International Journal of Computer

(IJC)

ISSN 2307-4523 (Print & Online)

© **_Global Society of Scientific Research and Researchers_**

http://ijcjournal.org/

Weight-based Word Sense Disambiguation Method for Myanmar-to-English Language Translation

Aye Mya Nyein[a]*,May Zin Oo[b]

[a,b]*Department of Information Technology, Mandalay Technological University, Myanmar*

[a]*ayemyanyein.it@gmail.com*

[b]*mayzinoo.mz@gmail.com*

Abstract

In many natural language processing (NLP) techniques, machine translation is a popular and useful technique. Machine translation technique is a translation process from one to another language. This technique is thus very useful for people around the world. While translating the languages, ambiguity is a big challenge because many words have several meanings. Ambiguous words have damaging effects on the precision of machine translation. To solve this problem, word sense disambiguation (WSD) method is useful for automatically identifying the correct meaning of an ambiguous word. In order to have a better precision, weight-based WSD method is proposed by taking advantage of a Minkowski distance method. As the proposed method considers the weight values of each sense of training and input vectors while observing the ambiguous words, it is more effective than the simple translation system. Experimental results show that the weight-based WSD method gives a better precision approximately 51% when compared to the simple machine translation method.

Keywords: Natural Language Processing; Word Sense Disambiguation; Machine Translation; Ambiguity

1. Introduction

In the computational linguistics field, machine translation is a computer-aided human translation that investigates the use of software to translate text to speech from one language to another. Machine translation can use a method based on linguistic rules, which means that words will be translated in a linguistic way – the most suitable words of the target language will replace the ones in the source language. The success of machine translation requires the problem of natural language understanding to be solved first.

----

* Corresponding author.

In this situation, ambiguity in natural language is a detrimental effect on the performance of machine translation (MT) process. To resolve the ambiguity, WSD is required since a word in the source language may have more than one possible translation in the target language. There are many WSD approaches. These are knowledge-based, semisupervised-based, supervised-basedand unsupervised-based WSD approaches. Among them, the proposed system uses knowledge-based WSD and supervised-based WSD methods.

This system is proposed as the Myanmar-to-English MT system by using both the WSD method and rule-based MT method. To support the WSD method, this system uses the Bilingual corpus as the knowledge resource that is created by using English and Myanmar languages. To know the better performance of weight-based WSD method, the proposed system compares its performance to the simple machine translation method.

2. Related Work

In 2007, D. Chiang and Y. S. Chan [1] presented conflicting evidence on whether word sense disambiguation systems that can help to improve the performance of statistical machine translation (MT) systems. This paper successfully integrates a state-of-the-art WSD system into a state-of-the-art hierarchical phrase-based MT system. They show for the first time that integrating a WSD system improves the performance of a state-of-the-art statistical MT system on an actual translation task.

In 2007, M. Carpua and D. Wu [2] described for the first time that incorporating the predictions of a word sense disambiguation system within a typical phrase-based statistical machine translation (SMT) model that consistently improves translation quality. They addressed a new strategy for integrating WSD into an SMT system that performs fully phrasal multi-word disambiguation. Instead of directly incorporating a senseval-style WSD system, they redefined the WSD task to match the exact same phrasal translation disambiguation task faced by phrase-based SMT systems. The results provide the first known empirical evidence that lexical semantics are indeed useful for SMT systems.

In 2013, M. Carpuat [3] presented the National Research Council Canada (NRC) submission to the Spanish Cross-Lingual Word Sense Disambiguation task at SemEval-2013. Since this WSD task uses Spanish translations of English words as gold annotation, it can be cast as a MT problem. They submitted the output of a standard phrase-based system as a baseline, and investigated ways to improve its sense disambiguation performance. Using only local context information and no linguistic analysis beyond lemmatization, the proposed machine translation system surprisingly yields top precision score based on the best predictions.

According to literature and concepts pointed out from the previous works, the proposed system intends to present the weight-based WSD Myanmar-English machine translation system. Moreover, this system compares the weight-based WSD MT system and simple MT system.

