Machine Translation: Concepts and Techniques

Harjeet Singh
Assistant Professor, P.G. Department of Computer Science
Mata Gujri College, Fatehgarh Sahib
zrjeet@gmail.com

Abstract

A machine translation (MT) is a process in when a text is inputted in one language (called the source language(SL)), then the system converts it into another language (called the target language(TL)). The SL and TL are natural languages like Hindi and Punjabi. Machine translation is an very important application of artificial intelligence (AI). Although a lot of research has been done in the past for the development of accurate Machine Translation Systems in India but still there are many problems that need to be solved. In this paper I discuss about the basic concepts and techniques used in the Machine Translation.

Introduction

The Machine Translation systems with the help of computer programs produce the output text which after some post-editing results in high quality translation. Although the translation of literary or cultural text is still a challenging task but Machine Translation proves to be a magic tool for those who require translation in bulk on daily basis for example translation of technical and scientific documents, reports of newspaper or legal documents etc. The requirement of automated system is there because the translation process in most of these fields is repetitive and requires lot of human efforts. The increased demand for such translations is not fulfilled by humans at economical cost.

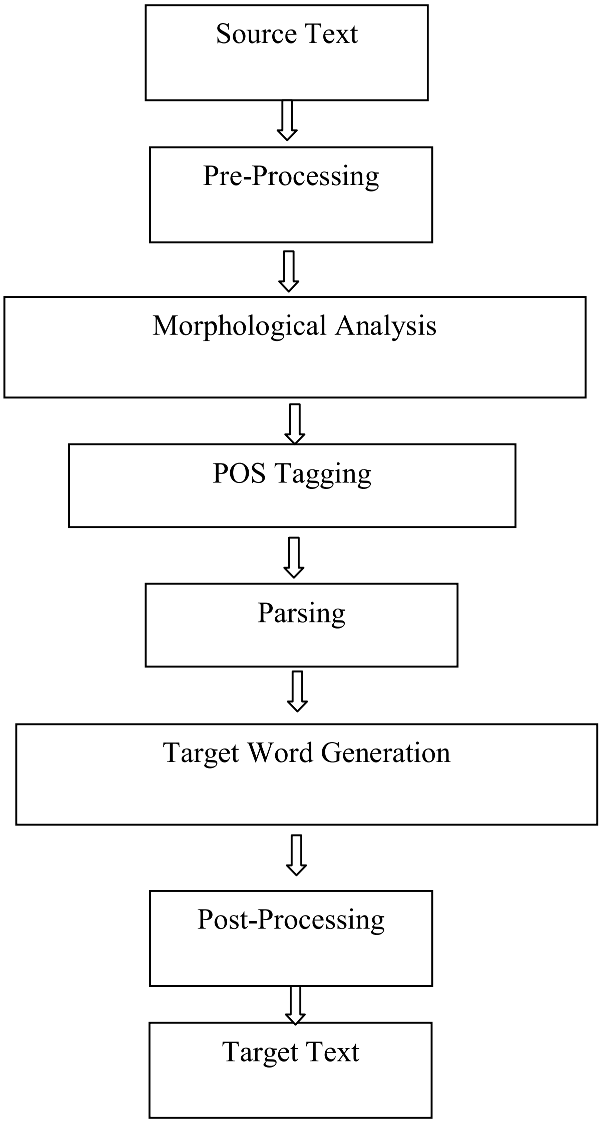
MT is now a field which is very important commercially, scientifically, philosophically and in many more domains of present life. Earlier it was assumed that the research on Machine Translation is a waste of time, the quality of translation is generally low, it is a threat for the human translators. But with time the developments in the field of MT have proved all these conceptions as false. The advancements in the architectures of MT i.e. from linguistic architectures to corpus based architectures improve the quality of MT.

The MT systems can be bilingual i.e. which are developed for specific language pair or multilingual means which are developed for more than two languages. Further the MT systems can be uni-directional means the translation takes place only in one direction (For example from

Punjabi to Malwai and not from Malwai to Punjabi) or bi-directional means the translation takes place in both directions.

