ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

Malayalam to English Machine Translation: A Hybrid Approach

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

P.G. Student, Department of Computer Engineering, MEA Engineering College, Perinthalmanna, Kerala, India[1]

Assistant Professor, Department of Computer Engineering, MEA Engineering College, Perinthalmanna, Kerala, India[2]

ABSTRACT: Machine Translation is a process of translation of text or speech from one natural language to another. Statistical based approach towards MT is the dominant approach of machine translation, but the outcome is highly depends upon the capacity of training data that we are used. When we are integrating some of the rules along with this training data the translation could yield high result. Such a hybrid approach is taken into account in this work. Malayalam is an agglutinative language in which words of different category combines with one or more words to form new word. Out of all the Dravidian languages Malayalam is rich with this property and the compound word formation leads to an inefficient translation. In order to resolve this issue a compound word splitter is integrated with the statistical part of Machine Translation. So, before the input sentence is move on to the statistical part, it gets modified as individual words by splitting the compound words. This translator is implemented and then tested by using BLEU score and it is found that the system get high performance with the new hybrid approach.

KEYWORDS: Natural Language Processing, Machine Translation, Statistical Machine Translation, Hybrid Machine Translation, Compound word Splitter.

I. INTRODUCTION

The language is a medium for communication that conveys the ideas in human brain and there are more than 5000 languages are for communication around the world. Knowing all these languages for effective means of communication is not an easy task for human beings. So it requires the help of an automated system which could convert between different languages. Natural Language Processing (NLP) is an area of research that explores how computers can be used for the conversion purpose. Machine Translation, speech recognition, artificial intelligence etc. are effective areas of Natural Language Processing.

Machine Translation (MT) is a process of translation of text or speech from one natural language to another, using computers. It is one of the interesting and the hardest problem in the field of NLP. India is a multilingual country, i.e., many of the states have their own native language and only 5 percent of the population knows to speak in English. So, it must require a translator which is capable of translating from their native language to English and vice versa for efficient communication and knowledge sharing. The research scenario in India is relatively young and various translators are developed for Indian language to English, English to Indian languages and Indian language to Indian language. Machine Translation has its application in various domains like health, education, information technology, business and various government agencies [1].

Due to the lack of proficiency in English, information access is very difficult for the common people. This can be avoided with machine translators. The machine translation system can be integrated with various Natural Language Processing applications such as information retrieval, summarization etc.

The two challenges in Machine Translation are adequacy and fluency [2]. 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. 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. As the number of rules increases, the system becomes very complicated. In the second approach, large parallel and monolingual corpora are used as source of knowledge. With the rapid proliferation of internet and

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

increasing availability of data, Statistical Machine Translation (SMT) a form of corpus based approach 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.

Inflection, derivation, compounding and concatenation are the major morphological behavior in Malayalam. Eighty percentage of the vocabulary in Malayalam is derived from Sanskrit [3]. According to grammar rules, Malayalam words are divided into two main types: vaachaka and dyothaka. Dhyothaka denotes relation and it has no individual meaning. Vaachaka is split into 3 types noun, verb and qualifying words [4].

In NLP Malayalam is still in the nascent stage because this piece of technology is not much popular among the scholars and computing people. 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. are possible in Malayalam. None of the training data can identify such compound words, so it leads to a meaningless translation. In order to resolve this problem we use a compound splitting module. The output of this splitter is then move on to SMT part. This hybrid based approach is a novel approach towards Malayalam to English machine translation.

The rest of this paper is organized as follows. Section II covers the related works. Section III gives an overview about the proposed system and section IV covers implementation. Section V gives the experimental result and section VI concludes the paper.

II. RELATED WORK

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 are 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 [5]. 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 [6]. 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 [7]. 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 [8], 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 [9] and 2012 [10] 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 [11], statistical based [12] and hybrid [13] approaches are also used in splitters. Here also hybrid approach gives high result.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

III. PROPOSED SYSTEM

Malayalam is a language in which the tense, mood, aspect, negation attachments etc. all are fully concentrated on verbs [4]. 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 by using a statistical based method. 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 words and makes it individual words and passes 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. The regenerated input sentence is then move on to the statistical translator module. Where the translation module identifies the phrases and converts it into English then the language module gives fluency to the sentence thus created. At last the decoder output the highly probable sentence.