3. Background Theory

In this section, this system describes the background theories about the word sense disambiguation method and machine translation method. For the simple MT system, this system uses only the MT method. But, this system uses both WSD method and MT method for weight-based WSD machine translation system.

3.1. Word Sense Disambiguation

Word sense disambiguation (WSD) is used to find the correct meaning of the sense or the word. WSD is usually performed on one or more texts although in principle bags of words, i.e., collections of naturally occurring words might be employed [4]. WSD can be viewed as a classification task: word senses are the classes, and an automatic classification method is used to assign each occurrence of a word to one or more classes based on the evidence from the context and from external knowledge sources such as Thesauri, Ontology, Machine readable dictionaries (MRD), WordNet and Bilingual Corpus. Among them, this system uses Bilingual corpus within WSD for finding semantically related words [5, 6].In the WSD approach, there are four main sub- approaches.

- Knowledge-based WSD approach,
- Semi-supervised based WSD approach,
- Supervised based WSD approach and
- Unsupervised based WSD approach [6].

Among them, the proposed system uses knowledge- based WSD and supervised-based WSD approaches. In the knowledge-based WSD approach, external knowledge resources are used to disambiguate the ambiguous word. Bilingual corpus is also used for the knowledge-based WSD approach. In the supervised based WSD approach, the classifier with training data is used for disambiguation. In this system, Minkowski classification method is modified as the weight-based WSD method to contribute a better precision capability.

3.1.1. Weight-based WSD Method

In this system, the weight method is used to compute the degree of relevance between the training vector and input vector. The proposed weight method is as follows:

$$d(i, j) = ((|w_{i1}-w_{j1}|)^q + (|w_{i2}-w_{j2}|)^q + \ldots\ldots + (|w_{ip}-w_{jp}|)^q)^{1/q} \qquad (1)$$

where d(i,j) is distance between training vector *i* and input vector *j*.w_{ip} is weight of the term *p* within training vector *i*.$w_{jp}$is weight of the term *p* within input vector *j*.

For the term weight within training vector, the weigh calculation equation is as follows:

$$w_{ip} = tf_{ip} \times idf_p \qquad (2)$$

where w_{ip} is weight of the term *p* in the training vector *i*.$tf_{ip}$is normalize term frequency of term *p* in training vector *i*.idf_p is inverse training vector frequency of term *p*.

$$tf_{ip} = \frac{f_{ip}}{max\{ f_{i1}, f_{i2}, \ldots, f_{i|V|}\}} \qquad (3)$$

where f_{ip} is raw frequency count of term *p* in the training vector *i*.

$$idf_p = log \frac{N}{df_p} \tag{4}$$

where df$_p$ is number of training vector in which term *p*appears at least once.Nis the total number of training vector in the system.

For the term weight within input vector, the weigh calculation equation is as follows:

$$w_{jp} = \left[0.5 + \frac{0.5 f_{jp}}{max\{ f_{j1}, f_{j2}, ..., f_{j|V|} \}} \right] \times log \frac{N}{df_p} \tag{5}$$

where w$_{jp}$is weight of term *p* in the input vector *j*. f$_{jp}$is the raw frequency count of term *p* in the input vector *j*.

3.1.2. Bilingual Corpus

Bilingual corpus is used as the external knowledge source. Knowledge is a fundamental component of WSD. Knowledge sources provide data which are essential to associate senses with words. Bilingual corpus is used since different senses of some words often translate differently in another language. This system uses Myanmar-English Bilingual corpus for Myanmar and English languages [6]. This corpus stores various English senses about ambiguous Myanmar words. Table 1 shows the sample bilingual corpus.

Table 1: Sample Bilingual Corpus

ID	Myanmar Word	No of English Sense	English Sense1	English Sense2	English Sense3
1	တူ	3	chopsticks	nephews	Hammer
2	ဈေး	2	price	market	-
3	နာရီ	3	clock	Hour	Watch

3.2. Machine Translation (MT)

Machine Translation system is based on a direct translation scheme with a limited syntactic grammar. The mechanism is quite simple replacing the equivalent words in the source language to produce the target language. It works for the language pairs that have similar grammar and word formation rules. Many new MT approaches have been proposed such as syntax or semantic transfer method, inter-lingual method implemented in rule-based, example-based and statistical-based paradigms [7].

