Ambiguous Myanmar Word Disambiguation System for Myanmar-English Statistical Machine Translation

Nyein Thwet Thwet Aung
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

In Statistical Machine Translation (SMT), there are many source words that can present different translations or senses. Word Sense Disambiguation (WSD) system is designed to determine which one of the senses of an ambiguous word is invoked in a particular context around the word. It is an intermediate task essential to many natural language processing problems, including machine translation, information retrieval and speech processing. There is not any cited work for resolving ambiguity of words in Myanmar language. This paper presents a new WSD method for ambiguous Myanmar words. It is based on supervised learning approach, Nearest Neighbor Cosine Classifier. The system uses Myanmar-English Parallel Corpus as a training resource. As an advantage, the system can overcome the problem of translation ambiguity from Myanmar to English language translation.

General Terms

Natural Language Processing, Statistical Machine Translation, Word Sense Disambiguation.

Keywords

Myanmar Language, ambiguous Myanmar words, supervised learning, Nearest Neighbor Cosine Classifier, Myanmar-English Parallel Corpus.

1. INTRODUCTION

Word Sense Disambiguation (WSD) is always an important and difficult problem that requires to be solved in Natural Language Processing. It refers to the process of selecting the most appropriate meaning or sense to a given ambiguous word within a given context. Resolving the word ambiguity is considered as the major bottleneck for large scale language understanding applications and their associate tasks such as machine translation (MT), information retrieval (IR), natural language understanding (NLU) and others. These various range applications of natural language processing need knowledge of word meaning to select the correct word sense in a context [3].

Generally, there are two types: polysemy-a single word form having more than one meaning; synonymy- multiple words having the same meaning are both important issues in natural language processing or artificial intelligence. In this paper, we present an application of WSD in machine translation (MT), where the system has to select the correct translation equivalent in the target language of an ambiguous item in the source language. For example, the ambiguous Myanmar noun "တူ (tu)" would translate to three different English meanings, **chopsticks** (each of a pair of small, thin, tapered sticks held in one hand and used as eating utensils by the Chinese and Japanese sense), **nephew** (a son of one's brother or sister) and **hammer** (used for breaking things and driving in nails) in the following three sentences:

(1)သူသည်တူဖြင့်ခေါက်ဆွဲစားသည်။
He eats the noodle with **chopsticks**.

(2)သူ့မှာတူသုံးယောက်ရှိသည်။
He has three **nephews**.

(3)လက်သမားသည်တူကိုသုံးသည်။
Carpenter uses the **hammer**.

Therefore, we propose an approach to disambiguate senses of several ambiguous Myanmar words for Myanmar-English statistical machine translation. Our method is based on Nearest Neighbor Cosine classifier. All the processes in our system are developed by Java Programming.

The remainder of this paper is organized as follows: We discuss the related work and the Ambiguity of Myanmar Language in section 2 and section 3. Section 4 and 5 show the overview of Statistical Machine Translation System and Disambiguation approaches. Section 6 and 7 describe Nearest Neighbor Cosine Classification and Myanmar-English parallel corpus. The overview of the proposed system is presented in section 8. Execution of Proposed WSD Algorithm is shown in section 9. Section 10 shows the implementation of the system. Experimental result is described in section 11and the paper is concluded in section 12.

2. RELATED WORK

Many researchers have been work for word sense disambiguation in English Language. For the research reported in this paper, we will emphasis on the ambiguity of the Myanmar words because it is still now open in Machine Translation. In the following paragraphs, we discuss briefly some of the related work and history in the area of Word Sense Disambiguation.

Phil Katz (2005) proposed supervised word sense disambiguation using Python [1].He implements five different context based classifiers: a Naive Bayes classifier, a decision list classifier, a nearest neighbor cosine classifier, a k-Nearest-Neighbor cosine classifier and a classifier based on Latent Semantic Analysis. The system also includes a meta-classifier that combines the outputs of the stand-alone systems into one classification. He showed that nearest neighbor cosine classifier is the most precise classifier in his system. Mohammad Teduh Uliniansyah and Shun Ishizaki (2006) performed a word sense disambiguation system using modified Bayesian algorithms for Indonesian language [2]. Sunee Pongpinigpinyo and Wanchai Rivepiboon (2006) presented distributional semantics approach to Thai word sense disambiguation [3]. Samir Elmougy, Taher Hamza and Hatem M.Noaman (2008) discussed rooting algorithm with Naïve Bayes Classifier for Arabic Word Sense Disambiguation [4]. Farag Ahmed and Andreas Nurnberger (2008) proposed Arabic/English Word translation disambiguation using parallel corpora and matching schemes [5].

