Ambiguous Myanmar Word Disambiguation for Myanmar-English Machine Translation

Nyein Thwet Thwet Aung, Khin Mar Soe, Ni Lar Thein
University of Computer Studies, Yangon, Myanmar
thwet.nyein@gmail.com, nilarthein@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. There is not any cited work for resolving ambiguity of words in Myanmar language. 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 to disambiguate ambiguous Myanmar words with part-of-speech 'noun' and 'verb', which uses topical feature that represent co-occurring words in bag-of-word feature. The system also uses Myanmar-English Parallel Corpus as training data. The WSD module developed here will be used as a complement to improve Myanmar-English machine translation system. As an advantage, the system can improve the accuracy of Myanmar to English language translation.

1. Introduction

Word sense disambiguation (WSD) is one of the most critical and widely studied Natural Language Processing tasks, which is used in order to increase the success rates of NLP applications like machine translation, information search and information extract, natural language understanding (such as man-machine conversation system, interrogator-responder system), text auto-proofreading, speech recognition, sound-character transformation, syntax structure recognition and the language study etc [8].

WSD can be defined as the process of identifying the correct sense or meaning of a word in a particular context. When a human being is encountered with a word with multiple senses, he easily identifies the exact sense of the word with the help of context without giving a single thought to the other senses. But when the same situation is provided to a computer it is not an easy task to correctly identify the desired sense. WSD process helps in resolving such ambiguity issues [1]. Sometimes a word differs in meaning when its Part- of-Speech (POS) is different. For example butter can be a verb or a noun as it can be seen in the following example:
Will you spread butter [Noun] on toast?
Don't think you can butter [Verb] me up that easily.

In one sentence butter as a noun means "a solid yellow food made from milk or cream" [3], while in the other sentence butter as a verb means "to say nice things to someone so that they will do what you want" [3]. As such ambiguities can easily be resolved with the help of POS, WSD does not entertain such words. The word with different meanings having same POS needs some WSD process to conclude the accurate sense. For example Chair in English can be "a separate seat for one person" or "the person in charge of a meeting or an organization".

Three main approaches have been applied in the WSD field. These are knowledge-based approaches, corpus based approaches and hybrid approach. Knowledge based approaches use Machine Readable Dictionaries (MRD). It relies on information provided by 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. The classification approach of WSD makes use of statistical approaches either referring lexicons or using corpus for training. Thesauri, lexicons and corpus are the main source of training in the supervised approach. Unsupervised leaning approaches determine the class membership of each object to be classified in a sample without using sense-tagged training examples. Hybrid approach combines aspects of fore mentioned methodologies [11].

All approaches mentioned above have been used by different researchers for different languages. 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. In this paper, we focus on implementing WSD process for Myanmar language. We aim an application of WSD for 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 current work is an initial step to resolve the ambiguity of words in Myanmar context. The technique that is implemented to resolve ambiguity is Bayesian Classification.

The remainder of this paper is organized as follows: We discussed the related work in section 2. Section 3 showed the ambiguity of Myanmar Language and section 4 presented the overview of Statistical Machine Translation System. Section 5 described about the parallel corpus and section 6 showed Naïve Bayesian Classifier for WSD. The overview of the proposed system is presented in Section 7. Section 8 discussed the Execution of Proposed WSD Algorithm and section 9 showed Experimental Result. The paper is concluded in Section 10.

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.

Cuong Anh Le and Akira Shimazu (2004) performed to obtain High WSD accuracy using Naive Bayesian classifier with rich features [2]. Ishizaki (2006) performed a word sense disambiguation system using modified Bayesian algorithms for Indonesian language [9]. Samir Elmougy, Taher Hamza and Hatem M.Noaman (2008) discussed rooting algorithm with Naïve Bayes Classifier for Arabic Word Sense Disambiguation [10]. Farag Ahmed and Andreas Nurnberger (2008) proposed Arabic/English Word translation disambiguation using parallel corpora and matching schemes [4].

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 [13]. Asma Naseer and Sarmad Hussain (2009) proposed Supervised Word Sense Disambiguation for Urdu Using Bayesian Classification [1]. In 2009, Zhang Zheng and Zhu Shu presented A New Approach to Word Sense Disambiguation in MT System [14]. Laroussi Merhbene, Anis Zouaghi and Mounir Zrigui (2010) discussed Ambiguous Arabic Words Disambiguation [7]. They used context matching algorithm.

