ISSN 2229-5518

Building Bilingual Corpus based on Hybrid Approach for Myanmar-English Machine Translation

Khin Thandar Nwet

Abstract—Word alignment in bilingual corpora has been an active research topic in the Machine Translation research groups. In this paper, we describe an alignment system that aligns English-Myanmar texts at word level in parallel sentences. Essential for building parallel corpora is the alignment of translated segments with source segments. Since word alignment research on Myanmar and English languages is still in its infancy, it is not a trivial task for Myanmar-English text. A parallel corpus is a collection of texts in two languages, one of which is the translation equivalent of the other.Thus, the main purpose of this system is to construct word-aligned parallel corpus to be able in Myanmar-English machine translation. The proposed approach is combination of corpus based approach and dictionary lookup approach. The corpus based approach is based on the first three IBM models and Expectation Maximization (EM) algorithm. For the dictionary lookup approach, the proposed system uses the bilingual Myanmar-English Dictionary.

Index Terms— EM Algorithm, IBM Models, Machine Translation, Word-aligned Parallel Corpus, Natural Language Processing

1 INTRODUCTION

PROCESSING Myanmar texts is difficult in its computation because sentences in Myanmar texts are represented as strings of Myanmar characters without spaces to indicate word boundaries. This cause problem for Machine Translation, Information Retrieval, Text Summarization and many other Natural Language Processing. Bilingual word alignment is the first step of most current approaches to Statistical Machine Translation or SMT [2]. 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.

The remainder of the paper is formed as follows. Section 2 describes some related work. Alignment Model is presented in section 3. Section 4, discuss about Proposed Alignment Model. In section 5, we describe Overview of System. In section 6, we present experimental results. Finally, section 7 presents conclusion and future work.

• *Khin Thandar Nwet is currently pursuing Ph.D degree program in University of Computer Studies, Yangon, Myanmar, PH-0973176035 I got M.C.Sc from University of Computer Studies, Mandalay in 1997. I am also an assistant lecturer. My research interest is Natural language processing. Email: khin.thandarnwet@gmail.com*
.

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 [6] 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. Bing Xiang, Yonggang Deng and Bowen Zhou [1] proposed Diversify and Combine: Improving Word Alignment for Machine Translation on Low-Resource Languages. This approach on an English-to-Pashto translation task by combining the alignments obtained from syntactic reordering, stemming, and partial words. 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 [4] 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) [3] 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%.

A hybrid approach to align sentences and words in English-Hindi parallel corpora[12] presented an alignment system that aligns English-Hindi texts at the sentence and word level in parallel corpora. They described a simple sentence length approach to sentence alignment and a hybrid, multi-feature approach to perform word alignment. They use regression techniques in order to learn parameters which characterize the relationship between the lengths of two sentences in parallel text. They used a multi-feature approach with dictionary lookup as a primary technique and other methods such as local word grouping, transliteration similarity (edit-distance) and a nearest aligned neighbors approach to deal with many-to-many word alignment. Their experiments are based on the EMILLE (Enabling Minority Language Engineering) corpus. They obtained 99.09% accuracy for many-to-many sentence alignment and 77% precision and 67.79% recall for many-to-many word alignment.

3 ALIGNMENT MODEL

Essential for building parallel corpora is the alignment of tanslated 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.

3.1 IBM Alignment Models 1 through 3

In their systematic review of statistical alignment models (Och and Ney ,2003[5]), 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.

$$\begin{gathered} m^J_1 = m_1, \ldots m_j, \ldots, m_J \\ e^I_1 = e_1, \ldots, e_i, \ldots, e_I \end{gathered} \quad (1)$$

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).

$$\begin{aligned} & A \subseteq \{(j,i) : j = 1,\ldots,J; i = 1,\ldots,I\} \\ & j \rightarrow i = a_j \\ & i = a_j \end{aligned} \quad (2)$$

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

$$\boxed{\begin{aligned} &\Pr(e_1^I \mid m_1^J) \\ &\Pr(e_1^I, a_1^I \mid m_1^J) \\ &\Pr(e_1^I \mid m_1^J) = \sum_{a_1^J} \Pr(e_1^I, a_1^I \mid m_1^J) \end{aligned}} \quad (3)$$

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:

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

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:

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

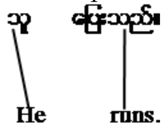
$$\boxed{\begin{aligned} a_{j=1}^{m}[i,j,l,M] = \arg\max (D_i \mid j,l,M) \times T(e_i \mid m_j) \times \\ D_{rev}(j \mid i,l,m) \times F(\phi_i \mid m_j) \end{aligned}} \quad (5)$$

3.2 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 an SL word, irrespective of positions.