Farag Ahmed and Andreas Nürnberger (2009) showed Corpora based Approach for Arabic/English Word Translation Disambiguation [6]. Yu Zheng-tao et al. (2009) discussed word sense disambiguation based on Bayes model and information gain [7]. Zhang Zheng and Zhu Shu (2009) proposed a new approach to WSD in machine translation [8]. Asma Naseer and Sarmad Hussain (2009) proposed Supervised Word Sense Disambiguation for Urdu Using Bayesian Classification [9]. Laroussi Merhbene, Anis Zouaghi and Mounir Zrigui (2010) discussed Ambiguous Arabic Words Disambiguation [10]. They used context matching algorithm. Zheng_Yu Niu et al. (2004) proposed Optimizing Feature Set for Chinese Word Sense Disambiguation [11]. They used supervised Naïve Bayes classifier. An optimal feature set was selected by maximizing the cross validated accuracy of supervised Naive Bayes classifier on sense-tagged data. Their system achieved 60.40% precision and recall in Chinese lexical sample task. Nancy Ide and Jean Véronis (1998) described Word Sense Disambiguation: The state of the art [12].

Cuong Anh Le and Akira Shimazu (2004) discussed High WSD accuracy using Naive Bayesian classifier with rich features [13]. They show that by adding more rich knowledge, represented by ordered words in a local context and collocations, the NB classifier can achieve higher accuracy in comparison with the best previously published results. The features were chosen using a forward sequential selection algorithm. Their experiments obtained 92.3% accuracy for four common test words (interest, line, hard, serve). Guo Jiang and Zhang Yangsen (2010) presented study on multiple classifiers for Chinese Word Sense Disambiguation [14]. In this paper, a new method of multiple layer classifiers integration based on single classifier is proposed which called Auto Weight Adjust. They chose Maximum Entropy (ME) and Naïve Bayesian (NB) as single classifiers and use the ME classifier result and the NB classifier result to fuse the final result. They use People Daily News (PDN) datasets to test their model, according to experiments their algorithm leads to less error and better performance than other algorithms. Jong-Hoon Oh and Key-Sun Choi (2002) proposed Word Sense Disambiguation using Static and Dynamic Sense Vectors [15]. This paper reports on word sense disambiguation of English words using static and dynamic sense vectors. The English SENSEVAL test suit is used for this experimentation and their method produces relatively good results.

3. AMBIGUITY OF MYANMAR LANGUAGE

Myanmar is an official language of the Union of Myanmar. It is written from left to right and no spaces between words, although informal writing often contains spaces after each clause. It is syllabic alphabet and written in circular shape. It has sentence boundary mark. It is a free-word-order language, which usually follows the subject-object-verb (SOV) order. In particular, preposition adjunctions can appear in several different places of the sentence. Unlike Myanmar, English Language has a rigid subject-verb-object (SVO) order.

However, Myanmar language has semantic ambiguity problem like English. Although using statistical methods has been very successful for some of important problems in Myanmar Natural Language Processing such as Part-Of-Speech tagging, segmentation and alignment of parallel translation, an effective method for solving semantic ambiguity problem does not exist yet. Table 1and 2 show some examples of Myanmar ambiguous nouns and verbs and its English translation meanings.

Table 1. Some Ambiguous Nouns and their Senses

Ambiguous Word	**No: of Sense**	**Sense 1**	**Sense 2**	**Sense 3**
တူ (tu)	3	Hammer	Chopsticks	Nephew
ငါး (ngar)	2	Five	Fish	-
ဘာသာ (barthar)	3	Language	Religion	Subject
လ (la)	2	Month	Moon	-
လက်မ (latma)	2	Inch	Thumb	-
ငွေ (Ngwe)	2	Money	Silver	-
ဈေး (zay)	2	Market	Price	-

Table 2. Some Ambiguous Verbs and their Senses

Ambiguous Word	**No: of Sense**	**Sense 1**	**Sense 2**	**Sense 3**	**Sense 4**
တက်သည် (tatthe)	4	Climb	Increase	Get	Mount
ခူးသည် (kuthe)	2	Pluck	Ladle	-	-
ခံသည် (khanthe)	4	Catch	Enjoy	Last	Resist
ဆူသည် (hsuthe)	3	Boil	Scold	(be) Noise	-
နင်းသည် (ninthe)	3	Tread	Pedal	Follow	-
တူးသည် (tuthe)	2	Burn	Dig	-	-
သောက်သည် (thoutthe)	3	Drink	Take	Smoke	-

