Machine Translation (MT) Techniques for Indian Languages

S.Anbukkarasi, Dr. S.Varadhaganapathy

Abstract: ***Machine Translation (MT) is the process of converting the text from one language (source) to another language (TL). MT draws the idea of linguistics, computer science, artificial intelligence, sociology, psychology etc. The linguistically rich country like India has the demand to develop a full-fledged MT system to convert the text across different languages. Though the research has been made on MT for the past 60 years, still it is considered to be a challenging task. Building a fully automatic MT system is extremely difficult. This paper deals with the various ideas in MT systems for Indian Languages. Advantages and limitations of some of the important Dravidian Language translation systems developed using MT techniques are discussed.***

Index Terms:Machine Translation, Rule based, Transfer based.

I. INTRODUCTION

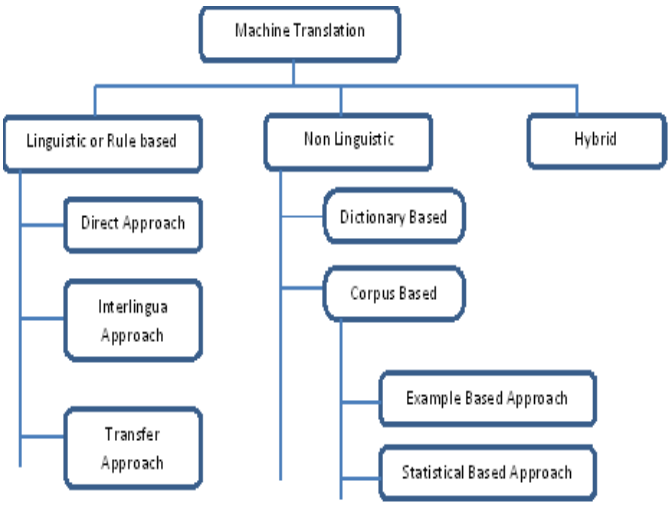
MT is the sub branch of Natural Language Processing which is the oldest, yet important research area. Machine Translation is defined as the application of computer to translate one language to another language. The research in MT has been started during the year 1940s and the various techniques, methods have been tried over the past decades. India is linguistically rich and diverged country. So there will always be a requirement for translation of one language to another language. India has 22 languages which areconstitutionally approved and written in 10 different kinds of scripts. In India, 73% of people speak Indo-Aryan languages and second major language called Dravidian language is spoken by 23 % of people. Hindi is known as the official language of India. English is used in many domains such as media, technology, commerce, education. Only 5% of the people could speak and write English. As many of the states have their own regional language, their documents and works will be in their own language whereas the work of the State government and its documents will be either in Hindi or English. Hence there is a requirement for translating documents from one language to another language. The manual translation is a time consuming task. Hence the demand for Machine translation grows rapidly.

II. BRIEF HISTORY

In India, late 80's, research in the field of Computational Linguistics and Artificial Intelligence provided a good development in translation technology. It helps in developing machine translation system in few of the domains. In 1991, the machine translation system called Anglabharti which is designed for translating English to Indian languages. It uses a pseudo-interlingua approach. In 1995, Anusaaraka Mt was developed at IIT Kanpur. Later, the Language Technology. Research Center (LTRC) at IIT Hyderabad was developing English-Hindi translation system. It is not only focuses on Machine Translation but also on Language access between Indian Languages. It develops language accesors for Kannada, Punjabi, Marathi, Bengali, Telugu to Hindi. In 1995, Anubharathi MT system was developed using EMBT paradigm for Hindi to English translation. The MANTRA (Machine Assisted Translation tool) MT translates text from the English language text to Hindi language text in a domain of administration which translates documents such as government appointments, circulars, office orders. It is developed based on the TAG formalism from PennsylvaniaUniversity. A Machine Assisted Translation system MAT for translating English texts into kannada text has been developed by Dr. K. Narayana Murthy in University of Hyderabad in 2002. This approach is based on the Universal Class Structure Grammar (UCSG) structure. Research on Machine Translation process takes its full speed after 2000. In 2009, a hybrid machine translation system was designed by IIT Kharagpur for translating from Bengali language to Hindi language. It uses multi-engine Machine translation approach.

III. APPROACHES FOR MACHINE TRANSLATION

There are many approaches in Machine Translation. But it is mainly classified into the three approaches as depicted in the Fig. 1.



Revised Manuscript Received on July 05, 2019.
S.Anbukkarasi, Assistant Professor in Nandha Engineering College, Erode
S.Varadhaganapathy, Professor, Department of Information Technology in Kongu Engineering College, Erode.

Figure 1: Machine Translation Techniques

Figure 1 specifies the basic classification in Machine Translation approach.

