Machine Translation

Preprint · October 2017

1 author:

Muhammad Irfan
Bahria University
1 PUBLICATION **0** CITATIONS

SEE PROFILE

Machine Translation

Muhammad Irfan
Department of Computer Science
Bahria University Islamabad
Email: muhammad.Irfan@outlook.com

***Abstract*—Language is core medium of communication and translation is core tool for the understand the information in unknown language.Machine translation helps the people to understand the information of unknown language without the help of Human translator. This study is brief introduction to machine translation.**

***Keywords*—Machine Translation, Rule-base Machine Translation, Example-base Machine Translation, Statistical Machine.**

I. Literature Review

The idea of universal language was found in 17th century [3].The concept of translation was base on word-to-word translation.
With modern computer and advance computation linguistics man has able to develop automatic machine translation. The Era of modern machine translation started in 1940s and upto 1960 much work was done in advance linguistics[3]. In 1966 the Automatic Language Processing Advisory Committee (ALPAC) reported that Machine Translation could not produce quality translations as human translators[3], [2].
In 1980s IBM started work in statistical machine translation and 1990s the parallel text availability had increased the interest in statistical machine translation. In 2006 an open-source Statistical Machine Translation tool called Moses was released and it currently the most complete Statistical Machine Translation software available[1].
In 2008 a text/SMS translation service for mobiles was introduced in Japan. In 2009 speech-to-speech translation were provided in mobile phone for English, Japanese and Chinese.The Neural machine translation is a new approach to machine translation proposed by Kalchbrenner and Blunsom in 2013[4].

II. Introduction

Language is medium in which human can express his/her idea. The total estimated language in world are between three thousand to eight thousand. It is very difficult to understand every language but the languages are the only medium of communication in human society. The idea of translation introduces to communicate messages from one language to others.

Machine Translation (MT)

Machine Translation means automatic translation, It the field of Artificial Intelligence. Machine translation is computer program which is design to translate text from one language (source language) to another language (target language) without the help of human. The aim of Machine Translation is to provide a system that translate text of source language into target language and translation express the same meaning as it in source language.

General Translation step

Any human translation can be described in following steps

- Decode the Source Text
- Re-encode it into Target language

Both decoding and encode require deep knowledge of source language(SL) and target language (TL). This includes grammar, semantic syntax understanding of both languages. Natural Language are very complex in term words meaning, grammar rules etc.

Machine Translation Approaches

The machine translation approaches are rule-based approach and corpus-based approach. For rule base machine translation, Human expert define set of rules for the translation process while in corpus base approach rules are automatically extracted.

1) Rule-based Machine Translation(RBMT) Approach: The Rule-based Machine Translation works on the morphology, syntax and semantic of both languages. So, we required the syntax analysis, semantic analysis of Source text and to generate the text in target language we need syntax generation and semantic generation. We also need the bilingual dictionary of source and target languages. General Steps of Rule-based Machine Translation are described in figure 1.

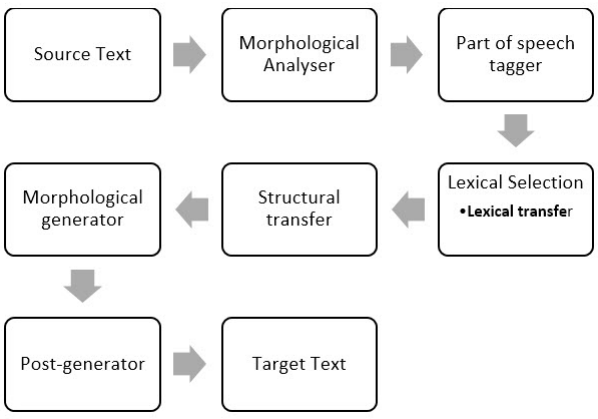

Fig. 1. Rule-based Machine Translation

Sub approaches in RBMT

The sub approaches in rule-based Machine translation are direct, transfer-based, interlingual Machine Translation approaches.

Direct Machine Translation Approach:
This is oldest approach and translation is performed at word level. There is no additional intermediary representation between source and target languages. Words of source language text is directly translated into the target language. This is uni-directional bilingual translation system. Direct machine translation approach involves the word by word translation with some modification at grammar level. The translation is not good as it is just the replacement of words from target language into source language text i-e word by word meaning replacement.