4. OVERVIEW OF STATISTICAL MACHINE TRANSLATION SYSTEM

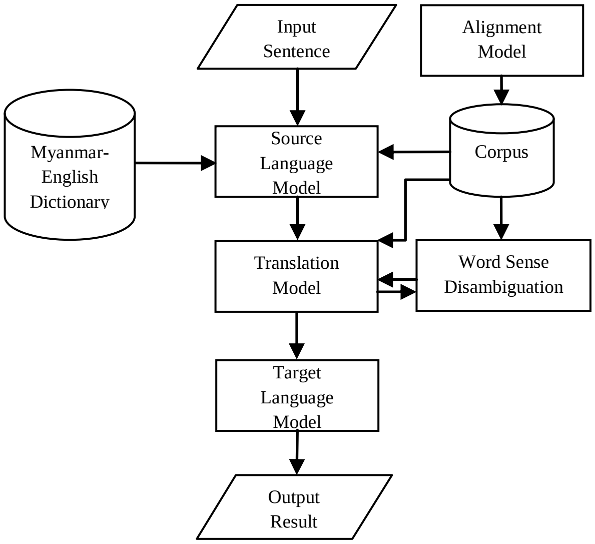


Figure 1. Myanmar-English Statistical Machine Translation System

To implement a Myanmar-English translation system, there are various problems that need to solve. This includes Source Language Model, Alignment Model, Translation Model and Target Language Model. Our work focuses on Word Sense Disambiguation process used in translation model. This phase is the most difficult stage with respect to the level of possible ambiguities. It is even more problematic when it comes to deal with two very divergent languages such as Myanmar and English. A word can have many senses and each of those senses can be mapped into many target language words. As an advantage, the proposed system can improve the accuracy of Myanmar to English language translation. The system is the first attempt to solve ambiguity in Myanmar language. It is also a part of the Myanmar to English Statistical machine translation project.

5. DISAMBIGUATION APPROACHES

Lexical ambiguity is **syntactic or semantic**. A word's syntactic ambiguity can be resolved by applying part-of-speech taggers which predict the syntactic category of a word in texts with high levels of accuracy. The problem of resolving semantic ambiguity, which is generally known as WSD, has proved to be more difficult than syntactic disambiguation. The only way to determine the meaning of a word in a particular usage is to examine its context. Word Sense Disambiguation (WSD) can be defined as the process of identifying the correct sense or meaning of a word in a particular context.

One could envisage building a WSD system using handcrafted rules or knowledge obtained from linguists. Such an approach would be highly labor-intensive, with questionable scalability. Another approach involves the use of dictionary or thesaurus to perform WSD. There are also three ways to approach WSD: a knowledge-based approach, which uses an explicit lexicon, corpus-based disambiguation, where the relevant information about word senses is gathered from training on a large corpus, or, third alternative, a hybrid approach combining aspects of aforementioned methodologies. On average, supervised methods yield better performance results. Supervised and unsupervised WSD tends to use a machine learning algorithm. During training on a disambiguated corpus probabilistic information about context words as well as distributional information about an ambiguous word is collected. In the testing phase, the sense with the highest probability computed on the basis of the training data (context words is chosen). Unfortunately, large sense annotated corpora are expensive and labor intensive to create, and the data acquisition bottleneck is particularly severe when moving to less studied languages and genres. A number of bootstrapping methods have been proposed to reduce the sense-tagging cost. For training, a possible solution is the use of an unsupervised approach, but for evaluation purposes sense-tagged material is still needed.

There are various information sources or feature types used in WSD regardless of the type of the approach. To disambiguate a word, a diversity of information, including syntactic tags, word frequencies, collocations, semantic context, role-related expectations, and syntactic restrictions can be considered. Many WSD algorithms rely on contextual similarity to help choose the proper sense of a word in context. Several important methodological issues come up in the context of word sense disambiguation. These are all words approach or unsupervised and supervised or lexical sample approach. Many Word Sense Disambiguation approaches use Dictionaries and thesauri, Word Net, Automatic corpus-based, apply heuristics, Variation or combination of above.

