Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 00 (2011) 000–000

Procedia Technology

www.elsevier.com/locate/procedia

2nd International Conference on Communication, Computing & Security

A framework for translating English text into Malayalam using statistical models

Mary Priya Sebastian[a], Sheena Kurian K[b], G. Santhosh Kumar[a,b,]a*

[a]Asst. Professor, Dept. of Computer Science, Rajagiri School of Engg. & Technology, Kochi-682039,Kerala, India
[b]Asst. Professor, Dept. of Computer Science, KMEA College of Engg. & Technology, Kochi-682039,Kerala, India
[a,b] Professor, Dept. of Computer Science, Cochin Unniversity of Science and Technology, Kochi-682039,Kerala, India

Abstract

A methodology for translating text from English into the Dravidian language, Malayalam using statistical models is discussed in this paper. The translator utilizes a monolingual Malayalam corpus and a bilingual English/Malayalam corpus in the training phase and generates automatically the Malayalam translation of an unseen English sentence. Various techniques to improve the alignment model by incorporating the morphological inputs into the bilingual corpus are discussed. 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 producing better alignments. Difficulties in translation process that arise due to the structural difference between the English Malayalam pair is resolved in the decoding phase by applying the order conversion rules. The 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. 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

* Mary Priya Sebatsian. Tel.: +91-484-2427835; fax: +91-484-2426241
E-mail address: marypriya_s@rajagiritech.ac.in

1. Introduction

Statistical Machine Translation (SMT) is one of the upcoming applications in the field of Natural Language Processing and it treats translation as a machine learning problem. In SMT as discussed in [1], a learning algorithm is applied to huge volumes of previously translated text usually termed as parallel corpus. Any numbers of previously unseen sentences are translated by the SMT system automatically after examining the sample corpora. The statistical machine translator proposed in this paper is developed to translate a sentence in English into Malayalam.

Since Malayalam is an agglutinative language, its grammar rules are too complex in nature. In most cases Malayalam words contain a lexical root to which one or more affixes are fitted. Malayalam affixes are commonly suffixes that are derivational or inflectional. The span of agglutination is really long in Malayalam and hence it results in lengthy words with great number of suffixes. In Malayalam the verb comes at the end of the sentence and it follows a typical word order of Subject Object Verb (SOV). However, Malayalam language permits word order to be altered and makes it a relatively word order free language. Due to the morphological richness and intricate nature of Malayalam, very few attempts have been made to translate texts from other languages into Malayalam.

Due to the fact that 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 a certain extent. The training results are improved by subjecting the Malayalam corpus to certain pre-processing techniques like suffix separation and stop word elimination. Various handcrafted rules based on the typical '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. The alignment model is refined by removing the insignificant alignments from the bilingual corpus with the aid of PoS Tags and by incorporating the knowledge of cognates, name entities and the predictable Malayalam words. In 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 by adding the missing suffixes with the help of the mending rules.

1.1. Related Work

SMT based on statistical method was first proposed by IBM in the early nineties [1].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 like a parallel corpus 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 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 rest of this paper is organized as follows: In Section 2 a brief overview of the proposed architecture of the English Malayalam SMT is done. Section 3 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 4. Observations and results achieved from the experiments conducted on a sample English/Malayalam corpus is discussed in Section 5. Finally, the work is concluded in Section 6.

2. Developing the corpora

The first challenging task in building the SMT is the requirement of huge volumes of translated text of English and Malayalam. Huge volumes of translated text of English and Malayalam are required to build the SMT. Malayalam corpus can be built from Malayalam data that is found on the web. Several Malayalam newspapers and magazines of different editions are available online. But the difficulty lies in identifying its equivalent line by line English translation, which is very rarely found in the web. 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. Therefore to build the English/Malayalam translation corpus, the work has to be started from the scratch.