Machine Translation Process

The Machine translation is the process in which various steps are executed in sequence but depending upon the language pair concerned some of the steps may be skipped. The General architecture of the MT system is



Machine Translation Process

Pre-Processing: There are a number of tasks to be performed in this step depending on the language pair involved. It is also called pre-editing of the input text. Some of the tasks that are performed during this phase are

i) The major task of this phase is to break the source text into morphemes depending upon the boundaries of language involved.

ii) Standardization of certain word in the inputted text is required as more than one writing styles may exist for the same words.

iii) To detect proper nouns for example names of persons, rivers, places etc., collocations that need not to be translated but to be transliterated.

Morphological analysis and generation

During the morphological analysis all the tokens of the input text are analyzed and it returns with root word and grammatical information about that word means the word class it belongs to. The role of the Morphological generator is just the reverse of morphological analyzer.

Part-of-speech tagging: The morphological analyzer sometimes creates ambiguity as the single word may have more than one part-of-speech tag. This is caused as the same word can come as noun, verb, or postposition etc. in the same sentence. This ambiguity is removed by the part-of-speech tagger as it uses the context information for that.

Phrase chunking: The task of the POS tagging is at the word level and tree structure of the sentence is made after the grammatical analysis. The purpose of the Phrase chunking is to put the tags to word sequences. This Phrase Chunking process is also known as shallow parsing.

Parsing: The full syntactic analysis of the source text is done by the parser. The purpose of the parser is to return a parse tree which describes its syntactic components and their associations with each other.

Translation: The process of translation starts when all the required information regarding the words of a sentence in the input text is obtained. Depending upon which approach of Machine Translation is being used, accordingly the different steps are performed in this phase. The different databases, Grammatical rules etc. are applied for conversion.

Transliteration: Certain words like technical terms that cannot be translated by the MT System, so these are transliterated.

Rearrangement of word order: If the source language and target language are not closely related and these have different word orders then we need a mechanism by which output is produced according to the word order of target language.

Post Processing: Sometime there is need of post processing to achieve more accuracy. It depends upon the output produced by the system whether the Post-processing is required or not.

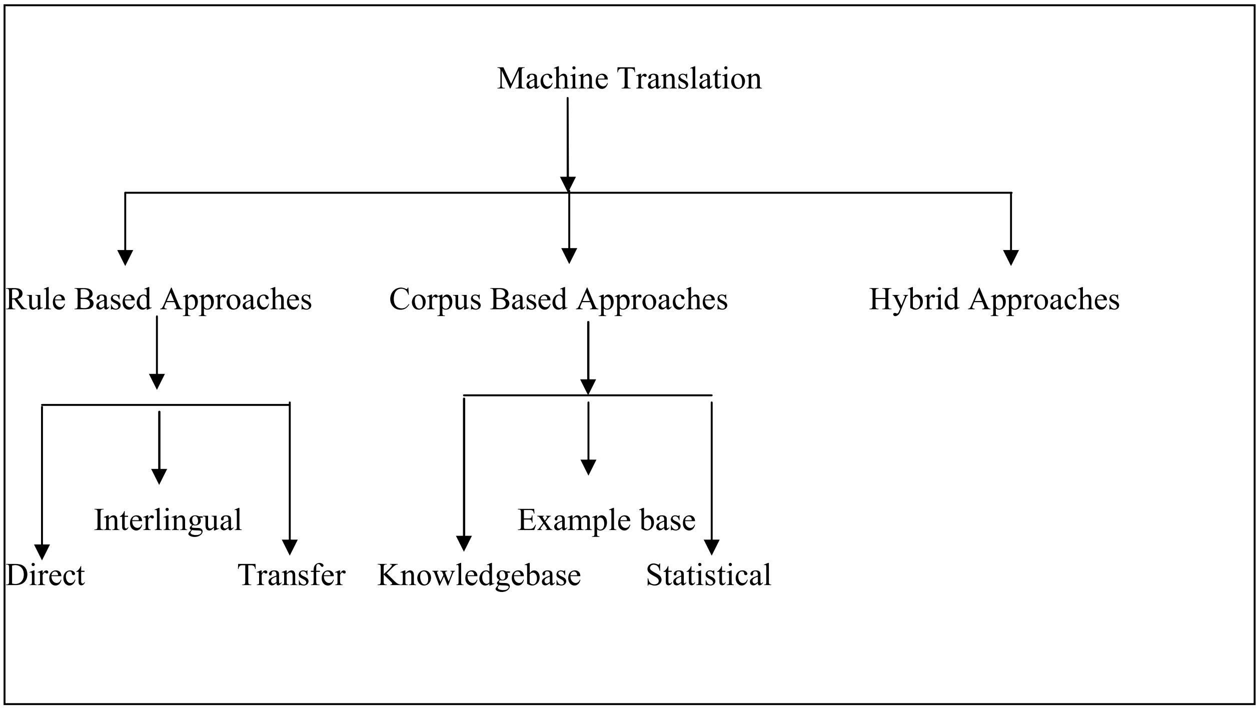
If the output is not as per the standard then more efforts are required in the post-processing to improve it.