6. NEAREST NEIGHBOR COSINE CLASSIFICATION

The nearest neighbor cosine classifier is a supervised corpus-based approach. It uses the context vectors created for each sense during training and for the ambiguous instance during testing. The context vectors are created for each sense as shown in figure 2. The cosines between the ambiguous vector and each of the context vectors are calculated, and the sense that is the "nearest" (largest cosine/smallest angle) is selected by the classifier. In this method, the distance between two examples is computed by summing the distances between the features values associated with those examples. In our system, the context vectors are created by using Myanmar-English parallel corpus and the ambiguous vectors are created from the input sentences.

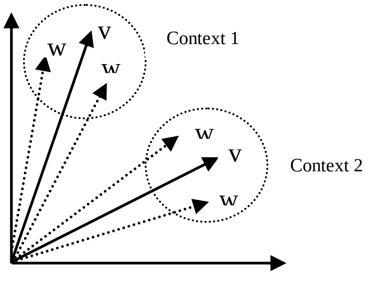


Figure 2. The word and context vectors for two contexts

The similarity between context vectors and for the ambiguous instances is computed through the Cosine distance as below:

$$\cos\theta = \frac{A.B}{\|A\|.\|B\|} = \frac{\sum_{i=1}^{n} A_i . B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \qquad (1)$$

where A represents ambiguous vector
B represents context vector
A_i means each word in ambiguous vector
B_i means each word in each context vector

7. MYANMAR-ENGLISH PARALLEL CORPUS

Parallel Corpora are also called bilingual corpora, one serving as primary language, and the other working as a secondary language. A bilingual corpus was used since different senses of some words often translate differently in another language. In our experiments, we use Myanmar-English parallel corpus as training. There is no Myanmar-English parallel corpus which contains Myanmar ambiguous words in public. So, we create Myanmar-English parallel corpus that contain ambiguous words manually. It contains various sense meanings of ambiguous Myanmar words. We present the following aligned sentences as part of the training corpus. The corpus structure of the following example sentences are as follows.

(1) ရထားသည်ဥမင်လိုဏ်ခေါင်းထဲသို့**ဝင်သည်**။
The train **enters** into tunnel .
ရထား ဝင်သည် ထဲသို့ ဥမင်လိုဏ်ခေါင်း ။

(2) ကျွန်မအဖေသည်အသက်ခြောက်ဆယ်သို့**ဝင်သည်**။
My father **reaches** the age of sixty .
ကျွန်မ အဖေ ဝင်သည် အသက် သို့ ခြောက်ဆယ် ။

(3) သူသည်တပ်ထဲသို့**ဝင်သည်**။
He **joins** to the army.
သူ ဝင်သည် ထဲသို့ တပ် ။

(4) နေသည်အရှေ့ဘက်မှထွက်ပြီးအနောက်ဘက်သို့**ဝင်သည်**။
The sun rises from the east and **sets** to the west .
နေ ထွက် မှ အရှေ့ဘက် ပြီး ဝင်သည် သို့ အနောက်ဘက်။

(5) □□□□□□□□□□□□□□□□**ဝင်သည်**။
The oil box **holds** five gallons.
□□□□□□ □□□□□□ □□□ □□□□ □

- [0]ရထား/[0]train[NN]	[1]ဥမင်လိုဏ်ခေါင်း/[3]tunnel[NN]
[2]ထဲသို့/[1]to[TO]	[3]**ဝင်သည်**/[1]**enter**[VB]
- [0]ကျွန်မ/[0]my[PP$]	[1]အဖေ/[1]father[NN]
[2]အသက်/[3]age[NN]	[3]ခြောက်ဆယ်/[5]sixty[CD]
[4]သို့/[4]of[IN]	[4]**ဝင်သည်**/[2]**reach**[VB]
- [0]သူ/[0]he[PP]	[1]တပ်/[3]army[NN]
[2]ထဲသို့/[2]to[TO]	[3]**ဝင်သည်**/[1]**join**[VB]
- [0]နေ/[0]sun[NN]	[1]အရှေ့ဘက်/[3]east[NP]
[2]မှ/[2]from[IN]	[3]ထွက်/[1]rise[VB]
[4]ပြီး/[4]and[CC]	[5]အနောက်ဘက်/[7]west[NP]
[6]သို့/[6]to[TO]	[7]**ဝင်သည်**/[5]**set**[VB]
- [0]ဆီပုံး/[0]oil box[NN]	[1]ငါး/[2]five[CD]
[2]ဂါလံ/[3]gallons[NNS]	[3]**ဝင်သည်**/[1]**hold**[VB]

Figure 3. The Structure of Myanmar-English Parallel Corpus

As it is clear, the Myanmar word "**ဝင်သည်**" are mapped into five different English words "**enter**", "**reach**", "**join**", "**set**" and "**hold**" based on its sense.