3. Overview of English Malayalam SMT

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 corpus that is considered 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. 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. 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 overall architecture of the English Malayalam SMT is given in Fig 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. To make the process of training less complex, different features are added in the training technique. The decoder is also modified for obtaining better Malayalam sentences by incorporating certain post editing techniques. The details are given in the following section.

3.1. Preprocessing the corpora in the training phase

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. 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.'ഉടെ'(ude), 'ഇല്'(il), 'കള്'(kal) etc are examples of suffixes separated from Malayalam corpus.

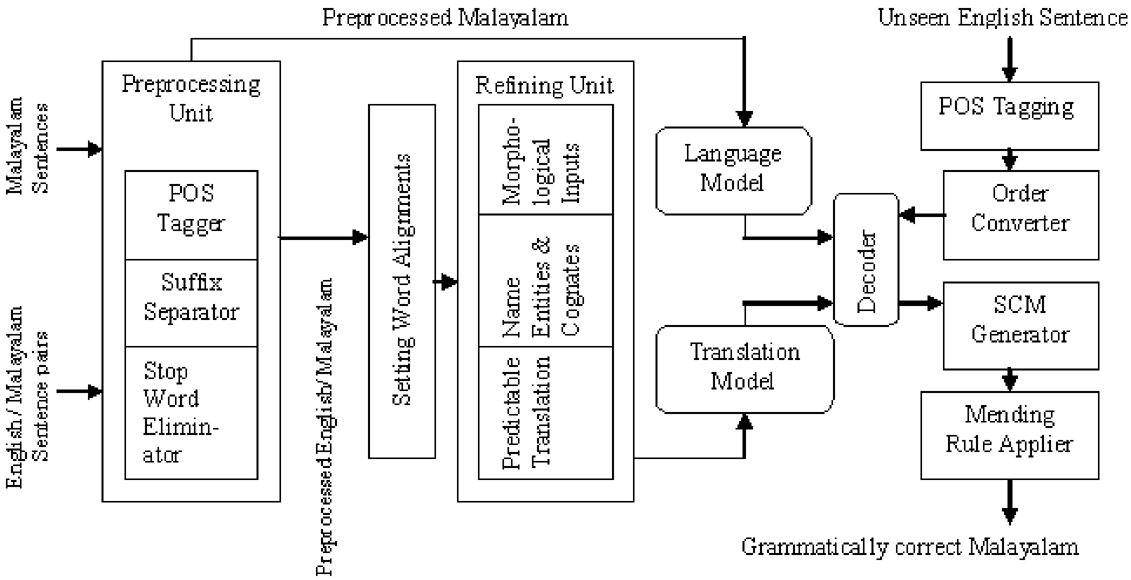

Fig. 1. Architecture of English Malayalam SMT

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. Suffixes separated from the Malayalam corpus and certain words in English like 'of', 'by' etc are useless in the translation process and therefore they are eliminated from the corpus before subjecting to training.

3.2. Decoding phase

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 using the POS tagger. 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 'അവന് വേഗത്തില് ഓടി'(avan vegathil oodi) since adverbs are always placed before verbs in Malayalam sentences.

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. 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’, 'ഞാന് അവള് കണ്ടു'(njan aval kandu) is the statistical output though 'ഞാന് അവളെ കണ്ടു'(njan avale kandu)is its correct translation. Mending Rule Applier rejoins the suffix and the word 'അവള്'(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 ‘എ ‘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.

4. Building and refining the Alignment Model

The EM Algorithm [3] defines a method of estimating the parameter values of translation for IBM model1[4]. 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 we set this value 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.

Hopefully the new values achieved will be better since they take into account the correlation data in the parallel corpus. Equipped with these better parameter values, we can again compute new alignment probabilities for the sentence pairs. From these values a set of even-more-revised fractional counts for word pairs is obtained. Repeating this process over and over will help the fractional counts to converge to better values. The translational probability of the English word given a Malayalam word is found to determine the best translation of a Malayalam word. It is achieved by comparing the translation probabilities of English words associated with it and picking the one with highest probability value.

