Alignment Model and Training Technique in SMT from English to Malayalam

Mary Priya Sebastian, Sheena Kurian K and G. Santhosh Kumar

Department of Computer Science,
Cochin University of Science and Technology, Kerala, India
marypriyas@gmail.com
sheenakuriank@gmail.com
san@cusat.ac.in

Abstract. This paper investigates certain methods of training adopted in the Statistical Machine Translator (SMT) from English to Malayalam. In English Malayalam SMT, the word to word translation is determined by training the parallel corpus. Our primary goal is to improve the alignment model by reducing the number of possible alignments of all sentence pairs present in the bilingual corpus. Incorporating morphological information into the parallel corpus with the help of the parts of speech tagger has brought around better training results with improved accuracy.

Keywords: Alignment, Parallel Corpus, PoS Tagging, Malayalam, Statistical Machine Translation.

1 Introduction

In SMT [1], by using statistical methods, 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 from English to Malayalam as discussed in [2], uses statistical models to acquire an appropriate Malayalam translation for a given English sentence.

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 is trained to obtain word to word alignments from the non-aligned sentence pairs [6].

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 is 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 has been 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 is enhanced. In this paper a discussion is done about the alignment model which uses morphological information for removing the irrelevant alignment pairs. The training technique adopted to find the word to word translation in SMT is also discussed. The paper also highlights the changes occurring in the training process when morphological knowledge is introduced into the corpus.

The rest of this paper is organized as follows: In Section 2 the motivation to initiate this work is portrayed. Section 3 presents the details of the training performed in the parallel corpus. In Section 4, the method of incorporating morphological knowledge into the corpus and the modified alignment model is presented. The observations and the outcomes achieved by adopting the new alignment model is discussed in Section 5. Finally, the work is concluded in Section 6.

2 Motivation

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 SMT. The issues start off with the scarcity in the availability of English/Malayalam translations required for training 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 [8], 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 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 increases 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 brings down the word count of the sentences and thereby the number of alignments too.

Similarly many insignificant alignments is 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. Methods were identified to strengthen the parallel corpus with more information so that only the relevant alignment is included in calculating translation probabilities. It is found that when the corpus is linked with a parts of speech tagger, many irrelevant alignments are eliminated. Also, training technique is further enhanced and better results are achieved.

3 Training the parallel corpus

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 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. 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.

3.1 Alignments and alignment vectors

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 the English sentence, only one to one alignment vectors are considered.

The Malayalam corpus is pre-processed and suffix separation [7] is done before pairing with English sentences. The suffix separated Malayalam sentence aligned with its English translation is given in Fig 1.

നമ്മൾ	ഇന്ത്യ	ഇൽ	താമസിക്കുന്നു
(nammal	india	il	thammasikkunnu)
m0	m1	m2	m3
We	live	in	India
e0	e1	e2	e3

Fig 1. Malayalam English sentence pair

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 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 depends on its word count.

In the above example, the Malayalam word 'നമ്മൾ' (nammal) is aligned with any of the English word in the sentence. The word നമ്മൾ is positioned as [m0,-,-,-], [-,m0,-,-], [-,-,m0,-] and [-,-,-,m0] in the alignment vector where m0 denotes the Malayalam word in the 0^{th} position. The alignment vector of the alignment shown in Fig 2 is [m0, m3, m2, m1].

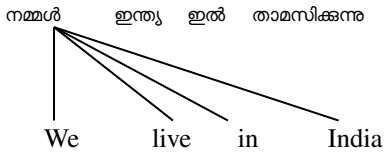

Fig 2. Alignments of the word നമ്മൾ (Nammal)

3.2 Finding translation probability

The EM Algorithm, as discussed in [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 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 is 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 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 computed. From these values a set of even-more-revised fractional counts for word pairs is obtained. Repeating this process over and over helps 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. The method of collecting fractional counts and setting alignment probabilities is illustrated with corpus1 having two sentence pairs (SP).

