Developing Word-aligned Myanmar-English Parallel Corpus based on the IBM Models

Khin Thandar Nwet
University of computer Studies, Yangon, Myanmar

Khin Mar Soe
University of computer Studies, Yangon, Myanmar

Ni Lar Thein
University of computer Studies, Yangon, Myanmar

ABSTRACT
Word alignment in bilingual corpora has been an active research topic in the Machine Translation research groups. Corpus is the body of text collections, which are useful for Language Processing (NLP). Parallel text alignment is the identification of the corresponding sentences in the parallel text. Large collections of parallel level are prerequisite for many areas of linguistic research. Parallel corpus helps in making statistical bilingual dictionary, in supporting statistical machine translation and in supporting as training data for word sense disambiguation and translation disambiguation. Nowadays, the world is a global network and everybody will be learned more than one language. So, multilingual corpora are more processing. Thus, the main purpose of this system is to construct word-aligned parallel corpus to be able in Myanmar-English machine translation. One useful concept is to identify correspondences between words in one language and in other language. The proposed approach is based on the first three IBM models and EM algorithm. It also shows that the approach can also be improved by using a list of cognates and morphological analysis.

General Terms
Natural Language Processing, Machine Translation

Keywords
Word-aligned Parallel Corpus, IBM Models, EM Algorithm

1. INTRODUCTION
In the writing system of many Asian languages, such as Myanmar, Japanese, and Thai, words are not delimited by spaces[12]. There are no blanks in Myanmar text for word boundary. 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. This system 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 [7] 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.

The remainder of the paper is formed as follows. Section 2 describes some related work. Segmenting Myanmar sentence into words is presented in section 3. Overview of statistical machine translation for Myanmar to English is presented in section 4. Section 5, discuss about IBM alignment models. In section 6, we describe proposed alignment model. Mining of noun and verb affixes are presented in section 7. Section 8 explains parallel corpus. In section 9, we present experimental results. Finally, conclusion and future work is presented.

2. RELATED WORK
A vast amount of research has been conducted in the alignment of parallel texts with various methodologies. 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. Li and Chengqing Zong [11] addressed the word alignment between sentences with different valid word orders, which changes the order of the word sequences (called word reordering) of the output hypotheses to make the word order more exactly match the alignment reference.

K-vec algorithm [13] makes use of the word position and frequency feature to find word correspondences using Euclidean distance. Ittycheriah and Roukos [8] proposed a maximum entropy word aligner for Arabic-English machine translation. Martin et al. [9] have discussed word alignment for languages with scarce resources. R. Harshawardhan , Mridula Sara Augustine and Dr K. P. Soman [17] proposed a simplified approach to word alignment algorithm for English-Tamil translation. The word alignment problem is viewed as a simple assignment problem and is formulated as an Integer Linear Programming problem. Jamie Brunning, Adria de Gispert and William Byrne proposed Context-Dependent alignment models for statistical machine translation [10]. This models lead to an improvement in alignment quality, and an increase in translation quality when the alignments are used in Arabic-English and Chinese-English translation.

Most current SMT systems [14] use a generative model for word alignment such as the one implemented in the freely available tool GIZA++ [16]. GIZA++ is an implementation of the IBM alignment models [15]. These models treat word alignment as a hidden process, and maximize the probability of the observed (e, f) sentence pairs using the Expectation Maximization (EM) algorithm, where e and f are the source and the target sentences. In [3] all the conducted experiments prove that the augmented approach, on multiple corpuses, performs better when compared to the use of GIZA++ and NATools individually for the task of English-Hindi word alignment. D.Wu, (1994) [2] has developed Chinese and English parallel corpora in the Department of Computer Science and University of Science and Technology in Clear Water Bay, Hong Kong. Here two methods are applied which are important once. Firstly, the gale's methods is used to Chinese and English which shows that length-based methods give satisfactory result even between unrelated languages which is a surprising result. Next, it shows the effect of adding lexical cues to a length – based methods. According to these results, using lexical information increases accuracy of alignment from 86% to 92%.

3. SEGMENTING MYANMAR SENTENCE INTO WORDS

Myanmar language is mother language of all Myanmar national people. This is the mother tongue for Myanmar people who speak Myanmar language as their native tongue. Myanmar writing has developed onward from 12th century to date. Myanmar sentences are clearly eliminated by a unique sentence boundary maker "။" which is called " ။" [pou ma.]. Myanmar language is like Chinese, Japanese, India, and Thailand and so on in Asian Languages. The words are not separated by the space. Therefore, i is considerable more difficult than for Western Languages. there is no character based or word based n-grams. It has only syllable level. The syllabification is done by using Myanmar Word Segmentation using Syllable level Maximum Matching [18].