Since for a sentence pair all the possible alignments were considered in the training process, huge number of alignment is produced due to this approach. Also, 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 Fig. 3.

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.

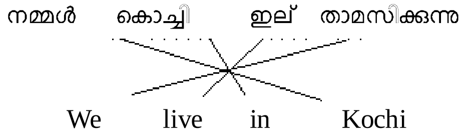

Fig.3. Insignificant Alignment

The bilingual corpus is tagged and then subjected to training. Tagging is done by considering the parts of speech entities of a sentence. 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, a word that does not have an exact translation in Malayalam may be removed 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

$$NA = factorial(Ws) . \qquad (1)$$

where Ws is the number of words in the sentence. The same corpus when tagged produces less number of alignments than factorial(Ws). By tagging we are finding the number of categories present in a sentence. There may be many words with the same tag in a sentence. Categorizing the sentence will lead 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. Therefore the total number of alignments of a sentence formed by tagging (NAT)will result in

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

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

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

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.

$$IA = factorial(Ws) - \prod_{i=1}^{m} factorial(Wci) - (NNE + NC) \qquad (4)$$

where NNE is the number of word pairs aligned with name entity and NC is the number of word pairs aligned with cognates.

On setting the alignment vectors to find English to Malayalam word translations, it is observed that certain words in the English sentence carry less sense when treated as single word units. Meaningful translations are generated only on considering group of words rather than individual ones. Representation of numbers in English is mainly expressed as a cluster which includes more than one word. For example, the number ‘22’ is denoted as ‘twenty two’ in English. But in Malayalam the equivalent word translation of ‘twenty two ‘is given by a single word as ‘ഇരുപത്തിരണ്ടു’ (irupathirandu).

Sentences are analyzed and grouping rules are applied to frame word sets in English. A match of the Malayalam word to be aligned with the English word set is identified from the ‘predictable words’ database which consists of Malayalam words and its corresponding English translations. The number of insignificant alignments is further brought down by this approach where words of Malayalam are identified whose English translations are predicted with ease. Equation 4 is modified by incorporating the knowledge of predictable Malayalam words and is given as

.

$$\mathrm{IA = factorial(Ws)} - \prod_{i=1}^{m} \mathrm{factorial(Wci)} - \{(\mathrm{NNE + NC}) + \mathrm{NPM}\} \qquad (5)$$

where NPM is the number of alignments identified between Malayalam and its predicted English translation.

5. Developing suffix separation rules

The requirement of a pre-processing 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 a word ‘ഇന്ത്യ ‘may appear 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‘. The word ‘India’ in an unseen English sentence, when subjected to decoding, produces a translation that definitely mismatches the expected results. All predictions of ‘India’ getting translated as 'ഇന്ത്യ ‘prove to be wrong since the word ‘India’ has probability to get translated as ഇന്ത്യയില്, ഇന്ത്യയുടെ and so on. The word translation chosen by the decoder, by analyzing the translation probabilities of different English Malayalam word pair, may not be an apt one to fit into 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. A post editing technique of rejoining the suffixes, as discussed in [6], is also applied to fill up the missing suffixes thereby bringing back the right meaning expressed by a sentence.

Various sandhi rules are defined in Malayalam for joining two words to form a new one. On applying these rules, the original appearance of the words taking part in this process is altered. Rules are applied by observing the ‘sounds’ of the end syllable of the first word and the start syllable of the second word. In Malayalam grammar, a classification of sandhi rules is done based on whether a word ends with a vowel (swaram) or a consonant (vyanjanam) and is discussed in [7]. This classification of sandhi rules along with an example is listed in Table 2.

