Handling OOV Words in Phrase-Based Statistical Machine Translation for Malayalam

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

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

Abstract

Statistical Machine Translation (SMT) is one of the potential applications in the field of Natural Language Processing. The translation process in SMT is carried out by acquiring translation rules automatically from the parallel corpora. However, for many language pairs (e.g. Malayalam- English), they are available only in very limited quantities. Therefore, for these language pairs a huge portion of phrases encountered at run-time will be unknown. This paper focuses on methods for handling such out-of-vocabulary (OOV) words in Malayalam that cannot be translated to English using conventional phrase-based statistical machine translation systems. The OOV words in the source sentence are pre-processed to obtain the root word and its suffix. Different inflected forms of the OOV root are generated and a match is looked up for the word variants in the phrase translation table of the translation model. A Vocabulary filter is used to choose the best among the translations of these word variants by finding the unigram count. A match for the OOV suffix is also looked up in the phrase entries and the target translations are filtered out. Structuring of the filtered phrases is done and SMT translation model is extended by adding OOV with its new phrase translations. By the results of the manual evaluation done it is observed that amount of OOV words in the input has been reduced considerably.

Keywords—SMT, OOV words, out-of-vocabulary, unknown words, phrase translation, Machine Translation, Malayalam Translation

I INTRODUCTION

A machine translation system clearly would save enormous amount of human power and time for the translation of one language text into another. Statistical Machine Translation has proven to be a significant approach in this field for a long time. The triggering component for the statistical approach arises from the rapidly growing availability of bilingual machine readable text [1]. SMT based on statistical method was first proposed by IBM in the early nineties [2]. The IBM Models are word-based models and represent the first generation of SMT models. But the problem with word-based models is that the concept of a word must be precisely defined in order to correctly tokenize the sentence. Although this is adequate for languages such as English, it is somewhat more problematic for morphologically complex languages such as Malayalam and Tamil. When translating such languages with word-based models, tokenization becomes a critical issue. Also, in word-based translation models, substitution and reordering decisions are all made independently for each individual word. These observations have motivated the phrase based translation model. In phrase-based models the assumption that the unit of translation is a single word is discarded [3]. Instead, a unit of translation may be any contiguous sequence of words, called a phrase.

Phrase-based SMT systems train their statistical models using parallel corpora and base themselves on training data. In the training phase, the entire corpus is examined and statistical methods are adopted to extract the appropriate meaning for the words in the source language. An alignment model is defined in training which sets all the possible alignments between the source and target sentence pairs in the parallel corpus [1]. Since this is a corpus based approach there are certain issues associated with translation. System finds it difficult to translate words that have not been

encountered in the training phase. If the source sentence contains words that are not present in the training corpus, its translation becomes difficult. Words that are not seen in the training corpus may not be translated and are either discarded or left as it is in the output. The words that belong to this category are termed as out of vocabulary words or unknown words [4].

Unknown word problem increases when the available bilingual data is scarce. Even though it is said that the availability of the parallel corpora is growing rapidly, a fully fledged parallel corpora for many language pairs especially for Indian languages is still not available. A limited sized corpus may not reflect all the features of the language it represents. Therefore the probability of the occurrences of the unknown words increases. Also, languages that are rich in morphology have different inflected forms for a word. The inflected form of a word in Malayalam can have various suffixes appended to its root [5]. For example the word 'ഇന്ത്യ' has different inflected 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

The objective of this work is to develop a system that finds possible translations of these unknown words. The OOV words are preprocessed and various inflected forms of the OOV words are generated. Matches for these inflected forms are looked up in the phrase translation table. New phrase translation entries corresponding to the word variants identified is generated and the phrase table is appended with the OOV and its new phrase translations.

The rest of this paper is organized as follows: In Section 2 the related works in this field is portrayed. Section 3 presents the details some of the morphological aspects of Malayalam. In Section 4, the method of generating new phrase translations for OOV words is detailed. The observations and the outcomes achieved are discussed in Section 5. The work is concluded in Section 6.

II. RELATED WORKS

A number of works have tackled the unknown word problem. A heuristic based identification and translation method is discussed in [6]. A model based on morphological analysis and large amounts of lexical information contained in a dictionary is proposed in [7]. In [8] a method to estimate Part-Of-Speech information of unknown words using a statistical model of morphology and context is demonstrated. In [9], external bilingual dictionaries are used to obtain target language words for unknown proper nouns. In [10], orthographic features are utilized to identify lexical approximations for OOV words. A method to translate unknown words by using lexical approximation techniques to identify known variant word forms and adjust the input sentence accordingly is discussed in [11]. Also, a method to increase the translation coverage by extending the original phrase-table with phrase translation pairs for source vocabulary words without single word entries in the original phrase-table is also mentioned in the same paper.

In contrast to these previous approaches, this paper proposes a method of handling OOV words by finding the word variants of OOVs

and its corresponding phrase translations. The best among these phrase translations from the training corpus are chosen and restructured to extend the phrase translation table of the translation model of SMT.