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. Ambiguity of Myanmar Language

Myanmar language is the 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. However, English Language has a rigid subject-verb-object (SVO) order.

Like English, Myanmar language has semantic ambiguity problem. 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. Consequently, this problem is frequently cited as one of the most important problems in natural language processing research today.

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 a polysemous item in the source language. For example, the polysemous Myanmar noun “**ကျွန်း**”(kjun) would translate to two different English words (**island** for the land surrounded by water sense, or **teak** for the kind of hard wood sense) in the following two sentences:

a. “ပင်လယ်ထဲတွင်**ကျွန်း**များစွာရှိသည်။”
(There are many islands in the sea.) and
b. “ကျွန်ုပ်၏အိမ်ကိုကျွန်းဖြင့်တည်ဆောက်ထားသည်။”
(My house is built by teak.)

In table 1 and 2, show some examples of Myanmar ambiguous nouns and verbs and their senses.

Table 1. Some Ambiguous nouns and their senses

Ambiguous Word	No: of Sense	Sense 1	Sense 2	Sense 3	Sense 4
တူ	3	Hammer	Chopsticks	Nephew	-
ဈေး	2	Market	Price	-	-
ကျွန်း	2	Island	Teak	-	-
နာရီ	4	O’clock	Hour	Watch	Clock
ငွေ	2	Silver	Money	-	-

Table 2. Some Ambiguous nouns and their senses

Ambiguous Word	No: of Sense	Sense 1	Sense 2	Sense 3	Sense 4
ကပ်သည်	4	Stick	Offer	Approach	Come
ခေါက်သည်	4	Knock	Play	Fold	Strike
ချုပ်သည်	2	Sew	Stitch	-	-
စားသည်	3	Eat	Divide	Exceed	-
ခူးသည်	2	Pluck	Ladle	-	-

4. Overview of Statistical Machine Translation System

There are two types of machine translation approaches: Rule-based Machine translation and 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.

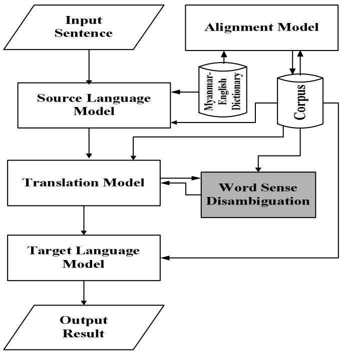

Figure 1. Myanmar-English Statistical Machine Translation System

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. As an advantage, the proposed system can improve the accuracy of Myanmar to English language translation. This 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. Parallel Corpus

Parallel Corpora are also called bilingual corpora, one serving as primary language, and the other working as a secondary language. Bilingual corpora were used since different senses of some words often translate differently in another language. By using a parallel aligned corpus, the translation of each occurrence of such words can be used to determine their correct senses automatically. In our experiments, we use Myanmar-English parallel corpus as training corpus for Naïve Bayesian classifier.

There is no available Myanmar-English parallel corpus, which contains Myanmar polysemous words in public. So, we create Myanmar-English parallel corpus that contain ambiguous words manually. The corpus consists of approximately 45963 words. It contains various sense meanings of ambiguous Myanmar word. We present the following aligned sentences as part of the training corpus. The corpus structure of the following example sentences are as follow.

(1) သူသည်တူဖြင့်ခေါက်ဆွဲစားသည်။
He eats the noodle with **chopsticks**.
သူ စားသည် ခေါက်ဆွဲ ဖြင့် **တူ** ။

(2) သူ့မှာတူသုံးယောက်ရှိသည်။
He has three **nephews**.
သူ မှာရှိသည် သုံးယောက် **တူ** ။

(3) လက်သမားသည်တူကိုသုံးသည်။
Carpenter uses the **hammer**.
လက်သမား သုံးသည် **တူ** ။

- [0]သူ/[0]he[NN][1]တူ/[4]**chopsticks**[NN][2]ဖြင့်/[3]with[IN][3]ခေါက်ဆွဲ/[2]noodle[NN][4]စားသည်/[1]eats[VBZ]
- [0]သူ/[0]he[NN][1]**တူ**/[3]**nephew**[NN][2]သုံးယောက်/[2]three[CD][3]မှာရှိသည်/[1]has[VBZ]
- [0]လက်သမား/[0]carpenter[NN][1]**တူ**/[2]**hammer**[NN][2]သုံးသည်/[1]uses[VBZ]

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 sentence by sentence. As it is clear, the Myanmar word "တူ" are mapped into three different English words "chopsticks", "nephew" and "hammer" based on its sense. From the corpus, we extract the possible English meanings of ambiguous Myanmar word.

