Developing Word to Phrase Alignment for Myanmar-English Machine Translation

Khin Thandar Nwet
khinthandarnwet@ucsy.edu.mm

Abstract

Efficient estimation and alignment procedures for word and phrase alignment are developed for the alignment of parallel text, and this is shown in Myanmar-English translation. Word alignment based on the combination of corpus based approach and dictionary lookup approach. This word alignment process is based on the IBM model 1 to 3. For the dictionary lookup approach, the proposed system uses the bilingual Myanmar-English Dictionary. The system also uses a list of cognates and morphological analysis to get better alignment accuracy. After word alignment, phrase alignment is extended using PoS Pattern. Phrase alignment for machine translation is to improve quality in translation using PoS Filtering. Phrase alignment uses English Part-of-Speech (PoS).

Keywords: Word Alignment, Phrase Alignment, IBM Models

1. Introduction

Alignment is one of the central modeling problems in statistical machine translation (SMT). Given a collection of parallel text - text in one language accompanied by its translation in another language - the process of alignment identifies translation equivalence between documents, paragraphs, sentences, and, within sentences, between words and phrases [1], [8], [11], [12], [3], [4], [5].

Word and phrase alignment is the task of identifying correct translation relationships among words in a bilingual parallel corpus. Our focus is on word alignment of Myanmar-English data. In many cases, Myanmar and English show different word alignment. Consider the example reported in Figure 2. In Myanmar-English word alignment, there are multiple cross links, apart from one-to-one and one-to-many mappings.

These phenomena occurring here are due to the fact that in English the verb follows the subject, while in Myanmar the verb appears at the end of the sentence. This is only a simple example, but the characteristics of the two languages often yield too long-span word movements. This would give us a general idea about the possible complexities and challenges involved in this task. In order to capture such aspects of the translation in a general manner, an effective word alignment technique is required.

Word and phrase alignment has various end-users. It is an essential step for statistical machine translation [1][6][11]. Word and phrase aligned corpora are useful in automatic extraction of bilingual lexica and terminology [4]. Ambiguities in word sense are distributed differently in language translation. Therefore, word sense disambiguation is another area to be looked into in Statistical Natural Language Processing [3]. Word aligned corpora can also help in transferring language tools to new languages. Text analysis tools such as morphologic analyzers or POS tagger are

extended to languages where such resources do not exist [4]. Ways of exploiting statistical word alignment for grammar induction are given [5]. Many NLP applications are enhanced when word alignment is of better-quality [11].

In this paper, we propose a word to phrase alignment using Part-of-Speech tag patterns filter for machine translation.

The remainder of the paper is formed as follows. Section 2 describes word alignment in Myanmar-English corpus. Phrase alignment with PoS filtering is described in section 3. In section 4, we present experimental results and discussion and this results effect on machine translation. Finally, conclusion and future work is presented.

