Techniques to Improve the word alignments in Statistical Machine Translation from English to Malayalam

Mary Priya Sebastian
Department of Computer Science and Engineering,
Rajagiri School of Engg. & Technology,
Rajagiri Valley,Kakkanad,
Kochi,Kerala, India
marypriyas@gmail.com

Sheena Kurian K
Department of Computer Science and Engineering,
KMEA Engineering College, Edathala,
Aluva,Kerala, India
sheenakuriank@gmail.com

G. Santhosh Kumar
Department of Computer Science,
Cochin University of Science and Technology, Kerala, India
san@cusat.ac.in

Abstract—**In Statistical Machine Translation from English to Malayalam, an unseen English sentence is translated into its equivalent Malayalam translation using statistical models like translation model, language model and a decoder. A parallel corpus of English-Malayalam is used in the training phase. Word to word alignments has to be set up among the sentence pairs of the source and target language before subjecting them for training. This paper is deals with the techniques which can be adopted for improving the alignment model of SMT. Incorporating the parts of speech information into the bilingual corpus has eliminated many of the insignificant alignments. Also identifying the name entities and cognates present in the sentence pairs has proved to be advantageous while setting up the alignments. Moreover, reduction of the unwanted alignments has brought in better training results. Experiments conducted on a sample corpus have generated reasonably good Malayalam translations and the results are verified with F measure, BLEU and WER evaluation metrics.**

***Keywords*—alignment, training, machine translation, English Malayalam translation,**

I Introduction

Statistical Machine Translation (SMT) is one of the upcoming applications in the field of Natural Language Processing that treats language translation as a machine learning problem. As discussed in [1], during the training phase of SMT a learning algorithm is applied to huge volumes of previously translated text termed as parallel corpus. By examining these samples, the SMT system automatically translates previously unseen sentences. In [2] a methodology of statistical machine translator that translates a sentence in English into Malayalam is discussed.

Since English and Malayalam belong to two different language families, various issues are encountered when English is translated into Malayalam using SMT. As a part of resolving the issues, the basic underlying structure of the SMT is modified to an extent. The training results are improved when the Malayalam corpus is subjected to certain pre-processing techniques like suffix separation and stop word elimination. Various handcrafted rules based on 'sandhi' rules in Malayalam are designed for the suffix separation process and these rules are classified based on the Malayalam syllable preceding the suffix in the inflected form of the word. A technique to remove the insignificant alignments from the bilingual corpus using a PoS Tagger is also employed. While decoding a new unseen English sentence, the structural disparity that exists between the English Malayalam pair is fixed by applying order conversion rules. The statistical output of the decoder is further furnished with the missing suffixes by applying mending rules.

For a statistical machine translator from English to Malayalam, the appropriate Malayalam translation for the given English sentence is acquired using the statistical models. A very large corpus of translated sentences of English and Malayalam is required to achieve this goal. In the current scenario there exist only very few numbers of such large corpora and the sad part is that they do not come with word to word alignments. However, there are techniques by which the large corpora can be trained to obtain word to word alignments from the non-aligned sentence pairs [3].

In training the SMT, sentence pairs in the parallel corpus are examined and alignment vectors are set to identify the alignments that exist between the word pairs. A number of alignments can be present between any pair of sentence. As the size of the corpus and the length of the sentence vary, the process of building the alignment vectors for sentence pairs becomes a challenging task. Moreover in training, representing the alignments using alignment vectors takes up major part of the working memory.

It is observed that many of the alignments in a sentence pair are insignificant and carry little meaning. By removing these insignificant word alignments from the sentence pairs the quality of training can be enhanced. Moreover word alignments can be refined by aligning words based on the word categories like named entities and cognates. In this paper we discuss about an alignment model with morphological knowledge which enables to filter the irrelevant alignment pairs in the corpus. Furthermore a discussion on certain techniques that aids in improving the word to word alignments that exist between the English-Malayalam sentences is done. Also we examine the changes occurring in the training process when morphological knowledge as well as the category information is introduced into the corpus.

The rest of this paper is organized as follows: The related work done in this area is presented in Section 2. In Section 3, an overview of SMT from English into Malayalam is discussed. The alignment model and the training technique adopted in SMT are explained in Section 4. Section 5 presents the details about the techniques adopted in improving the word alignments. Some observations and results obtained from the experiments conducted on a sample English/Malayalam corpus is discussed in Section 6. Finally, the work is concluded in Section 7.