8. OVERVIEW OF THE PROPOSED SYSTEM

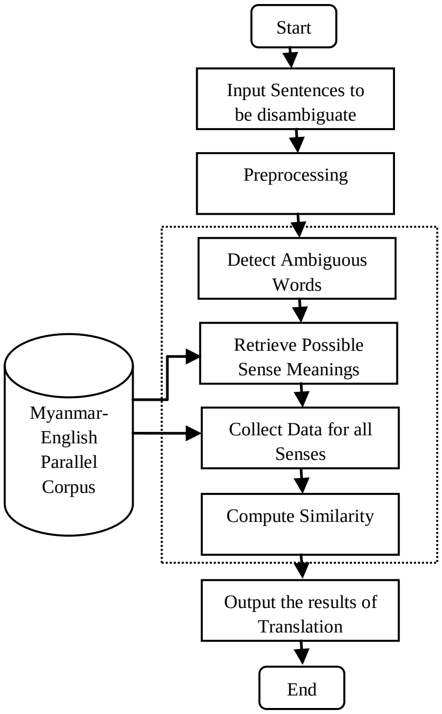

Start
Input Sentences to be disambiguate
Preprocessing
Detect Ambiguous Words
Retrieve Possible Sense Meanings
Collect Data for all Senses
Compute Similarity
Myanmar-English Parallel Corpus
Output the results of Translation
End

Figure 4. The Overview of the Proposed System

Figure 4 describes overview of the proposed system. The proposed system uses the idea of the Nearest Neighbor Cosine Classifier. Firstly, the system takes Myanmar Sentences that contains ambiguous words as input. In the preprocessing step, the system performs word segmentation by using Myanmar word segmenter and removing stop words such as prepositions, conjunctions and particles. Then, the ambiguous vectors are created. Secondly, the system detects the ambiguous words from the input. Then, all possible sense meanings of the target ambiguous words are retrieved from the training corpus. Thirdly, it also collects data concerning with each sense meaning of the ambiguous words to create context vectors. The system uses topical feature that represent co-occurring words in bag-of-word feature. It removes the stop words again from the context vectors and it might include the kinds of stop words such as prepositions, conjunctions and particles since they come from the training corpus. The system also removes the redundant words from each context vector. The process of making ambiguous vector and context vectors is described in detail in step 1 to 3 of the proposed algorithm.

Fourthly, the cosines values between ambiguous vector and each of the context vectors are calculated, and the sense that is the "nearest" (largest cosine/smallest angle) is selected as a correct sense. Finally, the system generates the correct English meanings for the target ambiguous words as output. This process is described in detail in step 4 and 5 of the proposed algorithm. The proposed algorithm is shown in the following figure 5.

```
Step 1:Preprocessing
    -Segment input sentences
    -Remove stop words from input sentences and create
    Ambiguous vectors
Step 2:Multi-sense Look-up
    -Detect ambiguous words from input sentences
    -Retrieve all possible sense meanings of ambiguous
    Words from training corpus
    -Collect training data concerning with these sense
    from corpus
Step 3:Build context vectors for each sense based
on Collected training data
    For all context vectors do
            -Remove stop words
            -Remove redundant words
    End For
Step 4:Calculate the cosines between ambiguous vector
and each of the context vectors
    $\cos\theta = \frac{A.B}{||A||.||B||} = \frac{\sum_{i=1}^{n} A_i . B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$
    where   A means ambiguous vector
            B means context vector
            $A_i$ means each word in ambiguous vector
            $B_i$ means each word in each context vector
Step 5: Choose correct sense of the target word.
    $s' = arg\ max\ score(s_i)$
    where  $s_i$ means the similarity value of each sense
```

Figure 5. Proposed algorithm for Myanmar Ambiguous Words Disambiguation

9. EXECUTION OF THE PROPOSED ALGORITHM

We give an example of the execution of our system. For example:
Input sentences:

သူသည်တူသုံးယောက်နှင့်အတူထမင်းစားသည်။

သူသည်တူဖြင့်ခေါက်ဆွဲစားသည်။

လက်သမားသည်တူကိုသုံးသည်။

In the preprocessing, we first segment the input sentences by using existing Myanmar word segmenter. This segmenter uses maximum matching. After segmentation: we get the following sentences.