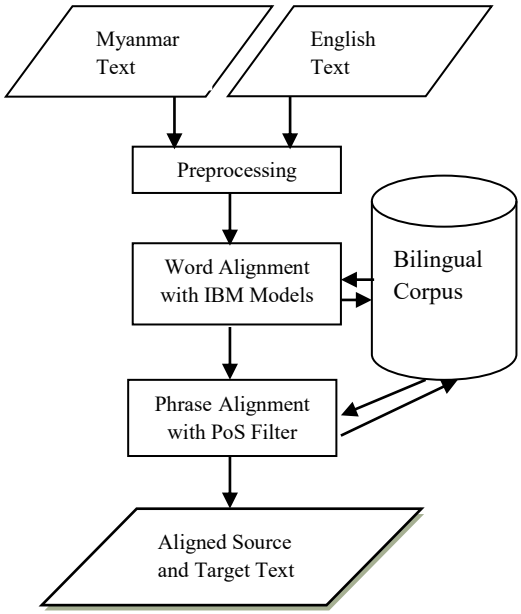


Figure 1 Proposed System

2. Word Alignment

One of the central modeling problems in machine translation is alignment between parallel texts. The duty of alignment methodology is to identify translation equivalence between sentences, words and phrases within sentences. Most current SMT systems [5, 13] use a generative model for word alignment such as the one implemented in the freely available tool GIZA++ [14]. The proposed system reduces the number of inappropriate alignments. Myanmar language is SOV (subject-object-verb) and English is SVO (subject-verb-object) language. Frequently, it is difficult for a human to judge which words in a given target string correspond to which words in its source string. Especially problematic is the alignment of words within idiomatic expressions, free translations, and missing function words. The problem is that the notion of "correspondence" between words is subjective. It is important to keep this in mind in the evaluation of word alignment quality.

Paper [9-10] presented word alignment based on the combination of corpus-based approach and dictionary lookup approach. For the dictionary lookup approach, the bilingual Myanmar-English Dictionary is used. Based on the combination of these proposed approaches, Myanmar-English word-aligned parallel corpus is also proposed to apply in the machine translation and word sense disambiguation. The word alignment process is done by using IBM word alignment models and it produces the possible aligned words. Our combined approach illustrated in the following stages.

Step 1: Accept pair of well-formed Myanmar and English sentences.

Step 2: English is well-developed, and there are many freely available resources for that language. English sentence is passed to Parser using TreeTagger [15] and it will produced Part-of-speech tagged output and root word output.

Step 3: Segment the words in Myanmar sentence using Longest Matching method[7], and remove the stop words. In this step, Myanmar sentence is morphological rich. Noun affixes "များ","တွေ" are . eg: ငှက်များ(birds), ငှက်တွေ(birds). Therefore, using Tri-Grams method,

analysis the noun and verb affixes (morphological analysis). Each word is calculated backward in Table 2.

Table 1 Mining Affixes

Verb Affixes	Adverb Affixes	Adjective Affixes	Noun Affixes
သည်	စွာ	သည့်	များ
ကြသည်	လျက်	သော	ခြင်း
ခဲ့သည်	ဆုံး	ဖွယ်	တွေ
...etc	...etc	...etc	...etc

Table 2 Examples N-gram for *စားခဲ့ကြသည်* Word

N gram (N=1,2,3)	Affixes
Unigram	သည်
Bigram	ကြသည်
Trigram	ခဲ့ကြသည်

Step 4: The output from Step 2 and Step 3 are aligned based on the first three IBM models using parallel corpus. The result from this step is the aligned words[10]. Translation probability, distortion probability and fertility are calculated by IBM models [1]. Then, candidate translation pair of greatest likelihood of connection is chosen.

Step 5: After Step 4, the remaining unaligned words are aligned using Myanmar-English bilingual dictionary (general domain). The lookup approach uses Myanmar root word and English PoS in the dictionary to get the English word. Parallel corpus is used as training data set and also the output of the system. In Figure 2, the last remaining unaligned words are Myanmar name "စုစု" and "Su Su" in English. Therefore, the last remaining unaligned words in Myanmar and English are aligned to solve Name Entity problem.

An example of the word-alignment result is as shown in Figure 2.

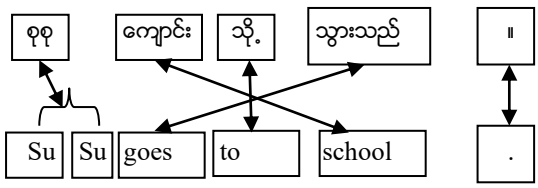


Figure 2. An Example of Word Alignment

3. Phrase Alignment with PoS Filtering

This paper focuses on acquiring translations of many to many alignment: contiguous sequences of a few words that encapsulate enough context to be translatable, but recur frequently in large corpora. This paper develops the word alignment to phrase alignment (many to many alignment) to improve machine translation accuracy using Part-of-Speech filter. In previous work [10], Myanmar words "ကော်ဖီ တစ်ခွက် " and English words "a cup of coffee " , "ကော်ဖီ " can be aligned with "coffee" and "တစ်ခွက်" can be aligned with "a cup" and " NULL " can be alignment with "of ". This paper develops to phrase alignment as shown in Figure 4. In table 1 for phrase alignment, Myanmar words "ကော်ဖီ တစ်ခွက် " can align with English words "