II Related Work

Experiments on statistical machine translation were carried out among many foreign languages and English. For SMT, development of statistical models as well as resources for training is needed. Due to the scarcity of full fledged bilingual corpus, works in this area remain almost stagnant. Therefore accomplishment of an inclusive SMT system for Indian languages still remains a goal to be achieved. A work on English to Hindi statistical machine translation [4] which uses a simple and computationally inexpensive idea for incorporating morphological information into the SMT framework has been reported. Another work on English to Tamil statistical machine translation is also reported in [5]. The morphological richness and complex nature of the Malayalam language account for the very few attempts made to translate texts from other languages into Malayalam. A pure statistical machine translation from/in the Malayalam language is yet to be published. The ideas integrated from the similar works in machine translation have been the source of motivation and the inputs gathered from the related methodologies has facilitated in outlining the framework of the proposed SMT from English to Malayalam.

III Overview of English Malayalam SMT

In SMT from English to Malayalam, a bigram estimator [6] is employed as the language model to check the fluency of Malayalam. For the translation model, which assigns probabilities to English-Malayalam sentence pairs, IBM Hidden Markov Model (Model 1) training technique [7] is chosen. A variation of Beam Search method [8] is used by the decoder to work with the statistical models. In the training process the translations of a Malayalam word is determined by finding the translation probability of an English word for a given Malayalam word. The method used for finding the translation probability estimate in SMT is the EM algorithm discussed in [3].

As discussed in [9], Malayalam language is enriched with enormous suffixes and the words appear mostly with multiple suffixes The Suffix separator is employed to extract roots from its suffixes. By incorporating a lexical database(a collection of noun roots and verb roots), a suffix database(suffixes in Malayalam) and a ‘sandhi’ rule generator, the functioning of the suffix separator is further enhanced, resulting in a Malayalam corpus comprising only of root words and suffixes. Certain Malayalam words, which are not in root form, still have equivalent meaningful translations in English. The word 'അവന്റെ'(avante) is semantically equivalent to the word ‘his’ in English. Even though ‘അവന്റെ'(avante) has a suffix appended, it need not be suffix separated.

The Malayalam corpus after suffix separation will contain many suffixes extracted from root words that have no meaningful word translation in English. Most of them are the suffixes of nouns and verbs in Malayalam. Since these words are useless in the translation process, they are not included in the corpus. The deletion of these stop words will bring down the complexity of the training process as well as improve the quality of the results expected from it. Similarly stops words in English language are also identified and are eliminated from the corpus before subjecting it to training.

On obtaining the estimates for the translation parameter from the training phase, an unseen English sentence can be translated by the decoder by applying Bayes rule [6]. In the decoder different syntactic tags are used to denote the syntactic category of English words. Since English and Malayalam belong to two different language families, they totally differ in their subject verb order. Order conversion rules are framed to reorder English according to the sentence structure and the word group order of Malayalam. To obtain Statistically Correct Malayalam (SCM), the end product of the decoder, the order converted English sentence is split into phrases and a phrase translation table with different options of Malayalam translations is developed. Various hypotheses are created by choosing translation options and the best translation is determined by extending the hypotheses and picking the one with maximum score. Since SMT is trained with root words in Malayalam, the statistical outcome of the decoder lacks the required suffixes in the words generated. Hence SCM fails to convey the complete meaning depicted in a sentence. This undesirable result has been set right by applying various mending rules which helps in converting SCM into Grammatically Correct Malayalam.

IV Training the Parrallel Corpus

In the training phase the corpus used is a sentence aligned one where a sentence in Malayalam is synchronized with its equivalent English translation. The aligned sentence pairs are subjected to training mechanism which in turn leads to the calculation of translation probability of English words. The translation probability is the parameter that clearly depicts the relationship between a word in Malayalam and its English translation. It also shows how closely a Malayalam word is associated with an English word in the corpus. The translation probability for all the English words in the corpus is estimated. This results in generating a collection of translation options in English with different probability values for each Malayalam word. Of these translation options the one with the highest translation probability is selected as the word to word translation of the Malayalam word.

