ISSN: 2321-7782 (Online)
Volume 3, Issue 2, February 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

A Review of Various Approaches Used for Machine Translation

Mamta
Department of Computer Science and Engineering
Career Point University Hamirpur
Himachal Pradesh, India

Abstract: ***Machine translation has been a topic of research or interest from the past many years. Many techniques and methods for various languages have been proposed and developed using hybrid based, statistical based as well as rule based approaches. At present, a number of government and private sector projects are working towards developing a full pledged MT for Indian language. Machine translation is an important branch of artificial intelligence. Artificial intelligence is very useful in providing people with a machine, which understands diverse language spoken around the world. Machine translation helps people from different places to understand an unknown language without the need of human aided translator. Machine translation is the process using a software application that translate the one language (source language) into the another language (target language) without human intervention. This paper gives a brief survey on various approaches of machine translation and gives a comparative view of HBMT, RBMT and SMT.***

Keywords: machine translation, hybrid based machine translation, statistical based machine translation, rule based machine translation.

I. INTRODUCTION

Language is an effective medium of communication to represents ideas and expressions of human mind. There are minimum of 30 different languages and 2000 dialects used for the communication by the Indian peoples. There are 22 major languages in India, written in 13 different scripts, with over 720 dialects and used for administration work and communication purpose for different states. The official Indian languages are Hindi (with approximately 420 million speakers) and English, which is also widely spoken.

Due to rapid industrialization and a bustling influence of multinationals in the economy, English has become the most common language both in India as well as world. It is defacto language for two key area: education and administration. Internet is media for information retrieval and information is available in English on internet and also people from different states have different languages and different culture, so there is a big need of inter language translation to transfer their information, share ideas and communicate with one another. Peoples of different states perform their work in respective regional languages where as the work at the Union Government offices is performed in English language which is assumed to be one of the most speaking languages in the world or Hindi Language. So to synchronize between state government and the central / Union government there is a need for translation from regional languages to English language and vice versa. From the above discussion it is clear that there is large scope of translation of text from English to Indian Languages and vice versa. The initial work on Indian Machine Translation (in the beginning of 90's) was performed at various locations by different persons like IIT Kanpur, Computer and Information Science department of Hyderabad , NCST Mumbai, CDAC Pune, Department of IT, Ministry of Communication and IT Government of India. In the mid 90's and late 90's some more machine translation projects also started at IIT Bombay, IIT Hyderabad, Department of computer science and Engineering Jadavpur University, Kolkata, JNU New Delhi etc.

Machine translation is one of the most important applications of Natural Language Processing. Machine translation helps the people from different places to understand an unknown language without the aid of a human translator. The module present concerns with the Machine Translation domain of Natural Language Processing. This area of Artificial Intelligence is very useful in providing people with a machine, which understands diverse languages spoken by common people. The Source Language (SL) is the language which is to be translated & the Target Language (TL) is language in which it is translated. While translating, the syntactic structure and semantics structure of both source language and target language should be considered. There are different techniques for machine translation is hybrid based, statistical based and rule based technique. The rest of the paper is organized as follows. Section II will provide the overview of machine translation and its working, the review of various proposed techniques will be discussed in Section III, in Section IV we will provide the a comparative study of above techniques, and in last section we will conclude our study.

II. OVERVIEW

Machine Translation:

Machine translation is one of the most important applications of computational linguistics that uses the computer software or web to translate text from one language to another language. Machine translation helps people from different places to understand an unknown language without the aid of a human translator.

Machine Translation (MT) is automated translation from one language to another by using computer software. Machine translation is often perceived as low quality based an outdated perception created by older translation technologies or freely available generic translation tools from Google or Bing that have not been customized for a specific purpose. Many technology advances have been made in recent years that are changing this perception, with customized machine translation engines [12].

Machine translation is the process of translating from source language text into the target language. The following diagram shows all the phases involved.

