A Classification of Sandhi Rules for Suffix Separation in 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: Suffix separation plays a vital role in improving the quality of training in the Statistical Machine Translation from English into Malayalam. The morphological richness and the agglutinative nature of Malayalam make it necessary to retrieve the root word from its inflected form in the training process. The suffix separation process accomplishes this task by scrutinizing the Malayalam words and by applying sandhi rules. In this paper, various handcrafted rules designed for the suffix separation process in the English Malayalam SMT are presented. A classification of these rules is done based on the Malayalam syllable preceding the suffix in the inflected form of the word (check_letter). The suffixes beginning with the vowel sounds like ആൽ, ഉടെ, ഇൽ etc are mainly considered in this process. By examining the check_letter in a word, the suffix separation rules can be directly applied to extract the root words. The quick look up table provided in this paper can be used as a guideline in implementing suffix separation in Malayalam language.

Keywords: suffix separation, sandhi rules, English Malayalam translation, vowels, consonants

1 Introduction

Machine Translation is one of the budding applications in the field of Natural Language Processing. A machine translation system clearly would save enormous amount of human power and time for the translation of one language text into another. In the past few years different methodologies are emerging in the field of machine translation. One among them is the statistical approach where translation process is carried out by acquiring word translations automatically from the parallel corpora that contain large amounts of bilingual text documents. Most SMT systems today employ linguistic knowledge and operate productively by incorporating the morphological details of the source and target language. Many works, both in foreign as well as Indian languages, are in progress in the field of Statistical Machine Translation (SMT). A similar work of translating English into Malayalam using statistical methods is discussed in [1].

As discussed in [2], morphologically rich languages need preprocessing in SMT. Since Malayalam language is morphologically rich and is having an agglutinative

nature, preprocessing process helped in improving the quality of the translations and the results are discussed in [1]. Experiments done on the Malayalam corpus proved that separation of suffixes adds on to the quality of the training process in SMT. In this paper certain guidelines which helps to speed up the implementation of suffix separation phase is discussed. On exploring the structure of words and by studying the sandhi rules in Malayalam, a number of rules to separate the suffixes from its root forms are developed. A classification of these rules based on check_letters (CL) is also done to simplify the task of suffix separation. The burden of building a suffix separator from the scratch is definitely reduced by adopting these rules.

The rest of this paper is organized as follows: The related work done in this area is presented in Section 2. In Section 3, the role of suffix separation in the field of machine translation is discussed. The steps involved in the suffix separation process in SMT from English to Malayalam are explained in Section 4. Section 5 presents the details about the classification of the suffix separation rules. 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.

2 Related Work

Due to the morphological richness and complex nature of the language, thorough preprocessing is needed for Malayalam. Suffix separation is the most important preprocessing technique adopted in many of the NLP projects in Malayalam. Various works in machine translation are in progress in different organizations all over India. A morphological processor for Malayalam language is discussed in [3]. Here a processor that deals with the processing of nouns, pronouns, verbs and modifiers is explained. Another method of morphological analyzer for Malayalam based on suffix separation is discussed in [4]. The method of incorporating suffix separation in the works related to parts of speech tagging and morph analyzer is explained in [8] and [9] respectively.

3 Necessity of Suffix Separation in Machine Translation

In the first phase of SMT, the parallel corpus of English Malayalam and the monolingual corpus of Malayalam are subjected to training [5]. The result expected from this phase is a set of English word translations for the different Malayalam words in the parallel corpus. 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 and is illustrated in Table 1.

Table 1. Word 'ഇന്ത്യ' and its various inflected forms

Root word	Different inflected forms
ഇന്ത്യ	ഇന്ത്യയുടെ
	ഇന്ത്യക്ക്
	ഇന്ത്യയോട്
	ഇന്ത്യയിൽ
	ഇന്ത്യയെ
	ഇന്ത്യയാണ്

On setting the word to word alignments in the English Malayalam sentence pair, the inflected Malayalam word is aligned with the English word 'India'. Fig. 1 highlights three sentences of a sample corpus in which the alignment of the word 'ഇന്ത്യ' is clearly depicted.

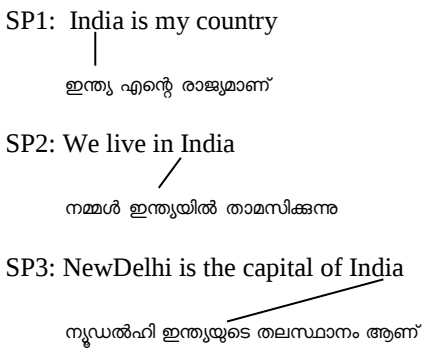

Fig. 1. Alignments of word 'ഇന്ത്യ'

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 ഇന്ത്യയിൽ, 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.

4 Suffix Separation in SMT

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	തേൻ + ഇല്ല = തേനില്ല
IV	Vyanjana sandhi	vyanjanam + vyanjanam	നെൽ + 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.