The corpus with aligned sentence pairs needs to be pre-processed to obtain the word to word alignments. In each sentence pair all the possible alignments of a Malayalam word is identified. The nature of the alignment truly depends on the characteristics of the language chosen. Since Malayalam with suffix separation holds a one to one mapping with words in

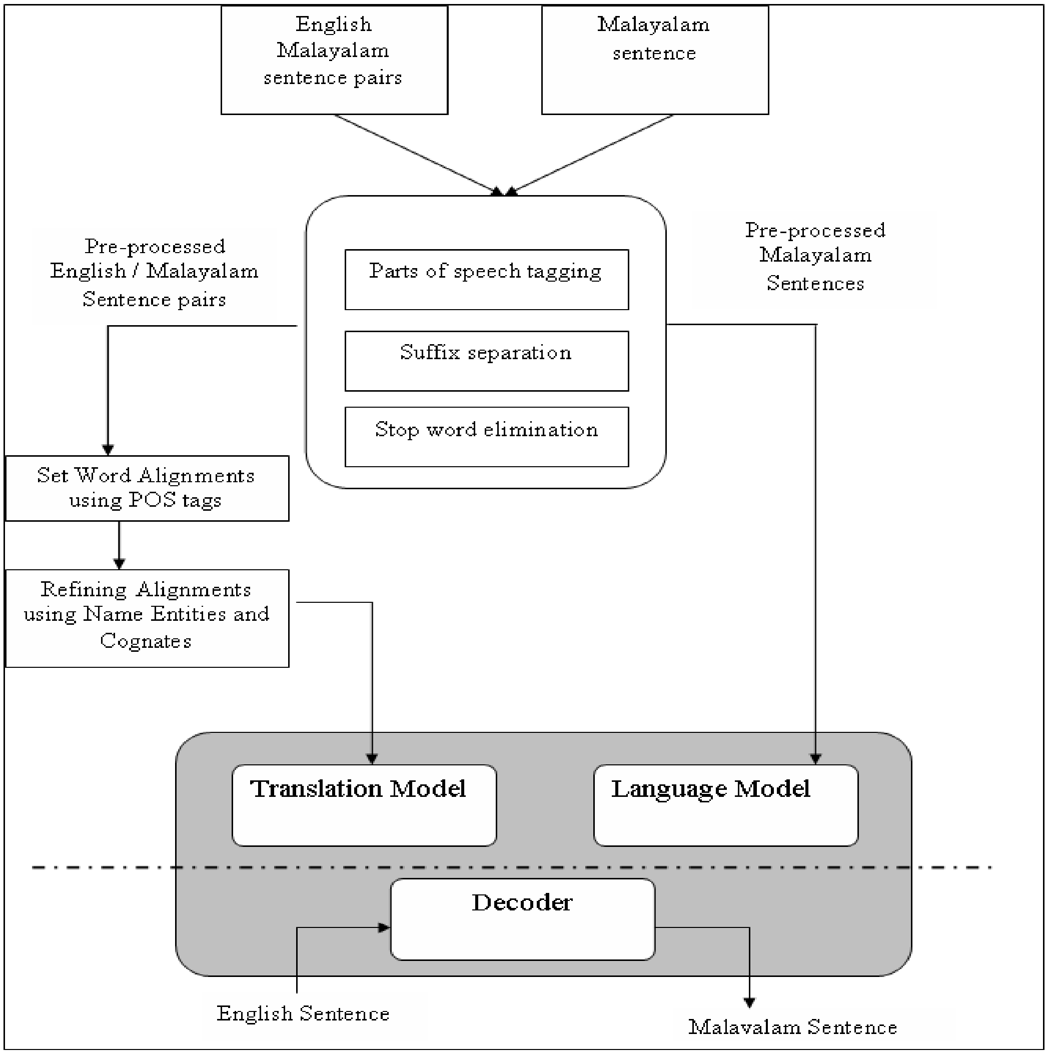
English Malayalam sentence pairs
Malayalam sentence
Pre-processed English / Malayalam Sentence pairs
Parts of speech tagging
Suffix separation
Stop word elimination
Pre-processed Malayalam Sentences
Set Word Alignments using POS tags
Refining Alignments using Name Entities and Cognates
Translation Model
Language Model
Decoder
English Sentence
Malayalam Sentence

Figure1. Training in SMT from English to Malayalam

the English sentence, only one to one alignment vectors are taken into consideration during training.

For a sentence pair all the possible alignments have to be considered in the training process. Depending upon the word count of the Malayalam sentence, the number of alignments will vary. The number of alignments generated for any sentence pair is equal to the factorial of the number of words in the sentence. The alignment vector for the sentence pair is obtained by placing the position of the aligned Malayalam word in place of the corresponding English word in the sentence pair. The length of the alignment vector of an English sentence will depend on its word count.

