Sentence and Word Alignment System for Myanmar-English Machine Translation

Khin Thandar Nwet, Khin Mar Soe, Ni Lar Thein
University of Computer Studies, Yangon
khin.thandarnwet@gmail.com

Abstract

This paper describes an alignment system that aligns Myanmar-English texts at the sentence and word level in parallel text. Sentence and word level alignment is the identification of the corresponding sentence and word in the parallel text. The paper is also to construct word-aligned parallel corpus to be able in Myanmar-English machine translation. It is also resource for machine translation. This paper describes a simple sentence length approach to sentence alignment and a combination of corpus based approach and dictionary lookup approach to perform word alignment. Corpus based approach is based on the first three IBM models and dictionary lookup approach uses Myanmar-English dictionary. The system also uses a list of cognates and morphological analysis to get better alignment accuracy.

Keywords*: Word alignment, statistical machine translation, IBM model, Myanmar-English Dictionary*

1. Introduction

Text alignment is not only used for the tasks such as bilingual lexicography or machine translation but also in other language processing applications such as multilingual information retrieval and word sense disambiguation. While resources like bilingual dictionaries and parallel grammars help to improve Machine Translation (MT) quality, text alignment, by aligning two texts at various levels (i.e. documents, sections, paragraphs, sentences and words), helps in the creation of such lexical resources (Manning & Schütze, 2003)[7].

If it makes too many errors in paragraph alignment, which is a rare case, it gives continuous blocks of wrong alignment beads. This paper presents a simple sentence length approach to align Myanmar-English sentences and a combination of corpus based approach and dictionary lookup approach to align word. Alignment can be roughly categorized into five levels: paragraph, sentence, phrase, word and character levels.

The remainder of the paper is formed as follows. Section 2 describes some related work. Sentence segmentation is presented in section 3. Section 4 discuss about sentence alignment. In section 5, we describe word alignment model. The proposed system is discussed in section 6. In section 7 and 8, we present testing results and experimental results. Finally, section 9 presents conclusion and future work.

2. Related Work

In this section, previous works in sentence and word alignment for statistical machine translation different languages are reviewed. Various researchers have improved the quality of statistical machine translation system by using different methods on different language. Length-based approaches are computationally better, while lexical methods are more resource hungry. Brown et al. and Gale and Church are amongst

the most cited works in text alignment work. Purely length-based techniques have no concern with word identity or meaning and as such are considered knowledge-poor approaches. The method used by Brown et al.[2] measures sentence length in number of words. Their approach is based on matching sentences with the nearest length. Gale and Church [4] used a similar algorithm, but measured sentence length in number of characters. Their method performed well on the Union Bank of Switzerland (UBS) corpus giving a 2% error rate for 1:1 alignment. Hla Hla Htay[6] used Gale & Church method and obtained an alignment accuracy of about 90%. G. Chinnappa and Anil Kumar Singh [5] 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. R. Harshawardhan, Mridula Sara Augustine[10] proposed the new objective function defined is tested for obtaining optimal alignment for English-Tamil translation pair. This alignment is necessary for creating the probabilistic bilingual dictionary and is also required for automatic machine translation. They have used this objective function to align words in 25 sentences of English-Tamil parallel corpora and is solved using the open source LP-Solver. Ahmet Mustafa Güngör[1] uses the location information of sentences and paragraphs as well as the lengths of them for aligning the bilingual texts. When the paragraph alignment is successful, if the text is easy (90% 1-1 beads) it has 96.1% accuracy.

3. Sentence Segmentation

In Myanmar script, we have "။" as a unique sentence boundary marker. Therefore segmenting paragraphs into sentences is trivial. In case of English language, however, detecting sentence boundary is not entirely trivial. Even though there are explicit sentence boundary markers such as the period(.), the question mark (?) and the exclamation mark(!), the same symbols can be used for other purposes.

4. Sentence Alignment

Sentence alignment is the task of finding correspondences of sentences in one language and another. It is a first step before the more ambitious task called word alignment. In our method, we use the length based method as well as the lengths of them for aligning the bilingual texts. This method is quite easy to implement and independent of the languages of the bilingual texts. The align program is based on a very simple statistical model of word lengths. The model makes use of the fact that longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences. However if it makes too many errors in paragraph alignment, which is a rare case, it gives continuous blocks of wrong alignment beads.