Table 3. List of Suffixes

Malayalam suffixes									
ഉടെ	ഓടെ	ഉള്ള	ആയ	ആയി	എ	ഉം	അ	ഇൻ	ഓട്
ഓ	ആൽ	ഏ	അല്ല	ഇല്ല	എന്ന്	ഉക	ഇൽ	ആണ്	ആൻ

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 check_letter. Based on these check_letters, suffix separation rules are drafted to split the words. Suffixes starting with vowel sounds are considered in this process and are listed in Table 3.

The inflected form of a word doesn't 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 'അവൾ + ഉടെ = അവളുടെ', the suffix 'ഉടെ' is present in an

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

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

Table 5. Suffix_labels

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

With check_letter, suffix_keys and suffix _labels the suffixes are separated from the roots. 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 in this category are identified and are listed in the split exception category. Table 6 shows the words belonging to the split exception category.

Table 6. Split Exception Category

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

Algorithm of suffix separation

begin
for all words in Malayalam corpus
choose a new word from the corpus to analyze
check whether the word is in the lexical database or in the split exception category
if present
exit
else
a. scan the word from right hand side
b. look up the suffix_ key table and identify the suffix present in the word
c. look up the suffix _label table to get the suffix_label
d. identify the check_letter that precedes the suffix_key in the word
e. choose the suffix separation rule from the look up table with check_letter and suffix_label as index
f. apply the suffix separation rule and remove the suffix
g. re-initialize the word and exit if it is a root word
h. repeat steps a to g until the root word is encountered.
end

5 Classification of Suffix Separation Rules

The rules classified based on the check_letter's are listed in Table 6. 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'. The suffixes considered in each category in the example in Table 6 are listed below.

- AA : ആണ്
- EE : ഇൽ
- UU : ഉള്ള
- EI : ഏ
- OO : ഓട്

For each category of the suffix_label, the suffix rules along with its functionality are explained in Table 7.

Table 7. Look up table

Rule No.	Check_letter (CL)	Suffix Label	Examples	Suffix separation rules	Function
1	ള	AA	വാളാണ്	prev_ള +ൾ+ suffix	Can retrieve words ending with ൾ Roots extracted : വാൾ, അവൾ
		EE	വാളിൽ		
		UU	വാളുള്ള		
		EI	അവളെ	prev_ളെ +ൾ+ suffix	
		OO	അവളോട്	prev_ളോ+ൾ+ suffix	
2	ന	AA	തേനാണ്	prev_ന +ൻ+ suffix	Can retrieve words ending with ൻ Root extracted : തേൻ
		EE	തേനിൽ		
		UU	തേനുള്ള		
		EI	തേനേ	prev_നേ+ൻ + suffix	
		OO	തേനോട്	prev_നോ +ൻ + suffix	
3	ല	AA	പാലാണ്	prev_ല + ൽ+ suffix	Can retrieve words ending with ൽ Root extracted : പാൽ
		EE	പാലിൽ		
		UU	പാലുള്ള		
		EI	പാലേ	prev_ലേ+ ൽ+ suffix	
		OO	പാലോട്	prev_ലോ+ ൽ+ suffix	

Rule No.	Check_letter (CL)	Suffix Label	Examples	Suffix separation rules	Function
4	റ	AA	കയറാണ്	prev_റ + ർ + suffix	Can retrieve words ending with ർ Root extracted : കയർ
		EE	കയറിൽ		
		UU	കയറുള്ള		
		EI	കയറേ	prev_റേ+ ർ + suffix	
		OO	കയറോട്ട	prev_റോ+ ർ + suffix	
5	ണ	AA	കടിഞ്ഞാണാണ്	prev_ണ+ൺ+ suffix	Can retrieve words ending with ൺ Root extracted :കടിഞ്ഞാൺ
		EE	കടിഞ്ഞാണിൽ		
		UU	കടിഞ്ഞാണുള്ള		
		EI	കടിഞ്ഞാണേ	prev_ണേ+ൺ+ suffix	
		OO	കടിഞ്ഞാണോ	prev_ണോ+ൺ+ suffix	
6	യ	AA	മേശയാണ്, മിഴിയാണ്	prev_യ + suffix	Can retrieve words ending with അ and ഇ sound. Roots extracted : മേശ, മിഴി
		EE	മേശയിൽ, മിഴിയിൽ		
		UU	മേശയുള്ള , മിഴിയുള്ള		
		EI	മേശേ മിഴിയേ	prev_ യേ+ suffix	
		OO	മേശയോട് മിഴിയോട്ട	prev_ യോ + suffix	