IV. LINGUISTICS OR RULE BASED APPROACHES

Rules based approach uses grammar and computer programs to translate a word or lexicon from a source language to a word or lexicon to a target language. Grammar and computer programs help to extract features and information for the words to be translated. This method requires the linguistic knowledge of the language. The rules can be formed as given below.

A. Direct MT System

Direct MT system uses two way dictionaries. It translates each word in a sentence from source language to target language. It makes use of grammar rules. It performs translation on one language pair; hence we call it as monolithic approach. Direct MT system needs less parsing, relies on large two-way dictionary and it requires little knowledge on source language. The first step in Direct MT system is morphological analysis. It extracts the root words from the words in a source language. The second step in direct MT system is to look up in bilingual dictionary. A bilingual dictionary has the match words for the target language words. The final step is to syntactic rearrangement of the words. It means the word order from the source language is rearranged to match the sentence in the target language.

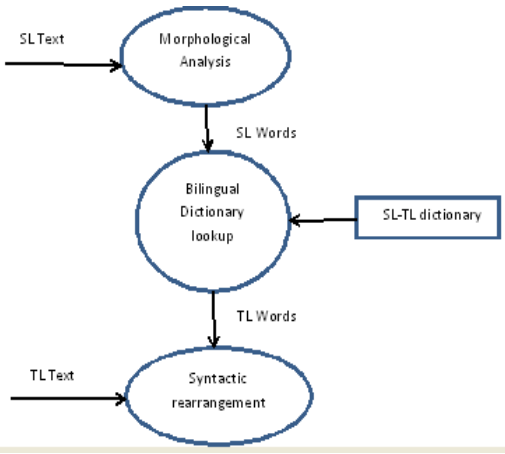


Figure 2: Direct MT system

B. Interlingua Machine Translation

Interlingua Machine Translation is used to convert the sentences into more than one language. This system translates the language into intermediate language called the universal language that helps the system to translate into multiple languages.

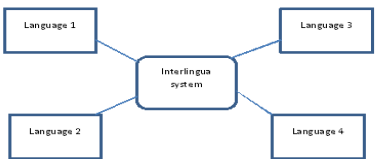


Figure 3: Interlingua MT System

C. Transfer based MT

In this system, a database with translation rules helps to translate text from source language to target language. Based on the rules in the dictionary, sentences from source language to target languages are translated. This translation is done in three phases.

1. Analysis

In analysis phase, morphological and syntactic information of source language is analysed and produce the base form of the source language. It produces the source language structure. A hierarchical syntax tree for the source language is generated by grammar rules. The sample rules for generating syntax tree for Tamil language is given below:

S -> NP VP | VP
NP -> N | NP VP
VP -> V NP | PP (Adv)

2. Transfer

In this phase, base form of the source language text is translated into target language. This modules needs rearrangement rules so that source language sentence can be transferred into target language sentence. A set of rules for English to Tamil MT system is given as below.

English Structure	**Tamil Structure**
VP -> V NP	VP -> NP V
PP-> P NP	PP -> P NP
VG-> ADV V	VG -> V ADV

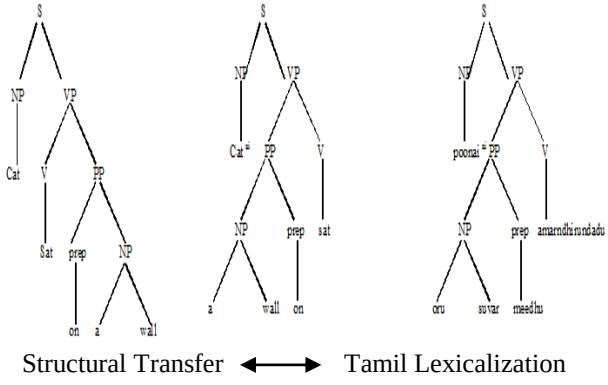


Figure 4: Structural Transfer English-Tamil MTS

In generation phase, the exact translation of source text to target language is taken place.

Figure 5: Transfer based Translation system

V NON-LINGUISTIC APPROACHES

Non-Linguistic approach doesn't require any linguistic knowledge to translate from source language to target language. It relies on dictionary for dictionary based approach and monolingual or bilingual corpus for corpus based approach.

A. Dictionary Based Approach

In Dictionary based approach, the dictionary is used to translate the text from the source language to target language. This approach is used to translate words rather than translating the sentences. It requires some pre-processing steps to morphological analysis and lemmatize the words of a source text to be translated.

B. Corpus Based Approach

This approach doesn't require explicit linguistic knowledge to translate from source language to target language like dictionary based approach. The system is trained using bilingual corpus and the monolingual corpus of the target language to translate a sentence.