N-Gram	Phrases
Unigram	ခွေး၊ များ၊ သည်၊ အသား၊ စား၊ သည်
Bigram	ခွေးများ၊များသည်၊သည်အသား၊အသားစား၊စားသည်
Trigram	ခွေးများသည်၊များသည်အသား၊သည်အသားစား၊ အသားစားသည်

Fig 1: N-gram based Phrases

4. OVERVIEW OF THE STATISTICAL MACHINE TRANSLATION OF MYANMAR TO ENGLISH

Figure 2 shows overview architecture of the statistical machine translation of Myanmar to English. The source language model includes Part-of-Speech (POS) tagging and finding grammatical relations. The translation model includes phrase extraction, translation by using bilingual Myanmar to English corpus. The translation model also interacts with WSD (Word Sense Disambiguation) to solve ambiguities when a phrase has with more than one sense. The target language model includes reordering the translated English sentence and smoothing it by reducing grammar errors. In this Myanmar to English machine translation system, we focus on Alignment model. The main goal is to construct Myanmar-English word-aligned parallel corpus. Alignment model is central components of any statistical machine translation system. The result corpus will be used in most parts of the Myanmar-English machine translation.

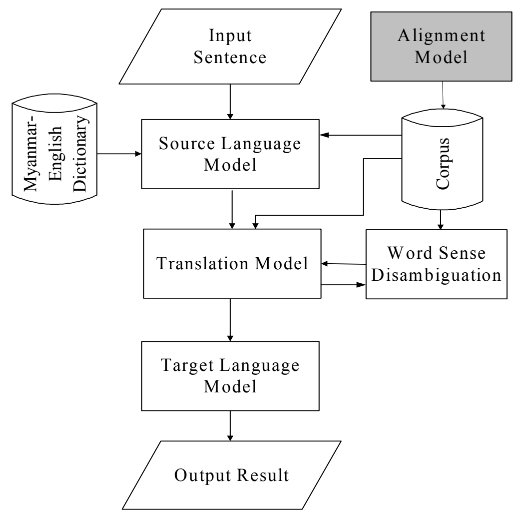


Fig 2: Machine Translation System of Myanmar- English

5. ALIGNMENT MODEL

Essential for building parallel corpora is the alignment of translated segments with source segments. 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).

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.

5.1 The IBM Alignment Models 1 through 3

In their systematic review of statistical alignment models (Och and Ney ,2003[4]), 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.

$$\boxed{\begin{array}{c} m^J_1=m_1,\ldots m_j,\ldots,m_J \\ e^I_1=e_1,\ldots,e_i,\ldots,e_I \end{array}} \qquad (1)$$

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).

$$\boxed{\begin{array}{l} A \subseteq \{(j,i) : j = 1,\ldots,J; i = 1,\ldots,I\} \\ j \rightarrow i = a_j \\ i = a_j \end{array}} \qquad (2)$$

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

$$\boxed{\begin{array}{l} \Pr(e_1^I \mid m_1^J) \\ \Pr(e_1^I, a_1^I \mid m_1^J) \\ \Pr(e_1^I \mid m_1^J) = \sum_{a_i^J} \Pr(e_1^I, a_1^I \mid m_1^J) \end{array}} \qquad (3)$$

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(l,M), 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:

$$\boxed{d^m_{j=1}[i,j,l,M] = \underset{i}{\operatorname{argmax}}\, d(i \mid j,M,l) * t(e_i \mid m_j)} \qquad (4)$$

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. 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:

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

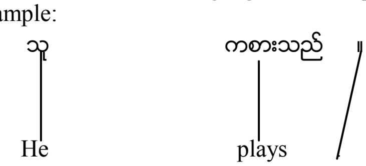
$$\boxed{\begin{array}{c} a^m_{j=1}[i,j,l,M] = \operatorname{argmax}(D_i \mid j,l,M) \times T(e_i \mid m_j) \times \\ D_{rev}(j \mid i,l,m) \times F(\phi_i \mid m_j) \end{array}} \qquad (5)$$

5.2 Problem Statement 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 an SL word, irrespective of positions.

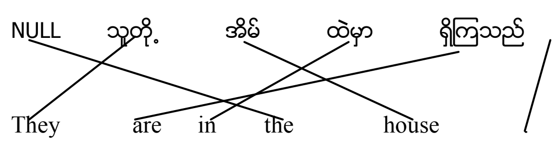
Solution: This part is taken care of by the translation model. This model describes the mathematical relationship between two or more languages. The main thing is to predict whether expressions in different languages have equivalent meanings. For example:



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:

NULL ကလေးများ ခုန်နေကြသည် ။

The children are jumping .