SP1:	നമ്മൾ	ഇന്ത്യ	ഇൽ	താമസിക്കുന്നു
	(nammal	india	il	thammasikkunnu)
	We	live	in	India
SP2:	ഇന്ത്യ			
	India			

Table 1. IFC of the word 'ഇന്ത്യ'(India)

Id	Possible translations	IFC
e0	We	0.25
e1	live	0.25
e2	in	0.25
e3	India	0.25

SP1 has 4 words and the total number of alignments is 4!. SP2 is a sentence with a single word and hence there is only one alignment defined for it. The total number of distinct English words in the corpus is four and initially any of these English words can be the translation of any Malayalam word in the corpus. Consider the word ഇന്ത്യ (India) that happens to appear in both the sentence pairs. To find the translation probability of English words in the corpus given the word ഇന്ത്യ(India), t(English word| ഇന്ത്യ), the IFC of ഇന്ത്യ is to be calculated. The IFC of ഇന്ത്യ is given in Table 1.

Table 2. A view of alignment vectors and the alignment probabilities

SP	Alignment vectors corresponding to [e0,e1, e2,e3]	Alignment Weights of e_i given m_i				AP	NAP
SP1	[m0,m1, m2,m3]	t(e0/m0) = 0.25	t(e1/m1) = 0.12	t(e2/m2) = 0.25	t(e3/m3) = 0.25	0.0019	0.02
	[m0,m1, m3,m2]	t(e0/m0) = 0.25	t(e1/m1) = 0.12	t(e2/m3) = 0.25	t(e3/m2) = 0.25	0.0019	0.02
	[m0,m2, m1,m3]	t(e0/m0) = 0.25	t(e1/m2) = 0.25	t(e2/m1) = 0.12	t(e3/m3) = 0.25	0.0019	0.02
	[m0,m2, m3,m1]	t(e0/m0) = 0.25	t(e0/m2) = 0.25	t(e0/m3) = 0.25	t(e0/m1) = 0.63	0.0098	0.11
	...	...	...	...	...	...	...
SP2	[m0]	t(e0/m0) = 0.25	-	-	-	0.25	1

Table 3. Revised fractional count of the word 'ഇന്ത്യ'(India)

Id	Possible translations	RFC
e0	We	ΣNAP of t(e0|ഇന്ത്യ) from SP1 = 0.12
e1	Live	ΣNAP of t(e1|ഇന്ത്യ) from SP1= 0.12
e2	in	ΣNAP of t(e2|ഇന്ത്യ) from SP1= 0.12
e3	India	ΣNAP of t(e3|ഇന്ത്യ) from SP1 + ΣNAP of t(e0|ഇന്ത്യ) from SP2= 1.66

The alignment probabilities of the sentence pairs are then calculated. The values of AP for all alignments of a sentence pair are equal in the first iteration and later it varies with revised fractional counts. The alignment probabilities and the normalized values calculated for the sample corpus after second iteration is shown in Table 2. Revised fractional count for the word ഇന്ത്യ(India) after second iteration is given in Table 3 and its normalized fractional count is given in Table 4. By doing this process

again and again, the translation of 'ഇന്ത്യ' converges to the word 'India' as t(India|ഇന്ത്യ) has the highest probability value.

Table 4. Normalized fractional count of 'ഇന്ത്യ'(India)

Id	Possible translations	NFC
e0	We	RFC(e0|ഇന്ത്യ)/ Σ{RFC(e0|ഇന്ത്യ),RFC(e1|ഇന്ത്യ), RFC(e2|ഇന്ത്യ),RFC(e3|ഇന്ത്യ)} = 0.06
e1	Live	RFC(e1|ഇന്ത്യ)/ Σ{RFC(e0|ഇന്ത്യ),RFC(e1|ഇന്ത്യ), RFC(e2|ഇന്ത്യ),RFC(e3|ഇന്ത്യ)} = 0.06
e2	in	RFC(e2|ഇന്ത്യ)/Σ{RFC(e0|ഇന്ത്യ),RFC(e1|ഇന്ത്യ), RFC(e2|ഇന്ത്യ),RFC(e3|ഇന്ത്യ)} = 0.06
e3	India	RFC(e3|ഇന്ത്യ)/Σ{RFC(e0|ഇന്ത്യ),RFC(e1|ഇന്ത്യ), RFC(e2|ഇന്ത്യ),RFC(e3|ഇന്ത്യ)} = 0.82