Sentence alignment techniques vary from simple character-length or word-length techniques to more sophisticated techniques which involve lexical constraints and correlations or even cognates (Wu 2000)[11]. The sentence alignment algorithm takes as input a pair of aligned paragraph. The output will be two separate aligned files with line to line correspondence.

4.1 Sentence Alignment Approaches

In the task of sentence alignment there are many papers proposing different methods but as far as the methodology we use is considered, we can group these approaches into 3 classes: length-based approaches, location-based approaches and lexical approaches.

4.1.1) Length-Based Approaches: In these approaches, content of the text in terms of semantics is not considered. These approaches use statistical methods for the task of alignment. In other words, they only consider the length of sentences while making the decision for alignment. Short sentences match with short sentences, long sentences match with long sentences. Despite their simplicity, these methods have very high accuracy. They are especially useful between texts in similar languages such as German, English and French.

4.1.2) Location-Based Approaches: These approaches resemble the length-based approaches in respect that location-based approaches are based on statistical information. They use the fact that most of the times, beads of sentences in the two texts have similar positions. For example, if a sentence in source text is in the middle of the text, its conjugate in the target text is probably in the middle of text too.

4.1.3) Lexical Approaches: These methods take into account the lexical information about texts. For example, in most of them a bilingual corpus is used to match the content words in one text with their correspondences in the other text and use these matches as anchor points in the sentence alignment process. In some methods, instead of these content word pairs cognates (words in language pairs that resemble each other phonetically, ex. doctor-doktor) are used for determining the beads of sentences.

4.1.1 Length-Based Approaches

Goal: Find alignment A with highest probability given the two parallel texts S and T.

$$\max_A P(A, S, T)$$

S: source text, T: target text, A: alignment

- To estimate the probability above, aligned text is decomposed in a sequence of aligned sentence beads where each bead is assumed to be independent of others.
- The question is determining the right formula and parameters for estimating the probability of a certain type of alignment bead such that the sentences in that bead are given.

5. Word Alignment

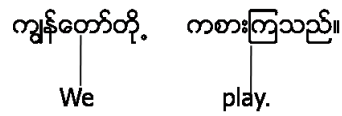
Extending sentence alignment to word alignment is a process of locating corresponding word pairs in two languages. In some cases, a word is not translated, or is translated by several words. A word can also be a part of an expression that is translated as a whole, and therefore the entire expression must be translated as a whole (Manning & Schütze). The word alignment algorithm takes as input a pair of aligned sentences and groups words in sentences of both languages. We have observed a few facts about the Myanmar Languages. Since there are no determiners and subordinate conjunctions in Myanmar, determiners are aligned to null.

5.1. Problem Statements and Solutions

In approaches based on IBM models, the problem of word alignment is divided into several different problems.

The first problem: is to find the most likely translations of a source word, irrespective of positions.

Solution: This part is taken care of by the translation model. This model describes the mathematical relationship between two or more languages. The main thing is to predict whether expressions in different languages have equivalent meanings. For example:

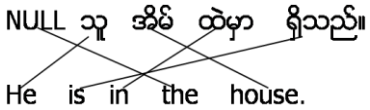


Translation (one to one alignment)

The second problem: is to align positions in the source language (SL) sentence with positions in the target language (TL) sentence.

Solution: This problem is addressed by the distortion model. It takes care of the differences in word orders of the two languages. A novel metric to measure word order similarity (or

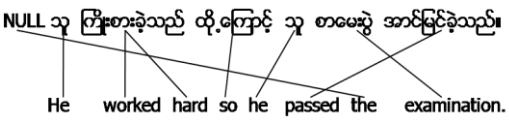
difference) between any pair of languages based on word alignments. For example:



Distortion (word order) and NULL Insertion (spurious words)

The third problem: is to find out how many TL words are generated by one SL word. Note that an SL word may sometimes generate no TL word, or a TL word may be generated by no SL word (NULL insertion).

Solution: The fertility model is supposed to account for this. For example:



Fertility (one to many alignment)

6. Proposed Alignment Model

This system consists of the following steps:

Step 1: Accept pair of 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 and it will produced Part-of-speech tagged output and root word output.

Step 3: Segment the words in Myanmar sentence using Myanmar Stop word list file, and remove the stop words. In this step, Myanmar sentence is morphological rich. After that, using Tri-Grams method, analysis the noun and verb affixes (morphological analysis). Each sentence is calculated backward.