3.2.1. MT Approach

In the MT approach, there are three main sub-approaches. These are syntax or semantic transfer method, inter-lingual method implemented in rule-based and example-based, and statistical-based paradigms [7]. Among them, the proposed system uses the second MT rule-based approach.

3.2.2. Inter-lingual Method Implemented in Rule-based

This method has to do with the morphological, syntactic and semantic information about the source and target language. This method consists of collection of rules that are called grammatical rules and, lexicon and software program to process the rules. Grammatical rules are written with linguistic knowledge gathered from linguists. Linguistic (grammatical) rules are built over source and target language information. Also millions of bilingual dictionaries for the language pair are used [7]. Table 2 shows the Myanmar linguistics rules and the English linguistics rules.

Table 2: Linguistic Rules for Machine Translation

Rule No	Myanmar Linguistic Rule	English Linguistic Rule
1	Subject + Object + Verb	Subject + Verb + Object
2	Adjective + Noun + Article + Verb	Article + Adjective + Noun + Verb
3	Adjective + Noun + Article + Adverb + Verb	Article + Adjective + Noun + Verb + Adverb
4	Pronoun + Noun1 + Preposition1 + Noun2 + Preposition2 + Verb	Pronoun + Verb + Preposition1 + Noun1 + Preposition2 + Noun2
5	Pronoun + Noun + Preposition + Verb	Pronoun + Verb + Preposition + Noun

3.2.3. Reordering Process

Reordering is essential to the translation of languages with different word orders. It is needed to get correct order in Myanmar-English translation system. Reordering is the movements of the word by their part-of-speech (POS) position by using the rules of movement POS. In this system, the raw English sentence from translation process will be reordered by using English Linguistic rules [8].

4. Proposed System Design

In this section, the proposed system design, the explanation, implementation and experimental results of the proposed system are described.

4.1. Overall Proposed System Design

The proposed system translates from the Myanmar sentence to the English sentence. Myanmar language has many ambiguous words. As it can face the problem for Myanmar-to-English translation, this system proposes the weight-based machine translation process based on WSD.Moreover, this system compares the performance between the weight-based machine translation process and simple machine translation process. In this system, there are two types of process. These are weight-based machine translation process and simple machine translation process.

