# Myanmar-English Word Translation Disambiguation using Parallel Corpus

Nyein Thwet Thwet Aung
*University of Computer Studies, Yangon, Myanmar*
*thwet.nyein@gmail.com*

## Abstract

*Word Sense Disambiguation (WSD) has always been a key problem in Natural Language Processing. WSD is defined as the task of finding the correct sense of a word in a specific context. WSD systems can help to improve the performance of statistical machine translation (MT) systems. It is crucial for applications like Machine Translation and Information Extraction. Using Naïve Bayesian (NB) classifiers is known as one of the best methods for supervised approaches for WSD. In this paper, we use Naïve Bayesian Classifier for solving the ambiguity of words in Myanmar language. This system acquires the linguistic knowledge from an annotated corpus and this knowledge is represented in the form of features. As an advantage, the system can overcome the problem of translation ambiguity from Myanmar to English language translation.*

## 1. Introduction

Machine Translation (MT) is one of the longest standing problems in Natural Language processing (NLP), but still presents many challenges so far. One of the main challenges is that of lexical choice in the case of semantic ambiguity, i.e., the choice for the most appropriate word in the target language for a word in the source language when the target language offers more than one option for the translation and these options have different meanings, all of them having the same part of speech.

Word sense disambiguation(WSD) is one of the most critical and widely studied NLP tasks, which is used in order to increase the success rates of NLP applications like machine translation, information retrieval etc. WSD can be defined as the process of selecting the correct or intended sense of a word, occurring in a specific context [4].

Word sense disambiguation is an intermediate task which is not an end in itself, but rather is a necessary for some other natural language processing tasks such as text categorization, machine translation, information retrieval, grammatical analysis, speech processing, and text processing [9].

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 related to fields [4].

WSD must be able to choose the correct translation from possible items for any source lexical. Usually distinguishing between the candidates that are closely related conceptually is a hard task in MT applications. There can be many distinctions between the meaning of a lexical item in one language and its counterpart in another language. These distinctions are sometimes critical to selecting the correct lexical item in the target language.

Two main approaches have been applied in the WSD field. These are knowledge-based approaches and corpus based approaches. Knowledge based approaches use Machine Readable Dictionaries (MRD). It relies on information provided by Machine Readable Dictionaries (MRD). Corpus based approaches can be divided into two types, supervised and unsupervised learning approaches. Supervised learning approaches use information gathered from training on a corpus that has sense-tagged for semantic disambiguation. A major obstacle of this approach is the difficulty of manual sense-tagged in a training corpus that impedes the applicability of many approaches to domains. Unsupervised leaning approaches determine the class membership of each object to be classified in a sample without using sense-tagged training examples [7].

Among them, corpus based approaches select a target word using statistic information that is automatically extracted from corpora. Corpus based method is one of the successful lines of research on WSD. Many supervised learning algorithms have been applied for WSD, Bayesian learning (Leacock et al., 1998), exemplar based learning (Ng and Lee, 1996), decision list (Yarowsky, 2000), neural network (Towel and Voorheest, 1998), maximum entropy method (Dang et al., 2002), etc [11].

In this paper, we employ Naïve Bayes classifier to perform WSD in Myanmar polysemous words. 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 a polysemous item in the source language.

The remainder of this paper is organized as follows: We discussed the related work in section 2. Section 3 showed the overview of Machine Translation System. Section 4 describes Naïve Bayesian Classifier. The proposed approach is presented in Section 5.The paper is concluded in Section 6.

## 2. Related Work on Word Sense Disambiguation in Other Languages

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. The focus will be on the use of supervised methods for WSD.

There are many methods on disambiguation senses of a polysemous word. In the following paragraphs, we discuss briefly some of the related work and history in the area of Word Sense Disambiguation.

