English to Malayalam Translation: A Statistical Approach

Mary Priya Sebastian
PG Student
Department of Computer Science,
Cochin University of Science and
Technology, Kerala, India
+91 9446128116
marypriyas@gmail.com

Sheena Kurian K
PG Student
Department of Computer Science,
Cochin University of Science and
Technology, Kerala, India
+91 9746798804
sheenakuriank@
gmail.com

G. Santhosh Kumar
Lecturer
Department of Computer Science,
Cochin University of Science and
Technology, Kerala, India
+91 9447305879
san@cusat.ac.in

ABSTRACT

This paper underlines a methodology for translating text from English into the Dravidian language, Malayalam using statistical models. By using a monolingual Malayalam corpus and a bilingual English/Malayalam corpus in the training phase, the machine automatically generates Malayalam translations of English sentences. This paper also discusses a technique to improve the alignment model by incorporating the parts of speech information into the bilingual corpus. Removing the insignificant alignments from the sentence pairs by this approach has ensured better training results. Pre-processing techniques like suffix separation from the Malayalam corpus and stop word elimination from the bilingual corpus also proved to be effective in training. Various handcrafted rules designed for the suffix separation process which can be used as a guideline in implementing suffix separation in Malayalam language are also presented in this paper. The structural difference between the English Malayalam pair is resolved in the decoder by applying the order conversion rules. 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, English Malayalam Translation, PoS Tagging, Statistical Machine Translation, Suffix Separation

1. INTRODUCTION

Statistical Machine Translation (SMT), which treats language translation as a machine learning problem is one of the upcoming application in the field of Natural Language Processing. In SMT as discussed in [8], a learning algorithm is applied to huge volumes of previously translated text usually termed as parallel corpus. By examining these samples, the system automatically translates previously unseen sentences. The statistical machine translator proposed in this paper translates a sentence in English into Malayalam. 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.

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.

The rest of this paper is organized as follows: The related work done in this research area is presented in Section 2. In Section 3 a brief overview of the proposed architecture of the English Malayalam SMT is done. Section 4 highlights the method of incorporating morphological knowledge into the corpus and the details of modified alignment model. The role of suffix separation in machine translation and details about the classification of the suffix separation rules is discussed in Section 5. Observations and results achieved from the experiments conducted on a sample English/Malayalam corpus is discussed in Section 6. Finally, the work is concluded in Section 7.

2. RELATED WORKS

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 [1] 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 [2]. The ideas integrated from these works 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.

3. OVERVIEW OF ENGLISH MALAYALAM SMT

The overall architecture of the English Malayalam SMT is given in Figure 1. In SMT, a bigram estimator [4] 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 Model 1 training technique [3] is chosen. A variation of Beam Search method [7] is used by the decoder to work with the statistical models.

3.1 Training Phase

In the training process the translations of a Malayalam word is determined by finding the translation probability of a English word for a given Malayalam word. The corpus that we consider is a sentence aligned corpus 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. To make the process of training less complex, different features are added in the training technique. The details are given in the following section.

3.1.1 Setting up the Corpora

Huge volumes of translated text of English and Malayalam are required to build the SMT. Malayalam corpus can be built from online Malayalam newspapers and magazines. Since it is hard to find the equivalent line by line English translation, building English/Malayalam corpus is a difficult task. 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. A linguist when asked to translate sentences into Malayalam, have a wide range of options to apply. The words “daily life” is translated as "നിത്യേനയുള്ള ജീവിതം" (nithyenayulla jeevitham) or “നിത്യജീവിതം" (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 one to one mapping between its word pair.

3.1.2 PoS tagging the bilingual corpus

The method used for finding the translation probability estimate in SMT is the EM algorithm [6] but a large number of insignificant alignments are generated when this method is adopted. Hence an alignment model with PoS tagging [10] is used in diminishing the set of alignments for each sentence pair. Here, category tags of the same type are used in tagging the words of both languages.

3.*1.3 Suffix separation from Malayalam corpus*

As discussed in [12], 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. Examples of suffixes separated from Malayalam corpus is given in Table 1. 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. A list of such words is given in Table 2.