Table 2. Types of sandhi rules

Category	Type of sandhi	Rule	Example
I	Swarasandhi	Swaram + swaram	മഴ + ഉണ്ട്= മഴയുണ്ട്
II	Swaravyajana sandhi	swaram + vyanjanam	താമര + കുളം = താമരക്കുളം
III	Vyanjanaswara sandhi	vyanjanam + swaram	തേന് + ഇല്ല = തേനില്ല
IV	Vyanjana sandhi	Vyanjanam + vyanjanam	നെല് + മണി = നെന്മണി

Out of this broad classification, words belonging to category I and III are of major concern and splitting up such words have more significance in the training process of SMT from English to Malayalam. The words under category II and IV are split into meaningful units prior to the suffix separation phase.

Separating the suffixes from its base form is a reverse process of ‘sandhi‘ where the essence of sandhi rule is applied in the reverse direction. For implementing suffix separation in Malayalam, the word structure is thoroughly analyzed to identify the preceding_syllable. Based on these preceding_syllable, suffix separation rules are drafted to split the words. Suffixes starting with vowel sounds like ഓടെ, ഉള്ള , എന്ന് etc. are considered in this process.

The inflected form of a word does not have the suffix present in its original form. To implement suffix separation, the category of suffix to be separated has to be identified. In the example 'അവള് + ഉടെ = അവളുടെ', the suffix ‘ഉടെ ‘is present in the abbreviated form as ‘ുടെ‘. These abbreviated forms are the keys to identify the suffixes and a few examples of these suffix_keys are listed in Table 4. 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. Table 5 illustrates some examples of the suffix labels.

Table 4. Suffix_keys and Suffix labels

Suffix_key	Suffix	Suffix_ label	Examples
ിൽ	ഇൽ	EE	ഇൽ, ഇന്, ഇല്ല
ാണ്	ആണ്	AA	ആണ്, അല്ലെ, ആൻ
ുള്ള	ഉള്ള	UU	ഉണ്ട്,ഉള്ള,ഉന്നു

To implement suffix separation in SMT a lexical database which is a depository of noun and verb roots in Malayalam is used. Also, certain Malayalam words which are not in root form still have equivalent meaningful translations in English. The word ‘അവന്റെ‘is semantically equivalent to the word ‘his’ in English. Even though 'അവന്റെ' has a suffix appended, it need not be suffix separated. Numerous words like in this category are identified and are listed in the split exception category.

For any 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‘. Picking up the appropriate rule with the help of the identified preceding _syllable, suffix_keys and suffix _labels and applying it on Malayalam words separates the suffixes from its roots. An example of the sandhi rule generated is given in below. 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.

Rule No.	Check_letter (CL)	Suffix Label	Examples	Suffix separation rules	Function
1	ള	AA EE UU	വാളാണ് വാളിൽ വാളുള്ള	prev_ള +ൾ+ suffix	Can retrieve words ending with ൾ Roots extracted : വാൾ, അവൾ

6. Observations and results achieved

The English Malayalam sample corpus used for training includes 250 sentences pairs with 1800 words. The experimental Malayalam corpus is built based on www.mathrubhumi.com, a news site providing local news on Kerala. It has been observed that better training results are achieved by selecting a corpus that is adequate enough to represent all the characteristics of the source and target languages chosen for translation. Also, the strength and correctness of the corpus is a necessity to achieve the desired output.

Table 6. Summary of 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
	Baseline + suffix separation + refined alignment model	0.17732	0. 81	0.74
Unseen sentences	Baseline + with suffix	0.6083	0. 26	0.22
	Baseline + suffix separation	0.4444	0.44	0.38
	Baseline + suffix separation + refined alignment model	0.3461	0.52	0.43

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 2. The criteria used for the evaluation are discussed below. Word Error Rate (WER) is a metric is based on the minimum edit distance between the target sentence and the sentences in the reference set. F measure is 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 is a metric that 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 2. 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 in this paper. The alignment model with the knowledge of category tags, name entities, cognates and predictable Malayalam translations 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. Also, post editing techniques like order conversion and mending rules for suffix rejoining enhances 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.

References

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

[2] 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.

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

[4] 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.

[5] 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

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

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

[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] 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.