In 2008, Samir Elmougy, Taher Hamza and Hatem M.Noaman discussed rooting algorithm with Naïve Bayes Classifier for Arabic Word Sense Disambiguation [6]. Farag Ahmed and Andreas Nurnberger (2008) proposed Arabic/English Word translation disambiguation using parallel corpora and matching schemes [2]. Cuong Anh Le and Akira Shimazu (2004) performed High WSD accuracy using Naive Bayesian classifier with rich features [1]. They use forward sequential selection algorithm for feature selection and obtain 92.3% accuracy for four common test words. Farag Ahmed and Andreas Nürnberger (2009) showed Corpora based Approach for Arabic/English Word Translation Disambiguation [3].

Zheng-Yu Niu, Dong-Hong Ji and Chew-Lim Tan (2004) proposed Optimizing Feature Set for Chinese Word Sense Disambiguation [10]. This classifier utilizes an optimal feature set, which is determined by maximizing the cross validated accuracy of NB classifier on training data. The optimal feature set includes part-of-speech with position information in local context, and bag of words in topical context. Ishizaki (2006) performed a word sense disambiguation system using modified Bayesian algorithms for Indonesian language [5]. Yu Zheng-tao, Deng Bin, Hou Bo, Han Lu and Guo Jian-yi (2009) discussed word sense disambiguation based on Bayes model and information gain [10].

After performing extensive reading on methods for disambiguation senses, we choose Naïve Bayesian method to be implemented in our system because it is reportedly as having good results and relatively simple.

## 3. Overview of Machine Translation System

The machine translation is the process by which computer software is used to translate a text from one natural language to another. It is not a mere word-for-word substitution must interpret and analyze all of the elements in the text and know how each word may influence another. There are two types of machine translation approaches:

- Rule-based Machine translation
- Statistical Machine translation

The Statistical Machine Translation (SMT) is to learn how to translate from a large corpus of pairs of equivalent source and target sentences. SMT models take the view that every sentence in the target language is a translation of the source language sentence with some probability. The process of Myanmar-English statistical machine translation is describing in the following figure.1.

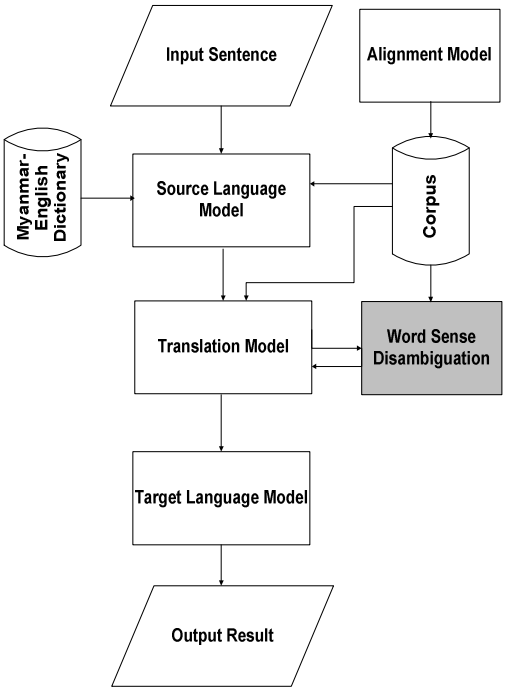


**Figure.1. Myanmar-English Statistical Machine Translation System**

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.

For example, the two Myanmar sentences such as "**ကျွန်း** များ ပတ်လည်တွင်ရေရှိသည်။" and "ကျွန်ပ်၏အိမ်ကို **ကျွန်း** ဖြင့်တည်ဆောက်ထားသည်။". In these sentences, the word **"ကျွန်း"(kjun)** have two possible senses such as: **ကျွန်း**=ရေပတ်လည်ရှိသောကုန်းမြေအရပ် (**island**: The land surrounded by water.) and **ကျွန်း**=အဖိုးတန်သောသစ်မာပင်တစ်မျိုး (**teak**: One kind of wood.)

The word "**ကျွန်း**" must be translated into "**island**" for the first sentence and "**teak**" for the second sentence. In order to translate this word to corrective English word, we will perform Word Sense Disambiguation by adjusting the meaning of neighboring words. We present an application of WSD in machine translation (MT), where the system has to select the correct translation equivalent in the target language. Therefore, WSD is very important for translation model in statistical machine translation.