(1)သူ_သည်_တူ_သုံးယောက်_နှင့်အတူ_ထမင်း_စားသည်_

(2)သူ_သည်_တူ_ဖြင့်_ခေါက်ဆွဲ_စားသည်_

(3)လက်သမား_သည်_တူ_ကို_သုံး_သည်_

Then, we remove all the function words (stop words). Stop words include pronouns, prepositions, conjunctions, particles, etc.

For example:

Stop word list= [ကြောင့်, က, ကျွန်တော်, ကျွန်မ, ၍, ကျွန်ုပ်, ကျုပ်, ငါ, သင်, မင်း, ကျွန်ုပ်တို့, နှင့်, ၏, သို့, ခင်ဗျား, ရန်, ညည်း, နင်, သူ, သင်း, ဤ, ထို, ဟို, ယင်း, ၎င်း, သည်, မည်သည်, သင့်ကို, သင့်အား, မှာ, မှ, ၌, မည်သူ, ဘယ်သူ, ကျွန်ုပ်၏, ဘယ်, အဘယ်, ဖြင့်, အား, တွင်, ကို, သူ့, သူ၏, သူမ, သူမ၏, ဘာ, သူ့ကို, သူ့အား, ကျွန်တော့်, သူတို့]

After removing stop words: we create the ambiguous vectors for each input sentence. Ambiguous vectors:

(1)=[စားသည်, ထမင်း, တူ, နှင့်အတူ, သုံးယောက်],

(2)=[စားသည်, တူ, ခေါက်ဆွဲ],

(3)=[တူ, လက်သမား, သုံးသည်]

Secondly, the system detects possible ambiguous words from each input sentence. The first sentence has two ambiguous words, the second sentence has also two ambiguous words and the third one has only one ambiguous word. Therefore, we get

(1)[စားသည်, တူ]

(2)[စားသည်, တူ]

(3)[တူ]

Then, we find all possible English meanings of Myanmar ambiguous words by using Myanmar-English parallel corpus. The word "**စားသည်**(sarthe)" has three senses, **exceed**, **eat** and **divide** and the word "**တူ** (tu)" has also three senses, **nephew**, **hammer** and **chopsticks**.

စားသည်	[exceed, eat, divide]
တူ	[nephew, hammer, chopsticks]

The system also collects data concerning with above senses from the training corpus. Thirdly, we construct the context vectors for each sense using the collected data. We remove stop words and redundant words from each context vector. So, we get the following six context vectors:

nephew=[တစ်ယောက်, တူမ, တူသည်, ထမင်း, တော်သည်, သုံးယောက်, စားသည်, နှင့်အတူ, တွင်ရှိသည်]

chopsticks=[ခေါက်ဆွဲ, စားသောက်, ပြုလုပ်သည်, စားသည်, ဝါး, ဂယ်လာသည်]

hammer=[သံ, ထုသည်, လက်သမား, ရိုက်သောအခါ, သုံးသည်]

exceed=[ထက်, အရပ်အမောင်း],

eat=[အသီး, ကျွန်တော်တို့, ငှက်များ, မာလကာသီး, သကြား, ထမင်း, တူ, လျှင်မြန်စွာ, ရောစားသည်, သုံးယောက်, မြက်, နှင့်အတူ, စားကြသည်, ကြောင်များ, ချင်သည်, ထောပတ်သီး, ပန်းသီး, ခေါက်ဆွဲ, သလား, အသား, ပွသော, နွားများ, စား, တစ်စုံတစ်ခုမျှ, ချင်, တစ်စုံတစ်ခု, မုန့်, သကြားလုံး, ညစာ, မနက်စာ, ထမင်း, ခြသေ့်, အသား, နေ့လည်စာ, အများအပြား]

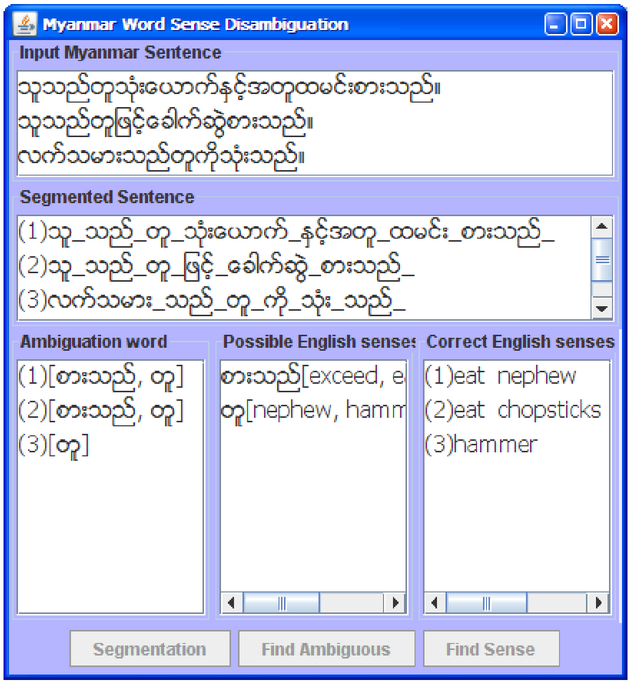
divided=[ငါးခန်း, သီးသန့်, တစ်ခု, ငါး, ယောက်ျား, မိန်းမ, ခွဲထားသည်, ကျမ်းစာအုပ်, အိမ်သာ, အခန်း, ကိန်း]