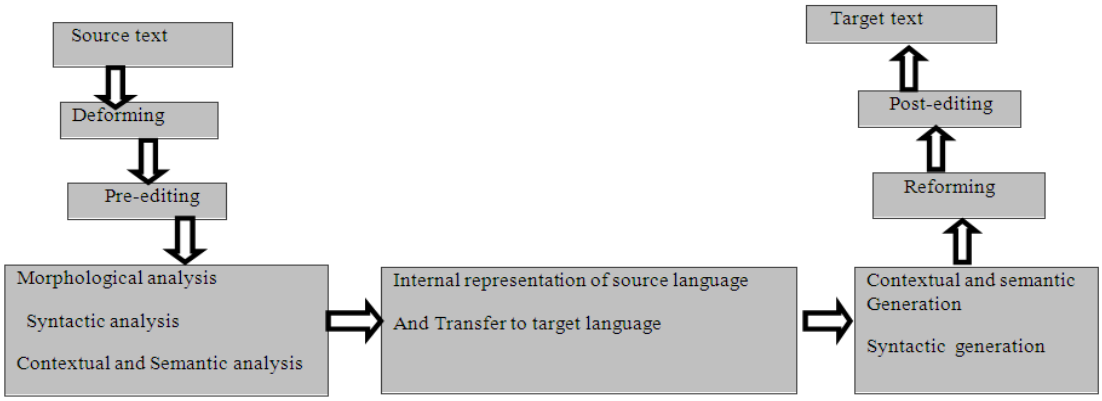

Fig. 1 Machine Translation Process

Text: Source text and target text comes in text, source text is the first phase in the *machine translation process*. The sentence can be classified that have relations, expectations, assumptions, and conditions make the MT system understand very difficult. World knowledge and commonsense knowledge could be required. Target text is the last phase in which required output comes.

Deforming and Reforming: To make the machine translation process easier Deforming and Reforming are used. The source text may contain figures, flowcharts, diagrams, etc that do not require any translation and only the translation portions should be identified by the deforming. Once the text is translated, the target text is to be reforming after post-editing to see that the target text also contains the non-translation portion.

Pre-editing and Post-editing: During pre-editing, fixing up the punctuation marks and blocking material that does not require translation. Post editing is done to make sure that the quality of the translation is upto the mark. Post-editing should continue till the MT systems reach to the target output.

Analysis, Transfer and Generation: Morphological analysis determines the word form such as tense, number, part of speech (POS), etc. Syntactic analysis determines whether the word is subject or object. Semantic and contextual analysis determines a proper interpretation of a sentence from the results produced by the syntactic analysis. Syntactic and semantic analysis is often executed simultaneously and produces syntactic tree structure and semantic network respectively. This results in internal structure of a sentence (source text). The sentence generation phase is just reverse of the process of analysis**.**

Parsing and Tagging: Parsing is the assessment of the functions of the words in relation to each other. And Tagging means the identification of linguistic properties of the individual words.

Semantic and Contextual analysis and Generation: The semantic analysers use lexicon and grammar to create context independent meanings. The source of knowledge consists of meaning of words, meanings associated with grammatical structures, knowledge about the discourse context and commonsense knowledge [14].

III. Review of Various Proposed Techniques in Machine translation

MT Approaches:

There are number of approaches in Machine Translation, but here we take only three approaches which are mostly used like:

1) Hybrid Based Machine Translation (HBMT)

2) Statistical Based Machine Translation (SMT)

3) Rule Based Machine Translation (RBMT).

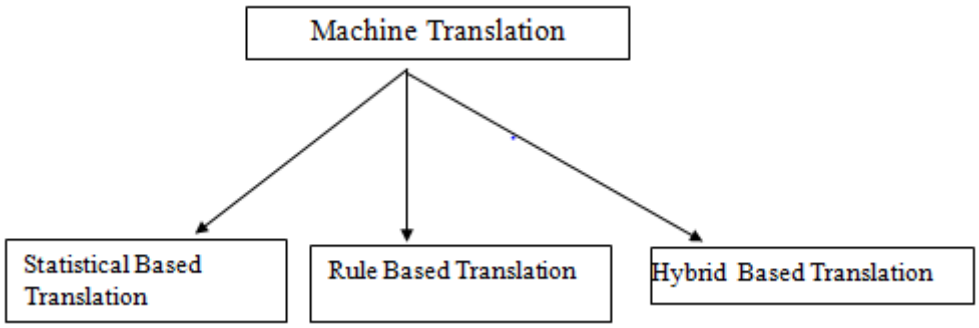

Fig. 2 Approaches of machine translation

1) Rule Based Machine Translation Technique (RBMT):

The rule-based paradigm is one of the important technique to Machine Translation. It translates the source text into target text by linguistic rules. There are three techniques of rule based translation- direct based, transfer based and Interlingua based approach. Methodology RBMT uses a set of linguistic rules in three different phases: analysis, transfer and generation. Rule based system requires: syntax analysis, semantic analysis, syntax generation and semantic generation. The analysis produces a complete parsing of a source language sentence. In the analysis and generation stages, most systems have clearly separated components with different levels of linguistic description: morphology, syntax and semantics etc. Analysis is divided into morphologic analysis, POS tagging, parsing, chunking, dependency analysis. Transfer phase consists of local reordering and long distance reordering. In the final, generation phase have lexical transfer, mapping and agreement. After these, systems

generate the target output. Rule based machine translation system is developed by hand coded rules for translation and the system requires special programs, good linguistic knowledge to write linguistic rules and bilingual dictionary also needed. There are no human interventions during the conversion from one language to another language. Human intervention only takes place, if at all, after translation: errors in the machine translation output are manually corrected.