III. MALAYALAM MORPHOLOGY

Malayalam is one of the 22 official languages of India, spoken predominantly in the state of Kerala, in Southern India, by around 37 million people. It belongs to the Dravidian family of languages (Tamil, Malayalam, Kannada, and Telugu). The origin of Malayalam as a distinct language may be traced to the last quarter of 9th Century A.D. Throughout its gradual evolution the most important influence on Malayalam has been that of Sanskrit and Prakrit brought into Kerala by Brahmins. In modern Malayalam also a good part of vocabulary is of Sanskritic origin. It has evolved mainly from Tamil and is most related to Tamil when compared to other Dravidian languages [12]. Malayalam is morphologically rich and highly agglutinative like any other Dravidian languages. But it differs from other Dravidian languages in that the personal endings on verbs are absent. The verb in Malayalam takes tense, aspect, mood but does not take person, number, and gender marker [13], Malayalam has an unmarked SOV word order, yet word order is relatively free. It has a vast and extensive grammatical structure. It also features a rich case marking system, with nominative, accusative, dative, sociative, locative, instrumental, and genitive case suffixes.

IV. HANDLING OOV WORDS

The decoding unit of SMT [14] is modified to handle the OOVs present in the Malayalam sentence. The overview of the SMT with the modified decoder is given in Figure 1. The input sentence is tokenized to obtain the words in the sentence. These words are passed in to a Match Locator module that checks the presence of the input words in the Phrase Translation (PT) table.

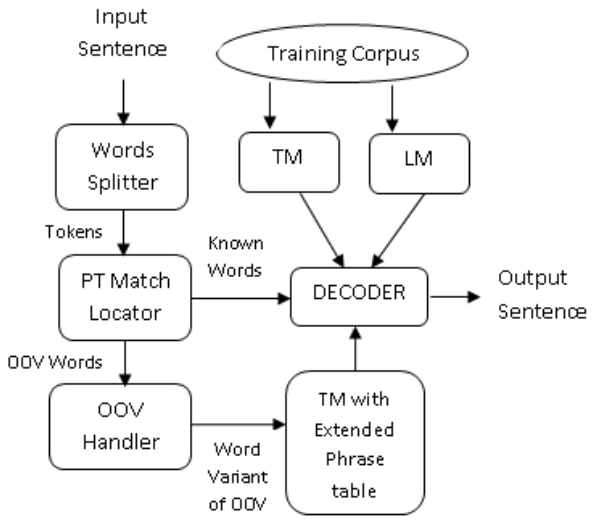

FIGURE 1. OVERVIEW OF THE SMT TO HANDLE OOV WORDS

SMT is based on statistical methods and since it is a corpus based approach only the words present in the PT table are translated. If the entries are not present SMT will consider it as an unknown source word that can never be translated. Such words will be listed as OOV and are passed into the OOV handler module for identifying their word variants, if present. The outcome of the OOV handler module is a list of new phrase translations for the OOV word. The PT table is appended with this information and the decoding process is restarted with the Translation Model(TM) with extended phrase translations.

The details of the OOV handler module are given in Figure 2. The input words that are identified as OOV are passed onto the root extractor. Morphological analysis of the OOV words is done and the root word is extracted. In the next phase all inflectional word forms are generated from the root word according to the inflectional attributes

of the respective word class. The module generates word variants for verbs, nouns, and adjectives separately.

These inflected words are then looked up in the PT table for a match. If a match is not found the OOV is rejected. Otherwise, a list of word variants along with their translations from the PT table, which is obtained as the result of the training phase, is generated. The list is processed by a vocabulary filter to identify the most frequently occurring phrase among the word variant translations. This module finds vocabulary weight of a phrase by calculating the product of the TM probability TMwt($Pharse_i$) and the unigram count of a phrase (UNIwt($Phrase_i$)). It then filters out the phrases with a probability less than a predetermined threshold.

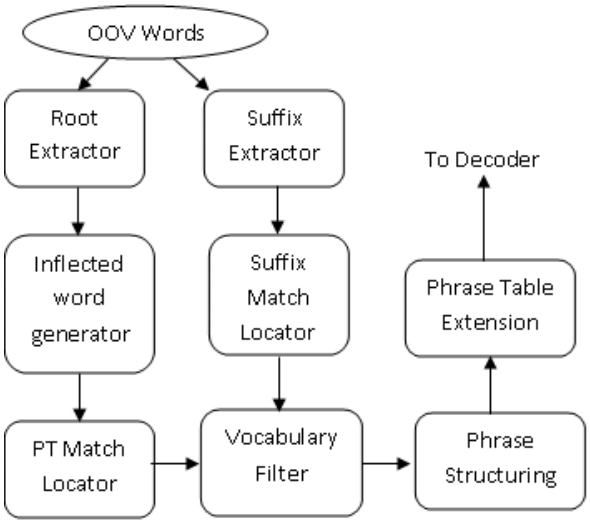

FIGURE 2. OVERVIEW OF THE OOV HANDLER

The phrases that maximize these values are considered as the best translations of the word variants. TMwt(Pharsei) is calculated as the probability of Pharsei to get translated as the inflected OOV and this value is obtained as the result of the training phase.