1. Example Based Approach

In this approach huge bilingual corpus of the language pair of source and target language is used. It works based on the human being approach of problem solving. The main problem is divided into sub problems. Each sub problem is solved based on the experience gathered in the previous of solving the same kind of problem and finally integrating the solutions of the sub problems to solve the main problem. The example based approach is depicted in Figure 6.

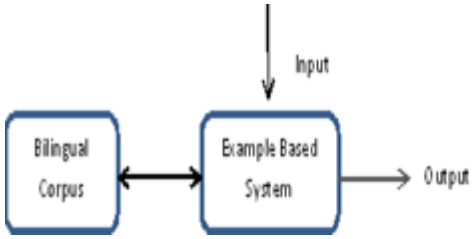


Figure 6: Example Based Translation System

2. Statistical Based Approach

In this approach, bilingual corpora and statistical methods are used to translate the sentence from source language to target language. The parameters from bilingual corpora and monolingual corpus is analysed and determined for creating translation and language models.

It requires more than 2 million words to design the translation system for a particular domain. This systemrequires elaborated hardware configuration to create translation models. The statistical based system is depicted in Figure.7.

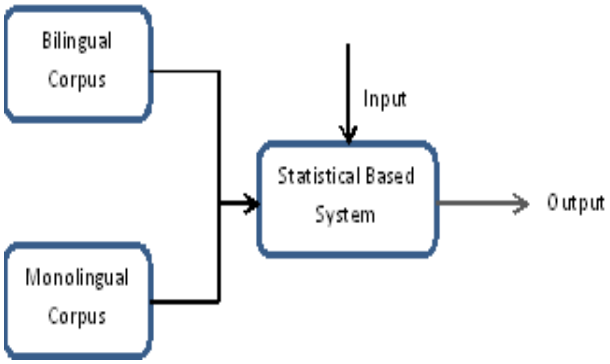


Figure 7: Statistical Based Approach

Statistical translation involves 3 steps.
1. Estimating the language probability p(t)
2. Estimating the translational model probability p(s/t)
3. Devise a product of them to achieve an efficient search of target text

Sentence T is found, for which p(s,t) is maximum.

$$P(s,t) = \arg\max_{t} p(s,t) = \arg\max_{t} p(t)p(s/t)$$

In the model, "s" is considered to be the source language sentence, "t" is considered to be the target language sentence.

VI. HYBRID MACHINE TRANSLATION APPROACH

Hybrid machine translation is the combination of Rule based as well as Statistical based translation systems. One of the examples for hybrid machine translation approach is Rule based system with post-processing by statistical approach. The overall system for hybrid approach is depicted in the following Figure 7.

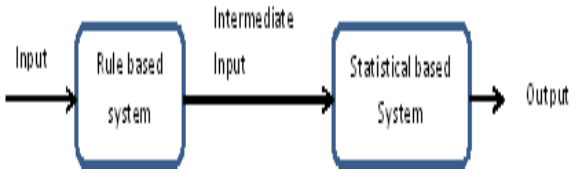


Figure 8: Hybrid Approach

VII. MACHINE TRANSLATION APPROACHES – A COMPARISON

As discussed above there are various approaches in Machine Translation and each one has its own advantages as well as disadvantages. Different languages use different MT systems based on their need and language structure. The merits and demerits of each approach are described in the Table 1.

Table 1: Difference between MT Approaches

S.no	Machine Translation Approach	Advantage	Disadvantage
1.	Direct Approach	1. Translation can be easily understood by the reader with minimal effort.	1. It does not consider only in lexical structure of the word and the relationship between the words as it involves analysis. 2. It mainly focuses on developing specific language pairs and not suitable for different language pairs. 3. More expensive for multi lingual implementation.
2.	Interlingua Approach	1. Well suitable for domain based approach. 2. It is very much useful for multilingual systems.	1. Less time efficiency when compare to direct machine translation. 2. Difficult to bring the intermediate representation which gains the meaningful sentence. 3. Representing many languages is difficult as the culture or structure of a language differs.
3.	Transfer based Approach	1. The system can handle the ambiguities in different languages. 2. It defines the modular structure.	1. Loss of meaning of the text may be lost at the end.
4.	Dictionary based Approach	1. It helps in speed up the human translation process by correcting syntax and grammar of the sentence.	1. Less useful in translating sentences.
5.	Example based Approach	1. It is well suitable for the languages which are similar in structure.	1. It requires a huge bilingual corpus of the language pair in which the translation has to be performed.
6.	Statistical based approach	1. It is language independent as it is not designed for a particular language pair. 2. It can be generalized for any language pair. 3. Less expensive when compare to rule based system. 4. Translations used to be natural as it is trained by the real time texts.	1. It requires extensive hardware configuration. 2. It requires words at least in millions in corpora for a domain based translation.