## 4. Naïve Bayesian Classifier

The Naïve Bayesian Algorithm was first used for general classification problems. Our approach is based on the idea of the Naïve Bayesian Algorithm. We exploit the distribution of words and related words in parallel corpus. It is based on the assumption that all features representing the problem are conditionally independent given the value of classification variables. For a word sense disambiguation tasks, giving a word w, candidate classification variables $S=(s_1, s_2, ..., s_k)$ that represent the sense of the ambiguous word, and the feature $F=(f_1, f_2, ..., f_n)$ that describe the context in which an ambiguous word occurs, the Naïve Bayesian finds the proper sense $s_i$ for the ambiguous word w by selecting the sense that maximizes the conditional probability $P(w=s_i|F)$.

Suppose C is the context of the target word w, and $F=(f_1, f_2, ..., f_n)$ is the set of features extracted from context C, to find the right sense $s'$ of w given context C, we have:

$$s' = \arg\max_{s_i} P(w = s_i | F)$$

$$= \arg\max_{s_i} \frac{P(F | w = s_i)}{P(F)} P(w = s_i)$$

$$= \arg\max_{s_i} P(F | w = s_i) P(w = s_i)$$

The NB classifier works with the assumption that the features are conditional independent, so that we have:

$$s' = \arg\max_{s_i} \prod_{f_j \in C} P(f_j | w = s_i) P(w = s_i)$$

$$= \arg\max_{s_i} [\sum_{f_j \in C} \log(P(f_j | w = s_i)) + \log P(w = s_i)]$$

The features for WSD using a NB algorithm are words which are extracted from the context of the ambiguous word. The probability of sense $s_i$, $P(s_i)$, and the conditional probability of feature $f_j$ with observation of sense $s_i$,, $P(f_j|s_i)$, are computed via Maximum-Likelihood Estimation:

$$P(s_i) = C(s_i) / N$$

$$P(f_j | w = s_i) = C(f_j, s_i) / C(s_i)$$

Where $C(f_j, s_i)$ is the number of occurrences of $f_j$ in a context of sense $s_i$ in the training corpus, $C(s_i)$ is the number of occurrences of $s_i$ in the training corpus, and N is the total number of occurrences of the ambiguous word w or the size of the training dataset. To avoid the effects of zero counts when estimating the conditional probabilities of the model, when meeting a new feature $f_j$ in a context of the test dataset, for each sense $s_i$ we set $P(f_j | w = s_i)$ equal 1/N.

## 5. General Overview of proposed System

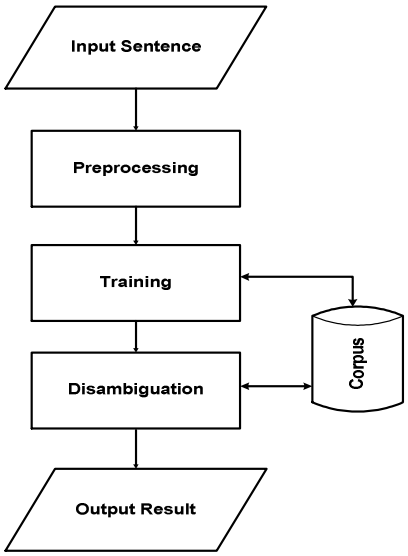


**Figure.2. General Overview of proposed System**

Our approach is based on the idea of the Naïve Bayesian Algorithm. The proposed system consists of three main parts. These are

1) Preprocessing

2) Training and
3) Disambiguation

**1) Preprocessing**
a. Segmentation
b. Remove stop words
**2) Training**
for all senses $s_i$ of W do
for all words $f_i$ in the vocabulary do
$P(f_i|s_i) = C(f_i,s_i)/C(s_i)$
end
end
for all senses $s_i$ of W do
$P(s_i) = C(s_i) / N$
end
**3)Disambiguation**
for all senses $s_i$ of W do
$score(s_i) = \log P(s_i)$
for all words $f_i$ in the context window c do
$score(s_i) = score(s_i) + \log P(f_i|s_i)$
end
end
Choose s' = arg max $score(s_i)$

**Figure.3. Naïve Bayes algorithm for WSD**

### 5.1 Preprocessing

In the text preprocessing step, it has two parts. The input sentence is segmented first by using existing segmentor. Segmentation is the necessary step to morphological and syntactic parsing since these analyses consider words or sentences most of the time. The task of the segmentor is the separation of words and punctuation. For instance, the source sentence to be disambiguated is given as "ထိုအိမ်ကိုကျွန်းဖြင့်တည်ဆောက်ထားသည်။" and includes some punctuation marks. The segmentor segments the input sentence as in the following Figure 4.