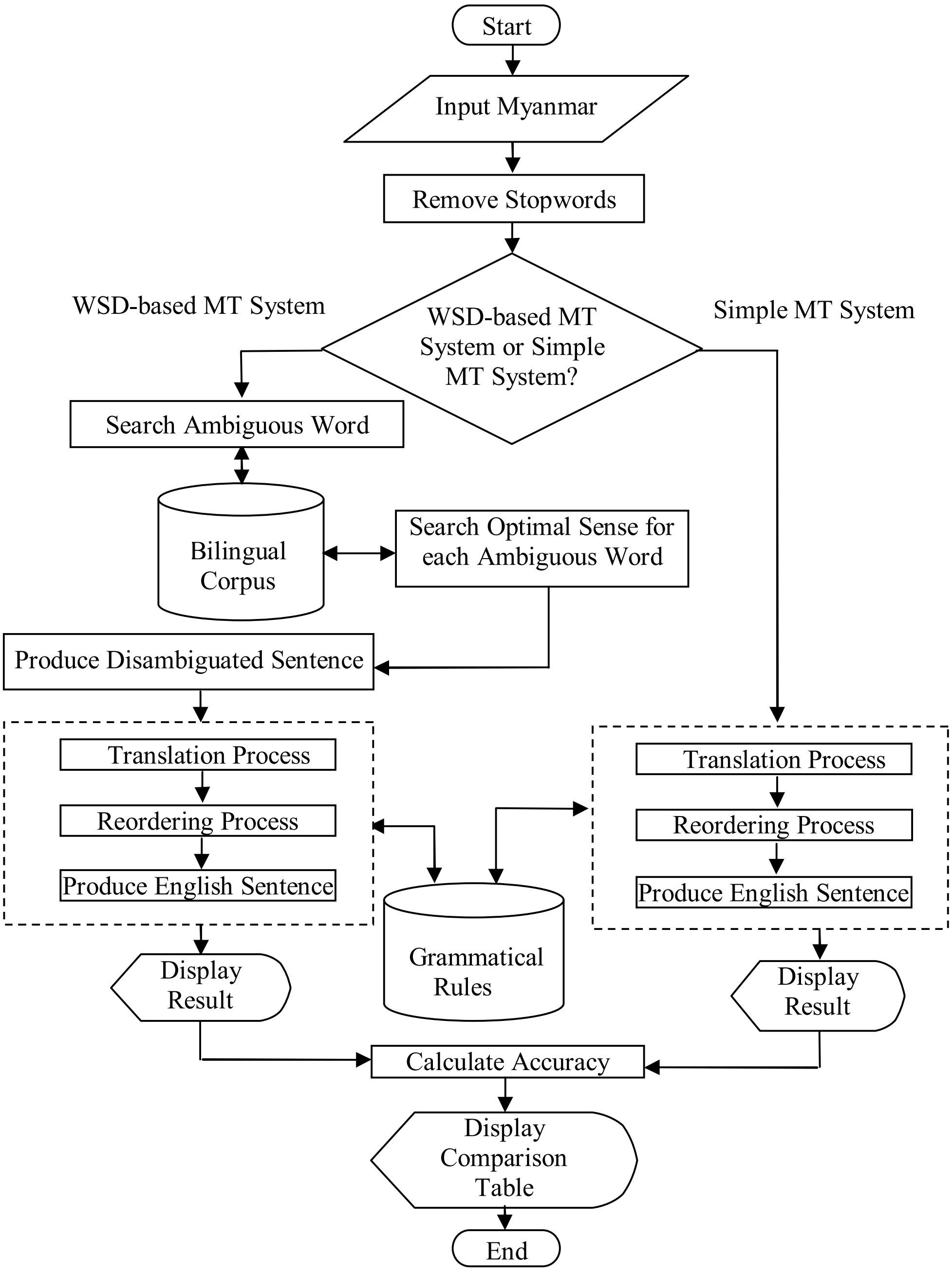


Figure 1: Overall Proposed System

In the weight-based machine translation process, the user must firstly input the Myanmar sentence. And then, this system removes the stopwords. There are two types of word that are stopwords and keywords. Stopwords are frequently occurring and insignificant words in a language that help to construct sentences but do not represent any content of the sentence. Keywords are meaningful words in the sentence.After removing the stopwords, this system searches the ambiguous word from the Myanmar sentence. If one Myanmar word has more than one English meaning, this word is ambiguous word. Otherwise, this word is disambiguous word. For ambiguous word searching process, this system uses the weight-based WSD algorithm and Bilingual corpus. This corpus stores the word and its meanings that are created by using Myanmar and English language.

After disambiguating ambiguous word, this system generates the disambiguated sentence. And then, the rule-based machine translation method is used to translate from the Myanmar to English sentences. According to the rule-based method, this system creates the Myanmar and English grammatical rules. Using these rules enablesEnglish words to reorder. After that, this system produces the English sentence as the translated sentence. In the simple machine translation method, this system uses only the rule-based MT on process. If the Myanmar words are ambiguous words, the simple MT method can face the problem that cannot translate the correct sense for ambiguous words. The proposed system is expected to achieve a better performance using weight-based WSD MT process. Experiment sections are proved by comparing the proposed method to the MT system, in terms of the precision. The overall proposed system design is shown in Figure 1. As a simple translation system, the Google translator is used to ascertain the performance differences of between two MT systems.

4.2. Explanation of the Proposed WSD-based MT System

As a sample, this system is explained using one Myanmar sentence. This sentence is "ကျွန်းသည်အလွန်အသုံးဝင်သောသစ်မာဖြစ်သည်". After inputting a sentence, the system removes stopwords from this sentence and searches the ambiguous word from this sentence. If the Myanmar word has more than one English meaning, this word assumes the ambiguous word. In this sample, "ကျွန်း" is ambiguous word because it has two meanings: "Teak" and "Island".