VIII. CONCLUSION

This paper describes various MT approaches used in Indian Languages. Each approach has its own Pros and cons. Most of the Indian languages MT system uses Rule based and Statistical approaches. The comparison among various MT approaches is given in a table format. Various MT systems use different approach depends on their application. For multilingual environment, statistical approach is best suited as well more flexible. For structurally similar languages, direct translation can be used. The interlingua based system is appropriated for multi lingual translation. In the past few years, data driven approach comes into the picture for its success in robustness. Example based approaches as well as Statistical based approaches lead to the other data-driven approaches such as Maximum entropy data driven modelling. Statistical techniques use the linguistic knowledge in the model and its performance can be much improved by using large parallel corpus. Hybrid systems showed the better performance while compare to the other models. In India most of the MT systems are developed for Hindi language and very few importance has been given to south Indian languages. Hence it is necessary to develop various MT systems for south Indian languages to overcome the language barriers in India.

REFERENCES

1. Dubey P., Overcoming the Digital Divide through Machine Translation, Translation Journal, Volume 15, 2011, http://translationjournal.net/journal/55mt_india. htm [Dec 12, 2011].
2. Murthy, B. K., Deshpande W. R., Language technology in India: past, present and future, 1998, http://www.cicc.or.jp/english/hyoujyunka/mlit3/7-12.html [Dec 11,2011]
3. Dave S., Parikh J. and Bhattacharyya P., Interlingua Based English Hindi Machine Translation and Language Divergence, Machine Translation, 2002, Volume 16, pp.251-304.
4. Vijayanand K., Choudhury S.I., P., Ratna P., VAASAANUBAADA - Automatic Machine Translation of Bilingual Bengali-Assamese News Texts, Language Engineering Conference,2002, Hydrabad, India, pp.183-188.
5. Bandyopadhyay S., State and Role of Machine Translation in India. Machine Translation Review, 2000, pp.11: 25.
6. R.M.K. Sinha et. al., ANGLABHARTI: A Multi-lingual Machine Aided Translation Project on Translation from English to Hindi, 1995 IEEE International Conference on Systems, Man Cybernetics, Vancouver, Canada, 1995, pp 1609-1614.
7. Bharati A., Chaitanya V., Kulkarni A.P., and Sangal R., Anusaaraka: Machine Translation in Stages, Vivek: A Quarterly in Artificial Intelligence, Vol. 10, No.3, 1997, pp. 22-25.
8. GoutamKumar S., The EB-Anubad translator: A hybrid scheme, Journal of Zhejiang University SCIENCE 2005, ISSN 1009-3095, pp.1047-1050.
9. G. Vasuki, S.Rajendran, English To Tamil Machine Translation System Using Parallel Corpus, 2013.
10. Sugata Sanyal & Rajdeep Borgohain, (2013) "Machine Translation Systems in India", Cornel University Library,

arxiv.org/ftp/arxiv/papers/1304/1304.7728.pdf
11. Antony P. J., (2013) "Machine Translation Approaches and Survey for Indian Languages", International journal of Computational Linguistics and Chinese Language Processing Vol. 18, No. 1, pp. 47-78.
12. Manoj Jain & Om P. Damani, (2009), "English to UNL (Interlingua) Enconversion", in proceedings of 4th Language and Translation Conference (LTC-09).
13. Smriti Singh, Mrugank Dalal, Vishal Vachhani, Pushpak Bhattacharyya & Om P. Damani, (2007), "Hindi Generation from Interlingua (UNL)", in proceedings of MT Summit, 2007.
14. Shachi Dave, Jignashu Parikh & Pushpak Bhattacharyya, (2002) "Interlingua-based English-Hindi Machine Translation and Language Divergence", Journal of Machine Translation, pp. 251-304.
15. Sudip Naskar & Shivaji Bandyopadhyay, (2005) "Use of Machine Translation in India: Current status" AAMT Journal, pp. 25-31.
16. Sneha Tripathi & Juran Krishna Sarkhel, (2010) "Approaches to Machine Translation", International journal of Annals of Library and Information Studies, Vol. 57, pp. 388-393.
Lata Gore & Nishigandha Patil, (2002) "English to Hindi - Translation System", In proceedings of Symposium on Translation Support Systems. IIT Kanpur. pp. 178-184.

AUTHORS PROFILE

S.Anbukkarasi, working as an Assistant Professor in Nandha Engineering College, Erode. She completed her B.Tech in Kongu Engineering College and M.E in Vidyaa Vikas College of Engineering and Technology.

S.Varadhaganapathy,working as Professor in the Department of Information Technology in Kongu Engineering College, Erode.