000). Petr Petrovich Troyanskii (1854-1950): A forgotten pioneer of mechanical translation. *Machine translation*, 15(3), 187-221.

IBM. (1954). *701 Translator.* IBM Archives online: Press release January 8th 1954, http://www-03.ibm.com/ibm/history/exhibits/701/701-translator.html.

Idicula, S. M. (1999). *Design and Development of an Adaptable Frame-based System for Dravidian Language.* Ph.D thesis, Department of Computer Science, COCHIN University of Science and Technology.

Jain, A. (2009). Machine Aided Translation Systems: *The Indian Scenario*. 2(6), 2009. www.iitk.ac.in/infocell/Archive/dirnov2/ techno_machine.html.

Koehn, P. & Hoang, H. (2007). Factored translation models. In Proceedings of the 2007 Joint Conference on Empirical Methods. In *NLP and Computational Natural Language Learning*, 868-876.

Mahesh, R., & Sinha, K. (2009). Developing English-Urdu Machine Translation Via Hindi. In *Third Workshop on Computational Approaches to Arabic Scriptbased Languages (CAASL3), MT Summit XII*, Ottawa, Canada.

Manning, C., & Schutze, H. (2003). Foundations of Statistical NLP. *Proceedings of HLT/NAACL*.

Mishra, S. K. (2007). *Sanskrit Karaka Analyzer for Machine Translation*. PhD. Thesis, Jawaharlal Nehru University.

Nalluri, A., & Kommaluri, V. (2011). SMT using Joshua: An approach to build 'enTel' system. *Language in India, Special Volume:Problems of Parsing in Indian Languages*, 11(5), 1-6. www.languageinindia.com.

Naskar, S., & Bandyopadhyay, S. (2005). Use of Machine Translation in India: Current Status. In *Proceedings of MT SUMMIT X*; September 13-15, 2005, Phuket, Thailand.

Noone, G. (2003). *Machine Translation - A Transfer Approach*, A project report, www.scss.tcd.ie/undergraduate/bacsll/bacsll_web/nooneg0203.pdf.

Oepen, S., Velldal, E., Lønning, J. T., Meurer, P., Rosen, V., & Flickinger, D. (2007). Towards hybrid quality-oriented machine translation on linguistics and probabilities in MT. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation*, 144-153.

Poornima, C., Dhanalakshmi, V., Kumar M. A., & Soman, K. P. (2011). Rule-based Sentence Simplification for English to Tamil Machine Translation System. *International Journal of Computer Applications* (0975 - 8887), 25(8), 38-42.

Rahul, C., Dinunath, K., Ravindran, R., & Soman, K. P. (2009). Rule-based Reordering and Morphological Processing For English-Malayalam SMT. *International Conference on Advances in Computing, Control, and Telecommunication Technologies*, 458-460.

Ramanathan, A., Bhattacharyya, P., Hegde, J., Shah, R. M., & Sasikumar, M. (2008). Simple Syntactic and Morphological Processing Can Help English-Hindi SMT. In *IJCNLP 2008*.

Rao, M. D. (2000). *Machine Translation in India: A Brief Survey*. www.elda.org/en/proj/scalla/SCALLA2001/SCALLA2001Rao.pdf.

Renganathan, V. (2002). An Interactive Approach to Development of English-Tamil Machine Translation System on the Web. *Tamil Internet 2002*, California, USA. 68-73. www.infitt.org/ti2002/hubs/ conference/papers.html.

Simard, M., Ueffing, N., Isabelle, P., & Kuhn, R. (2007). Rule-based translation with statistical phrase-based post-editing. In *Proceedings of the Second Workshop on SMT*, 203-206.

Singh, T. D., & Bandyopadhyay, S. (2010). Manipuri-English Bidirectional SMT Systems using Morphology and Dependency Relations. In *Proceedings of SSST-4, Fourth Workshop on Syntax and Structure in Statistical Translation*, 83-91, COLING 2010, Beijing.

Sinha, R. M. K. & Jain, A. (2003). AnglaHindi: An English to Hindi Machine-Aided Translation System. In *MT Summit IX*, New Orleans, Louisiana, USA, September, 2003.

Sinha, R. M. K., Sivaraman, K., Agrawal, A., Jain, R., Srivastava, R. & Jain, A. (1995). ANGLABHARTI: a multilingual machine aided translation project on translation from English to Indian languages. *IEEE International Conference on: Systems, Man and Cybernetics, 1995. Intelligent Systems for the 21st Century*, 1609-1614.

Unnikrishnan, P., Antony, P. J., & Soman, K. P. (2010). A Novel Approach for English to South Dravidian Language SMT System. *International Journal on Computer Science and Engineering (IJCSE)*, 02(08), 2749-2759.

Vijayanand, K., Choudhury, S., & Ratna, P. (2002). Vaasaanubaada Automatic Machine Translation Of Bilingual Bengali - Assamese News Texts. *Language Engineering Conference*, University of Hyderabad, India.

Weaver, W. (1999). Warren Weaver Memorandum, July 1949. *MT News International*, no. 22, July 1999, 5-6, 15.

Weerasinghe, R. (2011). A SMT Approach to Sinhala-Tamil Language Translation. citeseerx.ist.psu.edu /viewdoc/summary?doi= 10.1.1.78.7481, 2011.

Zhang, Y. (2006). Chinese-English SMT by Parsing. www.cl.cam.ac.uk/~yz360/ mscthesis.pdf.