Improving Phrase-based Statistical Myanmar to English Machine Translation with Morphological Analysis

Thet Thet Zin
University of Computer Studies, Yangon, Myanmar

Khin Mar Soe
Natural Language Processing Laboratory, University of Computer Studies, Yangon, Myanmar

Ni Lar Thein
University of Computer Studies, Yangon, Myanmar

ABSTRACT
This paper presents Myanmar phrases translation model with morphology analysis. The system is based on statistical approach. In statistical machine translation, large amount of information is needed to guide the translation process. When small amount of training data is available, morphological analysis is needed especially for morphology rich language. Myanmar language is inflected language and there are very few creations and researches of corpora in Myanmar, comparing to other language such as English, French, and Czech etc. Therefore, Myanmar phrases translation model is based on syntactic structure and morphology of Myanmar language. Bayes rule is also used to reformulate the translation probability of phrase pairs. Experiment results showed that proposed system can improve translation quality by applying morphological analysis on Myanmar language.

General Terms
Natural Language Processing, Machine Translation, Phrase-based SMT.

Keywords
Statistical Machine Translation, Morphological Analysis, Syntactic Structure, Bayes Rules, Out-of-Vocabulary.

1. INTRODUCTION
Machine translation (MT) is the task of automatically translating a text from one natural language into another. There exist diff erent approaches to address the problem of machine translation. This paper presents translation model for statistical Myanmar to English machine translation system. Baseline system used target and source language model based on N-gram (trigram) and translation model based on Bayes' rule to reformulate translation probability P(f| e) .N-gram method (trigram) based source language model is used to extract phrases for segmented Myanmar sentence.

Myanmar language likes other Southeast Asia languages that do not place spaces between words. Therefore, the system used Myanmar Word Segmenter (MWS) which is implemented in UCSYNLP Lab and is available for research purpose. Baseline system used translation probabilities without additional morphology analysis of Myanmar language. Languages may be divided into three broad categories: isolating, agglutinative and inflective languages. Isolating languages, such as Chinese, have little or no morphology and thus do not benefit from morphologically analysis. Agglutinative languages, also known as agglomerative or compounding languages, are those in which basic roots and words can be combined to make new words. These languages, such as Turkish or Finnish, tend to have many morphemes. Inflectional morphemes are used to modify a word to reflect information such as tense.

Myanmar language may be agglutinative language and inflective language because Myanmar word can be combined to make new word. eg: Myanmar word ရေ yae 'water' and အိုး oh 'pot' is isolating word. But, if they are combined, ရေအိုး yae-oh 'water-pot' is new word for each of the word. Baseline system only used translation probabilities. There are unknown words in baseline translation system. When a form of a word does not occur in the training data, the system is unable to translate it. According to experimental result, the Out-of-Vocabulary (OOV) rate exceeds 50% for tested dataset with 2000 training sentences, which means that half of the words in test set are not present in the training set. Most of the OOV words appear in proper nouns, verb and noun phrases. Therefore, translation model used syntactic structure and morphological analysis of Myanmar language to improve in translation direction and to reduce the number of unknown words in translation. The rest of this paper is organized as follows: In Section 2, previous works in statistical machine translation is presented. Section 3 describes baseline translation model. Section 4 presents analysis of Myanmar language. The proposed system is presented in section 5. Finally, Section 6 and 7 discusses translation results and conclusion.

2. RELATED WORK
In this section, previous works in Statistical machine translation on different languages are reviewed. Various researchers have improved the quality of statistical machine translation system by using different methods on different language. In [1] Brown creates probabilistic, models for simulating the translation process, in the models using bilingual corpora and then decoding a test sentence by searching. In 1993, he took the translation process as a noisy-channel model. In terms of modeling [2] appended context-based information based on the Maximum Entropy principle to enrich the word-based models. In terms of training, EM algorithm [3] dominated the parameter estimating process by taking word-level alignment of a parallel sentence pair as the latent variable. In [4] Wang and Waibel first proposed an alignment model based on phrase structures, which were automatically acquired from parallel corpus. Beam search algorithm was used in [5], which could make use of pruning strategies for balancing efficiency and accuracy.