The EM algorithm defines a method of estimating the parameter values of translation for IBM model 1. By this algorithm there is equal chance for a Malayalam word to get aligned with any English word in the corpus. Therefore initially the translation probability of all English words is set to a uniform value. Suppose there is N number of English words in the corpus, the probability of all Malayalam words to get mapped to an English word is 1/N. To start with the training process this value is set as the Initial Fractional Count (IFC) of the translation probability. Alignment weight for a sentence pair is calculated by observing the IFC of all the word pairs present in the alignment vector. The Alignment Probability (AP) of all the sentences is calculated by multiplying the individual alignment weight of each word pair in the sentence pair. The calculated alignment probability of the sentence pairs is then normalized to get Normalized Alignment Probability (NAP).

Fractional count for a word pair can be revised from the normalized alignment probabilities. A word in Malayalam may be aligned to a same English word in many sentences. Therefore when the fractional count of a word pair is recomputed, all sentence pairs are analyzed to check whether it holds that particular word pair. If it is present in any pair of sentence, the alignment probabilities of the alignment vectors holding that word pair are added up to obtain the Revised Fractional Count (RFC). By normalizing the revised fractional counts (NFC) new values of translation probability is obtained.

Moreover the new values thus achieved are better since they take into account the correlation data in the parallel corpus. Equipped with these better parameter values, new alignment probabilities for the sentence pairs are again computed. From these values a set of even-more-revised fractional counts for word pairs is obtained. By repeating this process over and over, fractional count converges to extremely better values. The translational probability of the English word given a Malayalam word is calculated to determine the best translation of a Malayalam word. It is achieved by comparing the translation probabilities of all English words associated with it and choosing the one with highest probability value

V INTEGRATING MORPHOLOGICAL INFORMATION INTO PARALLEL CORPUS

On introducing the training method described earlier into the parallel corpus, a large number of alignment vectors are obtained. Out of it a major share belongs to the group of insignificant alignments. Presence of these unwanted alignments complicates the training mechanism. Most of these alignments hold little meaning and is useless in building up the fractional count. To get rid of the alignments which have no significance and to reduce the burden of calculating the fractional count and alignment probabilities for every alignment of sentence pairs, the morphological information is incorporated into the corpora. The bilingual corpus is tagged and then subjected to training. Tagging is done by considering the parts of speech entities of a sentence.

By tagging the corpus extra meaning is embedded into each word which definitely helps in the formation of reasonably good alignments. The structure of the Malayalam sentence is analyzed and the different Parts of Speech (PoS) categories are identified. In a sentence there may be many words belonging to the same PoS category. After the tagging process, words that don't have an exact translation in Malayalam may be deleted to improve the efficiency of the training phase. The English sentence is tagged in the same manner and paired with its tagged Malayalam translation. The word to word alignments are found only for the words that belong to the same PoS category of both languages. There is little chance for the words belonging to two different categories to be translations of each other and hence they need not be aligned. This helps to bring down the total number of alignments to a greater extent.

Without tagging, when all the words in a sentence is considered, the number of alignments (NA) generated is equal to the factorial of its word count and is shown as

$$NA = \text{factorial}(Ws) . \qquad (1)$$

where Ws is the number of words in the sentence. The same corpus when tagged produces number of alignments less than factorial (Ws). Categorizing the sentence leads to grouping of words that belong to the same tag. The number of alignments for words belonging to same category is factorial (Wc) where Wc is the number of words in a category and the total number of alignments of a sentence formed by tagging (NAT) is given by

$$NAT = \prod_{i=1}^{m} \text{factorial}(Wci) . \qquad (2)$$

alignment vectors, where m is the number of PoS categories in a sentence pair. The Insignificant Alignments (IA) eliminated can be represented as the difference between Equation 1 and 2 and is given below:

$$IA = \text{factorial}(Ws) - \prod_{i=1}^{m} \text{factorial}(Wci) . \qquad (3)$$

V INCORPORATING NAME ENTITIES AND COGNATES

The insignificant alignments are further reduced by identifying the name entities and cognates present in the English Malayalam sentence pair. Name Entity identification[12] is a process in which the atomic elements in a text is located and classified into different predefined categories. The categories may include name of persons, organizations, places, time units, monetary units, quantities etc. Since entity identification is a subtask of information extraction, it is implemented using local pattern-matching techniques. A Name Entity Database (NED) that contains a large set of name entities is employed for the dictionary look up. In linguistics, cognates are defined as two words having a common etymological origin. Cognates in two different languages are words that are pronounced in a similar way or with a minor change**.** For example the word car in English and the word കാർ in Malayalam are similarly pronounced. Transliteration similarity between the word pairs can be considered for identifying such words.

$$IA = \text{factorial}(Ws) - \prod_{i=1}^{M} \text{factorial}(Wci) -(N_{NE} + N_{C}) \qquad (4)$$

where N_{NE} is the number of word pairs aligned with name entity and N_{C} is the number of word pairs aligned with cognates.