Rule No.	Check_letter (CL)	Suffix Label	Examples	Suffix separation rules	Function
7	വ	AA	മഴവാണ്	prev_ വ + suffix	Can retrieve words ending with ഉ sound Root extracted : മഴ
		EE	മഴവിൽ		
		UU	മഴവുള്ള		
		EI	മഴവേ	prev_ വേ + suffix	
		OO	മഴവോട്	prev_ വോ + suffix	
8	ത്ത	EE	പാലത്തിൽ	prev_ത്ത + ം + suffix	Can retrieve words ending with അം sound. Root extracted : പാലം
		EI	പാലത്തേ		
		OO	പാലത്തോട്		
	മ	AA	പാലമാണ്	prev_മ + ം + suffix	
		UU	പാലമുള്ള		
9	Consonant like ക സ, ത,etc...	AA	കാതാണ്	prev_ CL +് +suffix	Can retrieve words ending with consonants followed by 'chandrakkala' Root extracted : കാത്
		EE	കാതിൽ		
		UU	കാതുള്ള		
		EI	കാതേ	prev_ േ(CL)+ CL +് +suffix	
		OO	കാതോട്ട	prev_ ോ(CL)+ CL+് +suffix	

Rule No.	Check_letter (CL)	Suffix Label	Examples	Suffix separation rules	Function
10	Conjunct consonant like ണ്ണ, ല്ല, പ്പ ,ള്ള,ച്ച,ട്ട etc…	AA	കണ്ണാണ്	prev_ CL +ാ +suffix	Can retrieve words ending with conjunct consonant followed by 'chandrakkala' Roots extracted : കണ്ണ്
		EE	കണ്ണിൽ		
		UU	കണ്ണുള്ള		
		EI	കണ്ണേ	prev_ േ(CL) + CL +ാ +suffix	
		OO	കണ്ണോട്	prev_ോ(CL)+ CL+ാ +suffix	

6 Observations

Most of the words in Malayalam confine to the rules discussed in the previous section. But a few numbers of exceptions are also identified in the making of suffix separation rules. Words with check_letter ‘ര’ on splitting is changed into 'ർ' or 'ര്' . The word 'മലരിൽ' is split as മലർ where ര is changed as 'ർ' and in words like പാരിൽ it is transformed as 'ര്'. In the former case, the rules designed for ‘റ’ is also applicable for ‘ര’ and rule no. 4 is applied to extract the root words. But when the same rule is applied to the word 'പാരിൽ', it is split as പാർ + ഇൽ which is not the required form. For the latter one, rule no.9 is applied to retrieve the root word as പാര്. The ambiguity in choosing the rule is resolved by by making an appropriate selection with human aid. Similar issue arises in the case of 'ഴ'. Words with check_letter 'ഴ' can be spit as words ending with 'ൾ' and ഴ്.

The check_letters വ and ത്ത in a word are used to retrieve roots belonging to rule no. 7 and 8 respectively. But certain words like 'മാവിൽ' and 'കത്തിൽ' are split into meaningless units when this suffix rule is applied. On applying rule no. 7 and 8 , 'മാവിൽ' becomes 'മാ + ഇൽ' and 'കത്തിൽ' becomes 'കം + ഇൽ' respectively. Since check_letters വ and ത്ത are consonants and conjunct consonants respectively, rule no. 9 and 10 should be applied to retrieve the correct form of root words. Consonants and conjunct consonants that act as check_letters of other rules serve dual role and is termed as dual_check_letters. While splitting words with dual_check_letters, appropriate rules are chosen with human intervention.

Another issue is related to words with check_letter‘ട്’ . Words like 'കാട്ടിൽ', 'നാട്ടിൽ' etc. are some examples of words belonging to this category. On encountering these words, rule no.10 is applied to extract the roots and as per the rule കാട്ട് and നാട്ട് are the roots extracted. Here, കാട് and നാട് are the correct form of the roots required. Also, certian words in this category abide by rule no. 10, for example ‘പാട്ടിൽ’ on splitting gives പാട്ട് which is the correct form of the root. So in the case where check_letter ‘ട്’ is used the roots retrieved after applying the rule is searched in the lexical database. If the root is not present, it is not a valid one and to obtain its correct form the syllabe 'ട്' in the root is replaced with 'ട്'.

By the application of the suffix separation rules in the Malayalam 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 [1].

7 Conclusion

Suffix separation using sandhi rules have a crucial role in bringing good translation results in SMT from English to Malayalam. To simplify the task of implementing the suffix separator a scheme has been put forward in which various hand crafted rules are applied to separate the suffixes of Malayalam. A quick look up table that summarizes the classification of the suffix separation rules is provided. The work proposed here can be utilized as a guideline to separate suffixes beginning with vowel sounds from any word in the Malayalam language. This work can be further extended by incorporating rules to separate words beginning with consonants as well.

References

1. 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)
2. Lopez, A.: Statistical Machine Translation. ACM Computing Survey, 40, 3, Article 8 (2008)
3. Sumam M I, Peter S D. A Morphological Processor for Malayalam Language. South Asia Research, Volume27 (2). pp 173-186, 2008.
4. 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)
5. 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) (Accepted)
6. 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)

7. A R Rajaraja Varma. Keralapanineeyam, Eight edition, DC books, 2006.
8. 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
9. 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