Advances in Engineering Research, volume 178
Mizoram Science Congress 2018 (MSC 2018)

Machine translation and its approaches

Vanlalmuansangi Khenglawt[1]*, Lalṭanpuia[2]

[1]Department of Information Technology, [2]Department of Mathematics and Computer Science, Mizoram University, Tanhril 796004, Mizoram, India

Corresponding author: mzut208@mzu.edu.in

Language is a medium of communication for different cultures around the world. The communication between two cultures is blocked by language barrier. To solve the problem machine translation approach is widely used which is a software used to translate a text or speech from one language to another. It can be referred to as automatic or instant translation as it is much faster in comparison to human translation. Machine translation plays a very important role and in highly applicable in both individual and business areas. The translation system can be either bilingual or multilingual. Most of the machine translation is bidirectional translation which can be either translated from source language to target language or from target language to source language. Otherwise it is unidirectional which can be translated only from source language to target language. It has many approaches for automatic translation of language. This paper focusses on three different approaches, the rule-based machine translation, the corpus-based machine translation and the hybrid-based approach. Each of these approaches has its own pros and cons. We also present several challenges faced in the translation of languages. Also, the metrics for measuring the performance or the accuracy of the machine translation output have been mentioned. There are still several works required to develop completely automatic translation system. The performances of the translation system are still unsatisfactory especially in low resource language. As language is evolutionary in nature, there are still vase areas to cover in machine translation system.

Keywords: Machine translation, language, rule-based machine translation, corpus-based machine translation, hybrid-based translation.

INTRODUCTION

More than 6000 language exists in the world with a huge diversity in the structure of language. Therefore, it is impossible for human to understand all the languages. Hence, a simple dictionary lookup is not sufficient as languages are of different structure. So it is a challenging task for researcher to develop an automatic translation system called machine translation which translates a source language to target language. The objective of machine translation is to restore the meaning of the original text in the translated output (Tripathi and Sarkhel, 2010).

The issue of machine translation started since 1940 (Cheragui, 2012) with a massive improvement over the years. Machine translation is a subfield of computational linguistics that uses software to translate text or speech from one language to another. A simple substitution of word from one language to another cannot produce a good translation output because to translate a language, recognition of the whole phrase and the counterpart is required. Therefore, different approach has been developed through the years for a better translation.

MACHINE TRANSLATION

With a machine translation system there are two levels of translating a sentence: metaphrase and paraphrase (Tripathi and Sarkhel, 2010). Metaphrase is a literal translation that is word by word translation. The translated text may not convey the meaning of the original text. The semantics of the sentence may be differing from the original text.

Paraphrase on the other hand translate text contain a gist of the original text meaning but syntactic word order may or may not change and the approach results to dynamic equivalence of the text been translated (Benson *et al.,* 2016).

CHALLENGES OF MACHINE TRANSLATION

Translation of a language is not a simple task. There are several challenges to be dealt with when translating one language to another. The various challenges are as follows:

1. Lexical ambiguity

Lexical ambiguity is one of the main challenges in translating one language to another. The words in the source language can have more than one meaning as shown in Table 1. Also a group of words or a complete sentence can have more than one meaning as shown in Table 2. For a machine to understand and translate accurately the lexical ambiguity needs to be resolved and it is a great challenge to translate the language correctly.

Table 1: Word with different meaning.

English	Mizo
Read a **Book**	Lehkhabu chhiar
Book the flight ticket	Thlawhna ticket hauh rawh

Table 2: Sentence with more than one meaning.

Sentence	Meaning
I saw **bats**	1.**Bats** are animals which can fly
	2. Multiple cricket **bats**

2. Differing word order

Two languages may have different word order or different structures of word. For example, English language used a SVO structure (subject-verb-object) whereas Mizo language used OSV structure (object-subject-verb) as shown in Table 3.

Table 3: Differing word order.

Language	Sentence	Structure
English	I eat rice	SVO
Mizo	Chaw ka ei	OSV

3. Pronoun resolution

The pronoun is a substitution of a noun or noun phrase. It can refer to either the subject or the object of the sentence as shown in Table 4. So depending upon the sentence used the pronoun needs to be resolved which is very challenging task for machine translation.