Table 1. List of Suffixes

Malayalam suffixes						
ഉടെ	ഓടെ	ഉള്ള	ആയ	ആയി	എ	ഉം
ഓ	ആൽ	ഏ	അല്ല	ഇല്ല	എന്ന്	ഉക

Table 2. Split Exception Category

Split Exception Category		
നിന്റെ	അവിടെ	നമ്മുടെ
അവനെ	വാതിൽ	കരിക്ക്
മക്കൾ	കൌൺസില്	ഊഞ്ഞാൽ
കുമാർ	കോഴിക്കോട്	മിഠായി

3.1.4 Stop word elimination from the bilingual corpus

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 may not be 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. Some of the stop words in Malayalam and English are shown in Table 3.

Table 3. Stop words in Malayalam and English

Stop words in Malayalam and English					
ഏ	എ	ഉം	by	as	of
ഉ	ആൽ	ഓ	and	at	off

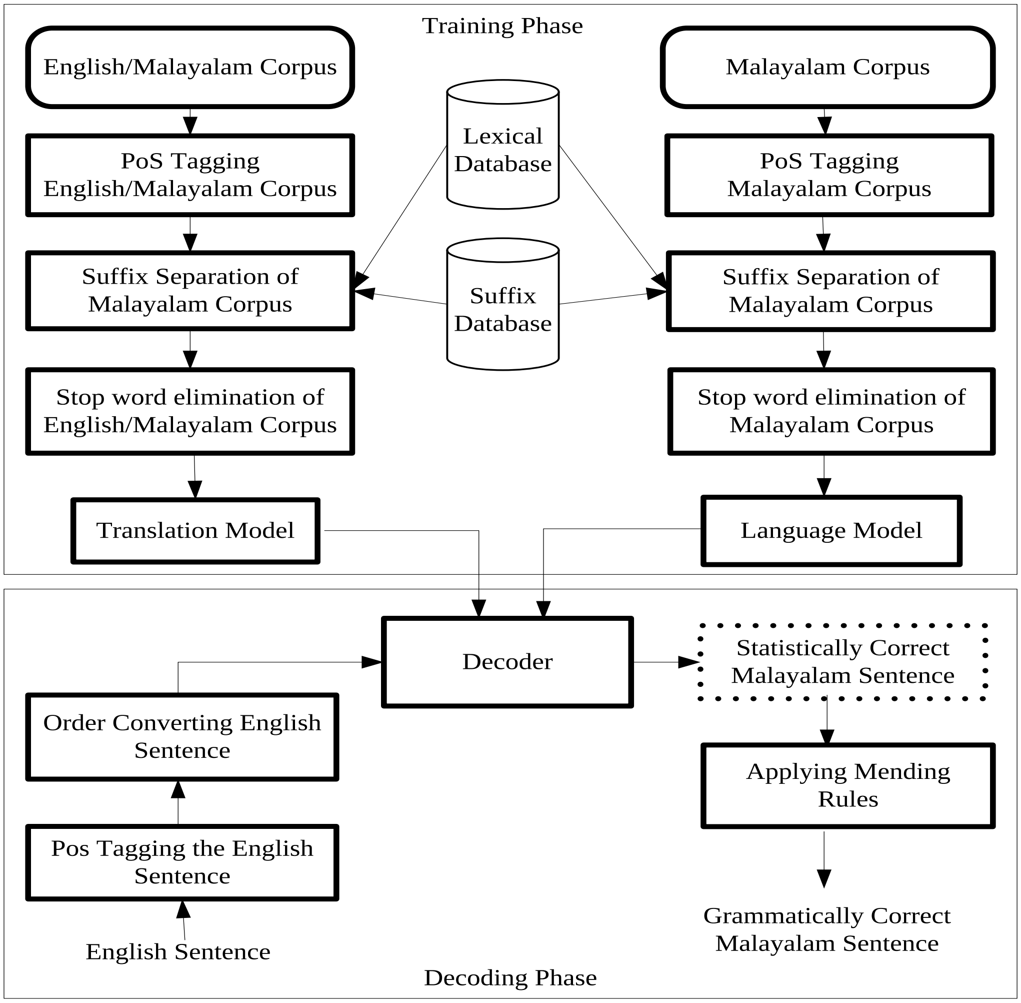
Training Phase
English/Malayalam Corpus
PoS Tagging English/Malayalam Corpus
Suffix Separation of Malayalam Corpus
Stop word elimination of English/Malayalam Corpus
Translation Model
Lexical Database
Suffix Database
Malayalam Corpus
PoS Tagging Malayalam Corpus
Suffix Separation of Malayalam Corpus
Stop word elimination of Malayalam Corpus
Language Model
Decoder
Statistically Correct Malayalam Sentence
Order Converting English Sentence
Pos Tagging the English Sentence
English Sentence
Applying Mending Rules
Grammatically Correct Malayalam Sentence
Decoding Phase