Finally, we compute the cosine similarity between the ambiguous vector and each context vector. After calculating the score of each sense, we can assign the sense with the highest similarity to the word.

(1)eat nephew
(2)eat chopsticks
(3)hammer

So, we choose "**eat**" for the ambiguous word "**စားသည်**(sarthe)"and **nephew** for "**တူ**(tu)" in the first sentence. For the second sentence, we choose "**eat**" and "**chopsticks**" and **hammer** for the third sentence. By the way, we can disambiguate a word with multiple senses in a given context.

10. IMPLEMENTATION OF THE SYSTEM



သူသည်တူသုံးယောက်နှင့်အတူထမင်းစားသည်။
သူသည်တူဖြင့်ခေါက်ဆွဲစားသည်။
လက်သမားသည်တူကိုသုံးသည်။

(1)သူ_သည်_တူ_သုံးယောက်_နှင့်အတူ_ထမင်း_စားသည်_
(2)သူ_သည်_တူ_ဖြင့်_ခေါက်ဆွဲ_စားသည်_
(3)လက်သမား_သည်_တူ_ကို_သုံး_သည်_

(1)[စားသည်, တူ]
(2)[စားသည်, တူ]
(3)[တူ]

စားသည်[exceed, e
တူ[nephew, hamm

(1)eat nephew
(2)eat chopsticks
(3)hammer

Figure 6. Execution of the system

The execution of our system is shown in figure 6. The system takes Myanmar sentences as input as shown in figure. The system shows segmentation results, ambiguous words for each sentence, the possible English meanings of each ambiguous word and the correct English translation of the target ambiguous word in each sentence as output.

11. EXPERIMENTAL RESULT

The experiments are conducted using data drawn from "Myanmar-English Parallel Corpus", which contains sentences used in various domains. We use Zawgyi-one Myanmar font. Our approach relies on supervised learning. The training set consists of 1500 sentence pairs and each sentence average length is 12 words. We had been collected 60 ambiguous nouns and 100 ambiguous verbs for experiment. We used only the pure text data, and not the speech transcriptions. The test sentences are collected from examples sentences in Dictionary and Myanmar websites. These sentences can contain at least one ambiguous word and three ambiguous words at most. For evaluation purpose, we group test sentences in three groups, the first group sentences are composed of words in the corpus. The second group sentences are composed of words in the corpus but not exactly the same sentences in corpus and the third sentence are composed of words not include in the corpus.

Table 3. Experimental results on test data set

Sentence Type	**Accuracy (%)**
Test Sentences in the training set	100%
Test Sentences are composed of words in the training sentences, but not exactly the same sentences in the training set	95.75%
Test Sentences that are not in the training set	89.25%

Table 3 shows the results of our experiment. The system gets 100% accuracy for the first group sentences, 95.75% for the second group sentences and 89.25% for the third group. Moreover, five nouns and five verbs were also selected as objects of our experiments. Table 4 shows the results of our experiment. The experiments show that disambiguation process by using the proposed method from the mentioned corpus received about 95% overall accuracy in detecting the correct translation of ambiguous words. The failure in disambiguation process is caused by the amount of training corpus, the different senses of words which may exist in the data set and the problem of segmentation.