To build a translation machine it requires having good linguistics knowledge to the language. However not only it requires grammar, linguistics and vocabulary it also requires the knowledge gathered from the past experienced.

Table 4: Pronoun resolution.

Sentence	Pronoun (it) resolution
The computer outputs the data, **it** is fast	**It** refers to the computer
The computer outputs the data, **it** is stored in ascii	**It** refers to the data

4. Approaches for machine translation

A machine translation system is broadly classified into three approaches. Rule based machine translation; Corpus based machine translation and Hybrid machine translation. The Rule based machine translation is further classified into direct translation, transfer based translation and Interlingua translation. The Corpus based machine translation is also further classified into statistical machine translation and example based machine translation. The Hybrid machine translation is also further classified into ruled based machine translation guided hybrid and statistical machine translation guided hybrid. The different approaches of machine translation have been summarized in Figure 1.

Translation of one language to another by a machine translation requires analysis of the source language and generates an output or target language by using anyone of the machine translation approaches.

RULE-BASED MACHINE TRANSLATION

The rule-based machine translation is also known as knowledge driven approach. This approach is the first approach developed in the field of machine translation and it is based on linguistic information. The translation system consists of a collection of grammar rules, a lexicon and software programs to process the rules (Antony, 2013; Benson *et al.,* 2013). It produces more predictable output for grammar since it deals with syntactic, semantic and morphological analysis in both source language and target language. Building ruled based machine translation is expensive as all the rules of the language need to be applied and there is a requirement of huge linguistic knowledge. But once it is built it can be deeply analysed at syntax and semantic level.

There are three different types of ruled based machine translation. They are direct translation, transfer based translation and Interlingua translation. The levels of analysis of the three types are shown through vauquois triangle in Figure 2 (Dorr *et al.,* 2004; Benson *et al.*, 2016).

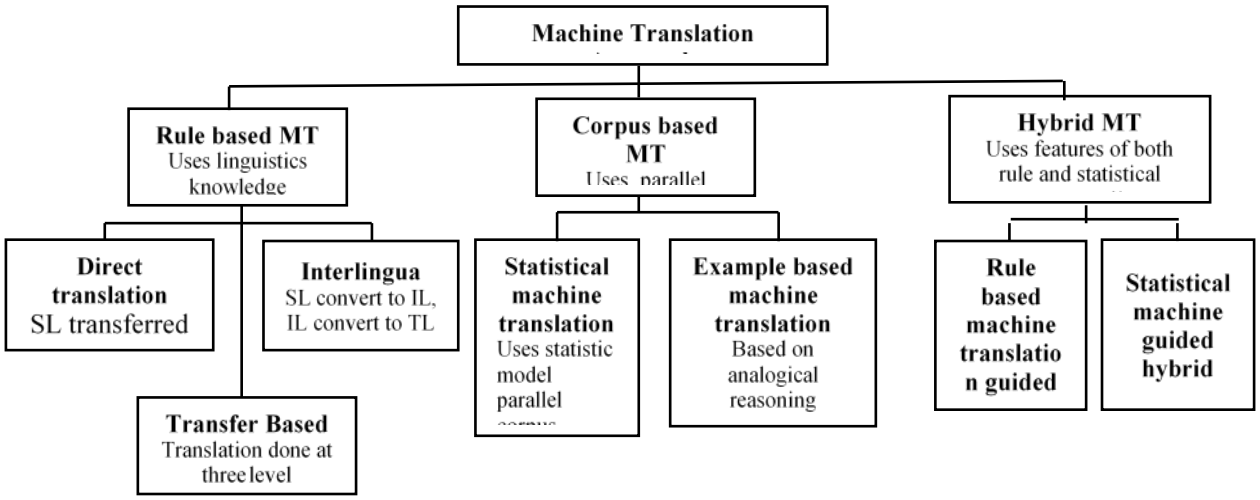

Figure 1: Different approaches of machine translation.