In [6] Och and Ney first introduced the log-linear model into SMT. Koehn suggested using features of lexical weighting. In 2004, the famous phrase-based decoder, Pharaoh [7], was released to be a free SMT toolkit by Philipp Koehn and further updated to Moses [8]. In [9] Koehn, Och and Marcu used noisy channel based translation model and beam search decoder. They achieved fast decoding, while ensuring high quality. They presented experiential result on many languages (English-German, French-English, Swedish-English, and Chinese-English). Zens and Ney proposed log-linear based statistical machine translation model. They solve search problem using dynamic programming and beam search with three pruning methods. A comparison with Moses showed that the presented decoder is significantly faster at the same level of translation quality.

A few researches investigated the use of morphology to improve translation quality. If source language is morphology rich language (such as German, Spanish, Czech), phrase-based model has limitations. When a form of a word does not occur in the training data, current systems are unable to translate it. Data sparseness problem can be overcome by using large training data or morphology analysis of source or/and target languages. In [10], Goldwater and McClosky used morphological analysis of Czech to improve a Czech-English statistical machine translation system. This system solve data sparse problem caused by the highly inflected nature of Czech. Their combine model achieved high BLEU score of development and test set. In [11] Nguyen and Shimazu proposed morphological transformational rules and Bayes' formula based transformational model to translate English to Vietnamese. The score of their system is better than baseline score.

An ideal system for machine translation would take advantage of both empirical data and linguistic analysis. Different authors have different objectives that they attempt to achieve high translation precision on many languages. Because of the lack of prior research on this task, we are unable to compare to our results to those of other researches; but the results do seem promising. Our translation model aims is to get correct translation phrases with very modest bilingual corpus for Statistical Myanmar to English machine translation. Moreover, the system reduced unknown words in translation system.

.

3. BASELINE TRANSLATION MODEL

3.1 Model

The system used Bayes' rule to reformulate the translation probability for translating Myanmar sentence into English sentence. Among all possible target language sentences, we will choose the sentence with the highest probability:

$$E = \arg\max_e \{\Pr(\underset{1}{\overset{I}{e}} \mid \underset{1}{\overset{J}{f}})\} \quad (1)$$

$$= \arg\max_e \{\Pr(\underset{1}{\overset{I}{e}}).\Pr(\underset{1}{\overset{J}{f}} \mid \underset{1}{\overset{I}{e}})\} \quad (2)$$

This allows for a language model Pr (e) and a separate translation model Pr (f |e). In this translation model, the system are not focus on English phrase reordering. Rearranging the English phrases is implemented in separate part as a subsystem of statistical Myanmar to English translation system.

3.2 N-gram Based Phrases Extraction from Corpus

Myanmar language does not place space between words. Thus, the proposed translation model use Myanmar Word Segmenter and phrase align Myanmar-English Bilingual corpus. The system created phrases by using N-gram method for input segmented sentence to search in the corpus. In this case, one segmented word is assumed one word.

Example:

Myanmar Sentence: သူသည်ကျောင်းသို့ဘတ်စ်ကားဖြင့်သွားခဲ့သည်

'He went to school by bus.'

After word Segmentation:

သူ_သည်_ကျောင်း_သို့_ဘတ်စ်ကား_ဖြင့်_သွားခဲ့သည်_။_

This sentence contains 7 words. Left-to-right tri-grams on segmented input sentence are used to create phrases for translation. If all trigram phrases have not been observed in the corpus, bigrams and unigram phrases are used. If unigram and trigram phrases have the same meaning, longer n-grams is selected. Therefore, the system generally gets less and less number of phrases. Phrases for input sentence according to the longest N-gram method are:

သူသည်၊ ကျောင်း၊ သို့၊ ဘတ်စ်ကားဖြင့်၊ သွားခဲ့သည်

3.3 Problems in Baseline System

In baseline system, the system is unable to learn translations of words that do not occur in the data, because they are unable to generalize. Translation model know nothing of morphology therefore it fail to connect different word forms. Baseline system faces this problem because Myanmar language is inflected language and we have only small amounts of training data. For example: the word (ပန်းများ:pan myar;flowers) appears in the training data, but system cannot translate (ပန်း:pan;flower). This problem occurs in number (singular/plural) categories of noun phrase. Myanmar verb can have many suffixes and some suffixes have the same meaning. This is difficult for translation.