In this situation, this system solves this ambiguous word into the disambiguous word. For solving, this system mustuse the gloss of senses about ambiguous word. There are two senses in "ကျွန်း" Myanmar word, which has one of the ambiguous words of Myanmar sentence. The gloss of "Teak" sense is "အိမ်ဆောက်လုပ်ရာတွင်အသုံးဝင်သည်။ပရိဘောဂများအတွက်လည်းအသုံးဝင်သည်။". And, the gloss of "Island" sense is "ရေပါတ်လည်ဝန်းရံနေသောကုန်းဖြစ်ပြီးပင်လယ်ထဲတွင်ရှိသည်။". This system removes stopwords from these glosses and extracts keywords. Stopwords and keywords are shown in Table 3.

Table 3: Stopwords and Keywords form Glosses

ID	Keywords	Stopwords
1	အိမ်	ရာတွင်
2	ဆောက်လုပ်	အတွက်
3	အသုံးဝင်	လည်း
4	ပရိဘောဂများ	နေသော
5	ရေ	ထဲတွင်
6	ပါတ်လည်ဝန်းရံ	သည်
7	ကုန်း	
8	ပင်လယ်	

After extracting keywords and stopwords from the glosses of senses and input Myanmar sentence, this system calculates the weight of each keyword and similarity between input sentence and gloss of senses. Sample calculation is shown as follows and Table 4. Table 4 shows the weight of each keyword.

$$Tf_{\text{sense1,အိမ်}} = f_{\text{sense1,အိမ်}} / \max\{f_{\text{sense1,အိမ်}}, f_{\text{sense1,ဆောက်လုပ်}}, f_{\text{sense1,အသုံးဝင်}}, f_{\text{sense1,ပရိဘောဂများ}}\}$$

$$= 1/ \max \{1, 1, 2, 1\} = 1/2 = 0.5$$

$$IDf_{\text{အိမ်}}= \log 2/1 = 0.30103$$

$$w_{\text{sense1,အိမ်}}= 0.5 * 0.30103 = 0.150515$$

$$w_{\text{inputted Myanmar sentence, အလွန်}}=[0.5 + (0.5 * 1)/ \max\{1, 1, 1, 1, 1\}] * \log 2/ 0 = 0.30103$$

Table 4: Weight of Each Keyword

ID	Keywords	TF_{ip}	IDF_p	Weight
1	အိမ်	0.5	0.30103	0.15051
2	ဆောက်လုပ်	0.5	0.30103	0.15051
3	အသုံးဝင်	1	0.30103	0.30103
4	ပရိဘောဂများ	0.5	0.30103	0.15051
5	ရေ	1	0.30103	0.30103
6	ပါတ်လည်ဝန်းရံ	1	0.30103	0.30103
7	ကုန်း	1	0.30103	0.30103
8	ပင်လယ်	1	0.30103	0.30103

Similarity results are as follows:

$$d(\text{sense 1, input sentence}) = ((|w_{\text{sense1,အလွန်}} - w_{\text{input sentence, အလွန်}}|)^2 + (|w_{\text{sense1,အသုံးဝင်}} - w_{\text{inputsentence, အသုံးဝင်}}|)^2 + (|w_{\text{sense1,သစ်မှာ}} - w_{\text{input sentence, သစ်မှာ}}|)^2 + (|w_{\text{sense1,ဖြစ်သည်}} - w_{\text{input sentence, ဖြစ်သည်}}|)^2)^{1/2}$$

$$= ((|0-0.30103|)^2+ (|0.30103-0.30103|)^2 + (|0-0.30103|)^2 + (|0-0.30103|)^2)^{1/2}$$

$$= (0.09061 + 0 + 0.09061 + 0.09061)^{1/2} =0.5213$$

$$d(\text{sense 2, input sentence}) = ((|w_{\text{sense2,အလွန်}} - w_{\text{input sentence, အလွန်}}|)^2 + (|w_{\text{sense2,အသုံးဝင်}} - w_{\text{inputsentence, အသုံးဝင်}}|)^2 + (|w_{\text{sense2,သစ်မှာ}} - w_{\text{input sentence, သစ်မှာ}}|)^2 + (|w_{\text{sense2,ဖြစ်သည်}} - w_{\text{input sentence, ဖြစ်သည်}}|)^2)^{1/2}$$

