WORD SENSE DISAMBIGUATION BASED MYANMAR-TO-ENGLISH MACHINE TRANSLATION SYSTEM

[1]AYE MYA NYEIN, [2]MAY ZIN OO

[1,2]Department of Information Technology, Mandalay Technological University
E-mail: [1]ayemyanyein.it@gmail.com, [2]mayzinoo.mz@gmail.com

Abstract- Today, word sense disambiguation (WSD) is an important technique for many natural language processing (NLP) applications such as grammatical analysis, content analysis, information retrieval and machine translation. Among them, the WSD technique is used for machine translation to find the correct sense of a word in a specific context. In machine translation, the input sentences in the source language are disambiguated in order to translate correctly in the target language which is Myanmar language that has many ambiguous words. Therefore, the distance-based WSD method is used for resolving ambiguity of words in Myanmar language. This system uses the Bilingual Corpus as the training data. This system can solve the semantic ambiguous problems that usually happen in Myanmar-to-English translation. After disambiguating, the rule-based method is used for language translation. This system intends to improve the precision of the Myanmar-to-English language translation system.

Index Terms- ambiguity, disambiguation, machine translation, word sense disambiguation, bilingual corpus.

I. INTRODUCTION

Word sense disambiguation (WSD) has been a hot topic for natural language processing (NLP) applications such as grammatical analysis, content analysis, information retrieval and machine translation. The most obvious application of WSD is machine translation. MT process requires for understanding the source language for translation and generating sentences in the target language. In this situation, ambiguity in natural language is a detrimental effect on the performance of machine translation 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. Among many languages such as Chinese language, Japan language, English language and Myanmar language, this system is implemented for Myanmar language and English language. By using distance-based WSD method and rule-based method, this system is proposed as the Myanmar-to-English machine translation system. To disambiguate ambiguous words in Myanmar language, this system also uses the Bilingual corpus as the lexical resources. In this proposed Myanmar-to-English machine translation system, there are two parts. In the first part, the ambiguous words in the source sentence that is written in Myanmar language are disambiguated using distance-based WSD method and Bilingual corpus. To produce English word sense, some Myanmar words have more than one English meaning. For example, Myanmar word " " has two English meanings such as "Teak" and "Island". So, it is needed to disambiguate each ambiguous word. In the second part, the English sentence is translated according to the rule-based method after finishing disambiguation process. By using WSD process, this system overcomes the barrier about machine translation. The rest of the paper is organized as follows: section 2 presents related works. Background theories about word sense disambiguation and machine translation are presented in section 3 and 4. The design of the system, explanation and experimental results are described in section 5. Finally, conclusion is drawn in section 6.

II. RELATED WORK

S. M. Fakhrahmad, A. R. Rezapour and M. H. Sadreddini [1] presented a supervised learning method for WSD, which s based on K-Nearest Neighbor algorithm. They extracted two sets of features; the set of words that have occurred frequently in the text and the set of words surrounding the ambiguous word. In order to improve the classification accuracy, they described a feature weighting strategy. The results are encouraging comparing to state-of-the-art. D. Chiang and Y. S. Chan [2] 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 ntegrated a state-of-the-art WSD system into a state-of-the-art hierarchical phrase-based MT system. They showed for the first time that integrating a WSD system mproved the performance of a state-of-the- art statistical MT system on an actual translation task. S. K. Srivatsa and P. Tamilselvi [3] implemented disambiguation system with three different set of features with three different distance measuring functions combined with three different classifiers for word sense disambiguation. Using Neural Networks with enormous number of features, accuracy was

achieved 97.40% for words with more than two senses and 75% for words with two senses. According to literatures and concepts from the previous works, this system intends to present the distance-based WSD method for Myanmar-English machine translation system.

III. WORD SENSE DISAMBIGUATION

Word sense disambiguation is used to find the correct meaning of the sense or the word. Word sense is a commonly-accepted meaning of a 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 with WSD for finding semantically related words [5, 6]. There are four WSD approaches. These are knowledge-based WSD, semi-supervised-based WSD, supervised-based WSD and unsupervised-based WSD approaches. Among them, the distance-based WSD method, based on knowledge based WSD and supervised based WSD approaches is used.