6. Naïve Bayesian Classifier for WSD

Naïve Bayes methods have been used in most classification work and were first used for WSD by Gale et al. (1992). Bayesian classifier for word sense disambiguation is that it looks at the words around an ambiguous word in a large context window. Each content wore contributes potentially useful information about which sense of the ambiguous word is likely to be used with it. The supervised training of the classifier assumes that we have a corpus where each use of ambiguous words is labeled with its correct sense. These context windows can be presented in two classes: Bag-of-word feature vectors – These are unordered set words with their exact position ignored. Collocation feature vectors – A collocation is a word or phrase in a position specific relationship to a target word.

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 si for the ambiguous word w by selecting the sense that maximizes the conditional probability P(w=si|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.

7. Overview of proposed System

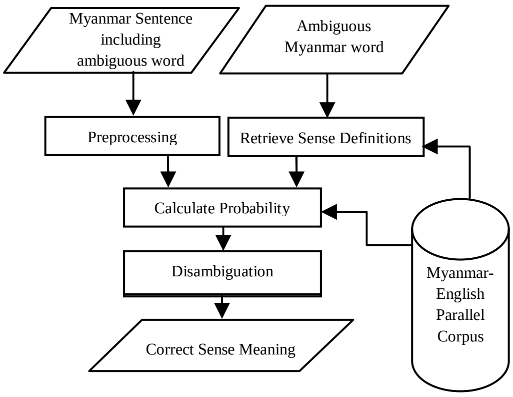
Myanmar Sentence including ambiguous word
Ambiguous Myanmar word
Preprocessing
Retrieve Sense Definitions
Calculate Probability
Disambiguation
Correct Sense Meaning
Myanmar-English Parallel Corpus

Figure 2. General Overview of proposed System

The proposed system consists of four main parts. These are:

1) Preprocessing
2) Multi-Senses look-up on Corpus
3) Calculating Probability based on Bayes Theorem
4) Disambiguation (Calculating maximum scores using Bayes decision rule)