$$= ((|0-0.30103|)^2+ (|0-0.30103|)^2 + (|0-0.30103|)^2 + (|0-0.30103|)^2)^{1/2}$$

$$= (0.09061 + 0.09061 + 0.09061 + 0.09061)^{1/2} = 0.60202$$

According to the similarity results, the “Teak” English sense is the most relevant sense with “ကျွန်း” ambiguous word. To translate from the Myanmar sentence to the English sentence, this system uses the Bilingual corpus and rule-based method. For this input Myanmar sentence, the sample Bilingual corpus is shown in Table 5.

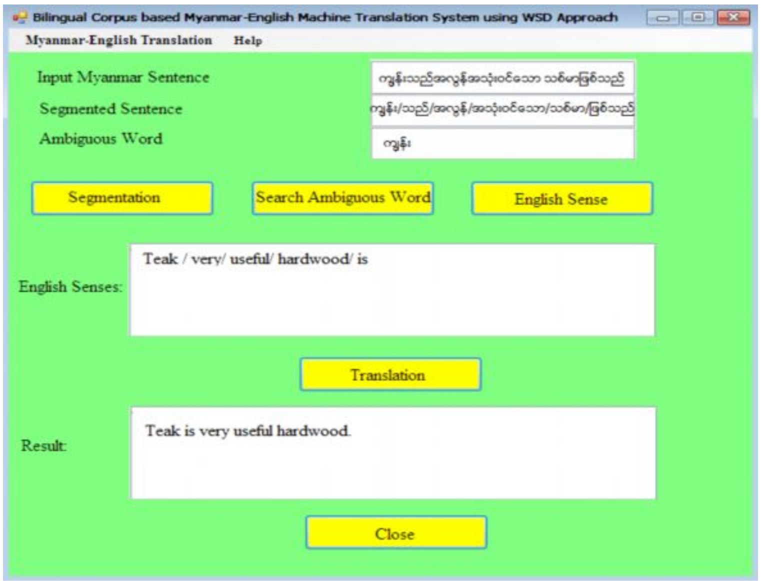
Table 5: Sample Bilingual Corpus

ID	Myanmar Keywords	English Keywords
1	ကျွန်း	Teak
2	အလွန်	Very
3	အသုံးဝင်	Useful
4	သစ်မာ	Hardwood
5	ဖြစ်သည်	Is

According to the grammatical rules, the Myanmar sentence is translated into the English sentence. The result of Myanmar sentence is “Teak is very useful hardwood”. The above explanation is about the process of weight-based WSD machine translation system. But, in the simple machine translation system, this system uses only the Bilingual corpus and grammatical rules. So, this simple machine translation system can face the problem about ambiguous word.

4.3. Implementation of the Proposed System

The WSD-based machine translation system is shown in Figure 2. The proposed translation system uses weight-based WSD method.



Bilingual Corpus based Myanmar-English Machine Translation System using WSD Approach
Myanmar-English Translation
Help
Segmentation
Search Ambiguous Word
English Sense
Translation
Close

Input Myanmar Sentence: ကျွန်းသည်အလွန်အသုံးဝင်သော သစ်မာဖြစ်သည်

Segmented Sentence: ကျွန်း/သည်/အလွန်/အသုံးဝင်သော/သစ်မာ/ဖြစ်သည်

Ambiguous Word: ကျွန်း

English Senses: Teak / very/ useful/ hardwood/ is

Result: Teak is very useful hardwood.

Figure 2: Weight-based WSD Machine Translation System

4.4. Experimental Results of the Proposed System

This system is tested by using different twenty-seven Myanmar sentences. To know the performance of the proposed system, the weight-based WSD-based machine translation system compares the simple machine translation system. As a sample, the results of weight-based WSD MT system are shown in Table 6.