4 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 belong to the group of insignificant alignments. An example of an unwanted alignment is shown in Fig 3.

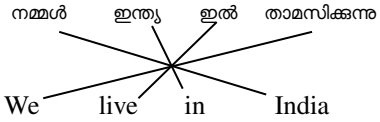

Fig 3. Insignificant alignments

Presence of these unwanted ones 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.

4.1 Tagging the corpus

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 doesn'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 are considered, the number of alignments(N_A) generated is equal to the factorial of its word count and is shown as

$$N_A = \text{factorial}(Ws) . \quad (1)$$

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, the number of categories present in a sentence are identified. There may be many words with the same tag in a sentence. 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. Therefore the total number of alignments of a sentence formed by tagging, N_AT, results in

$$N_AT = \prod_{i=1}^{m} \text{factorial}(Wc_i) . \quad (2)$$

alignment vectors, where m is the number of PoS categories in a sentence pair. The insignificant alignments, I_A, eliminated is represented as the difference between Equation 1 and 2 and is given below:

$$I_A = \text{factorial}(Ws) - \prod_{i=1}^{m} \text{factorial}(Wc_i) . \quad (3)$$

By this method the number of alignment vector for the sentence pair SP1 is just one instead of twenty four. The steps involved in eliminating the insignificant alignments from the corpus is given below.

Step1: Tag all sentences in English and Malayalam corpus based on its parts of speech category.

Step2: Separate the suffixes from the Malayalam corpus and remove the suffixes that doesn't have an equivalent English translation.

Step3: Generate the alignment vectors corresponding to each sentence pair by considering the category tags.

Step4: For the Malayalam word mw_i ,initialize the fractional count as 1/number of words present in mw_i's category .

Step5: For all sentence pair in the parallel corpus calculate the alignment probability and normalize it.
Step6: For all words find the revised fractional count and normalize it
Step7: Repeat steps 5 and 6 until convergence

The training technique with tagged corpus is analyzed with corpus2 containing two sentences.

SP1: നമ്മൾ ഇന്ത്യ ഇൽ താമസിക്കുന്നു
(nammal india il thammasikkunnu)
We live in India

SP2: ന്യൂഡൽഹി ഇന്ത്യയുടെ തലസ്ഥാനം ആണ്
(Newdelhi indiayude thalasthaanam aanu)
NewDelhi is the capital of India

In SP2 the tagged sentence structure is NewDelhi/NNP is/VBZ the/DT capital/ NN of/IN India/NNP. The sentence pairs with its category is shown in Table 5.

Table 5. PoS category of corpus2

Category	id	Word
Proper Noun	m0	ഇന്ത്യ (India)
	m1	ന്യൂഡൽഹി (Newdelhi)
Noun	m2	തലസ്ഥാനം (thalasthaanam)
Verb, 3rd ps.sing.present	m3	ആണ് (aanu)
Personal pronoun	m4	നമ്മൾ (nammal)
Verb,non-3rd ps.sing.present	m5	താമസിക്കുന്നു (thaamasikunnu)

Trying to find the word translation for the word ഇന്ത്യ(India), it is well understood that ഇന്ത്യ(India) may no longer be associated with all the English words in the corpus. It need to be associated only with the proper nouns (NNP) in the parallel corpus. To find the translation for the word ഇന്ത്യ t(Proper Nouns | ഇന്ത്യ) is considered and the IFC of ഇന്ത്യ is calculated as 1/ Total number of Proper Nouns in the corpus. Table 6 gives the IFC of word 'ഇന്ത്യ' in the tagged corpus. The alignment probabilities of the sentence pairs and the revised fractional count for the word ഇന്ത്യ is calculated with the new IFC.