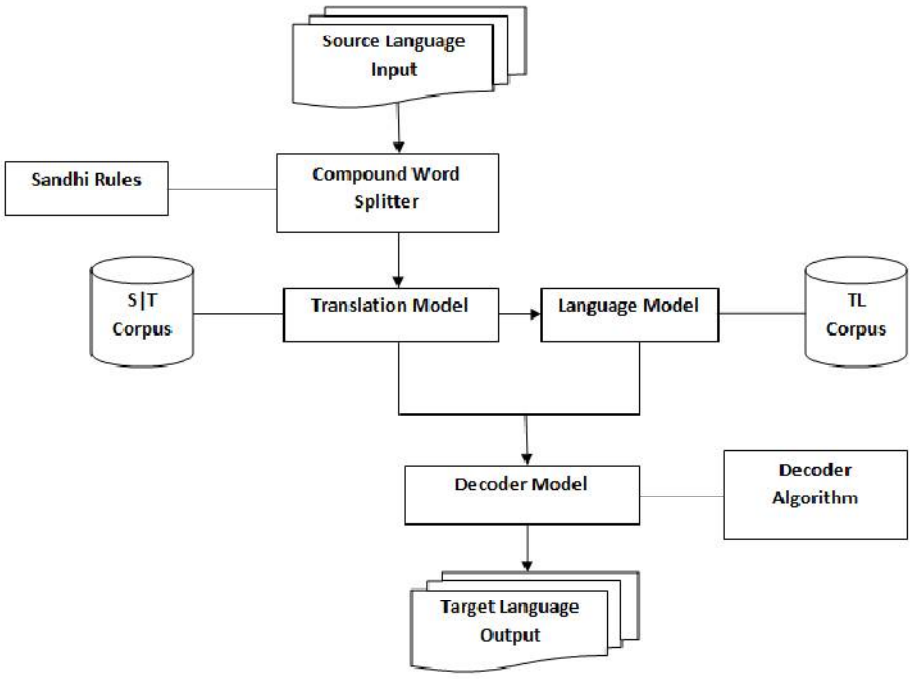

Fig 1: Module Design

IV. IMPLEMENTATION

First we identify the major rules that reqires to form a compound word in Malayalam and it is found to be that the words are formed with koottaksharam, yakaaram, makaaram, vyanjanam etc. So identifying these factors and splitting this into two or more independent words are generally performed in Compound Word Splitter module. Different weights are given to each rules and according to the weight the type of compound word generation are identified. Then by using the splitting parameter the process of separation is carried out. The output of splitter is then given to the statistical translator part where it identifies the phrases and converts it into English sentence. To train a translation system we take parallel data which is aligned at the sentence level. Before training the corpus it gets tokenized,

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

truecased and cleaned. Tokenization add spaces between the words and punctuations in a sentence, truecasing converts all the initial letter of the first word to capital letters and cleaning is a process for removing extra long sentences. The language model is used to ensure fluent output, so it is built with the target language i.e., English in this case. In order to estimate a Language Model, we first prepare the training corpus. 3-gram (trigram) language modeling is used in this project, by considering the 3 words at a glance. ARPA format output is created for the language modeling which help the evaluation more quick. After that a binary formatted output is generated for speedy extraction. Finally we train the translation model by running the word alignment, phrase extraction and scoring, create lexicalized reordering tables and create the configuration file. The last stage is tuning, and it requires a small amount of parallel data, separate from the training data, this particular corpus is tokenized, truecased and cleaned for the tuning purpose. Then we tune the system for better performance.