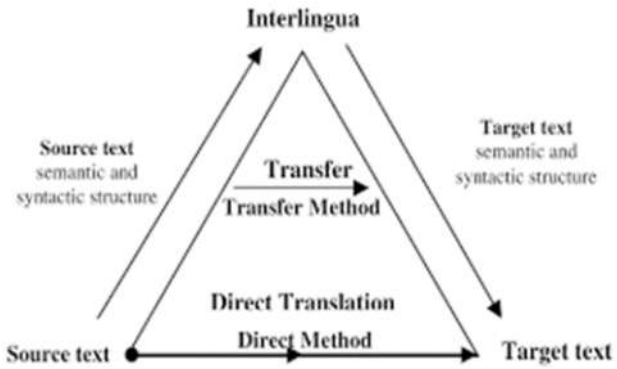

Figure 2: Vauquois triangle – Different methods of rule based machine translation.

1. Direct translation

Direct translation is a word by word translation approach. It directly translates the source language to the target language. It is unidirectional bilingual machine translation. It requires huge amount of morphological analysis but only a little syntax and semantic analysis is required.

2. Transfer-based translation

A transfer based translation involves three stages; analysis, transfer and generation (Sindhu, 2014). In analysis stage, the source language is analyse and converts it into syntactic representation of source language. In transfer stage, it transfers the syntactic representation of source language to a syntactic representation of target language. In generation stage, the target language is generated using morphological analyser.

3. Interlingua translation

Interlingua is a combination of two Latin words Inter and Lingua which means intermediary and Language respectively (Benson *et al.,* 2016). The source language is transformed into intermediate language then the intermediate language is transformed into target language. There is no language pair involves; therefore, it can be used in multilingual machine translation.

CORPUS-BASED MACHINE TRANSLATION

Corpus based machine translation is the most widely used areas in machine translation as it has a high level of accuracy as compared to the other approach (Tripathi and Sarkhel, 2010). It uses bilingual parallel corpora. It requires a large amount of bilingual content in the source language and the target language. These parallel data are used by the machine translation system for acquiring translation knowledge. The corpus based machine translation is further classified into two sub approached: statistical machine translation and example based machine translation.

1. Statistical machine translation

This approach treats the translation as a mathematical reasoning problem; it uses the statistical model built by analysis of bilingual corpus. (Sindhu, 2014). The statis-

tical machine translation consists of three models (Benson *et al*., 2016): language model, translation model and decoder model. The language model gives possible translation for each word or phrase in the input sentence with a probability assigned to each translation P(T). The translation model compute the conditional probability of target sentences by giving the source sentence P(T/S). The decoder model search for the best translation possible P(S,T) by maximizing the product of two probabilities, the language model and translation model as in the equation:

$$P(S,T)=argmax\ P(t)*P(T/S)$$

2. Example-based machine translation

The translation system uses the corpora to find analogous examples between the source language and the target language. It is also called memory based translation. It uses a point to point mapping with a similarity measures such as word, syntactic or semantic similarity to identify the approximately matching sentence. The translation system is categorised into two modules: retrieval module and adaption module (Sandeep, 2015). For a given input sentences, the retrieval module retrieves the similar sentences and it translation from the corpus. From the retrieval module the adaption module finds out the part of translation that can be reused. If the input sentence and the retrieval sentence match, then the correct translated output is given. If it does not match exactly, the relevant match correspond to the source language is used instead.

HYBRID MACHINE TRANSLATION

The hybrid approach uses a combination of rule based approach and statistical based approach (Benson *et al.,* 2016). Rule based approach has a high accuracy as it deeply analyse at syntax and semantic level but it is very expensive as it requires a huge linguistic rules. On the other hand, the statistical based approach is less expensive as it uses the mathematical reasoning problem but needs huge corpora which are a not available for low resourced languages. So, in a Hybrid machine translation the drawback of both the approaches were excluded to give a high efficiency (Sanjay, 2010). They are further classified into two approaches: Rule based machine translation guided hybrid and Statistical machine guided hybrid.

1. Rule-based machine translation-guided hybrid

The translation is preform using rule based engine. A corpus is introduced to reduce expensive development of linguistics rules. Statistical models are used to correct or adjust the output from the rules engine. It is also known as statistical smoothing and automatic post editing.