a cup of coffee" . This paper uses 20 tag patterns for phrase alignment. If the sentence patterns increased in this system, the tag patterns can be increased. A collocation is an expression consisting of two or more words that correspond to some conventional way of saying things. Collocations of a given word are statements of the habitual or customary places of that word. Collocations include noun phrases like *a beautiful girl*.

Collocations are characterized by limited compositionality. A natural language expression compositional is called if the meaning of the expression can be predicted from the meaning of the parts. Collocations are not fully compositional in that there is usually an element of meaning added to the combination. Collocations are important for a number of applications: natural language generation, computational lexicography, parsing and corpus linguistic research.

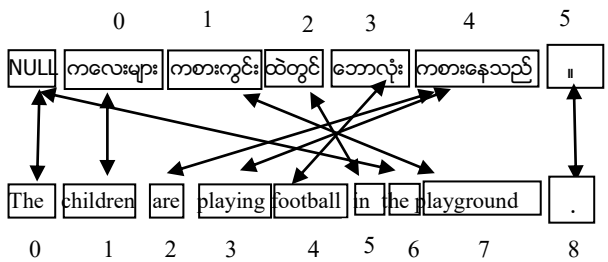
There are a number of approaches to finding collocations: selection of collocations M frequency, selection based on mean and variance of the distance between focal word and collocating word, hypothesis testing and mutual information. E

Surely the simplest method for finding collocations in a text corpus if counting. If two words occur together a lot, then that is evidence that they have a special function that is not simply explained as the function that results from their combination. In the proposed system, Part-of-speech filters which only lets through those patterns that are likely to be phrases. For English, this paper used Part-of-Speech tags are obtained with TreeTagger. When the sentence patterns increased, the Part-of-speech pattern will be improved. Justeson and Katz [3] suggest the patterns in Table 1. In these patterns DT refers to a determiner, JJ to an adjective, NN to a noun, IN to a preposition, RBS to an adverb and VBZ to verb. After applying the filter, the results are surprisingly good.

Table 3 Example Part of Speech Tag Patterns for Phrase Alignment

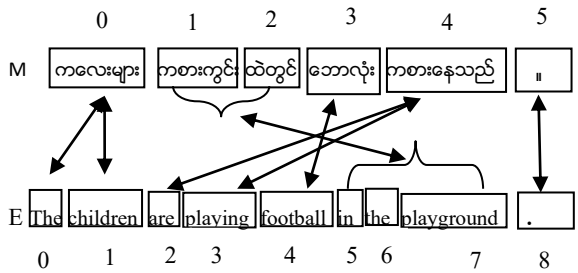
Tag Pattern	English	Myanmar
DT JJ NN	a beautiful flower	လှပသော ပန်း တစ်ပွင့်
DT NN IN NN	a cup of coffee	ကော်ဖီ တစ်ခွက်
IN DT NN	in the car	ကား ထဲတွင်
DT RBS JJ NN	the most beautiful city	အလှပဆုံး မြို့
PP$ NN	his nephew	သူ၏ တူ

Figure 3 becomes Figure 4 using part of speech tag pattern for phrase filtering such as (1, 2) in Myanmar sentence align with (5,6,7) in English sentence.



A {[NULL,0], [NULL,6], [0,1], [1,7], [2,5], [3,4], [4,(2,3)],[5,8]}

Figure 3. The Target Sentence E, Source Sentence M and one to one and one to many Word Alignment A



A {[0,(0,1)],[(1,2),(5,6,7)],[3,4],[4,(2,3) ,[5,8]]

Figure 4. The Target Sentence E, Source Sentence M and Phrase Alignment A

4. Experimental Results and Discussion

Data prepared for experimentation was 3,000 parallel sentences. This 3,000-sentences corpus is collected from many different resources of bilingual texts (such as local newspaper, dictionaries, middle school, etc.) in selected fields such as sport, health and general. Sentence length is at least 2 words long.