The main drawback of RBMT is the construction of such systems demands a great amount of time and linguistic resources which is very expensive. Moreover, to improve the quality of a RBMT, it is necessary to modify rules, which requires more linguistic knowledge. Modification of one rule cannot guarantee that the overall accuracy will be better.

2) Statistical Machine Translation technique (SMT):

According to [20], the statistical machine translation (SMT) is a machine translation where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. The SMT is a corpus based approach, where a massive parallel corpus is required for training the SMT systems. The SMT systems are built based on two probabilistic models: language model and translation model. The advantage of SMT system is that linguistic knowledge is not required for building them. The difficulty in SMT system is creating massive parallel corpus. SMT systems work well for machine translation of English to European languages because the word order is almost preserved in such translations. For machine translation of English to Indian languages, the parallel corpora have to be pre-processed (changing word-order) and trained in SMT. There are *word-based* and *phrase-based* models. Word-based models consider sentences as a combination of single words, ignoring the structural relations between them. Phrase-based models consider sentences as a combination of phrases or chunks. In both cases, the combination of elements is modeled purely statistically. Modern SMT systems are phrase-based rather than word-based, and assemble translations using the overlap in phrases.

Acc. to Alvi Syahrina, SMT is based on the concept of probability. The translation is chosen from the highest probability. The probability score is obtained by previous data from training the SMT with human translated document. The probability score is obtained from mathematical model, including language model and translation model. The source language text is pre-processed first before applying language model and global search model and preprocessed again for the final presentation in the target language text.

Drawback of SMT is that in the beginning of the establishment, not only it needs a lot of data, but also a number of repetitions of training. There is also no specific method quality control of corpora. Some languages also lacking in monolingual data or/and bilingual data.

3) Hybrid based translation:

Hybrid Machine Translation (HMT) was built due to the limitations of the two approaches and their possibility to be integrated. Statistical Machine Translation and Rule-Based Translation are two MT approaches which work oppositely. SMT did not need to learn about the language at all, while RBM□s basis is gathering language rules. Due to this difference, SMT and RMT give a different performance. There are several forms of hybrid machine translation such as Multi-Engine, statistical rule generation and multi-pass, the most common forms are:

» **Rules post-processed by statistics**:

Translations are performed using a rules based engine. Statistics are then used in an attempt to adjust/correct the output from the rules engine .This is also known as statistical smoothing and automatic post editing. This is more of a “Band-Aid” approach to machine translation where there is an attempt to improve lower quality output from an RBMT engine rather than addressing the root cause of issues.

» **Statistics guided by rules**:

Rules are used to pre-process data in an attempt to better guide the statistical engine. Rules are also used to post-process the statistical output to perform functions such as normalization. This approach has a lot more power, flexibility and control when translating. Many issues can be addressed at their root causes through rules that go beyond the capabilities on a statistical only approach.

[17]The drawback of HBMT is while hybrid solutions may successfully combine the benefits of both approaches they also combine the limitations of each approach. They maintain the high costs of Rule-Based MT while introducing additional complexities of managing side-by-side systems making their true commercial value questionable.

IV. Comparative Study of Machine Translation Approaches

We compare three approaches of machine translation on the basis of processing, benefits, limitations, languages, products and engine. And on the basis of comparision we show that which one of approach is best out of three. The comparision is explained as follow:

RBMT	SMT	HBMT
Core process is the bilingual dictionaries and rules for converting SL structures into TL structures.	Core process is the 'translation model' taking SL words or phrases as input and producing TL words or phrases as output.	Core process is the combining multi engine machine translation using ' black box integ-ration' taking SL words. Multiengine can be RBMT and SMT engine.
The preceding stage of analysis interprets input SL strings into appropriate 'translation units' (like canonical nouns and verb forms) and relations (like dependencies and syntactic units).	The preceding 'analysis' stage is repre-sented by the process of matching indivi-dual words or word sequences of input SL text against entries in translation model.	The preceding stage of Analysis there are two main task performed. First, identification of the correct function and meaning of word, phrase and clauses. Second, analysis is to capture and save information of subject and predicate of the sentence.
Succeeding stage of 'synthesis'deri-ves TL texts from TL representations produced by the core Process	Succeeding stage involve a 'language model' which synthesizes TL words as 'meaningful' TL sentences.	Succeeding stage is comparing the output of two engines for TL sentences.
Product: Desktop and Server Solution.	**Product:** Server-based solutions only	**Product:** Server-based solutions only
Engine Training: Source texts (at least, 100,000 words / 10,000 translation units)	**Engine Training:** Parallel corpora (at least, 5,000,000 words / 500,000	**Engine Training:** Parallel corpora (at least, 500,000 words / 50,000 translation units)
Languages: English, Russian, Ger--man, French, Spanish, Italian, Polish Portuguese, Chinese, Ukrainian, Kazakh, Turkish, Bulgarian and Latvian.	**Languages:** Any language pair, for which there are enough training data.	**Languages:** English, Russian, German, French, Spanish, Italian, Portuguese.
Limitation: Language-dependent (algorithms depend on source/target languages). High customization effort.	**Limitation:** Requires large and clean parallel corpora for training. Domain-specific (usually trained on/for specific texts) . Hard to customize the translation of a particular word/construction.	**Limitation:** Requires parallel corpora for training (but less than pure SMT). Domain-specific (usually trained on/for Specific texts).
Benifits: Full control over termi-nology and translation style. More accurate syntax and morphology Predictable and deterministic. Profiling (multiple profiles can be easily created in one engine.	**Benifits:** Fast and fully automated engine training (in most cases, language independent). More fluent and "human-like" MT output	**Benifits**: More customizable and predictable than pure SMT. More fluent and "human-like" MT output than pure RBMT Engine training is faster than pure RBMT.