Another module is designed to extract the suffix from the OOV word and it is done by applying the sandhi rules in the reverse direction [15]. In the PT table a substring search is done from the word endings to identify words with similar suffixes. If such words are found, the phrase translations corresponding to them are passed on to the vocabulary filter. The translation that has the highest unigram count is chosen as the translation of the suffix part.

When Malayalam is translated into English phrases there is an order change in the position of the suffixes. The word corresponding to the suffix is appended in the beginning of the English phrase. For example, the word 'ഇന്ത്യക്ക്' is translated as 'for India'. Therefore the filtered out phrases will be rejoined based on the pharse struturing rules where the pharses from the suffix module will be placed in the beginning followed by the root section.

The list of new phrase translations thus obtained is appended on to the TM phrase translation table by adding a new entry for the OOV word. After the completion of this process, the decoding is resumed to translate the input sentence. The OOVs identified earlier now have an entry in the PT table and hence the translation of OOV takes place without fail.

V. OBSERVATIONS AND DISCUSSIONS

Significant gains in coverage and translation quality can be had by integrating OOV handler into statistical machine translation. In effect, this method introduces some amount of generalization into statistical machine translation. In the earlier approaches, translations were made possible only if having observed a particular word or phrase in the training set. For translating them, the condition of having seen every word in advance need not hold any more. Knowledge of inflected forms to identify the word variants, which are words that have

similar meanings, is found to be useful in the process of translation. This method is particularly applicable to small data conditions, which are overwhelmed by sparse data problems. Translation of languages that are morphologically rich are found to be a critical when the data set is less and this approach is effective in reducing the unknown words in the source text.

Translation models suffer from sparse data. When only a very small parallel corpus is available for training, translations are learned for very few of the unique phrases in a test set. In contrast after expanding the phrase table using the translations of OOV word variants, the coverage of the unique test set phrases goes up dramatically. A manual evaluation by judging the accuracy of phrases for a small set of OOV handled translations using the manual word alignments is done. For the training corpus the coverage goes up from less than 50% of the vocabulary items being covered to 85%. The results are summarized in the Table 2.

TABLE 2. OOV WORD REDUCTION

OOV Handler	OOV reduction rate
with root extractor module	50-65%
with root & suffix extractor	50-85%

VI. CONCLUSION

Translating the unknown source words have always been difficult in SMT, since the phrase translation table generated after the training phase will not be having the translations for such source words. A method to find the possible translations of such OOV words using an OOV handler module resulting in an extension of phrase translation table is discussed in this paper. A manual evaluation is done and the result shows that the rate of reduction of the OOV words has improved. Also, this result has an impact on improving the quality of translations produced.

VI. REFERENCES

[1] P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer (1993). The mathematics of statistical machine translation: parameter estimation. Computational Linguistics, 19(2), 263-311.

[2] Lopez, Adam. "Statistical machine translation." ACM Computing Surveys 40.3 (2008): 1-49.

[3] Koehn P, Och F J, and Marcu D. Statistical Phrase-based Translation. In the Proceedings of HLT-NAACL, (2003)

[4] N. Habash. 2008. Online Handling of Out-of-Vocabulary Words for Statistical Machine Translation. CCLS Technical Report.

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

[6] R. Mahesh K. Sinha. 2005. Interpreting unknown words in machine translation from hindi to english. In Computational Intelligence, pages 278–282.

[7] Kiyotaka Uchimoto, Satoshi Sekine, and Hitoshi Isahara. 2001. The unknown word problem: a morphological analysis of japanese using maximum entropy aided by a dictionary. In Proceedings of EMNLP 2001, pages 91–99, Pittsburgh.

[8] Masaaki Nagata. 1999. A part of speech estimation method for japanese unknown words using a statisticalmodel of morphology and context. In in Proceedingsof ACL 1999, pages 277–284, University of Maryland.

[9] H. Okuma et al., "Introducing Translation Dictionary into phrase-based SMT," in Proc. of MT Summit XI, Copenhagen, Denmark, 2007, pp. 361–368.

[10] C. Mermer et al., "The TUBITAK-UEKAE SMT System for IWSLT 2007," in Proc. of the IWSLT, Trento, Italy, 2007, pp. 176–179.

[11] Arora, Karunesh, Michael Paul, and Eiichiro Sumita. "Translation of unknown words in phrase-based statistical machine translation for languages of rich morphology." Proceedings of SLTU (2008).

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

[13] Manju, K., S. Soumya, and Sumam Mary Idicula. "Development of a POS Tagger for Malayalam-An Experience." Advances in

Recent Technologies in Communication and Computing, 2009. ARTCom'09. International Conference on. IEEE, 2009.

[14] 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.

[15] Mary Priya Sebastian, Sheena Kurian K. and G. Santhosh Kumar: Suffix Separation in Statistical Machine Translation from English to Malayalam using Sandhi Rules. In: Proceedings of National Conference on Indian Language Computing, Cochin, Kerala(2011)

[16] Joseph Turian, Luke Shen, and I.Melamed, "Evaluation of machine translation and its evaluation," in Proc. of the MT Summit IX, New Orleans, USA, 2003, pp. 386–393.

[17] 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.