Table 6. IFC of the word 'ഇന്ത്യ' in tagged corpus

Id	Possible translations	IFC
e0	India	0.5
e1	NewDelhi	0.5

In corpus1 and corpus2 the word ഇന്ത്യ(India) occur twice and therefore they are compared and analyzed to identify the advantages of the tagging method. The

alignment probability and the normalized fractional count of the word 'ഇന്ത്യ'(India) after adopting the tagged corpus is given in Table 7 and 8 respectively.

Table 7. Alignment vectors and alignment probabilities of tagged corpus

SP	Align-ment vectors	Alignment Weights of e_i given m_i				AP	NAP
SP 1	[m4,m5, m0]	t(e0/m4) = 1	t(e1/m5) = 1	t(e3/m0) = 0.75	-	0.75	1
SP 2	[m0,m3, m2,m1]	t(e0/m4) = 0.25	t(e1/m4) = 1	t(e3/m4) =1	t(e5/m4) =0.5	0.13	0.25
	[m1,m3, m2, m0]	t(e0/m4) = 0.5	t(e1/m4) = 1	t(e3/m4) = 1	t(e5/m4) = 0.75	0.38	0.75

Table 8. RFC of words in Proper Noun category

Proper Noun	Possible translations	RFC
ഇന്ത്യ	India	0.88
	NewDelhi	0.13
ന്യൂഡൽഹി	India	0.25
	NewDelhi	0.75

5 Observations and results achieved by tagging the parallel corpus

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.

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. 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 2. Here the alignment vectors is 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.

By tagging the corpus better translations for English words are obtained and it has enhanced the final outcome of the SMT. These results are evaluated using WER, F measure and BLEU metrics and is discussed in [11].

6 Conclusion

An alignment model and a training technique mostly suited for statistical machine translators from English to Malayalam have been put forward. Using the parts of speech tags as an additional knowledge source, the parallel corpus is enriched and it contains more information to select 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. This technique helps to improve the quality of word translations obtained for Malayalam words from the parallel corpus.

References

1. Lopez, A.: Statistical machine translation. In: ACM Comput. Surv., 40, 3, Article 8(2008)
2. Mary Priya Sebastian, Sheena Kurian K. and G. Santhosh Kumar: Statistical Machine Translation from English to Malayalam. In: Proceedings of National Conference on Advanced Computing, Alwaye, Kerala(2010)
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. In: Computational Linguistics, 16(2), pages 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**.** In: Computational Linguistics, 19(2), pages 263–31(1993)
5. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood From Incomplete Data Via The EM Algorithm. In: Journal of The Royal Statistical Society, 39(B): 1-38(1999)
6. Knight, K.: A statistical MT tutorial work book. Unpublished, http://www.cisp. jhu. edu/ws99/projects/mt/wkbk.rtf(1999)
7. Ananthakrishnan, R., Hegde, J., Bhattacharyya, P., Shah R., Sasikumar, M.: Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation. In: International Joint Conference on NLP (IJCNLP08), Hyderabad, India(2008)
8. Sumam, M.I., Peter, S.D.: A Morphological Processor for Malayalam Language. In: South Asia Research, Volume 27(2): pages173-186(2008)
9. Ueffing, N., Ney, H.: Using POS Information for Statistical Machine Translation into Morphologically Rich Languages. In: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1,(2003)
10. Sanchis, G., S´nchez, J.A.: Vocabulary Extension via PoS Information for SMT. In: Proceedings of the NAACL ,(2006)

11. 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)(Accepted).
12. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition and Computational Linguistics, second Ed. Prentice-Hall(2008)
13. Allen, J.F.: Natural Language Understanding. Pearson Education, Second Edition (2002)