Table 4 Results of experimentation

Ambiguous Words	**No: of Sense**	**No: of training sentences**	**No: of testing sentences**	**Accuracy (%)**
တူ (tu)	3	55	30	98.45
နာရီ (naryee)	4	31	20	97.45
အဆက် (asat)	4	12	10	94.12
ကျွန်း (kjun)	2	25	15	100
အခန်း (akhan)	3	19	10	95.75
ပေါက်သည် (poukthe)	11	50	37	92.01
ဝင်သည် (winthe)	11	40	35	90.21
သိမ်းသည် (thaythe)	10	30	15	93.31
ကျသည် (kyathe)	8	25	17	94.21
ခတ်သည် (katthe)	8	18	9	93.45
Average				94.896

12. CONCLUSION AND FUTURE WORK

This research was the first attempt to create Myanmar ambiguous words Disambiguation system. We present an approach for solving the ambiguity of words in Myanmar language. We evaluate our approach through an experiment using the Myanmar-English parallel corpus aligned at sentence level. We ensured that the input sentence contained ambiguous word with multiple English translations. The system is achieved about 95% accuracy. Therefore, the system can improve the accuracy of Myanmar to English language translation.

As a future work, we plan to investigate the suitability of other algorithms for Myanmar word sense disambiguation such as Naïve Bayesian Classifier, Support Vector Machine, Decision Lists and Trees and various feature types. This system disambiguates the words with part of speech 'Noun' and 'Verb'. We would like to implement this system for words with other part of speech such as 'Adjective' and 'Adverb'. Our plan also is to use this work in the areas that must have word sense disambiguation algorithm before it such as grammatical analysis, speech processing and text processing. Hence, our proposed system of disambiguation senses can be considered to be useful and applicable for other research efforts in natural language processing.

13. ACKNOWLEDGMENTS

I would like to thank to my supervisor Prof. Dr. Ni Lar Thein, Rector, University of Computer Studies, Yangon and my co-supervisor Associate Prof. Dr. Khin Mar Soe, University of Computer Studies, Yangon whose provide helpful guidance, a lot of inspiration, motivation, suggestions and encouragement during my research work. I also wish to express my deepest gratitude to my parents and my relations for their encouragement, understanding and support throughout the period of doing my Ph-D candidate.

14. REFERENCES

[1] Phil, Kat. 2005. Supervised Word Sense Disambiguation using Python.

[2] Mohammad, T. U., and Shun, I. 2006. A Word Sense Disambiguation System Using Modified Naïve Bayesian Algorithms for Indonesian Language. Information and Media Technologies 1(1): pp.257-274.

[3] Pongpinigpinyo and Wanchai, R. 2006. Distributional Semantics Approach to Thai Word Sense Disambiguation. In Proceedings of the International Journal of Computational Intelligence 2:3.

[4] Samir, E., Taher, H., and Hatem M. N. 2008. Naïve Bayes Classifier for Arabic Word Sense Disambiguation. In Proceedings of the INFOS2008, Cairo-Egypt, March 27-29.

[5] Farag, A., and Andreas, N. 2008. Arabic/English Word Translation Disambiguation using Parallel Corpora and Matching Schemes. In Proceedings of the 12th EAMT conference, Hamburg, Germany, 22-23 September.

[6] Farag, A., and Andreas, N. 2009. Corpora based Approach for Arabic/English Word Translation Disambiguation. Speech and Language Technology, Volume 11.

[7] Yu, Z., Deng, B., Hou, B., Han, L., and Guo, J. 2009. Word Sense Disambiguation Based on Bayes Model and Information Gain. In the Proceedings of the International Journal of Advanced Science and Technology, Vol.3, February.

[8] Zhang, Z., and Zhu, S. 2009. A New Approach to Word Sense Disambiguation in MT System. World Congress on Computer Science and Information Engineering.

[9] Asma, N., and Sarmad, H. 2009. Supervised Word Sense Disambiguation for Urdu Using Bayesian Classification. Unpublished.

[10] Laroussi, M., Anis, Z., and Mounir, Z. 2010. Ambiguous Arabic Words Disambiguation. 11th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing.

[11] Zheng, Y. N., Dong, H. J., and Chew, L. T. 2004. Optimizing Feature Set for Chinese Word Sense Disambiguation. In Proceedings of the SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain, July.

[12] Nancy, I., and Jean, V. 1998.Word Sense Disambiguation: The State of the Art. Computational Linguistics, Department of Computer Science. Vassar College, Poughkeepsie, New York.

[13] Cuong, A. L., and Akira, S. 2004. High WSD accuracy using Naïve Bayesian classifier with rich features, In Proceedings of the PACLIC 18, Waseda University, Tokyo, December 8th-10th.

[14] Guo, J., and Zhang, Y. 2010. Study on Multiple Classifier for Chinese WSD. International Conference on Artificial Intelligence and Computational Intelligence.

[15] Jong, H. O., and Key, S. C., 2002. Word Sense Disambiguation using Static and Dynamic Sense Vectors. 19th International Conference on Computational Linguistics.