A. Distance-based WSD Method

To produce English senses for each ambiguous word in the Myanmar sentence, this system applies to Minkowski distance method [7] as shown in equation 1. Minkowski distance method is useful to measure the distance between two objects.

$$dist(x_i,x_j)=((|x_{i1}-x_{j1}|)^q+(|x_{i2}-x_{j2}|)^q+.....+(|x_{ip}-x_{jp}|)^q)^{1/q} \quad (1)$$

where dist(xi , xj) is the distance between the two objects. This method is based on 1, 0 logic. If the feature of object xi is occurred, this is defined as 1. Otherwise, this is defined as 0 [4]. However, the Minkowski method does not consider the weight of each feature in object xi. Thus, this system presents the distance-based WSD method is proposed by considering the weight results of each featuring to sense the ambiguous words. To produce English senses for each ambiguous word in the Myanmar sentence, this method computes the degree of relevance between the training vector and input vector. This WSD method is as follows:

$$d(i,j)=((|w_{i1}-w_{j1}|)^q+(|w_{i2}-w_{j2}|)^q+.....+(|w_{ip}-w_{jp}|)^q)^{1/q} \quad (2)$$

where d(i, j) is the distance between training vector i and input vector j. wip is the weight of the term p within training vector i. wjp is the weight of the term p within input vector j. For the weight within training vector, the weigh calculation equation is as follows:

$$w_{ip} = tf_{ip} \times idf_p \quad (3)$$

where wip is weight of the term p in the training vector i. tfip is the normalize term frequency of term p in training vector i. idfp is the inverse training vector frequency of term p.

$$tf_{ip} = \frac{f_{ip}}{\max\{f_{i1}, f_{i2}, \ldots, f_{i|V|}\}} \quad (4)$$

where fip is the raw frequency count of term p in the training vector i.

$$idf_p = \log \frac{N}{df_p} \quad (5)$$

where dfp is the number of training vector in which term p appears at least once. N is the total number of training vector in the system. For the term weight within input vector, the weigh calculation equation is as follows:

$$w_{jp} = \left[0.5 + \frac{0.5 f_{jp}}{\max\{f_{j1}, f_{j2}, \ldots, f_{j|V|}\}}\right] \times \log \frac{N}{df_p} \quad (6)$$

where wjp is the weight of the term p in the input vector j. fjp is raw frequency count of term p in input vector j [7].

B. Bilingual Corpus

As the external knowledge source, Bilingual corpus is used [8]. Knowledge is a fundamental component of word sense disambiguation. Knowledge sources provide data which are essential to associate because 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. This corpus stores various English senses about ambiguous Myanmar words.

IV. MACHINE TRANSLATION

Machine Translation (MT) 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 described such as syntax or semantic transfer method, inter-lingual method implemented in rule-based, example-based and statistical-based paradigms [9].

A. Rule-based Machine Translation Method

Rule-based machine translation method has to do with the morphological, syntactic and semantic information about the source and target languages. 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 pairs are used [9]. Some of linguistic rules are shown in Table I.

TABLE I. Linguistic Rules for Machine Translation

Rule No	Myanmar Linguistic Rules	English Linguistic Rules
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

B. Reordering

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 [10]. Sample reordering process is shown in Figure 1.

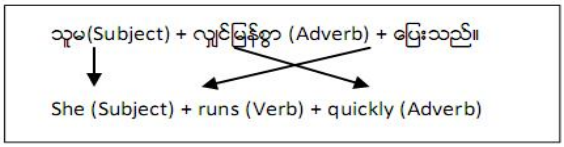

Fig 1: Sample Reordering Process

V. SYSTEM DESIGN

This system is implemented as the Myanmar-English machine translation system. The detail flow chart of the system is shown in Figure 2.

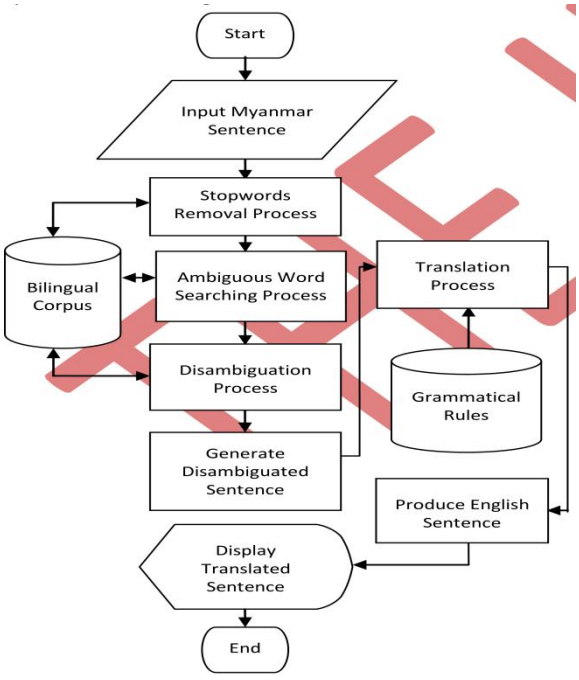

Fig 2: Flow chart of the System

In this system, a user firstly needs to input Myanmar sentence. After accepting, this system segments the input Myanmar sentence. As the stop-words are not important, this system removes the stop-words from this sentence. And then, this system searches the ambiguous word in the Myanmar sentence. If ambiguous words are found, this system disambiguates for each ambiguous word by using Bilingual corpus and distance-based WSD method. And then, this system produces the disambiguated Myanmar sentence. By using rule-based machine translation method, this system translates from the Myanmar sentence to the English sentence. According to the linguistic rules, this system produces the English sentence as the output.