Table 6: Sample Translation Results

ID	Input Myanmar Sentence	English Sentence from the WSD-based Machine Translation	English Sentence from the Simple Machine Translation
1	ကျွန်းသည်အလွန်အသုံးဝင်သောသစ်မာဖြစ်သည်။	Teak is a very useful heard-wood.	The island is very useful hardwood.
2	အင်္ဂလိပ်ဘာသာစကားသည်နိုင်ငံတကာဘာသာစကားဖြစ်သည်။	English language is international language.	English language is international language.
3	သူသည်ခေါက်ဆွဲကိုတူဖြင့်စားသည်။	He eats noodle with chopsticks.	He is like a noodle.

To measure the performance of this system, a precision method is used. Precision is the ratio of number of correctly translated sentences to the number of all translated sentences.

$$Precision = \frac{Number\ of\ Correct\ Translated\ Sentences}{Number\ of\ Proposed\ Translated\ Sentences} \times 100\% \qquad (6)$$

For experimental results, this system is tested with 50 ambiguous sentences and 50 simple sentences and the results are shown in Tables 7 and 8. In the simple machine translation system, the 32% correct results are obtained by testing ambiguous sentences. But, when the simple sentences are tested, the 64% correct results are obtained. When the sentences are tested using the proposed weight-based WSD method, this system can give the 98% correct results for ambiguous sentences. However, the 100% correct results are obtained about simple sentences that have no ambiguous words because the proposed system is well-trained. Although this system can give the 100% precision for trained sentences that have ambiguous words, its performance degrades approximately 50% when the unknown sentences are tested. Therefore, the proposed system needs to be well-trained to achieve a higher precision for those sentences whether they have ambiguous words or not. However, the proposed system can correctly distinguish the ambiguous words and translate the sentences due to the fact that the sentences are well-trained in the Bilingual corpus. The more many sentences are well-trained, the more the system mimics the simple MT system not only for the unknown sentences, but also for the ambiguous sentences.

Table 7: Testing Results about Simple Machine Translation System

ID	Simple Machine Translation System	Precision Result
1	Ambiguous Sentences	32 %
2	Simple Sentences	64 %

Table 8: Testing Results about Weight-based WSD Machine Translation System

ID	Weight-based WSD Machine Translation System	Precision Result
1	Ambiguous Sentences (Trained-data in Bilingual Corpus)	98 %
2	Simple Sentences (Trained-data in Bilingual Corpus)	100 %

5. Conclusion

A simple Myanmar-to-English machine translation system is not very effective to disambiguate the sentences that involve ambiguous words. Such problem influences on the effectiveness of the translation system. Several researchers have been trying to solve such problem. Although traditional word sense disambiguation method can disambiguate the ambiguous words, the precision is still a big challenge. Therefore, this paper proposes weight-based WSD method using Bilingual Corpus, which enables not only to search the correct sense of ambiguous Myanmar words, but also to give precision while translating the sentences of ambiguous words. Experimental results show that the proposed system performs very well while translating the ambiguous sentences and gives a better precision about 66% for ambiguous sentences as well as 36% for simple sentences.

6. References

[1] D. Chiang and Y. S. Chan, *Word Sense Disambiguation Improves Statistical Machine Translation*, Department of Computer Science National University of Singapore, 2007.

[2] M. Carpuat and D. Wu, "Improving Statistical Machine Translation using Word Sense Disambiguation", Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 61-72, 2007.

[3] M. Carpuat, "NRC: A Machine Translation Approach to Cross-Lingual Word Sense Disambiguation", Second Joint Conference on Lexical and Computational Semantics, pp. 188-192, 2013.

[4] S. ViswanadhaRaju, J. Sreedhar and P. Pavan Kumar, "Word Sense Disambiguation: An Empirical Survey", *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 2, pp. 494-503, May. 2012.

[5] R. Giyanani, "A Survey on Word Sense Disambiguation", *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol. 14, pp. 30-33, Sep. 2013.

[6] R. Navigli, "Word Sense Disambiguation: A Survey", *ACM Computing Surveys*, vol. 41, pp. 1-59, Feb. 2009.

[7] BushraJawaid, *MachineTranslation with Significant Word Reordering and Rich Target-Side Morphology*, Faculty of Mathematics and Physics, Charles University in Prague, UK, January 7, 2013.

[8] Jing He, *Word-reordering for Statistical Machine Translation using Trigram Language Model*, Tsinghua University, Beijing, China, November 8, 2011.