Fertility (one to many alignment)

6. PROPOSED ALIGNMENT MODEL

The proposed system is combination of corpus based approach and dictionary lookup approach. Alignment step uses corpus based approach as first and dictionary lookup approach. If the corpus has not enough data, the system uses 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 and Expectation Maximization (EM) algorithm. The Expectation-Maximization (EM) algorithm is used to iteratively estimate alignment model probabilities according to the likelihood of the model on a parallel corpus. In the Expectation step, alignment probabilities are computed from the model parameters and in the Maximization step, parameter values are re-estimated based on the alignment probabilities and the corpus. The iterative process is started by initializing parameter values with uniform probabilities for IBM Model 1. The EM algorithm is only guaranteed to find a local maximum which makes the result depend on the starting point of the estimation process. This system is implemented EM algorithm and deals with problem statements. The iterative EM algorithm corresponding to the translation problem can be described as:

Step-1: Collect all word types from the source and target corpora. For each source word m collect all target words e that co-occurs at least once with m.

Step-2: Initialize the translation parameter uniformly (uniform probability distribution), i.e., any target word probably can be the translation of a source word e. In this step, there are two main tasks for aligning the source and target sentences. The detail algorithm of each task is shown Figure 3 and Figure 4.The first task is pre-processing and the second task is the usage of the first three IBM models.

Pre-processing Phase

Accept Source Sentence;
Accept Target Sentence;
Remove Stop Word in Source Words (S) eg: သည်,ပြီး
For each Source Sentence S do
Separate into words;
Store Source Words Indexes;
End For
For each Target Sentence T do
Separate into words;
Store Target Words Indexes;
End For

Fig 3: Algorithm for Pre-processing

Step-1: Collect all word types from the source and target corpora.
For each source word *m* collect all target words *e* that co occurs at least once with *m*.
Step-2: Any target word (e) probably can be the translation of a source word (m) and the lengths of the source and target sentences are s and t, respectively.
Initialize the expected translation count Tc and Total to 0
Step-3: Iteratively refine the translation probabilities.
For i=1 to *s* do
Source Words with N-grams Method
Select Target Words FROM Bilingual corpus WHERE Source Similar m_i
total+=T(m_i) in corpus
For j=1 to t do
If e_j Found in Corpus
Tc($e_j|m_i$)+= T($e_j|m_i$)
Store Source Word Index and Target Word Index
Align Source Word and Target Word and Store in Corpus
Else if
Use the English Pattern (combine English words with N-grams method)
If T (m_i) with Target Word found in Corpus
Tc($e_j|m_i$)+= T($e_j|m_i$)
Store Source Word Index and Target Word Index
Align Source Word and Target Word and Store in Corpus
Else English Word with Null insertion
End If
End For
Calculate Probability T
End For

Fig 4.: The First Three IBM Models Based Algorithm

Myanmar Word	English Word
အိမ်	house
	home
	building
ရှိသည်	is
	exist
	are
	has
	have
ကျွန်း	island
	teak

Fig 5: Example of Ambiguity Words

6.2 Dictionary Lookup Approach

We have used dictionary (bilingual Myanmar-English dictionary) which consists of 10,000 word to word translations. The dictionary lookup approach algorithm for alignment is as below:

Let M_E be the set of English Meanings based on Myanmar word and its POS.
For each Myanmar word
  Begin
    Find M_E in Myanmar-English Dictionary
      If $|M_E|>1$ then
      Match each meaning in M_E 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 6: 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.

7. MINING OF NOUN AND VERB AFFIXES

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.

Having a list of salient affixes is not sufficient to parse a given word into stem and affix (es). For example, **sing** happens to end in the most salient suffix yet it is not composed of **s** and **ing** because crucially, there is no ***s, *sed** etc. Thus to parse a given word we have to look at additional evidence beyond the word itself, such as the existence of other inflections of potentially the same stem as the given word, or further, look at inflections of other stems which potentially share an affix with the given word [6].