Figure 1. Overall architecture of English Malayalam SMT

3.2 Decoding Phase

Once the estimates for the translation parameter are obtained from training, an unseen English sentence can be translated by the decoder by applying Bayes rule [4]. The outcome of the decoder is influenced by introducing some additional components and their details are discussed in the coming section.

3.2.1 Tagging the English sentence

In the decoder different syntactic tags are used to denote the syntactic category of English words. For example the sentence ‘He has a car ‘is tagged as He/PRP has/VBZ a/DT car/NN[1] using the POS tagger.

3.2.2 Order conversion

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. For example, ‘he ran quickly’ may be translated as 'അവൻ വേഗത്തില് ഓടി' (avan vegathil odi) since adverbs are always placed before verbs in Malayalam sentences. Some samples of order conversion rules are listed in Table 4.

Table 4: Order Conversion Rule Examples

English sentence & its structure	Order conversion rule
He/PRP is/VBZ a/DT boy/NN	PRP DT NN VBZ
Our/PRP$ car/NN is/VBZ white/JJ	PRP$ NN JJ VBZ
I/FW gave/VBD them/PRP sweets/NN	FW PRP NN VBD

[1] PRP, VBZ, DT and NN denote the personal pronoun, the verb in the present tense, the determiner and the noun categories respectively

3.2.3 Generating Statistically Correct Malayalam (SCM)

To obtain 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.

3.2.4 Generating Grammatically Correct Malayalam (GCM)

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 GCM. For the sentence ‘I saw her’, 'ഞാൻ അവൾ കണ്ടു'(njan aval kandu) is the statistical output though 'ഞാൻ അവളെ കണ്ടു' (njan avale kandu)is its correct translation. Mending Rule Applier rejoins the suffix and the word 'അവൾ' (aval) becomes 'അവളെ'(avale). For the sentence having the structure ‘I/PRP saw/VBD her/PRP$’, the mending rule is given as *If (PRP VBD PRP$) append the suffix ‘ഏ ‘to the translation of PRP$.* Equipped with a decoder having a complete set of hand crafted rules, capable of handling all types of sentence structures, better results are obtained.

3. ALIGNMENT MODEL

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 varies. The number of alignments generated for any sentence pair is equal to the factorial of the number of words in the sentence. 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 increases the number of word alignments. Also the training method based on EM Algorithm generates a large number of insignificant alignments. An example of an unwanted alignment is shown in Figure 2.

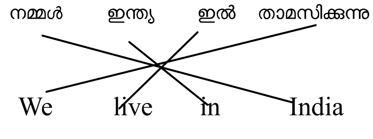

Figure 2. Insignificant alignments

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. 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 are considered, the number of alignments generated is equal to the factorial of its word count. By tagging, the number of categories present in a sentence is identified. There may be many words with the same tag in a sentence. The number of alignments for words belonging to same category is factorial of the number of words in a category. The insignificant alignments, $I_{A,}$ eliminated is represented as

$$I_A = \text{factorial}(Ws) - \prod_{i=1}^{m} \text{factorial}(Wc_i) \qquad (1)$$

where Ws is the number of words in the sentence, Wc is the number of words in a category and m is the number of PoS categories in a sentence pair.