V. EXPERIMENTAL RESULT

The BLEU evaluation is comparatively an inexpensive method of evaluation and it also offers speed and is language independent. It compares the n-grams of the candidate with the n-grams of the reference translation and counts the number of matches. This evaluation method is used to evaluate the accuracy of the system that we have developed. The output of both the baseline translator and the hybrid based translator are evaluated with the help of same data set and the result is found to be very high for the hybrid translator. For the baseline system the dataset is used with a total of 25% of compound words, the system could not find the translation for the compound words and hence the resulted output is incorrect for all the compound word sentences. The evaluation result for the statistical part and the hybrid part are as follows. Where the parameters are the following

- BLEU score
- Brevity Penalty (BP) = min(1, number of output words / number of reference words)
- hyp _len = number of output words in the data set
- ref _len = number of reference words in the data set

Table 1: Evaluation Result

PARAMETERS	STATISTICAL MACHINE TRANSLATOR	HYBRID MACHNE TRANSLATOR
BLEU Score	59.61	72.91
BP	0.897	0.963
Ratio	0.902	0.964
hyp_len	717	766
ref_len	795	795

VI. CONCLUSION

In order to resolve the inefficiency in translation due to the peculiarity nature of Malayalam, hereby we are using a hybrid based translation by adding a compound word splitter along with an unsupervised learning from training corpus. The system is generated and the performance of both the normal statistical based system and the newly developed hybrid based system are evaluated by using BLEU score. From the evaluation it is very much clear that the hybrid based system is good for translation. For any machine translation the hybrid based approach could yield high result. This is very much evident from the developed system. The hybrid based approach is a combination of corpus and rule based approach. Hence to an extent the accuracy of the system is depends upon the training data. Only 500 sentences were used in this system. If we are adding more sentences the system can give good result. So the accuracy can be

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

increased by adding more and more training data. Similarly adding more rule sets can also increase the performance of the system. The research is going to continue in this way.

REFERENCES

[1] Sugata Sanyal and Rajdeep Borgohain**, "**Machine Translation Systems in India".
[2] Nadeem Jadoon Khan, "Statistical Machine Translation of Indian Languages: A Survey".
[3] "Malayalam Literary Survey", Kerala Sahitya Akademi, Volume 27, 2005.
[4] 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.
[5] Remya Rajan, Remya Sivan, Remya Ravindran, and K. P. Soman, "Rule base machine translation from English to Malayalam," International conference on Advances in Computing, Control and Telecommunication Technologies, 2009.
[6] Anitha T. Nair and Sumam Mary Idicula, "Syntactic based machine translation from English to Malayalam," International Conference on Data Science and Engineering (ICDSE), 2012.
[7] Mary Priya Sebastian, Sheena Kurian K, and G. Santhosh Kumar, "English to Malayalam translation: A statistical approach," Proceeding of the 1st Amritha ACM-W Celebration on Women in Computing in India, 2010.
[8] B. Nithya and Shibily Joseph, "A hybrid approach to English to Malayalam machine translation," International Journal of Computer Applications, vol. 81, no. 8, November 2013.
[9] E. S. Anju and K. V. Manoj Kumar, "Malayalam to English machine translation: An ebmt system," IOSR Journal of Engineering, vol. 4, pp. 18-23, January 2014.
[10] Latha R Nair, David Peter, and Renjith P. Ravindran, "Design and development of a Malayalam to English translator - a transfer based approach," International Journal of Computational Linguistics (IJCL), vol. 3, 2012.
[11] V. V. Devadath, Litton J. Kurishinkel, Dipti Misra Sharma, and Vasudeva Varma, "A sandhi splitter for Malayalam."
[12] Divya Das, K. T. Radhika, R. R. Rajeev, and P. C. Reghu Raj, "Hybrid sandhi splitter for Malayalam using unicode," Proceedings of National Seminar on Relevance of Malayalam in Information Technology, 2012.
[13] Latha R Nair and S. David Peter, "Development of a rule based learning system for splitting compound words in Malayalam language," Recent Advances in Intelligent Computational Systems (RAICS), pp. 751-755, 2011.