A Hybrid Translator: From Malayalam to English

Anisree P G[1], Radhika K T[2]

[1]*PG Student, Dept. of Computer Science and Engineering, MEA Engineering College, Kerala, India*
[2]*Asst. Professor, Dept. of Computer Science and Engineering, MEA Engineering College, Kerala, India*

---***---

Abstract - *Machine translation is a process of translation from one language to another. The language from which we are translating is called the source language and the language to which we are translating is called the target language. So machine translation can be defined as a process of translation from a source language to a target language. Here the source language is chosen as Malayalam and the target language is English. Statistical based approach towards machine translation is commonly used for the translation purpose because it is superior to any other approaches. But due to the peculiarity nature of Malayalam it is not a good practice to develop a translator for this language by using the statistical method. Hence we are integrating some of the rules for the translation purpose which helps the system to understand how to translate from Malayalam in an easier way. Here we are proposing a translator which can translate from Malayalam to English by combining a rule based and a statistical based approache, hence it is a hybrid translator.*

Key Words: Machine Translation, Statistical Based Machine Translator, Rule based Machine Translator, Hybrid Translator, Sandhi Splitter.

1.INTRODUCTION

Machine Translation (MT), perhaps the earliest NLP application, is the translation of text units from one language to another, using computers. It is one of the most interesting and the hardest problem in the field of NLP. The two challenges in machine translation are adequacy and fluency [1]. The former is to develop a system that adequately represents the ideas expressed in the source language into the target language. The latter is to represent those ideas grammatically. India is a multilingual country, i.e., many of the states have their own native language and only 5% of the population knows English [2]. So, it must require a translator which is capable of translating from their native language to English for efficient communication and knowledge sharing. The common approaches to machine translation are the rule based approach and corpus based approach. In the rule based approach, a large number of rules are necessary to capture the phenomena of natural language. These rules transfer the grammatical structure of the source language into target language. As the number of rules increases, the system becomes very complicated. Formulation of a large number of rules is a tedious process and require years of effort and linguistic analysis. In the second approach, large parallel and monolingual corpora are used as source of knowledge. This approach can be further divided into statistical approaches and example based approach. Statistical machine translation (SMT) is superior to rule based and example based systems in that they do not require human interpenetration and can build a translation system in an unsupervised manner directly from the training data. Rule based systems are language dependent and require careful analysis of source and target languages. With the rapid proliferation of internet and increasing availability of data, SMT is currently the most popular and prevalent paradigm.

For an SMT system, a parallel corpus consisting of source and target language sentences and a monolingual corpus consisting of target language sentences are required. The SMT system is trained on these large quantities of parallel data and monolingual data. The statistical model learns the translation parameters from the corpus and performs the translation. SMT takes place in three phases, namely language modeling, translation modeling and decoding as shown in Fig 1. The language model determines the probability of the target language T which helps in achieving the fluency in the target language and choosing the right word in the translated language. It is generally denoted as P (T). The translation model, on the other hand helps to compute the conditional probability of the target language T given the source language S generally denoted as P (T|S). Finally, in the decoding phase, the maximum probability of product of both the language model and the translation model is computed which gives the statistically most likely probable sentence in the target language.

$$P\ (S,\ T) = argmax\ P\ (T)\ P\ (T|S)$$

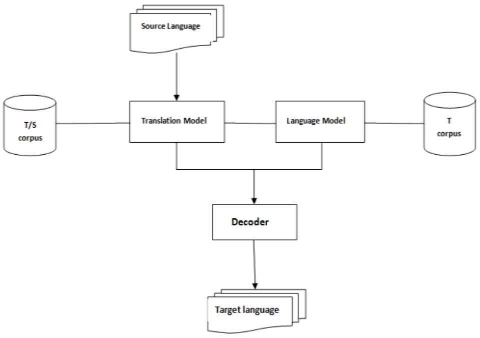

Fig 1: Statistical Machine Translation working

We are hereby proposing a system which can translate from Malayalam, a Dravidian language to English by a hybrid model. Malayalam is an agglutinative language where words of different syntactic categories are combined to form a single word. Formation of new words by combining a noun and a noun, noun and adjective, verb and noun, adverb and verb, adjective and noun etc. None of the training data can identify such compound words, so it leads to a meaningless translation. In order to resolve this problem we are proposing a compound splitting module before the input going to the SMT. Here in this sense it is a hybrid based machine translator.

The rest of this paper is organized as follows. Section 2 covers the related works. Section 3 gives an overview about the proposed system and section 4 concludes the paper.

2. RELATED WORKS

In India, researchers have been pursuing on machine translation since 1980. Different machine translation systems has been developed and is using in different parts of India. Out of all the 22 official languages some of the languages is not showing a good result in machine translation and not a tremendous research is focused on these languages. Malayalam is spoken by 38 million people in the south east state Kerala and is one such language. The peculiarity nature of Malayalam is the major reason for this.