Firstly, the system takes the Myanmar sentence including ambiguous word and the ambiguous Myanmar word that want to disambiguate as input. In preprocessing stage, it segments the input sentence by using Myanmar word segmentor and removes all words that can be stop words which are a list of common or general terms from the input sentence. Stop words include pronouns, preposition, conjunctions, particles etc. After gathering information in the preprocessing step, the system uses the remaining words in the input sentence as features. Secondly, the system also retrieves the possible English sense definitions of the ambiguous word from the corpus. Thirdly, the system also calculates prior probability and likelihood based on Bayes Theorem. Finally, disambiguation process is performed using Bayes decision rule. The disambiguation process compute the score of each sense of ambiguous word and decide the most appropriate sense for a given word in the test sentence. Our proposed algorithm is shown in the following figure 3.

```
1) Preprocessing
  a. Segment input sentence
  b. Remove stop words from input
2) Multi sense lookup
      -Lookup possible sense meanings of ambiguous
       word from the corpus
3) Calculating Probability
  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
4) 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 Myanmar WSD

8. Execution of Proposed WSD Algorithm

We give an example of the execution of our system and we try to disambiguate the word “ကျွန်း” (kjun) in the sentence: For example:

Input sentence:

“ကျွန်းသည်အလွန်အသုံးဝင်သောသစ်မာဖြစ်သည်။”

(Teak is a very useful hardwood.)

Input word: “ ကျွန်း(kjun)”

1) Preprocessing

In the preprocessing, we first segment the input sentence by using Myanmar word segmentor**.** After segmentation: we get the following sentence.

“ကျွန်း_သည်_အလွန်_အသုံးဝင်သော_သစ်မာ_ဖြစ်သည်_။_”.

Then, we remove all the function words (stop words). Stop words include pronouns, preposition, conjunctions and particles. After removing stop words: we obtain the following sentence.“ကျွန်း,အလွန်,အသုံးဝင်သော,သစ်မာ,ဖြစ်သည်” .

2) Multi-sense loop up

Secondly, we find the English meanings of Myanmar ambiguous word from the corpus. The word “ကျွန်း(kjun)” has two senses, **teak** and **island**. We have Bag-of-words: “အလွန်,အသုံးဝင်သော,သစ်မာ,ဖြစ်သည်”.

3) Calculating Probabilities for each sense

Thirdly, we find prior probabilities and likelihood of each sense. Assume the total word count of “ကျွန်း(kjun)” in corpus is 10 (4 times for teak and 6 times for island).

P(ကျွန်း=teak) = 0.4, P(ကျွန်း=island) = 0.6

For P(Fi/S=teak),

P(အလွန်/teak) =0.25,
P(အသုံးဝင်သော/teak) =0.25,
P(သစ်မာ/teak) =0.5,
P(ဖြစ်သည်/teak) =0.5,

For P(Fi/S=island),

P(အလွန်/island) =0.1,
P(အသုံးဝင်သော/island) =0.1,
P(သစ်မာ/island) =0.1,
P(ဖြစ်သည်/island) =0.33,

4) Disambiguation

Finally, we compute the score of each sense. For P(F1,F2,…Fn/S) , we multiply P(S) and P(Fi/S) for each sense :

P(F1,F2,…Fn/teak) = 0.4* 0.25 *0.25* 0.5 * 0.5
= **0.00625**

P(F1,F2,…Fn/island) = 0.6* 0.1 *0.1* 0.1 * 0.33
= 0.000198

After calculating the score of each sense, we can assign sense with highest probabilities to the word.So, we choose “teak” for the ambiguous word “ကျွန်း(kjun)”. By this way, we can disambiguate a word with multiple senses in a given context.

9. Experimental Result

The experiments are conducted using data drawn from “Myanmar-English Parallel Corpus", which contains sentences used in various domains. Our approach relies on supervised learning. The training set consists of 1000 sentence pairs, including 45963 words and test set contains 100 sentences and each sentence average long is 12 words. We have collected 60 ambiguous nouns and 100 ambiguous verbs for experiment. We used only the pure text data, and not the speech transcriptions.

For evaluation purpose, we group test sentences in three groups, first group sentences are composed of words in corpus. The second group sentences are composed of words in corpus but not exactly same sentences in corpus and the third sentence are composed of words not include in corpus. There are 50 sentences in first group, 30 sentences in the second group and 20 sentences in the third group. So there are altogether 100 sentences for evaluation.

Table.3. Experimental results on test data set

Sentence Type	**Accuracy (%)**
Test Sentences in training set	98.02%

Test Sentences composed of words in training sentences , but not exactly same sentences in training set	90.85%
Test Sentences that are not include in training set	82.91%
Average	90.6%

Table.3. shows the results of our experiment. The experiments show that disambiguation process by using the proposed method from the mentioned corpus, received about 91% overall accuracy in detecting the correct translation of ambiguous words. The 9% 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.

10. Conclusion and Future Work

This research was the first attempt to create a word sense disambiguation system for Myanmar Language. We use Naïve Bayesian Classifier 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 91% 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 Decision Lists and Trees, and various feature types. This system disambiguates the words with part of speech ‘Noun’ and ‘Verb’. We can also 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 machine translation, 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.

References

[1] A.Naseer and S.Hussain, “Supervised Word Sense Disambiguation for Urdu Using Bayesian Classification”, 2009.

[2] 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 8th-10th , 2004.

[3] Compaq Oxford Dictionary and Thesaurus.

[4] F. Ahmed and A. Nurnberger, “Arabic/English Word Translation Disambiguation using Parallel Corpora and Matching Schemes”, In Proceedings of the 12th EAMT conference, Hamburg, Germany, 22-23 September 2008.

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

[6] Ide and veronis, “Word Sense Disambiguation: The State of the Art.” Computational Linguistics, 1998.

[7] L.Merhbene, A.Zouaghi and M.Zrigui, “Ambiguous Arabic Words Disambiguation”, 11th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2010.

[8] N.Ide and J.Veronis, “Word Sense Disambiguation Computational Linguistics”, 1998,vol.24(1),1-42.

[9] 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).

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] Z.Zheng and Z.Shu, “A New Approach to Word Sense Disambiguation in MT System”, World Congress on Computer Science and Information Engineering, 2009.