Among all these activities which are to be used in the MT System, it depends upon the approach being used by the machine translation system.

Approaches to Machine Translation

The Machine Translation approaches are classified into generally three categories

1. Rule Based Approaches
2. Corpus Based Approaches
3. Hybrid Approaches



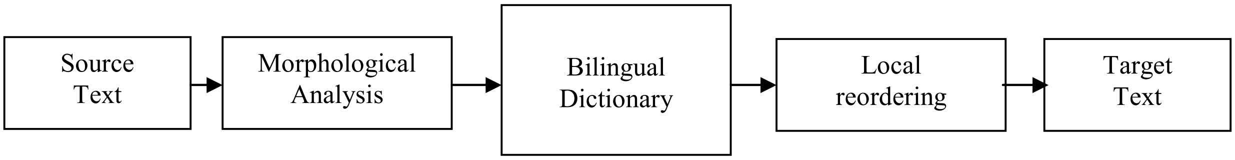
Approaches to Machine Translation

Rule-Based Machine Translation

The Rule based approaches are based on the bilingual dictionaries and database of linguistic rules of the source and target language. The Rule based approach is further categorized into three approaches:

1. Direct Approach
2. Interlingual Approach
3. Transfer Approach

Direct Approach: The direct approach of MT is also known as word for word translation. The text of the source language is converted into the target language with the help of bilingual dictionaries and some reordering is done using linguistic database. The process of Direct approach is as follows:



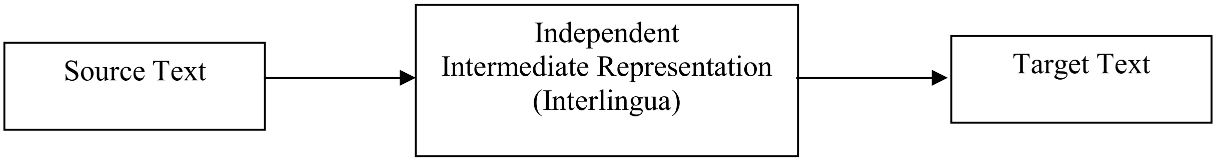
Direct MT System

GAT system (1954) developed by Georgetown University for Russian-English language pair was based on the direct approach.

The limitation of Direct approach is that 1) it is uni-directional which means that it works for the translation in one direction only (source language to target language and not vice-versa). 2) If there is complexity in the source text then the system does not produce quality output due to lack of capability to analyze the source text.

Interlingual Approach

In this approach the source text is represented in the intermediate language which is then converted into the target text. This intermediate representation is an abstract representation which contains the information required to get the target text.