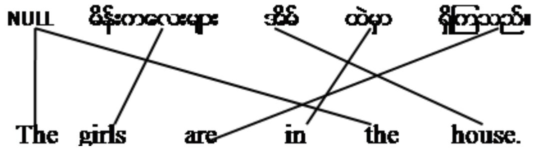
Solution: This part is taken care of by the translation model. This model describes the mathematical relationship between two or more languages. The main thing is to predict whether expressions in different languages have equivalent meanings. For example:



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.

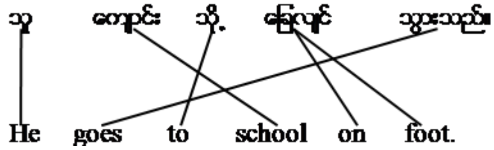
Solution: This problem is addressed by the distortion model. It takes care of the differences in word orders of the two languages. A novel metric to measure word order similarity (or difference) between any pair of languages based on word alignments. For example:



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:



Fertility (one to many alignment)

4 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.

4.1 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 1 and Figure 2. The first task is preprocessing 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: □ □ □,□ □ □ □
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

Figure 1. 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

Figure 2. The First Three IBM Models Based Algorithm

Myanmar Word	English Word
□ □ □ □ □	house
	home
	building
□ □ □ □ □ □	is
	exist
	are
	has
	have
□ □ □ □ □ □	island
	teak

Figure 3. Example of Ambiguity Words

4.2 Dictionary Lookup Approach

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

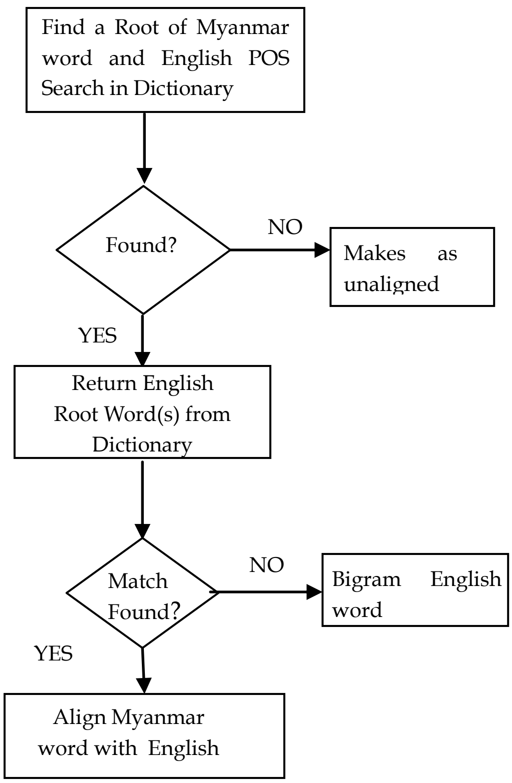


Figure 4. Dictionary Lookup Algorithm

[0] □ □/[0]he[PP] [1]□ □ □ □ □ □ □ □ □/[1]sat[VBD]
[0] □ □ □/[0]She[PP] [1]□ □ □ □ □ □ □ □/[1]ran[VBD]

[0] □ □ □ □ □/[0]They[PP] [1]□ □ □ □ □ □ □ □/[1]came[VBD]
[0] □ □ □ □ □/[1]boy[NN] [1]□ □ □ □ □ □ □ □/[2]ran[VBD]

[0]□ □ □ □ □ □ □ □/[1]boys[NNS]
[1]□ □ □ □ □ □ □ □ □ □/[2]ran[VBD]
[0] □ □/[0]He[PP] [1]□ □ □ □ □ □ □ □ □/[2]quickly[RB]
[2]□ □ □ □ □ □/[1]ran[VBD]
[0]□ □ □ □ □ □ □/[0]I[PP] [1]□ □ □ □ □ □/[2]slowly[RB]
[2]□ □ □ □ □ □ □ □ □ □ □/[1]walk[VBP]
[0] □ □/[0]He[PP] [1]□ □ □/[4]mat[NN]
[2]□ □ □ □ □ □/[3]a[DT] [3]□ □ □ □ □/[2]on[IN]
[4]□ □ □ □ □ □ □ □ □/[1]sat[VBD]
[0]□ □ □ □ □ □ □ □ □ □ □/[0]We[PP] [1]□ □ □ □ □/[4]tree[NN]
[2]□ □ □ □/[2]under[IN] [3]□ □ □ □ □ □ □ □/[1]play[VBP]
[0] □ □/[0]He[PP] [1]□ □ □/[1]is fat[VBZ]
[2]□ □ □ □ □ □ □/[3]but[CC] [3]□ □ □/[4]she[PP]
[4]□ □ □ □ □ □/[5]is thin[VBZ]
[0] □ □/[0]He[PP] [1]□ □ □ □ □ □ □ □ □ □ □/[1]worked hard[VBD]
[1] □ □/[0]he[PP] [2]□ □ □ □ □ □ □ □ □/[1]failed[VBD]
[0] □ □/[0]he[PP] [1]□ □ □ □ □ □/[3]examination[NN]
[2]□ □ □ □ □ □ □ □ □ □ □/[1]passed[VBD]
[0] □ □/[0]He[PP] [1]□ □ □ □ □ □/[1]was lazy[VBD]
[0]□ □ □ □ □/[0]They[PP] [1]□ □ □ □ □/[2]football[NN]
[2]□ □ □ □ □ □ □/[1]play[VBP]
[0] □ □/[0]He[PP] [1]□ □ □/[4]house[NN] [2]□ □ □ □/[2]in[IN]
[3]□ □ □ □ □/[1]is[VBZ]

