Building Word-Aligned Bilingual Corpus for Statistical Myanmar-English Translation

Khin Thandar Nwet
University of Computer Studies, Yangon, Myanmar
khin.thandarnwet@gmail.com

Abstract

In recent years statistical word alignment models have been widely used for various Natural Language Processing (NLP) problems. In this paper we describe our work in constructing an aligned English-Myanmar parallel corpus. Corpora are not available for Myanmar language and our work in developing parallel corpus will also hopefully be very useful in many natural language applications. Word alignment plays a crucial role in statistical machine translation, since word-aligned corpora have been found to be an excellent source of translation-related knowledge. If there were errors in alignment, this will cause subsequence failure NLP processes. The alignments produced when the training on word-aligned data are dramatically better than when training on sentence-aligned data. The main purpose of this system is to provide as part of translation machine in Myanmar-English machine translation. The proposed system is combination of corpus based approach and dictionary lookup approach. The corpus based approach is based on the first three IBM models.

1. Introduction

Bilingual word alignment is the first step of most current approaches to Statistical Machine Translation or SMT [1]. One simple and very old but still quite useful approach for language modeling is n-gram modeling. Separate language models are built for the source language (SL) and the target language (TL). For this stage, monolingual corpora of the SL and the TL are required. The second stage is called translation modeling and it includes the step of finding the word alignments induced over a sentence aligned bilingual (parallel) corpus. This paper deals with the step of word alignment.

Corpora and other lexical resources are not yet widely available in Myanmar. Research in language technologies has therefore not progressed much. In this paper we describe our efforts in building an English-Myanmar aligned parallel corpus. A parallel corpus is a collection of texts in two languages, one of which is the translation equivalent of the other. Although parallel corpora are very useful resources for many natural languages processing applications such as building machine translation systems, multilingual dictionaries and word sense disambiguation, they are not yet available for many languages of the world. Myanmar language is no exception.

Building a parallel corpus manually is a very tedious and time consuming task. A good way to develop such a corpus is to start from available resources containing the translations from the source language to the target language. A parallel corpus becomes very useful when the texts in the two languages are aligned. Here we have used the IBM models to align the texts at word level.

Many words in natural languages have multiple meanings. It is important to identify the correct sense of a word before we take up translation, query-based information retrieval, information extraction, question answering, etc. Recently, parallel corpora are being employed for detecting the correct sense of a word. Ng [4] proposed that if two languages are not closely related, different senses in the source language are likely to be translated differently in the target language. Parallel corpus based techniques for word sense disambiguation therefore work better when the two languages are dissimilar. It may be noted that English-Myanmar scores well here.

2. Related Work

G. Chinnappa and Anil Kumar Singh [3] proposed a java implementation of an extended word alignment algorithm based on the IBM models. They have been able to improve the performance by introducing a similarity measure (Dice coefficient), using a list of cognates and morph analyzer.

In 1991, Gale and Church's approach [9] has been widely adopted for the alignment of European languages and has subsequently been improved with complementary techniques. Their method is based on a simple statistical model of sentence length measured in terms of characters. This method uses the fact that longer sentences tend to be translated into longer sentences in the target language and shorter sentences tend to be translated into shorter sentences.

K-vec algorithm [7] makes use of the word position and frequency feature to find word correspondences using Euclidean distance. Ittycheriah and Roukos [5] proposed a maximum entropy word aligner for Arabic-English machine

translation. Martin et al. [6] have discussed word alignment for languages with scarce resources.

Most current SMT systems [8] use a generative model for word alignment such as the freely available GIZA++ [2], which is an implementation of the IBM word alignment models. These models treat word alignment as a hidden process, and maximize the probability of the observed sentence pairs using the expectation maximization (EM) algorithm.

3. Overview of the Statistical Machine Translation of Myanmar to English

In this system, input sentence (source language) is Myanmar language. Source Language Model uses this input sentence to segment and tag by using Myanmar-English Dictionary and corpus. Translation model use segmented sentence and then looking up at bilingual corpus for relevant target translation words. One source word can have more than one Target words. N-best translation list is sent to WSD (Word Sense Disambiguation) system to get the best translation. Translation model use the output of WSD to get target language sentence (English sentence). Target translation model checks sentence patterns and grammar patterns of target language sentence. In this Myanmar to English machine translation system, we focus on Alignment model. Alignment is a central issue in the construction and exploitation of parallel corpora.