Interlingual Machine Translation Approach:
This approach introduces an intermediary language representation between source and target languages. This intermediary language is called Neutral Language. Neutral language can represent any natural language. It is independent of source and target Languages. It is also useful for multilingual translation machine system.KANT system was developed on interlingual approach in 1992 by Nyberg and Mitamura[5]. Building interlingual language is not an easy job. Too much efforts are required to develop truly neutral language.

Transfer base Machine Translation Approach
In this approach the text of source language is converted into intermediary representation, it is then used to generate the target language text with help bilingual dictionary and grammar rules. Transfer based machine translation process is divided into three phases.

Analysis
In this phase source language text is analyzed on basis of linguistic information and heuristics to parser the text (syntactic representation)

Transfer
The syntactic representation of source language is converted into the syntactic form of target language.

Generation
The final text in target language is generate with help of morphological analysis. This approach heavily dependent on the grammar and structure of sentence and changes to a monolingual component affect all transfer modules for that language.

2) Corpus-based Machine Translation Approach: It is actually data driven machine translation. It was introduced an alternative approach to the rule-based approach. In this approach the bilanguage parallel corpus is used to extract the translation for new sentences. A large amount of raw data is collected in parallel corpora. The raw data is actually the translation between source and target languages and this data is used for translation. The sub-approaches of Corpus-based Machine Translation are Statistical Machine Translation and Example-based Machine Translation.

Statistical Machine Translation(SMT)
This approach is basis on statistical model. It has two statistical probabilities models: language model and translation model and massive parallel corpora of source and target languages. 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. We have to two models in SMT, one is Word-based and other is phrase-based. In word-bases MT sentences are consider as combination of single words and structure relation between the words are ignored while in phrase-based model consider sentences as combination of phrases or chunk. The basic concept in SMT is probability. The probability score of translations are generated from already available translated data (parallel corpus, translated by human), the translation having high probability is selected as final translation. The probability is calculated with help of language and translation models.

A huge amount of data is need for SMT and evolved many training repetition process. There is also no specific method quality control of corpora.

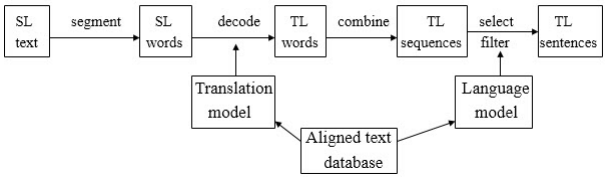

Fig. 2. Statistical Machine Translation

Example-based Machine Translation (EBMT)
Example-based machine translation contains the point to point mapping between the source and target language sentences i-e we have examples data that is translated between the source and target language[6]. This data is used for translation. The basic idea is if already translated sentence occur again it, the same translation is likely to be correct again. Basically, EBMT is memory-based translation and the concept of analogy is used for the translation.

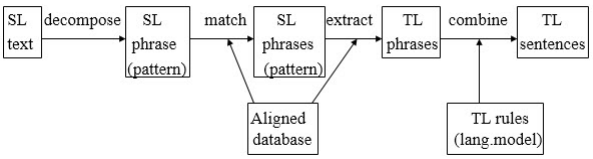

Fig. 3. Example-based Machine Translation

III. Conclusion

Many approaches have been proposed but none of them produce equivalent translation between target and source languages. It is difficult to achieve 100 percent accuracy in translation due the complexity nature of natural Languages. Every natural language having different sentence structures, grammar and lexicons. However, combining the best feature of different approaches into hybrid approach can improve the accuracy of translation. This quality improvement can increase the machine translation role in cross-language Information Retrieval System and Multilingual Information Retrieval systems.

Acknowledgment

The author would like to thank Dr. Arif UR Rehman Bahria University Islamabad.

REFERENCES

[1] Statistical Machine Translation. http://www.statmt.org/.

[2] Translation Automation Timeline. https://www.taus.net/academy/timelines/translation-automation-timeline.

[3] John Hutchins. Machine translation: History and general principles. *The encyclopedia of languages and linguistics*, 5:2322–2332, 1994.

[4] Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *EMNLP*, volume 3, page 413, 2013.

[5] Eric H. Nyberg and Teruko Mitamura. The kant system: Fast, accurate, high-quality translation in practical domains. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 3*, COLING '92, pages 1069–1073, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.

[6] MD Okpor. Machine translation approaches: issues and challenges. *International Journal of Computer Science Issues (IJCSI)*, 11(5):159, 2014.