Input: ထိုအိမ်ကို**ကျွန်း**ဖြင့်တည်ဆောက်ထားသည်။

Output:ထို_အိမ်_ကို_**ကျွန်း**_ဖြင့်_တည်ဆောက်ထားသည်_။_

**Figure.4.Output from Segmentation**

After segmentation, the stop words are removed. We remove all the function words (stop words). Stop words include preposition, conjunctions, particles, inflections etc which appear as suffixes added to other words.

### 5.2 Training

After gathering the formatted information in the preprocessing step, we use the words in the input sentence as features and training process is initiated.

### 5.3 Disambiguation

After finishing the training phase, disambiguation process is started. The disambiguation module uses the output of the training phase to compute the score of each sense of ambiguous word and to decide the most appropriate sense for a given word in the test sentence.

We summarize the algorithm in above figure 3.

## 6. Conclusions and Future Work

This research was the first attempt to create a word sense disambiguation system for Myanmar Language. We evaluated 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.

As a future work, we plan to investigate the suitability of other algorithms for Myanmar word sense disambiguation such as Decision Lists and Trees and various feature types. Our plan also is to use this work in the areas that must have word sense disambiguation algorithm before it such as machine translation, grammatical analysis, speech processing and text processing.

## 7. References


[1] C.A. Le and A. Shimazu, "High WSD accuracy using Naïve Bayesian classifier with rich features", In Proceedings of the PACLIC 18,Waseda University, Tokyo, December $8^{th}$-$10^{th}$ , 2004.

[2] F. Ahmed and A. Nurnberger, "Arabic/English Word Translation Disambiguation using Parallel Corpora and Matching Schemes", In Proceedings of the $12^{th}$ EAMT conference, Hamburg, Germany, 22-23 September 2008.

[3] F. Ahmed and A. Nurnberger, "Corpora based Approach for Arabic/English Word Translation Disambiguation", Speech and Language Technology. Volume 11.

[4] Ide and veronis, "Word Sense Disambiguation: The State of the Art." Computational Linguistics, 1998.

[5] M.T. Uliniansyah and S. Ishizaki, "A Word Sense Disambiguation System Using Modified Naïve Bayesian Algorithms for Indonesian Language", Information and Media Technologies 1(1): 257-274(2006).

[6] S. Elmougy, T. Hamza and H.M. Noaman, "Naïve Bayes Classifier for Arabic Word Sense Disambiguation",

In Proceedings of the INFOS2008, Cairo-Egypt, March 27-29, 2008.

[7] S. Pongpinigpinyo and W. Rivepiboon, “Distributional Semantics Approach to Thai Word Sense Disambiguation”, In Proceedings of the International Journal of Computational Intelligence 2:3 2006.

[8] T.M. Ma and N.L. Thein., “MASE Framework for Selecting Most Appropriate Sense of English Content Words in support of English-Myanmar Translation”, In Proceedings of the sixth international conference on Computer Applications, 2008.

[9] Wilks and M.Stevenson, "Sense tagging: Semantic tagging with a lexicon", Proceedings of the SIGLEX Workshop,1997.

[10] Y. Zheng-tao, D. Bin, H. Bo, H. Lu. and G. Jian-yi, “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, 2009.

[11] Z.Y. Niu, D.H. Ji and C.L.Tan, “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 2004.

[12] Z. Zhong and H. T. Ng, “It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text”, In the Proceedings of the ACL 2010 System Demonstrations, pages 78-83, Uppsala, Sweden, 13 July 2010.