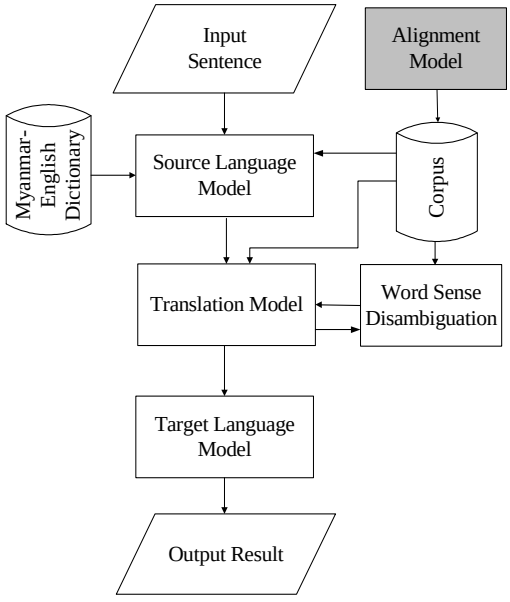

Figure 1. Machine Translation System of Myanmar- English

4. Alignment

Alignment is a central issue in the construction and exploitation of parallel corpora. One of the central modeling problems in statistical machine translation (SMT) is alignment between parallel texts. The duty of alignment methodology is to identify translation equivalence between sentences, words and phrases within sentences. In most literature, alignment methods are categorized as either association approaches or estimation approaches (also called heuristic models and statistical models). Association approaches use string similarity measures, word order heuristics, or co-occurrence measures (e.g. mutual information scores). For the latter, the idea is to find out if a cross-language word pair co-occurs more often than could be expected from chance.

The central distinction between statistical and heuristic approaches is that statistical approaches are based on well-founded probabilistic models while heuristic ones are not. Estimation approaches use probabilities estimated from parallel corpora, inspired from statistical machine translation, where the computation of word alignments is part of the computation of the translation model.

4.1 The IBM Alignment Models 1 through 3

In their systematic review of statistical alignment models (Och and Ney ,2003[2]), Och and Ney describe the essence of statistical alignment as trying to model the probabilistic relationship between the source language string f, and target language string e, and the alignment a between positions in f and e. The mathematical notations commonly used for statistical alignment models follow.

$$f_1^J = f_1, \ldots f_j, \ldots, f_J$$

$$e_1^I = e1, \ldots, e_i, \ldots, e_I$$

Equation 1.

Foreign and English sentences f and e, contain a number or tokens, J and I (Equation 1). Tokens in sentences f and e can be aligned, correspond to one another. The set of possible alignments is denoted A, and each alignment from j to i(foreign to English) is denoted by a_j which holds the index of the corresponding token i in the English sentence(see equation 2).

$$A \subseteq \{(j,i) : j = 1,..., J; i = 1,..., I\}$$
$$j \rightarrow i = a_j$$
$$i = a_j$$

Equation 2.

The basic alignment model using the above described notation can be seen in Equation 3.

$$\Pr(f_1^J \mid e_1^J)$$
$$\Pr(f_1^J, a_1^J \mid e_1^J)$$
$$\Pr(f_1^J \mid e_1^J) = \sum_{a_i^J} \Pr(f_1^J, a_1^J \mid e_1^J)$$

Equation 3.

From the basic translation model $\Pr(f_1^J \mid e_1^J)$, the alignment is included into equation to express the likelihood of a certain alignment mapping one token in sentence f to a token in sentence e, $\Pr(f_1^J, a_1^J \mid e_1^J)$. If all alignments are considered, the total likelihood should be equal to the basic translation model probability.

The above described model is the IBM Model 1. In model-1 word positions are not considered.

Model 2

One problem of Model 1 is that it does not have any way of differentiating between alignments that align words on the opposite ends of the sentences, from alignments which are closer. Model 2 add this distinction.

Model 3

Languages such as Swedish and German make use of compound words. Myanmar language also makes use of compound words. Languages such as English do not. This difference makes translating between such languages impossible for certain words, the previous models 1 and 2 would not be capable of mapping one Myanmar, Swedish or German word into two English words. Model 3 however introduces fertility based alignment, which considers such one to many translations probable.