4. SUFFIX SEPARATION RULES

The requirement of a preprocessing step in the training phase is solely attributable to the peculiar nature of Malayalam language. The inflected form of a word in Malayalam can have various suffixes appended to its root. This characteristic of Malayalam language reduces the probability of a word in the corpus to be present in its root form. For example the word ‘ഇന്ത്യ ‘appears in the corpus in different forms, for example ഇന്ത്യയുടെ, ഇന്ത്യക്ക്, ഇന്ത്യയോടു etc.

On setting the word to word alignments in the English Malayalam sentence pair, the inflected Malayalam word is aligned with the English word ‘India’. These alignments add on to the total alignment weight and in effect reduce the probability rate of the translation of ‘India’ as ‘ഇന്ത്യ‘. For the word ‘India’, the word translation chosen by the decoder is one among the inflected forms and it may not be an apt one that fits the context of the newly translated sentence. To resolve this issue, suffix separation is brought into picture and the corpus with root words is subjected to training. Suffix separation rules are formed by applying sandhi rules in Malyalam in the reverse direction. A classification of sandhi rules based on whether a word ends with a vowel (swaram) or a consonant (vyanjanam) is discussed in [9].

Table 5. Suffix_keys

Suffix_key	Suffix
◌ിൽ	ഇൽ
◌ാണ്	ആണ്
◌ുള്ള	ഉള്ള
◌ുണ്ട്	ഉണ്ട്
◌േ	ഏ
◌ോട്	ഓട്

Out of this classification, words belonging to Swarasandhi and Vyanjanaswara sandhi are of major concern and splitting up such words have more significance in the training process of SMT from English to Malayalam. To implement suffix separation, the category of suffix to be separated has to be identified. In the example ' അവൾ + ഉടെ = അവളുടെ‘, the suffix ‘ഉടെ ‘is present in an abbreviated form as '◌ുടെ'. These abbreviated forms are the keys to identify the suffixes. A few examples are listed in Table 5.

Table 6. Suffix_labels

Suffix_ label	Suffixes
AA	ആണ്, അല്ലെ, ആൻ
EE	ഇൽ, ഇൻ, ഇല്ല
UU	ഉണ്ട്, ഉള്ള, ഉന്നു
EI	എ, ഏ, എന്ന്
OO	ഓട്ടു, ഓടെ, ഓ

The suffixes are grouped together based on the vowel sound of the start syllable. The suffixes അല്ലെ and ആണ് starts with the same vowel sound 'അ'. Since the vowel sound in these two suffixes is same, the advantage is that a common rule can be applied to this category in the suffix separation process. Various labels are identified for this category by observing the vowel at the beginning of the suffix for example AA for ആണ്, അല്ലെ, ആന്. The categories and the suffix_labels are listed in Table 6.The word structure is thoroughly analyzed to identify the Malayalam syllable preceding the suffix_key in the inflected form of the word (check_letter).With check_letter, suffix_keys and suffix _labels, the suffixes are separated from the roots.

4.1 Classification of Suffix Separation Rules

A few of the rules classified based on the check_letter's are listed in the Table 7. For the check_letter 'x' in any Malayalam word W, the term prev_(x) denotes a substring that starts from the first syllable of W and ends on the syllable preceding x when scanned from the right hand side of W. In the word 'മലയാളമാണ്', prev_(മ) denotes the substring 'മലയാള'.

5. 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.