VI OBSERVATIONS AND RESULTS ACHIEVED

The sample corpus used for training includes 250 sentences with 1800 words. The experimental Malayalam corpus is built based on www.mathrubhumi.com, a news site providing local news on Kerala. For better training results, the corpus selected should be adequate enough to represent all the characteristics of the languages. Also, the strength and correctness of the corpus is a necessity to achieve the desired output. The process of extending the English/Malayalam corpus is still continuing.

Evaluation metrics proposed in [10] were applied on sentences present in the training set and on totally unseen sentences. Three reference corpora were used for testing. The

Type of sentence	Technique	Evaluation Metric		
		WER	F measure	BLEU
Sentences in training set	Baseline + with suffix	0.3313	0. 57	0.48
	Baseline + suffix separation	0.1863	0. 78	0.69
Unseen sentences	Baseline + with suffix	0.6083	0. 26	0.22
	Baseline + suffix separation	0.4444	0.44	0.38

Table 1. Summary of evaluation results

summary of the results are shown in Table 1. The criteria used for the evaluation are discussed below.

Word Error Rate (WER): This metric is based on the minimum edit distance between the target sentence and the sentences in the reference set.

F measure: A "maximum matching" technique where subsets of co-occurrences in the target and reference text are counted so that no token is counted twice.

BLEU: This metric is based on counting the number of n-grams matches between the target and reference sentence

Owing to the fact that English and Malayalam belong to two different language family, various issues are encountered when English is translated into Malayalam using a SMT. The issues start off with the scarcity in the availability of English/Malayalam translations required for the training of SMT. The functioning of SMT completely rely on the parallel corpus. Less number of these resources in the electronic form adds on to the difficulty of implementing SMT.

Moreover in the bilingual translations available, a one to one correspondence between the words in the sentence pair is hard to find. The reason behind this occurrence is solely the peculiarity of Malayalam language. Due to the agglutinative nature of Malayalam [11], a linguist when asked to translate sentences into Malayalam, have a wide range of options to apply. The words “daily life” can be translated as “nithyenayulla jeevitham” or “nithyajeevitham” according to the will of the linguist. Even though the two translations share the same meaning, there is a difference of latter being a single word. Scope of occurrence of such translations cannot be eliminated and hence certain sentence pairs may lack a one to one mapping between its word pair.

In training, the entire corpus is examined and statistical methods are adopted to extract the appropriate meaning for a word. An alignment model is defined in training which sets all the possible alignments between a sentence pair. The amount of memory required to hold these alignments is a problem which cannot be overlooked. Lengthy sentences worsen the situation since word count of the sentence is the prime factor in determining alignments. In the pre-processing phase suffixes are separated from the Malayalam words in the corpus. Suffix separation results in further increase of sentence length which in turn will increase the number of word alignments.

Certain suffix doesn’t have a correct translation when they stand alone in the corpus. Hence setting alignments for such suffixes doesn't have any significance in training. Eliminating them from the corpus before the training phase will bring down the word count of the sentences and thereby the number of alignments too.

Similarly many insignificant alignments are avoided by scrutinizing the structure of the sentence pair. Close observation reveals the fact that many words belonging to different categories are mapped together when alignment vectors are figured out. The English word that forms the subject of a sentence need not be aligned with the ‘kriya’ (verb) in Malayalam. Likewise verbs in English have little chance to get associated with words that forms ‘karthavu’(subject) and ‘karmam’ (object) in Malayalam.

Insignificant alignments take up time and space in training. Therefore the parallel corpus is strengthened with more information so that only the relevant alignment is included in calculating translation probabilities. It is observed that when the corpus is linked with a parts of speech tagger many irrelevant alignments are eliminated. It has been observed that the rate of generating alignment vectors have fallen down to a remarkably low value as shown by Equation 2. Here the alignment vectors are directly proportional to the number of words in the PoS category and not to the number of words in the sentence pair. Also, the training technique is further enhanced by aligning words based on the name entities and cognates identified. This method has brought down the insignificant alignments further as stated by Equation 4.