Figure 5. Bilingual Corpus Format

5 OVERVIEW OF SYSTEM

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 Maximun matching algorithm[17], 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.

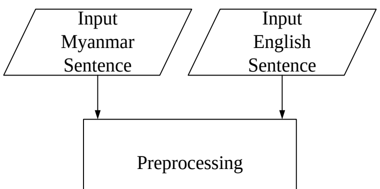


Figure 6. Proposed Alignment System

Input two sentences are

□ □
They go to pagoda.

After Segmentation
□ □ □ □ □_□ □ □_□ □ □ □ □_□ □ □ □_□ □ □ □ □ □ □ □
They go to pagoda.

After Stemming
[0]□ □ □ □ □ □ [1]□ □ □ □ □ [2]□ □ □ □
[3]□ □ □ □ □ □ □ □
[0]They[PP] [1]go[VB] [2]to[TO] [3]pagoda[NN].

Afer Align
[0]□ □ □ □ □ □/[0]They[PP]
[1]□ □ □ □ □/ [3]pagoda[NN].
[2]□ □ □ □/[2]to[TO]
[3]□ □ □ □ □ □/ [1]go[VB]

Figure 7. Example Alignment Procedure

6 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:

$$\text{Recall} = \frac{W_{correct}}{W_{Dtotal}} \times 100\%$$

$$\text{Precision} = \frac{W_{correct}}{W_{Stotal}} \times 100\%$$

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

Where, $W_{correct}$ is the number of correctly aligned words, W_{DTotal} is the number of words and W_{STotal} is the number of aligned words by the system. According to the experimental results, it shows in Table 1. By using combination of Corpus based approach and dictionary lookup approach, the precision increased.

Experiment
S_1 is Corpus Based Approach
S_2 is Dictionary Lookup Approach
S_3 is Corpus Based Approach + Dictionary Lookup Approach

Table 1. Results for Experinment

Experiment	S1	S2	S3
Precision (%)	88	91	94
Recall (%)	80	82	90
F-measure (%)	83	86	92

7 CONCLUSION AND FUTURE WORK

We have shown that building Myanmar-English parallel corpus can be improved by a combination of corpus based approach and dictionary lookup approach. Myanmar languages are morphologically rich. Thus, in future, the proposed model will be better result by using a list of cognates and morphological analysis. This system can be extended as phrase alignment model. We will work on many to many word alignments and have to test the algorithm for large bilingual corpora.

ACKNOWLEDGMENT

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.

REFERENCES

[1] Bing Xiang,Yonggang Deng,and Bowen Zhou, "Diversify and Combine: Improving Word Alignment for Machine Translation on Low-Resource Languages", *Proceedings of the ACL 2010 Conference Short Papers*, 2010, pages 22–26.

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

[3] 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

[4] Eknath Venkataramani and Deepa Gupta, "English-Hindi Automatic Word Alignment with Scarce Resources". *In International Conference on Asian Language Processing*, *IEEE*, 2010.

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

[6] 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.

[7] Helen 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, 2005, Pages 89–96.

[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 ,2005, Pages 65–74,.

[10] Jamie Brunning, Adria 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 Chengqing Zong, "Word Reordering Alignment for Combination of Statistical Machine Translation Systems", *IEEE*, 2008.

[12] Niraj Aswani and Rpbert Gaizauskas, "A hybrid approach to align sentences and words in English-Hindi parallel corpora". *In Proceedings of the ACL* Workshop on Building and Using Parallel Texts, June, 2005, page 57-64.

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

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

[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., 2003, Pages 1–6.

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

Khin Thandar Nwet is currently pursuing Ph.D degree program in University of Computer Studies, Yangon, Myanmar, PH-0973176035. I got M.C.Sc from University of Computer Studies, Mandalay in 1997. I am also an assistant lecturer. My research interest is Natural language processing. Email: khin.thandarnwet@gmail.com