Table 7. Look up table

Check_letter (CL)	Examples	Suffix separation rules	Function
ള	വാളാണ്	prev_ള + ൾ+ suffix	Can retrieve words ending with ൾ Roots extracted : വാൾ, അവൾ
ന	തേനാണ്	prev_ന + ൻ+ suffix	Can retrieve words ending with ൻ Root extracted : തേൻ
ല	പാലാണ്	prev_ല + ൽ+ suffix	Can retrieve words ending with ൽ Root extracted : പാൽ
റ	കയറാണ്	prev_റ + ർ + suffix	Can retrieve words ending with ർ Root extracted : കയർ
ണ	കടിഞ്ഞാണാണ്	prev_ണ+**ൺ**+ suffix	Can retrieve words ending with **ൺ** Root extracted :കടിഞ്ഞാ**ൺ**
യ	മേശയാണ്, മിഴിയാണ്	prev_യ + suffix	Can retrieve words ending with അ and ഇ sound. Roots extracted : മേശ, മിഴി
വ	മഴുവാണ്	prev_ വ + suffix	Can retrieve words ending with ഉ sound Root extracted : മഴു
ത്ത	പാലത്തില്	prev_ത്ത + ◌ം + suffix	Can retrieve words ending with അം sound. Root extracted : പാലം
മ	പാലമാണ്	prev_മ + ◌ം + suffix	
Consonant likeക സ, ത,etc...	കാതാണ്	prev_ CL +◌് +suffix	Can retrieve words ending with consonants followed by 'chandrakkala' Root extracted : കാത്
Conjunct consonant like ണ്ണ, ല്ല, പ്പ ,ള്ള,ച്ച,ട്ട etc…	കണ്ണാണ്	prev_ CL +◌് +suffix	Can retrieve words ending with conjunct consonant followed by 'chandrakkala'. Roots extracted : കണ്ണ്

Table 8. Summary of evaluation results

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

Evaluation metrics proposed in [11] were applied on sentences present in the training set and on totally unseen sentences. Three reference corpora were used for testing. The summary of the results are shown in Table 8. 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.

Imparting the parts of speech information into the parallel corpus has made it rich with more information which in turn helps in picking up the correct translation for a given Malayalam word. It has reduced the complexity of the alignment model by cutting short the insignificant alignments. Again eliminating the stop words in Malayalam and English corpus before the training phase has brought down the word counts of the sentences and thereby the number of alignments too.

The meaningless alignments have a tendency to consume more space and time thereby increasing the space and time complexity of the training process. It has been observed that the rate of generating alignment vectors have fallen down to a remarkably low value as shown by Equation 1. 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. Utmost care has to be taken while tagging the corpus, since wrong tagging leads to the generation of absurd translations. For the annotation of the corpus with morphological information, we use an in-house parts of speech tagger for Malayalam and the Stanford POS tagger for English.

By enhancing the training technique, it is observed that the translation probabilities calculated from the corpus shows better statistical values of translation probability. 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.

The effect of suffix separation is clearly depicted in Table 8. On evaluating the results of the corpus trained without suffix separation, it was found that the final translation included many number of unwanted insertions which reduced the quality of translation. It is noted that the results of suffix separated corpus is giving better score for WER, F measure and BLEU than the one with suffixes. Even though the translations produced depicts correct meaning of the English sentence, the expected score is not met. This is due to the large number of word substitutions rather than insertions and deletions occurring in the translated sentence when compared to the reference text.

6. CONCLUSION

A frame work to build a machine translation system from English to Malayalam using statistical models is presented. The alignment model with category tags eliminates the insignificant alignments and simplifies the complexity of the training phase in SMT. This technique helps to improve the quality of word translations obtained for Malayalam words from the parallel corpus. To simplify the task of implementing the suffix separator various hand crafted rules are designed to separate the suffixes of Malayalam. The quick look up table that summarizes the classification of the suffix separation rules can be utilized as a guideline to separate suffixes beginning with vowel sounds from any word in the Malayalam language.Also, post editing techniques like order conversion and mending rules for suffix rejoining enhanced the outcome of the decoder. 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. This method can be further extended and employed in translating any language into Malayalam by incorporating the corresponding bilingual corpus along with its order conversion rules.

7. REFERENCES

[1] 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.

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

[3] 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. *Comput. Linguistics, 16(2),* pp 79–85, 1990.

[4] 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

[5] Durgesh, R. Machine Translation in India: A Brief Survey. *In the Proceedings of SCALLA. Conference,* Bangalore, 2001.

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

[7] 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.

[8] Lopez, A.,.Statistical machine translation. *ACM Comput. Surv.,* 40, 3, Article 8, 2008.

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

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

[11] 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.

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