2. Statistical machine translation-guided hybrid

Rules are adapted in the corpus at the pre-processing stage to guide the statistical engine. Rules are also used at the post processing stage or at the core model of the system to normalize the performance (Costa-Jussa *et al.*, 2016). These approaches avoid the needs of creating set of linguistic rules by extracting those rules from the training corpus. The drawback is still the same as normal statistical machine translation that the accuracy of the translation depends solely of the similarity of the input test to the text of the training corpus. (Sandeep and Chang, 2015).

PERFORMANCE METRICS FOR EVALUATION

There are several evaluation methods which have been used for measuring the performance of the machine translation output. The evaluation of machine translation contains both manual and automatic evaluation methods. Some of the automatic evaluation metrics are as follows:

Word error rate (WER)

It works at the word level. WER can be obtained by using editing distance between both the sentences (Vidal, 1997). WER is the percentage of word that is to be changed in the translation to produce the desired sentence. However, the drawback is the dependency of the reference sentence. There can be multiple correct translation of sentence but the metric considers only the reference sentence to be correct.

Sentence Error Rate (SER)

It is the error metric of machine translation that measures the number of changes required to match the reference sentence.

BLEU Score

BLEU stands for bilingual evaluation understudy. It is one of the first metrics to claim a high correlation with human judgement of quality. The BLEU score is calculated in terms of sentences. It measures how many word sequences in the sentence under evaluation match the word sequences of some reference sentence. It also deals with the penalty for translation with sentences having a significantly high differs in length comparing to the reference translation.

The METEOR metric

The metric is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. It also has several features like stemming and synonymy which are not present in other metrics. It is design to produce a good correlation at the sentence or segment level

CONCLUSION

In this paper we have reviewed various machine translation approaches, the challenges faced in translation system and the performance metrics for evaluation. The survey shows that there are several approaches to cater the different needs of translation. However, with different approaches the translation performance of a machine is quite low in terms of fluency, fidelity, post edit and precision in comparison to human translation. Therefore, we conclude that there are still vast areas to improve the system to utilise a fully automatic translation without the need for human evaluation.

REFERENCES

Antony, P.J. (2013). Machine translation approaches and survey for Indian languages. *International Journal of computational Linguistics and Chinese Language Processing*. 18(1): 47-78.

Benson, K., Lawrence, M., Wanjiku, N. (2016). A review on machine translation approaches. *Indonesian Journal of Electrical Engineering and Computer Science* 1: 182-190.

Chang, J.S., Su, K.S. (1997). Corpus based statistics oriented machine translation researches in Taiwan. *AMTA.* 165-173.

Cheragui, M. A. (2012). Theoretical overview of machine translation. *Proceedings ICWIT.*

Costa-Jussa, M.R., Jose, A.R., Fonollosa (2015). Latest trends in hybrid machine translation and its applications. *Computer Speech and Language* 32(1): 3-10.

Dorr, B.J., Eduard, H., Lori, L. (2004). Machine translation: interlingual methods. In: *Encyclopedia of Language and Linguistics 2nd edition,* p. 939.

Harjinder, K., Vijay, L. (2013). A survey of machine translation approaches. *International Journal of Science, Engineering and Technology Research (IJSETR)* 2 Issue 3.

Mohamed, H.Z., Shafeen, N. (2017). A brief study of challenges in machine translation. *International Journal of Computer Science Issues*, 14 issue 2.

Sandeep, S., Vineet, S. (2015). A survey of machine translation techniques and systems for indian languages. *IEEE International Conference on Computational Intelligence and Communication Technology.*

Sanjay, K. D., Pramod, P. S. (2010). Machine translation system in Indian perspectives. *Journal of Computer Science*. 6, 1111-1116.

Sindhu, D.V., Sagar, B.M., Rajashekar, M. (2014). Survey on machine translation and its approaches. *International Journal of Advanced Research in Computer and Communication Engineering*. 3, Issue 6.

Tripathi, S., Sarkhel, K. (2010). Approaches to machine translation. *Annals of Library and information studies* 57: 388-393.

Vidal, E. (1997). Finite state speech to speech translation. In: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* Munich, Germany.