In the same way, Myanmar language can be mined verb affixes and noun affixes from any Myanmar sentences. Noun affixes are များ, တွေ. eg: ငှက်များ (birds), ငှက်တွေ (birds). 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**
ကစားသည်။	သည်။	play	play
ကစားကြသည်။	ကြသည်။	play	play
ကစားခဲ့သည်။	ခဲ့သည်။	played	play
ကစားနေသည်။	နေသည်။	playing	play
……..etc			

eg: In ကစားသည်။, ကစား is stem and သည်။ is affix and in ကစားခဲ့သည်။, ကစား is stem and ခဲ့သည်။ are affixes and they all are verb affixes.

8. PARALLEL CORPUS

Corpora are useful for natural language processing (NLP). Corpus is the body of text collections. Most text corpus collections are so called corpora. Text corpora are usually big. Corpora/corpus can be

1. Raw corpora, which are just running text
2. Word segmented corpus in which the corpus is segmented in to word level
3. part-of-speech tagged corpus in which each word in the corpus is tagged with its respective part-of-speech
4. Semantic tagged corpus in which each word in the corpus is tagged with the sense in the dictionary along with its respective part- of –speech
5. Parallel corpus in which sentence or sentences of a language are aligned with the equivalent translation of another language
6. Multilingual corpus in which sentence or sentences of a language are aligned with the equivalent translation of more than one language, etc.

Parallel corpus consists of one or more languages. Parallel corpus is the most convenient for NLP and other research. Parallel corpus helps in making bilingual dictionary, machine translation and word sense disambiguation (WSD). Parallel corpus is pairs of translation alignment of bitext. One useful step is to identify correspondences between sentences in one language and in other language. Sentence alignment problem is one seeks to say that some group of sentences in one language corresponds in content to some group of sentences in the other languages. Such a grouping is referred to as a sentence alignment or bead.

Myanmar Sentence	English Sentence
ကျွန်တော်တို့ ဘုရား သို့ သွားသည်။	We go to pagoda.
သူတို့ မှာ အိမ် တစ်လုံး ရှိသည်။	They have a house.
ကလေးများ ခုန်နေကြသည် ။	The children are jumping.
သူ အလွန် အရပ်မြင့်သည်။	He is very tall.

Fig 7: Input Myanmar and English Sentence

[0]ကျွန်တော်တို့/[0]We<PP> [1]ဘုရား/[3]pagoda<NN> [2]သို့/[2]to<TO> [3]သွားသည်/[1]go<VBP>

[0]သူတို့/[0]They[PP] [1]အိမ်/[3]house[NN] [2]တစ်လုံး/[2]a[DT] [3]မှာရှိသည်/[1]have[VBP]

[0]ကလေးများ/[1]children[NNS] [1]ခုန်နေကြသည်/[2]are jumping[VBG]

[0]သူ/[0]he[PP] [1]အလွန်/[2]very[RB] [3]အရပ်မြင့်သည်/[1]is tall[VBZ]

Fig 8: Bilingual Word-aligned Corpus Format

We first align the index of Myanmar word, Myanmar word and then followed by /, then the index of corresponding English word, English meaning and part of speech. Words are separated by tab. The sentences are aligned sentient by sentence.

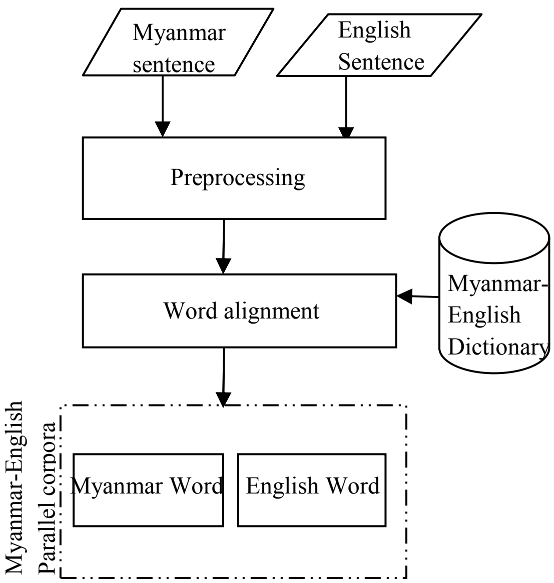


Fig 9: Proposed Alignment System

9. EXPERIMENTAL RESULT

This system used the Myanmar-English corpus (1000 sentence pairs) and 250 sentence pairs for testing. The sentences were at least 4 words long. We report the performance of our alignment Models in terms of precision and recall defined as:

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

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

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

According to the experimental results, it shows in Table 2. By using combination of Corpus based approach and dictionary lookup approach, the precision increased.