By enhancing the training technique, it is observed that the translation probabilities calculated from the corpus shows better statistical values. The end product of the training phase is obtained much faster. In the iterative process of finding the best translation, it takes less number of rounds to complete the training process.

VII CONCLUSION

Alignment model used in SMT from English to Malayalam results in many insignificant alignments which brings down the quality of translations obtained from the training phase. Techniques to improve the word to word alignments between the English-Malayalam sentence pairs are discussed in this paper. Using the parts of speech tags as an additional knowledge source, the parallel corpus is enriched to contain more information for selecting the correct word translation for a Malayalam word. The alignment model with category tags is useful in diminishing the set of alignments for each sentence pair and thereby simplifying the complexity of the training phase. The name entities and cognates located in the sentence pairs also have an important role in reducing the insignificant alignments. These techniques helps to improve the quality of word translations obtained for Malayalam words from the parallel corpus. The performance of the SMT is evaluated using WER, F measure and BLEU metrics and the results prove that the translations are of fairly good quality.

REFERENCES

[1] Lopez, A.: Statistical Machine Translation. ACM Computing Survey, 40, 3, Article 8 (2008)

[2] Mary Priya Sebastian, Sheena Kurian K. and G. Santhosh Kumar: A Framework of Statistical Machine Translator from English to Malayalam. In: Proceedings of Fourth International Conference on Information Processing, Bangalore, India ,2010.

[3] Knight, K.: A statistical MT tutorial work book. Unpublished,http://www.cisp.jhu.edu/ws99/projects/mt/wkbk.rtf(1999)

[4] Ananthakrishnan, R, Hegde, J, Bhattacharyya, P., Shah, R., Sasikumar, M. Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation. In the Proceedings of International Joint Conference on NLP(IJCNLP08), Hyderabad, 2008.

[5] Badodekar, S. A survey of Translation Resources,Services and Tools for Indian Languages. In the Proceedings of the Language Engineering Conference, Hyderabad, 2002.

[6] Brown P F, Pietra S A D, Pietra V J D, Mercer R L. The mathematics of statistical machine translation: Parameter estimation**.** *Comput. Linguistics, 19(2)*,pp263–31, 1993

[7] Brown, P.F., Pietra, S.A.D., Pietra, V.J.D., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S.: A Statistical Approach to Machine Translation. In: Computational Linguistics, 16(2), pages 79–85, (1990)

[8] Koehn P. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA), 2004.

[9] Sumam M I, Peter S D. A Morphological Processor for Malayalam Language. South Asia Research, Volume27 (2). pp 173-186, 2008.

[10] Stent A, Marge M, Singhai M. Evaluating evaluation methods for generation in the presence of variation. In Proceedings of CICLing 2005, Mexico City, pp 341-351, 2005.

[11] Rajeev R R, Elizabeth Sherly, A suffix Stripping based Morph Analyser for Malayalam Language, Proceedings of 20th Kerala Science Congress, p 482-484, 28- 31, (2008)

[12] Niraj Aswani and Robert Gaizauskas. 2005. Aligning words in English-Hindi parallel corpora. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts* (ParaText '05). Association for Computational Linguistics, Stroudsburg, PA, USA, 115-118.

[13] Mary Priya Sebastian, Sheena Kurian K. and G. Santhosh Kumar: Alignment Model and Training Technique in SMT from English to Malayalam. In: Proceedings of International Conference on Contemporary Computing, Noida,India,2010.

[14] A R Rajaraja Varma. Keralapanineeyam, Eight edition, DC books, 2006.

[15] Rajeev R.R, Jisha P Jayan, and Elizabeth Sherly, " Parts of Speech Tagger for Malayalam", IJCSIT International Journal of Computer Science and Information Technology, Vol 2, No.2, December 2009, pp 209-213

[16] Sanchis G, S´nchez J A. Vocabulary Extension via PoS Information for SMT. In the Proceedings of the NAACL ,2006.

[17] Rajeev R R, Rajendran N, Elizabeth Sherly Morph Analyser for Malayalam Language-Suffix Stripping Based Approach Dravidian Studies, a quarterly research journal, Vol V -VI, Nos3-4;1-2 pages 61-72, Oct 2008