5. The Proposed Model

Our proposed system consists of three steps. They are input, preprocessing and alignment. The inputs are Myanmar and English sentence. In preprocessing step, the two input sentences are segmented, stemmed and removed the stop words. Myanmar stop words are "သည်", "မြီး",etc.The last step is alignment. In alignment step, firstly we find English word in the bilingual Myanmar-English corpus by using Myanmar word and N-gram. If we found the search word in corpus, we align input Myanmar word and English word. If we do not find this Myanmar word in corpus, we look up in Myanmar-English dictionary. When we search in Myanmar-English dictionary, we use Myanmar root word and English POS. If we found, we get the English word. If the resulting English word is the same with the root word of input English word, we align input Myanmar word and English word and then store this alignment in parallel corpus. Parallel corpus is used as training data set and also the output of the system.

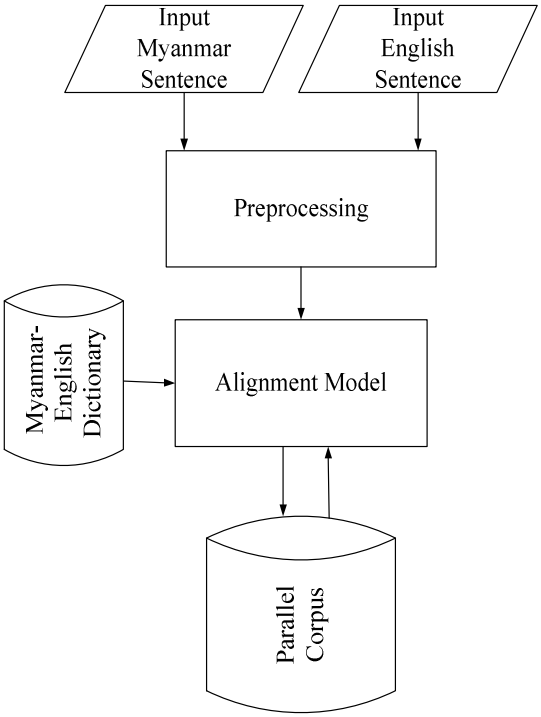

Figure 2. Proposed Alignment System

(1) Input two sentences are

ငှက်များသည်ပျံကြသည်။

Birds fly.

(2) After Segmentation

ငှက်_များ_သည်_ပျံ_ကြ_သည်_

Birds<NNS>fly<VBP> .<SENT>

(3) After Stemming and removing stop words

[0] ငှက် [1]ပျံ

[0]bird<n> [1]fly<v> [2].<SENT>

(4) After Align

[0]ငှက်များ/birds

[1]ပျံကြသည်/fly

6. Conclusion

The main goal of word alignment is to improve Myanmar-English translation. The main objective of this system is to provide students in learning English easily. The second objective is to build the standard system for Myanmar-English Bilingual Corpus and the standard translation system from Myanmar language to English. The proposed system is combination of corpus based approach and dictionary lookup approach. The corpus based approach is based on the first three IBM models.

7. References

[1] C. Callison-Burch, D. Talbot, and M. Osborne. 2004. Statistical machine translation with word- and sentence-aligned parallel corpora. In Proceedings of ACL, pages 175–182, Barcelona, Spain, July.

[2] F. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. Computational Linguistics, 29(1):19–52.

[3] G. Chinnappa and Anil Kumar Singh Language Technologies Research Centre, “A java implementation of an extended word alignment algorithm Based on the IBM models”,2007.

[4] Helen Langone, Benjamin R. Haskell, Geroge, A.Miller, “Annotating WordNet”, In Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL, 2004.

[5] Ittycheriah and S. Roukos. 2005. A maximum entropy word aligner for Arabic-English machine translation. In Proceedings of HLT-EMNLP. Vancouver, Canada. Pages 89–96.

[6] J. Martin, R. Mihalcea, and T. Pedersen. 2005. Word alignment for languages with scarce resources. In Proceedings of the ACL Workshop on Building and Using Parallel Texts. Ann Arbor, USA. Pages 65–74.

[7] Pascale Fung and Kenneth Ward Church. 1994. K-vec: a new approach for aligning parallel texts. In Proceedings of the 15th conference on Computational linguistics. Pages 1096-1102. Kyoto, Japan.

[8] P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase based translation. In Proceedings of HLT-NAACL. Edmonton, Canada. Pages 81–88.

[9] W. Gale and K. Church. , “A program for aligning sentences in bilingual corpora”. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL), Berkley, pp. 177–184, 1991.