Step 4: The output from Step 2 and Step 3 are aligned based on the first three IBM models and EM algorithm using parallel corpus. The result from this step is the aligned words. The high probability words are taken to insert to Parallel Corpus.

Step 5. After Step 4, the remaining unaligned words are aligned using Myanmar-English bilingual dictionary. 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.

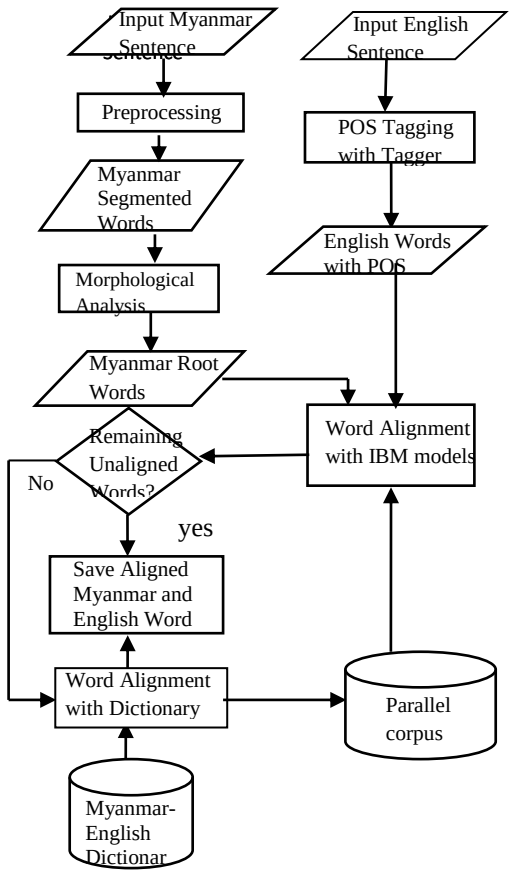


Fig. 1 Proposed Alignment System

The proposed system is combination of corpus based approach and dictionary lookup approach. The following sections explain each approach.

6.1 Corpus Based Approach

The corpus based approach is based on the first three IBM models.

6.1.1 The IBM Alignment Models 1 through 3

In their systematic review of statistical alignment models (Och and Ney ,2003[3]), Och

and Ney describe the essence of statistical alignment as trying to model the probabilistic relationship between the source language string m, and target language string e, and the alignment a between positions in m and e. The mathematical notations commonly used for statistical alignment models follow.

$$m^J_1=m_1,\ldots m_j,\ldots,m_J \qquad e^I_1=e_1,\ldots,e_i,\ldots,e_I \tag{1}$$

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

$$A \subseteq \{(j,i) : j = 1,\ldots,J; i = 1,\ldots,I\} \qquad j \rightarrow i = a_j \qquad i = a_j \tag{2}$$

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

$$\Pr(e_1^I \mid m_1^J) \qquad \Pr(e_1^I, a_1^I \mid m_1^J) \qquad \Pr(e_1^I \mid m_1^J) = \sum_{a_i^J} \Pr(e_1^I, a_1^I \mid m_1^J) \tag{3}$$

From the basic translation model $Pr(m_1^J|e_1^J)$, 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, $Pr(m_1^J,a_1^J|e_1^J)$. 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 this model, 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. Given source and target lengths (M,L), probability that i^{th} target word is connected to j^{th} source word. the distortion probability is given as D(i | j, l , m).The best alignment can be calculated as follow:

$$a_{j=1}^{m}[i,j,l,M] = \arg\max_i \; d(i \mid j,M,l) * t(e_i \mid m_j) \tag{4}$$

Model 3

Languages such as Swedish and German make use of compound words. Myanmar language also makes use of compound words. 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 word into two English words. Model 3 however introduces fertility based alignment, which considers such one to many translations probable. We uniformly assign the reverse distortion probabilities for model-3. Given source and target lengths(l,M), probability that i^{th} target word is connected to j^{th} source word. The best alignment can be calculated as follow:

F($|m\phi$) = probability that m is aligned with target words.