Different approaches are taken by a number of researchers in Kerala for the translation between Malayalam and English. Rule based, statistical based, syntactic based approaches are some of the commonly used approaches. A rule based translator is developed for English to Malayalam in 2009 [3]. The core process is done with bilingual dictionary of English-Malayalam pair and rules for converting source language structure to the target language structure. There are mainly two types of rules are used by them, one is transfer link rules and the other one is morphological rules. Where the transfer link rules are used for obtaining target structure and morphology rules are used for assigning morphological features. Syntactic based approach is used for the translation from English to Malayalam itself in 2012 [4]. For the translation purpose this system uses a bilingual English-Malayalam dictionary and a morphology generator. General rules are identified for certain sentences and these rules are used for translating new sentences. A statistical based translator is used for the same pair in 2010 [5]. POS tagging, suffix separation, stop word elimination and order conversion are used in their work. A hybrid based machine translator is developed for English to Malayalam in 2013 [6], here it is named as hybrid in the sense that it extent the statistical approach with a translation memory, where the translation memory is used as a cache which store the recent translation and hence avoid redundant translations. A corpus based and a transfer based translators are developed for the translation from Malayalam to English in 2014 [7] and 2012 [8] respectively. In corpus based approach the main idea used is of reusing the already translated examples. The transfer based approach is a rule based method which uses the rule of Malayalam language for the translation purpose. Various splitters are used in Malayalam for identifying the compound words and splitting them. Rule based [9], statistical based [10] and hybrid [11] approaches are also used in splitters. Here also hybrid approach gives high result.

3. PROPOSED SYSTEM

Malayalam is a language in which the tense, mood, aspect, negation attachments etc. all are fully concentrated on verbs [12]. By taking into consideration all the verbs in Malayalam it requires a total of 25 lakhs of parallel data for training purpose in order to develop a good translator. But using this much amount of data is not an easy task. At the same time Malayalam is also an agglutinative language where words of different syntactic categories are combined to form a single word. Even a full sentence may exist as a single string in Malayalam. None of the training data can identify such compound words, so it leads to a meaningless translation. A hybrid based approach towards machine translation is the method used in the construction of this Malayalam to English translator. Which is a combination of statistical based machine translation and any of the rules from the rule based machine translation. For the rule based part, a compound word splitter is used. The splitter identifies the compound word and makes it individual words and pass it to SMT part of translator. First the input source sentence is fed to the compound split module, where the system get identify the words that want to be separate. Then the splitter split the word into individual words. Then the regenerated input sentence is move on to the statistical translator module. Where the translation module and the language module along with the decoder perform the translation in a phrase based approach. Then the output English sentence is generated. The major compound words in Malayalam are formed with Koottaksharam, Yakaaram, Makaaram, Vyanjanam, Just Separate etc. So identifying these factors and splitting this into two or more independent words are generally performed in Sandhi Splitter.

3. CONCLUSIONS

Malayalam is spoken by 38 million people in the south east state Kerala. In Malayalam the information regarding tense, aspect, mood, negative attachment etc. are fully concentrated on verbs. If we are using a machine translator which uses an SMT approach, then in order to overcome this property it requires around 25 lakhs of sentences for training. But this is a tedious task and hence we can integrate some of the rules of Malayalam along with the training data. Also Malayalam has a property of compound word generation, i.e., new words can be formed by combining one or more words together. Both the two properties can leads to an inefficient translation. In order to resolve this, hereby we are proposing a compound word splitter along with an unsupervised learning from training corpus. And we are hoping that the proposal is going to be a good translator for Malayalam to English.

REFERENCES

[1] Nadeem Jadoon Khan, “Statistical Machine Translation of Indian Languages: A Survey”.

[2] Sudip Naskar, Jadavpur University and Sivaji Bandyopadhyay, Jadavpur University, “Use of machine translation in India: Current status,” 10 august 2015, 465 – 470 6 pages.

[3] R. Remya, S. Remya, R. Remya, and K. P. Soman, “Rule base machine translation from English to Malayalam," International conference on Advances in Computing, Control and Telecommunication Technologies, 2009.

[4] T. Anitha and Sumam, “Syntactic based machine translation from English to Malayalam," International Conference on Data Science and Engineering (ICDSE), 2012.

[5] S. Mary, K. Sheena, and G. Santhosh, “English to Malayalam translation: A statistical approach," Proceeding of the 1st Amritha ACM-W Celebration on Women in Computing in India, 2010.

[6] B. Nithya and J. Shibily, “A hybrid approach to English to Malayalam machine translation," International Journal of Computer Applications, vol. 81, no. 8, November 2013.

[7] E. S. Anju and K. V. Manoj, “Malayalam to English machine translation : An ebmt system," IOSR Journal of Engineering, vol. 4, pp. 18-23, January 2014.

[8] R. Latha, D. Peter, and R. P. Ravindran, “Design and development of a Malayalam to English translator - a transfer based approach," International Journal of Computational Linguistics (IJCL), vol. 3, 2012.

[9] V. V. Devadath, J. K. Litton, S. Dipti, and Vasudevan, “A sandhi splitter for Malayalam."

[10] D. Divya, K. T. Radhika, R. R. Rajeev, and P. C. Reghu, “Hybrid sandhi splitter for Malayalam using unicode," Proceedings of National Seminar on Relevence of Malayalam in Information Technology, 2012.

[11] R. Latha and S. David, “Development of a rule based learning system for splitting compound words in Malayalam language," Recent Advances in Intelligent Computatioanl Systems (RAICS), pp. 751-755, 2011.

[12] V. Jayan and V. K. Bhadran, “Difficulties in processing Malayalam verbs for statistical machine translation," International Journal of Artificial Intelligence and Application (IJAIA), vol. 6, no. 3, May 2015.