A. Explanation of the System

This system provides the better performance of Myanmar-English language translation. As a sample, the input Myanmar sentence is shown in Figure 3.

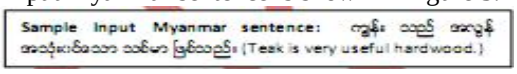

Fig 3: Sample Myanmar Sentence

This system tokenizes the input Myanmar sentence. Tokenization process is shown in Figure 4.

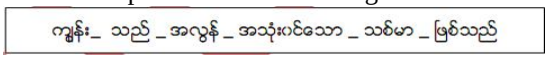

Fig 4: Tokenization (Segmentation) Process

And then, this system removes stop-words from the inputted Myanmar sentence. Sample stop-words are " " and " ". After removing stop-words, this system extracts the keywords from the input Myanmar sentence to translate the English word. Keywords in the input Myanmar sentence are " (Teak), (very), (useful), (hardwood) and (is)". By using Bilingual corpus, this system checks the keywords whether they are ambiguous words or not. Bilingual corpus stores various English senses about ambiguous Myanmar words.

If a Myanmar word has more than one English sense, this system assumes this Myanmar word as an ambiguous word. In this sample, the Myanmar word ( ) has two English senses that are "teak" and "island". So, this word is an ambiguous Myanmar word. This system disambiguates this ambiguous word by using distance-based WSD method and glosses from the Bilingual corpus. Sample sense and its glosses are shown in Table II.

TABLE II. Sample Sense and Its Gloss

Myanmar Word	English Sense and Its Gloss	
ကျွန်း	teak-	(အဖိုးတန်သော၊ အသုံးဝင်သော သစ်မာဖြစ်သည်။ တံတားများ ဆောက်ရာတွင် သုံးသည်။)
	island-	(ရေပါတ်လည် ပန်းရံနေသော နေရာ ကို ခေါ်သည်။ ပင်လယ်ထဲတွင် ရှိသည်။)

And then, this system creates the input vector and each training vectors to disambiguate the ambiguous Myanmar word. The input vector is created from the input Myanmar sentence and each training vector is also created from the gloss of each English sense. After creating each vector, this system calculates weight of each term in each vector. By using the weight results about each term, this system calculates the distance between the input vector and training vectors. After calculating, the training English sense that has the smallest distance value among other training senses is selected for ambiguous Myanmar word. There are two training vectors for Myanmar word " " as shown in Table II. The distance between input vector and first training vector is 0.306. The distance between input vector and second training vector is 0.912. Which this system thus chooses first sense of training vector "teak", which is the most relevant Myanmar word ( ). Then, this system produces each English sense that is relevant with each Myanmar word in the input Myanmar sentence. Myanmar words and English senses are shown in Table III.

TABLE III. Myanmar Words and English Senses

ID	Myanmar Word	English Sense
1	ကျွန်း	Teak
2	အလွန်	Very
3	အသုံးဝင်သော	Useful
4	သစ်မာ	Hardwood
5	ဖြစ်သည်	Is

Afterwards, this system uses the Linguistic rules and reordering process to translate from the Myanmar sentence to the English sentence. The translation process is shown in Figure 5.

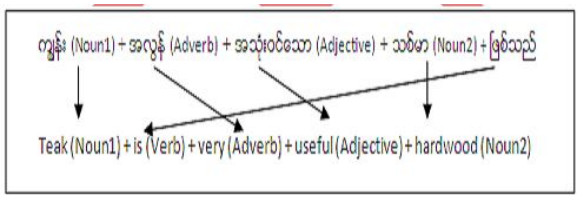

Fig 5. Translation Process

And then, this system produces English sentence ("Teak is very useful hardwood") from the Myanmar sentence

B. Implementation of the System

For machine translation, this system consists of three processing steps. These are preprocessing step, word sense disambiguation step and rule-based machine translation step. The WSD-based machine translation process is shown in Figure 6.