$$a_{j=1}^{m}[i,j,l,M] = \arg\max(D_i \mid j,l,M) \times T(e_i \mid m_j) \times D_{rev}(j \mid i,l,m) \times F(\phi_i \mid m_j) \tag{5}$$

```
for j=1 to M do
        set total to 0
  for i=1 to L do
        total += T(ei|mj)
      for i=1 to L do
        tc(ei|mj) += T(ei|mj)/total   (IBM 1to 3)
  end for
```

```
    end for
end for
```

Fig.2 Translation Algorithm Based on IBM Models

6.1.2 Dictionary Lookup Approach

Since bilingual dictionaries contain base forms, the system pre-process the text to find the base form for each word. So, this system uses part-of-speech tagger TreeTagger to obtain POS-tags and based form for English sentences and morphological analysis for Myanmar sentences. Morphological analysis is based on N-gram method.

We have used dictionary (bilingual Myanmar-English dictionary) which consists of 30,000 word to word translations. The dictionary lookup approach algorithm for alignment is as below:

```
Let ME be the set of English Meanings based on
Myanmar word and its POS.
For each Myanmar word
  Begin
    Find ME in Myanmar-English Dictionary
        If |ME|>1 then
        Match each meaning in ME with the input
        English word
        If the matching is found then
        Align these two words and
        Store these two words in corpus
        End if
        End if
  End
```

Fig. 3 Dictionary Lookup Algorithm

Both approaches can make alignment based on the exact match of two words. Sometimes, the words can be in varying morphological forms. Thus, the proposed approach considers to use morphological analysis to improve alignment.

6.1.3 Morphological Analysis

Unlike European languages, most of the Myanmar languages are morphologically rich and have the feature of compounding, thereby making the problem different in terms of SMT. For better word alignment of text in Myanmar languages, information about Morphological analysis is certainly needed. Affixes mining is the important task of morphological analyzer in NLP application such as same stem decision translate from one language to the cross-language, classify the word type from any language etc. In English, if we have the words governed, governing, government, governor, governs, and govern in that corpus, **govern** is (stem) verb and affixes are **ing, s, ment**, or but all affixes are not verb affixes. Because if **govern** and **ment** are combine, government is became but is not Verb. This is Noun. Thus, every combination of verb and affixes are not verb affixes. So, we uses part-of-speech tagger TreeTagger to obtain POS-tags and based form for English sentences.

In the same way, Myanmar language can be mined verb affixes and noun affixes from any Myanmar sentences. Noun affixes are များ, တွေ. eg: ကြောင်များ (birds), ကြောင်တွေ (birds). Myanmar morphological analysis is based on N-gram method calculated by backward. Examples of Verb affixes are shown in Table 1.

Table 1. Mining Affixes from Various Patterns of Verb

Various Patterns of Verb	Verb affixes	English Word	English Root Word
စားသည်။ စားကြသည်။ စားခဲ့သည်။ စားနေသည်။ ..etc	သည်။ ကြသည်။ ခဲ့သည်။ နေသည်။	eat eat ate eating	eat eat eat eat

eg: In စားသည်။, စား is stem and သည်။ is affix and in စားခဲ့သည်။, စား is stem and ခဲ့သည်။ are affixes and they all are verb affixes. The proposed system can analyze the noun and verb affixes (morphological analysis) using trigram method. Each sentence is calculated backward. We will

extract affixes from these sentences by using N-Grams method.

7. Testing Result

This system is tested based on corpus based and dictionary lookup approach. The alignment step by step is described as follows;

Input သူတို့ ကျောင်း သို့ သွားသည်။

They go to school.

Table 2. Index of Input Sentences

Source Words	Word Index	Target Words	Word Index
သူတို့	0	They	0
ကျောင်း	1	go	1
သို့	2	to	2
သွားသည်	3	school	3

Table 3. Calculate Probability for each Word

သူတို့

e	T(e|m)
they	**0.74074**
go	0.03703
to	0.07407
school	0.03703

ကျောင်း

e	T(e|m)
they	0.03333
go	0.06667
to	0.03333
school	**0.83333**

သို့

e	T(e|m)
they	0.03703
go	0.07407
to	**0.7778**
school	0.03703

သွား

e	T(e|m)
they	0.02857
go	**0.9142**
to	0.0285
school	0.05714

Then, alignment process iteratively refines the translation probabilities until values are good enough. The alignment values can be calculated by looking at the individual translation probability values. The best alignment can be calculated in a quadratic number of steps equal to *(sl+1)×tl*. 1 is used to add for the NULL value. For example for the above sentence pairs,

sl=4, *tl*=3; (*sl* + 1) x *tl* = (4+1) x 3 =15 steps,

where *sl=source sentence's length, tl=target sentence's length.*

The best alignment is shown in following table 4:

Table 4. Output Alignment Table

[0]သူတို့	[0]They [PP]
[1]ကျောင်း	[3]school[NN]
[2]သို့	[2]to[TO]
[3]သွားသည်	[1]go[VBP]

8. Experimental Results

We used the Myanmar-English corpus (1000 sentence pairs). We tested only on sentences which were at least 2 words long and used Zawgyi-one Myanmar font. We report the performance of our word alignment Models in terms of precision, recall and F-measure are defined as:

$$\text{Recall} = \frac{\text{Number of correctly aligned words}}{\text{Number of all words}} \times 100(\%)$$

$$\text{Precision} = \frac{\text{Number of correctly aligned words}}{\text{Number of aligned words}} \times 100(\%)$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100(\%)$$

Experiment

Trained on: 1000 sentences

Tested on: 250 sentences

S1 is Corpus based approach

S2 is Corpus based approach +Morphological analysis

S3 is Corpus based approach + Morphological analysis + Bilingual Dictionary

Table 5. Results for One to One Alignment Experiment

Experiment	S1	S2	S3
Precision (%)	80	89	95
Recall (%)	82	92	96
F-measure (%)	81	90.5	95.49

Table 6. Results for One to many Alignment Experiment

Experiment	S1	S2	S3
Precision (%)	79.34	87.23	90.72
Recall (%)	73	82	88
F-measure (%)	76.03	84.53	92.28

9. Conclusion and Future Work

In this paper, we proposed to align Myanmar-English texts at the sentence and word level. The main goal of word alignment is to improve statistical Myanmar-English machine translation. The sentence alignment is based on the length-based approach. Since the proposed word alignment approach is based on corpus based and dictionary based approaches, this system can generate correct alignment words. Most of the Myanmar languages are morphologically rich. Adding morphological processing improved translation results in both directions for both text types.

In future, we will work on many to many word alignments and have to test the algorithm for large bilingual corpora. The interested person can modify this system in order to apply for corpus size by using bilingual dictionary to improve alignment accuracies. Complex sentences can be extended. We can get better results with good accuracy; we have to test the algorithm for large bilingual corpora; the model can be also extended to multilingual word alignment.

10. References

[1] Ahmet Mustafa Güngör, "Turkish-English Sentence Alignment", Submitted to the Department of Computer Engineering in the Faculty of Engineering as CMPE 492 Senior Project, 2006.

[2]Brown, P., Lai, J. C., and Mercer, R., 1991, Aligning Sentences in Parallel Corpora, In Proceedings of ACL-91, Berkeley CA.

[3] F. Och and H. Ney, “A Systematic Comparison of Various Statistical Alignment Models”. Computational Linguistics, 29(1):19–52, 2003.

[4] Gale, W.A. and Church, K.W.: A program for aligning sentences in bilingual corpora, In Proc. of the 29th Annual Meeting of the ACL (1991) 177-184.

[5] G. Chinnappa and Anil Kumar Singh, “A java Implementation of an Extended Word Alignment Algorithm Based on the IBM Models”, 2008.

[6] H. H. Htay, G. Bharadwaja Kumar, Kavi Narayana Murthy, "Constructing English-Myanmar Parallel Corpora", In processing of ICCA 2006: International Conference on Computer Applications, Yangon, Myanmar, February 2006.

[7] Manning C. and Schütze H., 2003, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, Massachusetts.

[8] Myanmar-English dictionary, Department of the Myanmar Language Commission, Ministry of Education, Union of Myanmar.

[9] Neeraj Aswani and R. Gaizauskas. ‘A Hybrid Approach to Align Sentences and Words in English-Hindi parallel corpora’. In Proceedings of the ACL Workshop on ”Building and Exploiting Parallel Texts”, 2005.

[10] R. Harshawardhan, Mridula Sara Augustine, " A Simplified Approach To Word Alignment Algorithm For English-Tamil Translation", Indian Journal of Computer Science and Engineering (IJCSE), Vol. 2 No. 1, 2010.

[11] Wu, D.: Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico (1994) 80–87.