V. Conclusion

In this project, we compare the three approaches of machine translation RBMT, SMT and HBMT on the basis of benefits, limitations, languages, products and engine etc. It has been observed that SMT approach is better than the other approaches for translation of languages on the basis of its ability to translate all the languages and engine training having large number of words.

Acknowledgment

I take this opportunity to express a deep sense of gratitude to HOD of CSE Dept. Ms. Pratibha Sharma for her cordial support, valuable information and guidance, which helped me to do better than I can. Their guidance and motivation conceived a direction in me. I am obliged to all the faculty members of CSE Department of Career Point University Hamirpur, for the valuable information provided by them in their respective fields.

Last but not the least I shall thankful to my parents and all my friends for their constant encouragement and thoughts whenever I was in low spirits.

References

1. Sitender, Seema Bawa, "Survey of Indian Machine Translation Systems", Jan - March 2012.
2. D.D. Rao, "Machine Translation A gentle Introduction", RESONANCE, July 1998.
3. Antony P. J." Machine Translation Approaches and Survey for Indian Languages" Computational Linguistics and Chinese Language Processing Vol. 18, No. 1, March 2013, pp. 47-78.
4. W. John Hutchins "Machine translation: a concise history" [Website: http://ourworld.compuserve.com/homepages/WJHutchins]
5. "Competitiveness And Innovation Framework Programmatic" Policy Support Programme (ICT PSP) Project Acronym: MORMED, Project Full Title: Multilingual Organic Information Management in the Medical Domain, date 2010-11-05 version 0.12.s
6. Marta R. Costa-Juss's, Mireia Farr´us, Jos´e B. Mari˜no, Jos´e A.R. Fonollosa" Study And Comparison Of Rule-Based And Statistical Catalan-Spanish Machine Translation Systems" Vol. 31, 2012, 245–270.
7. Vishal Goyal, M.Tech. Gurpreet Singh Lehal, Ph.D."Advances in Machine Translation Systems", Volume 9 : 11 November 2009 ISSN 1930-2940.
8. Uday C. Patkar, P. R. Devale, S. H. Patil, 'Transformation of multiple English text sentences to vocal Sanskrit using Rule-Based technique', International Journal of Computers and Distributed Systems, Vol. No.2, Issue 1, December 2012.
9. Euro Matrix Statistical and Hybrid Machine Translation between All European Languages "Survey of Machine Translation Evaluation".
10. Jörg Tiedemann "Machine Translation Rule-based MT & MT evaluation" Department of Linguistics and Philology Uppsala University September 2009.
11. Alvi Syahrina (s104854)" Online Machine Translator System and Result Comparison" Year: 2011
12. http://www.asiaonline.net/EN/MachineTranslation/default.aspx?QID=1#QID1
13. http://www.asiaonline.net/EN/MachineTranslation/default.aspx?QID=14#QID14
14. http://language.worldofcomputing.net/machine-translation/machine-translation-process.html
15. http://language.worldofcomputing.net/machine-translation/challenges-in-machine-translation.html
16. http://www.promt.com/company/technology/compare/
17. http://www.safaba.com/machine-translation/machine-translation-technologies/hybrid-machine-translation
18. http://www.asiaonline.net/EN/MachineTranslation/default.aspx?QID=18#QID18
19. http://language.worldofcomputing.net/machine-translation/machine-translation-overview.html
20. R.Harshawardhan, Mridula Sara Augustine, K. P. Soman, "Advanced English – Malayalam Translation Memory for Natural Language Processing Applications", in Proc. of Nat. Conf. on Indian Language Computing (NCILC), February, 2011.

Author(s) Profile

Mamta, received the B. Tech degree in Computer Science and Engineering from Himachal Pradesh University, India in 2013. Currently, pursuing M.Tech in Computer Science and Engineering (Specialization in Mobile Computing) from Career Point University Hamirpur during 2013-15.