In *k*-fold cross-validation, the original sample is randomly partitioned into *k* equal sized subsamples. Of the *k* subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $k-1$ subsamples are used as training data. The 10-fold cross-validation is commonly used [16]. The whole data was divided into ten partitions for tenfold cross-validation. Each nine partitions are joined to be used as training data, thus running the proposed system ten times. Thus, in each run, nine-tenth of the whole data is used for training.

The results are reported for all the ten experiments with proposed system. Table 4 shows the accuracies achieved, in all the ten experiments, and their average.

Table 4. Tenfold Cross Validation Accuracy

No. of Experiments	Word Alignment Accuracy (%)	Phrase Alignment Accuracy (%)
1	92.4	90.4
2	90.4	89.4
3	95.9	87.9
4	89.6	89.6
5	94.5	84.5
6	97.9	97.9
8	96.5	86.5
9	98.9	80.9
10	90.5	90.5

The performances are also evaluated by precision, recall, and F-measure.

$$P_{Word} = \frac{W_{correct}}{W_{Atotal}} \times 100(\%)$$

$$R_{word} = \frac{W_{correct}}{W_{Stotal}} \times 100(\%)$$

$$F = \frac{2PR}{P+R} \times 100(\%)$$

Where,

W_{Stotal} = Total words in test

$W_{correct}$ = Number of correctly aligned words

W_{Atotal} = Number of aligned words

P_{word} = Precision

R_{word} = Recall

F =F-measure

Experiment

Trained on: 5000 sentences (general pair sentences)
Tested on: 1600 pair sentences
Myanmar- English dictionary (general dictionary): 10,000 words approximately

Experimental Setups:
E1 is Corpus Based Approach
E2 is Corpus Based Approach + Dictionary Lookup Approach
E3 is Corpus Based Approach + Dictionary Lookup Approach + Morphological Analysis

E4 Phrase Alignment (using PoS Pattern)

Table 5. Experiment for Alignment

Experiment	E1	E2	E3	E4
Precision (%)	80	83	93	90
Recall (%)	82	92	92	88
F-measure (%)	81	90	92	88

Experiment E1, E2 and E3 evaluates for word alignment using two approaches. Experiment E3 is better than other experiments. Experiment E4 evaluates for phrase alignment using PoS pattern. This experiment is based on E3 word alignment.

Table 6. Experiment for Alignment impact on Translation

Experiment	Word Alignment [9]	Phrase Alignment (proposed)
Accuracy(%)	80	83

In word alignment, some English word such as " article, conjunction, preposition " cannot aligned with Myanmar words. The quality of the resulting translation greatly depends on the quality of the alignment accuracy. In phrase alignment, unaligned words can be reduced. Therefore, Phrase alignment can improve machine translation accuracy.

Evaluation and error analysis of machine translation output are important. The experiments show that the proposed combination approach achieves better performance than other approaches. The proposed system also fails to attach the subordinate conjunctions ('which', 'that'). Some subordinate conjunctions are not present in Myanmar Language. Since there is no determiner in Myanmar, determiners are only existed on the English side. Three main types of unaligned word are observed. The first unaligned word type, in evaluating the proposed system, some errors are found which are made since the segmentation step. In Myanmar, if a word is incorrectly segmented, the alignment result is also incorrect. Thus, the segmentation errors in Myanmar may change the word meaning, which in turn cause alignment errors. Myanmar Language has separated and Name entity recognition which is another research issue in Asian Language. The second unaligned word type, the person name, organization name in Myanmar word cannot align with English word. The third unaligned word type, Myanmar sentences are omitted sometimes prepositions. However, prepositions are existed in English side. Therefore, Myanmar word can be missed the alignment with English word. For example, " သူ ကျောင်း သွား သည်။ ", the word သို့ (to) can be omitted in Myanmar sentence but it is valid in Myanmar sentence. The word omission can affect the accuracy of the alignment process.