For instance;စားသည်:sar ti; စား၏ :sar ei; စားပါသည် :sar par ti have different verb particles (သည်:ti; ၏:ei; ပါသည် :par ti). But they have the same meaning. The root verb is (စား:sar;eat) and they are present tense.

When translation model has learned multiple possible translations for a particular word or phrase, the choice of which translation to use is guided by conditional probability rather than by linguistic information. Sometimes linguistic factors like case marker, tense, or number categories of noun phrases are important determinants for what translation ought to be used in a particular context. Myanmar word ရှိသည်shi-ti have three different English align words 'am, is, are'. But they are different in usage according to subject of sentences. Baseline system selects one word according to conditional probability. Therefore, sometime translation result is incorrect.

There are unknown words in baseline translation system. Unknown words can be reduce by using large training data or morphology analysis of source or/and target language. The large scale Myanmar Corpus is unavailable at present. Morphology

analysis is complex and computationally expensive method. Firstly, we analyzed OOV words category for morphology analysis process. The system separately takes 215 parallel sentences as testing datasets and 12827 sentences are used as the training dataset. There is no overlap of parallel sentences between training and testing datasets. The effect of the rate of out-of-vocabulary words on translation quality, the training dataset is divided into several different smaller sizes. We measure OOV based on types (each word in the vocabulary) as well as tokens (each word in the text). "Fig.1" shows the OOV rate of Myanmar-English testing dataset. According to the "Fig.1", the OOV rate increases as the number of training sentences decreases. OOV words profile is shown in "Table. 1". Most of the unknown words occur in proper noun, noun and verb phrases. To reduce unknown words in noun and verb phrases, the system considers morphology analysis on number category of noun phrases, suffixes and particles of verb phrases for Myanmar language.

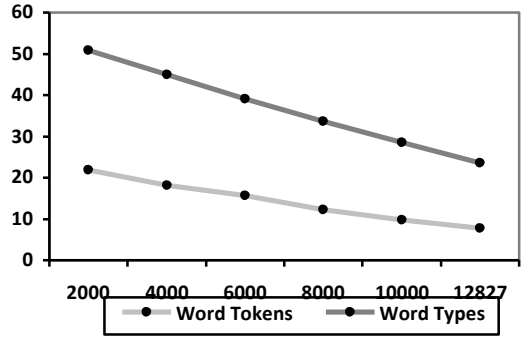

Fig. 1 OOV rate on Myanmar-English Test Set

Table 1. OOV words profile

Category of OOV words	OOV %	OOV words
Proper Nouns	30	146
Nouns	26.4	129
Verbs	29.6	145
Adjectives	10.1	49
Others	3.9	19

Some postpositional markers have ambiguous meanings in translation. Proposed translation model handle this problem by using syntactic structure of Myanmar language. Example of ambiguous in postpositional markers (PPM) is shown in table 2.

Table 2. Ambiguous in Postpositional Markers

Myanmar Sentences	Postpositional Markers	English translation
သူမတွင်ကားတစ်စီးရှိသည်။ 'She has a car.'	တွင်	တွင်ရှိသည်(has)
အမေသည်အိမ်၌ရှိသည်။ 'Mother is at home'.	သည်, ၌	သည်(omit) ၌ (at)
မောင်လှသည်ည၁၀နာရီ၌အိပ်ရောဝင်သည်။ Mg Hla goes to bed at 10 o'clock	သည်, ၌	သည်(omit) ၌ (at)

In the first sentence, PPM တွင် twin is subject-PPM. It combines with verb phrase ရှိသည် shi-ti and its meaning is တွင်ရှိသည် twin-shi-ti 'has'. But in the second and third sentences, PPM ၌ hnai is place-PPM and time-PPM respectively. Their meaning is "at". Therefore, translation of PPM is depended on sentence structure. To overcome these problems, proposed translation model considers syntactic structure of Myanmar language. Translation model performs analysis on suffixes and particles of verb phrases, postpositional markers such as nominative, accusative, dative and genitive and number category of noun phrases.

4. ANALYSIS OF MYANMAR LANGUAGE

The Myanmar Language is the official language of Myanmar. It is also the native language of the Myanmar and related sub-ethnic groups of the Myanmar, as well as that of some ethnic minorities in Myanmar like the Mon. Myanmar Language is spoken by 32 million as a first language and as a second language by 10 million, particularly ethnic minorities in Myanmar and those in neighboring countries. Myanmar language is a tonal and pitch-register, largely monosyllabic and analytic language, with a Subject Object Verb (SOV) word order. The language uses the Myanmar script, derived from the Old Mon script and ultimately from the Brāhmī script.