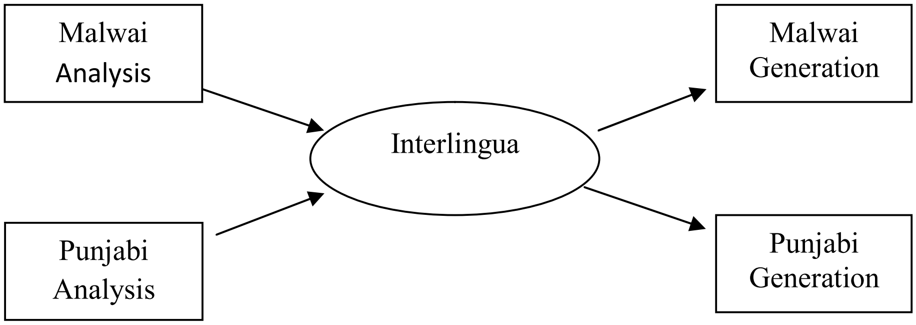


Interlingual MT System

In this approach, the translation is done in two stages:

1. from the source language to the Interlingua (IL) and
2. from the IL to the target language.

The following figure represents the interlingua Model for two languages:



Iinterlingua Model for Two Languages

The advantage of the Interlingual approach is that the addition of new language pair is easy and the problem with this approach was the difficulty in designing intermediate language independent Interlingua.

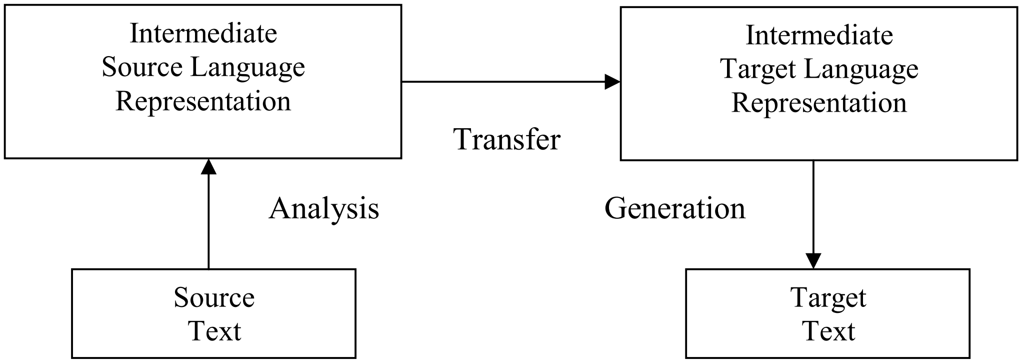
Mechanical Translation and Analysis of Languages (METAL) system for German-English pair was developed by Linguistic Research Center at University of Texas used interlingua approach of MT.

Transfer Approach

In this approach the source text is converted into intermediate representation which is dependent upon the source language and then this is converted into the representation which is dependent upon the target language and finally it is converted into the target text.

The Translation in Transfer approach takes three steps

1. Analysis: The source language is analyzed i.e. the syntactic analysis is performed.
2. Transfer: The syntactic structure of the source language is converted into the syntactic structure of the Target Language.
3. Generation: Target text is generated.


Intermediate Source Language Representation
Intermediate Target Language Representation
Transfer
Analysis
Generation
Source Text
Target Text

Transfer Approach

The reason behind the wide use of this approach over the Interlingua approach was due to the fact that it becomes easy to design language dependent representation as compared to language independent representation. The other foremost reason was the complex structure of the analysis and context grammar in Interlingua approach.

When a new language is to be added then there is need to add two modules of analysis and generation along with new transfer modules.