5. Conclusion and Future Work

The main goal of word alignment is to improve Myanmar-English machine translation. The proposed word alignment system to phrase alignment for machine translation can have better performance on sentence-based machine translation system. Since the proposed word alignment approach is based on a combination of corpus based approach and dictionary lookup approach, this research can generate correct alignment words [9]. Moreover, it can get the better result by using a list of cognates and morphological analysis. Phrase alignment uses frequency based Part-of-Speech filter. Therefore,

the proposed system can increase translation accuracy for Myanmar-English machine translation.

Since word and phrase aligned bilingual corpus is an essential step for further processing, the proposed approach helps for many natural language processing applications. The word-and phrase aligned corpus can also be used for other processes such as word sense disambiguation, machine translation and building bilingual dictionary. Improving the quality of alignment leads to the systems with the accurate translation and the better quality of output.

We aim to do this as part of a failure analysis of the algorithm in future. We also aim to improve our alignment results by using Part-of Speech information for the Myanmar texts. This can be used with other approaches such as selection based on mean and variance of the distance between focal word and collocating word, hypothesis testing and mutual information.

References

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

[2] C. Manning and H. Schütze, "Collocations", 1999

[3] Diab M., Resnik P. 'An Unsupervised Method for Word Sense Tagging Using Parallel Corpora.' In: Proc. of the Annual Meeting of the Association for Computational Linguistics. Philadelphia, PA, pp. 255–262, 2002.

[4] David Yarowsky G.N., Wicentowski, R. "Inducing Multilingual Text Analysis Tools via Robust Projection Across Aligned Corpora." In: Proc. of the 1st International Conference on Human Language Technology Research (HLT), pp. 161–168, 2001.

[5] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," Computational Linguistics, vol. 29, no. 1, pp. 19–51, 2003.

[8] Eknath Venkataramani, Deepa Gupta "English-Hindi Automatic Word Alignment with Scarce Resources", International Conference on Asia Language Processing, 2010.

[6] Gale, W. A. and Church, K. W. 'Identifying word correspondences in parallel texts'. In Proc. Fourth DARPA Workshop on Speech and Natural Language, pages 152–157. Morgan Kaufmann Publishers, Inc., 1991

[7] H. H. Htay, Kavi Narayana Murthy " Myanmar Word Segmentation using Syllable level Longest Matching", The 6th Workshop on Asian Languae Resources, 2008.

[8] Kuhn J. 'Experiments in Parallel-Text Based Grammar Induction.' In: Proc. of the 42th Annual Meeting of the Association for Computational Linguistics. Barcelona, Spain, pp. 470–477, 2004.

[9] K. T. Nwet, K. M. Soe, N. L. Thein," Word Alignment System for Myanmar-English Machine Translation", Proceeding of 9th International Conference on Computer Application, Myanmar, 5-6 May, 2011.

[10] K. T. Nwet, K. M. Soe, N. L. Thein, "Word Alignment System Based on Hybrid Approach for Myanmar-English Machine Translation", SICE Annual Conference 2011 September 13-18, 2011, Waseda University, Tokyo, Japan.

[11] Och F., Ney H. 'A Systematic Comparison of Various Statistical Alignment Models.' Computational Linguistics, 29(1), pp. 19–51, 2003.

[12] Smadja F.A., McKeown K.R., Hatzivassiloglou V. Translating Collocations for Bilingual Lexicons: A Statistical Approach. Computational Linguistics, 22(1), pp. 1–38, 1996.

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

[14] R. Mihalcea and T. Pedersen. 2003. An evaluation exercise for word alignment. In

Proceedings of HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond. Edmonton, Canada. Pages 1–6.
[15] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[16] McLachlan, Geoffrey J.; Do, Kim-Anh; Ambroise, Christophe. “Analyzing microarray gene expression data. Wiley”, 2004.