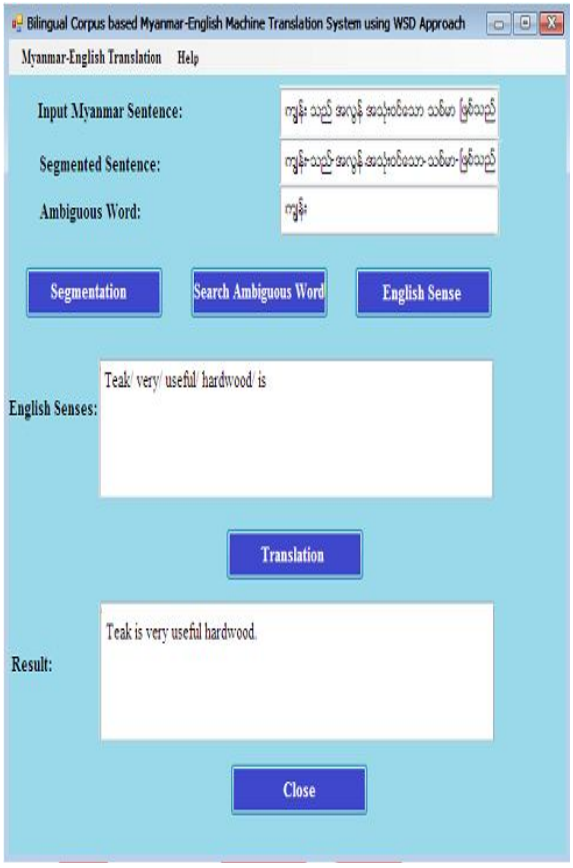

Fig 6. WSD-based Machine Translation Process

C. Experimental Results of the System

This system is tested by using different fifty Myanmar sentences. To know the performance, this system compares the simple machine translation system with the WSD-based machine translation system. The sample translations are shown in Table IV.

TABLE IV. Sample Translation Results

ID	Input Myanmar Sentence	English Sentence from the WSD-based Machine Translation	English Sentence from the Simple Machine Translation
1	ကျွန်း သည် အလွန် အသုံးဝင်သော သစ်မာ ဖြစ်သည်။	Teak is a very useful hardwood.	ERROR
2	အင်္ဂလိပ် ဘာသာစကား သည် နိုင်ငံတကာဘာသာစကား ဖြစ်သည်။	English language is international language.	English language is international language.
3	သူသည် ခေါက်ဆွဲကို တူဖြင့် စားသည်။	He eats noodle with chopsticks.	ERROR

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

$$\text{Precision} = \frac{\text{Number of Correct Translated Sentences}}{\text{Number of Proposed Translated Sentences}} \times 100\% \quad (7)$$

From the above sample translations, the performance analysis results of this translation system shown in Figure 7.

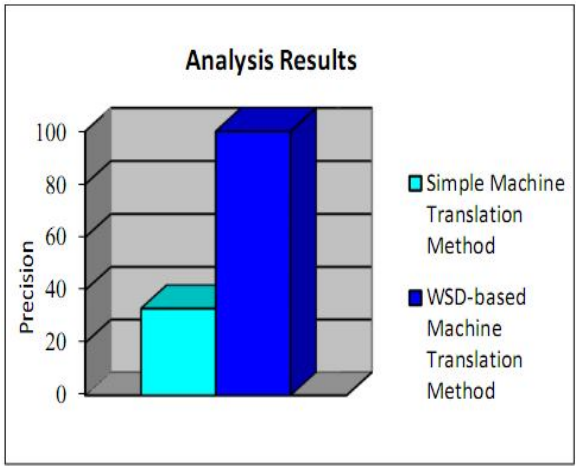

Fig 7. Experimental Result of the System

CONCLUSION

This research intends to improve the performance of Myanmar-English machine translation system by eliminating the ambiguous words in Myanmar Language. This system searches the correct sense of the ambiguous words by using distance-based WSD method and Bilingual Corpus. Afterwards, Myanmar language is translated to English language by using rule-based machine translation method. As this system can firstly disambiguate each ambiguous word in the Myanmar language, it helps to improve the precision of machine translation system. According to the experimental results, WSD-based Machine Translation method is more precise than simple MT methods.

REFERENCES

[1] S. M. Fakhrahmad, A. R. Rezapour and M. H. Sadreddini, "Applying Weighted KNN to Word Sense Disambiguation", Proceedings of the World Congress on Engineering, vol. 3, July 6-8, 2011.

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

[3] S. K. Srivatsa and P. Tamilselvi, "Case Based Word Sense Disambiguation Using Optimal Features", IACSIT, vol. 16, Singapore, 2011.

[4] S. Viswanadha Raju, 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] B. Liu, Web Data Mining, Department of Computer Science, University of Illinois at Chicago, USA, 2007.

[8] N. Ide and K. Suderman, "Integrating Linguistic Resources: The American National Corpus model", Proceedings of the 5th Language Resources and Evaluation Conference, 2006.

[9] Bushra Jawaid, Machine Translation with Significant Word Reordering and Rich Target-Side Morphology, Faculty of Mathematics and Physics, Charles University in Prague, UK, January 7, 2013.

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

★★★