Corpus Based Approach

Corpus based approaches are more efficient than the traditional approaches of Machine Translation as these approaches give more accuracy. In this approach a rich repository of bilingual parallel text is used. The larger is the size of parallel corpora, the higher level of accuracy will be given by the MT system. There are a number of approaches which are based on this approach:

1. Knowledgebase Approach
2. Example Based Approach
3. Statistical Approach

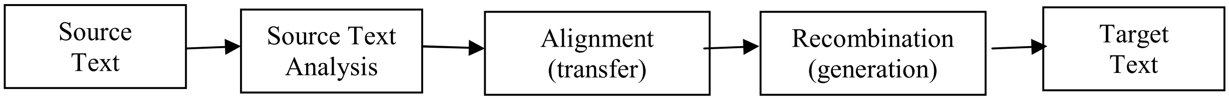
Knowledgebase Approach:

The focus of this approach is to analyze and understand the source language to higher level before the translation process. The idea is to collect as much linguistic information as possible and add to the knowledge base of translation system to get more accurate translation. To produce the high quality output, there is need to collect the domain knowledge base along with the dictionaries, knowledge of structure and rules of source and target language, word/sentences/paragraphs which are translated previously etc.

KANT (Knowledge-based, Accurate Natural-language Translation) Machine Translation system is the example of Knowledge Base Machine Translation System.

Example-Based MT

This approach is based on the bilingual corpora of parallel text in which example of the sentences of the source text and their corresponding sentences in the target language are given. The translation engine use these examples to translate the similar sentences of the source text to produce the target text.



Example-Based MT

The translation is divided into three steps in Example based Machine Translation. The steps are as follows

1. Source Language Analysis: When the source text is inputted in the system, the system tries to find the match the source strings with the examples in the example database.
2. Alignment: The matches of the source-target strings are extracted
3. Recombining: generate the Target translation by combining the fragments as per the rules of target language.

AnglaBharti-II and AnglaHindi systems are the MT systems which used this approach.

Statistical approach

This approach is also based on rich repository of bilingual text corpora but the translation is done using statistical methods. Baseline Statistical Model has main three components:

1. Translation Model: This contains target language monolingual data.
2. Language Model: This contains line-aligned parallel text of source - target language pair.
3. Decoding : It is a search process which finds possible target translations.

Conclusion

Machine translation has been an area of research which proves to be very useful in near future. But their are number of challenges during translation that need to be solved and for which more detailed study of various natural languages is required. So still a lot of work is required to develop a completely automatic translation system.

References

[1] Hutchins, W.J., Somers, L., (1992), "An Introduction to Machine Translation," Academic Press, London, from web site http://www.hutchinsweb.me.uk/IntroMT-0-Contents.pdf .

[2] Hutchins, J.., (1992), " Machine Translation: General Overview," from web site http://www.hutchinsweb.me.uk/Mitkov-2003.pdf .

[3] Goyal, V., (2010), “Development of a Hindi to Punjabi Machine Translation System,” Ph.D. Thesis, Department of Computer Science, Punjabi University, Patiala.

[4] Allen, J., (1995), "Natural Language Understanding. Second Edition. Benjamin Cummings".

[5] Seasly, J.,(2003), "Machine Translation: A Survey of Approaches," University of Michigan, Ann Arbor, 2003.

[6] Tripathi, S., Sarkhel, J.K., (2010), "Approaches to Machine Translation," Annals of Library and Information Studies, Vol. 57, December 2010, pp 388-93 from web site http://nopr.niscair.res.in/bitstream/123456789/11057/1/ALIS%2057(4)%20388-393.pdf

[7] Tiedemann, J., (2009), "Machine Translation: Rule-based MT & MT evaluation," from web site http://stp.lingfil.uu.se/~joerg/mt09/f2_RBMT_eval- 2x2.pdf

[8] Vauquois, B. and Boitet, C., (1985), "Automated translation at Grenoble University," Computational Linguistics 11, pp. 28-36.

[9] Slocum, J., (1985), "A survey of machine translation: its history, current status, and future prospects," Journal of Computational Linguistic - Special issues on Machine Translation, Volume 11, issue 1,pp. 1-17.

[10] John, H., (2007), "Machine translation: a concise history, Computer aided translation: Theory and practice," ed. Chan Sin Wai. Chinese University of Hong Kong, 2007. Citesser. pp. 1-21.