Experiment

A is Corpus based approach
B is Corpus based approach +Morphological analysis
C is Corpus based approach + Morphological analysis + Bilingual Dictionary

Table 2. Results for Experiment

Experiment	A	B	C
Precision (%)	80	89	95
Recall (%)	82	92	96
F-measure (%)	81	90.5	95.5

10. ACKNOWLEDGMENTS

I wish to offer my deep gratitude to my supervisor Professor Dr. Ni Lar Thein, Rector of the University of Computer Studies, Yangon and my co-supervisor Associate Prof. Dr. Khin Mar Soe, Computer Software Department, University of Computer Studies, Yangon for giving their valuable advice, helpful comments and precious time to my research.

11. CONCLUSION AND FUTURE WORK

The aligner also fails to attach the subordinate conjunctions ('which', 'that'). Some subordinate conjunctions are not present in Myanmar Language. Other examples of alignment errors are due to erroneous formatting. One fairly commonly occurring error is when the aligner does not attach the apostrophe (') with the following suffix (attached to proper nouns in Myanmar) to the English language due to tokenization problem. The main goal of word alignment is to improve statistical Myanmar-English machine translation. The second objective is to build the standard system for Myanmar-English parallel Corpus. Word alignments can have better performance on sentence-based SMT system. Since the proposed 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. This system uses Zawgyi-one Myanmar font. In future, we will work on many to many word alignments and have to test the algorithm for large bilingual corpora.

12. REFERENCES

[1] C. Callison-Burch, D. Talbot, and M. Osborne, “Statistical Machine Translation with Word- and Sentence-Aligned Parallel Corpora”. In Proceedings of ACL, pages 175–182, Barcelona, Spain, July 2004.

[2] D. Wu. “Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria” In: Proc. of the 32nd Annual Conference of the ACL: 80-87. Las Cruces, NM in 1994. http://acl.ldc.upenn.edu/P/P94/P94-1012.pdf

[3] E. Venkataramani and D. Gupta, "English-Hindi Automatic Word Alignment with Scarce Resources". In International Conference on Asian Language Processing, IEEE, 2010.

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

[5] G. Chinnappa and Anil Kumar Singh, "A java Implementation of an Extended Word Alignment Algorithm Based on the IBM Models". In Proceedings of the 3rd Indian International Conference on Artificial Intelligence, Pune, India. 2007.

[6] H. Hammarstrom, "Poor Man's Stemming: Unsupervised Recognition of Same-Stem Words". Chalmer University, 412 96 Gothenburg Sweden, 2007.

[7] H. Langone, Benjamin R. Haskell, Geroge, A.Miller, "Annotating WordNet", In Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL, 2004.

[8] Ittycheriah and S. Roukos, "A Maximum Entropy Word Aligner for Arabic-English Machine Translation". In Proceedings of HLT-EMNLP. Vancouver, Canada. Pages 89–96, 2005.

[9] J. Martin, R. Mihalcea, and T. Pedersen, "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, 2005.

[10] J. Brunning, A. de Gispert and William Byrne, "Context-Dependent Alignment Models for Statistical Machine Translation". The 2009 Annual Conference of the North American Chapter of the ACL, pages110–118, Boulder, Colorado, June 2009.

[11] Li and C. Zong, "Word Reordering Alignment for Combination of Statistical Machine Translation Systems", IEEE, 2008.

[12]"Myanmar Grammar", Department of the Myanmar Language Commission, Ministry of Education, Myanmar, 2005.

[13] P. Fung and K. Ward Church " K-vec: A New Approach for Aligning Parallel Texts". In Proceedings of the 15th conference on Computational linguistics. Pages 1096-1102. Kyoto, Japan, 1994.

[14] P. Koehn, F. J. Och, and D. Marcu, "Statistical Phrase based Translation". In Proceedings of HLT-NAACL. Edmonton, Canada. Pages 81–88, 2003.

[15] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L.Mercer "The Mathematics of Statistical Machine Translation: Parameter Estimation". Computational Linguistics, 19(2):263–311, 1993.

[16] R. Mihalcea and T. Pedersen, "An evaluation exercise for word alignment". In Proceedings of HLT-NAACL Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond. Edmonton, Canada. Pages 1–6, 2003.

[17] R. Harshawardhan , Mridula Sara Augustine and Dr K. P. Soman "A Simplified Approach to Word Alignment Algorithm for English-Tamil Translation". In Indian Journal of Computer Science and Engineering", 2008.

[18] W.P.Pa,N.L.Thein, "Disambiguation in Myanmar Word Segmentation",ICCA,February,2009.