4.1 Literary Language and Spoken Language

The language is classified into two categories. One is formal, used in literary works, official publications, radio broadcasts, and formal speeches. The other is colloquial, used in daily conversation and spoken. This is reflected in the Myanmar words for "language": စာ sa refers to written, literary language, and စကား sa-ka refers to spoken language. Therefore, Myanmar language can mean either "မြန်မာစာ mran-ma-sa" (written Myanmar language), or "မြန်မာစကား mran-ma-sa-ka:" (spoke Myanmar language). Much of the differences between formal and colloquial Myanmar language occur in grammatical particles and lexical items. Different particles (to modify nouns and verbs) are used in the literary form from those used in the spoken form. For example, the postposition after nouns is □ hnai in formal Myanmar language and မှာ hma in colloquial Myanmar language.

Example: အမေသည်အိမ်၌ရှိသည်။ Mother is at home. Formal form) အမေအိမ်မှာရှိတယ်။ Mother is at home. (Spoken form) The proposed system focuses on written Myanmar language.

4.2 Syntactic Structure of Myanmar Language

The syntactic structure of every language is organized in term of subject, object and other grammatical functions, most of which are familiar from traditional grammatical work. Myanmar is SOV language. One problem in Myanmar language processing is the lack of grammatical regularity in the language. This leads to very complex Myanmar grammar in order to obtain satisfactory results. Many postpositional markers can be used in Myanmar sentences. Nouns and verbs need the help of suffixes or particles to show grammatical relation. Myanmar verb affixes are at the end of sentences and verb (stem) is very complex to define. The system defined five types of adverb phrases, seventeen types of postpositional markers and thirteen types of

verb particles according to Myanmar grammar rules to detect verb phrases in the sentences.

4.3 Morphology for Machine Translation

The system also used syntactic structure of Myanmar sentence and Myanmar grammar to improve translation quality. The roots of Myanmar language verbs are almost always suffixed with at least one particle which conveys such information as tense, intention, politeness, mood, etc. These verb suffixes make us difficult in translation of Myanmar to English. Because some suffixes have the same tense and the same meaning. However, Burmese verbs are not conjugated in the same way as most European languages; the root of the Burmese verb always remains unchanged and does not have to agree with the subject in person, number or gender. The most commonly used verb particles and their usage are shown below with an example verb root ကစား ka-sa 'play'. The statement ကစား ka-sa is imperative.The suffix သည်: ti (literary form) can be viewed as a particle marking the present tense and/or a factual statement: ကစားသည် ka-sa-ti 'play'. The suffix ခဲ့: hkai denotes that the action took place in the past. The suffix သည်: ti in this case denotes a factual statement rather than the present tense. ကစားသည် ka-sa-ti 'play'. The particle နေ: nay is used to denote an action in progression. It is equivalent to the English '-ing'. ကစားနေသည် ka-sa-nay ti 'playing'. The system also used syntactic structure of Myanmar sentence and Myanmar grammar to improve translation quality. The roots of Myanmar language verbs are almost always suffixed with at least one particle which conveys such information as tense, intention, politeness, mood, etc. These verb suffixes make us difficult in translation of Myanmar to English. Because some suffixes have the same tense and the same meaning. However, Burmese verbs are not conjugated in the same way as most European languages; the root of the Burmese verb always remains unchanged and does not have to agree with the subject in person, number or gender. The most commonly used verb particles and their usage are shown below with an example verb root ကစား ka-sa 'play'. The statement ကစား ka-sa is imperative.The suffix သည်: ti (literary form) can be viewed as a particle marking the present tense and/or a factual statement: ကစားသည် ka-sa-ti 'play'. The suffix ခဲ့: hkai denotes that the action took place in the past. The suffix သည်: ti in this case denotes a factual statement rather than the present tense. ကစားသည် ka-sa-ti 'play'. The particle နေ: nay is used to denote an action in progression. It is equivalent to the English '-ing'. ကစားနေသည် ka-sa-nay ti 'playing'. The particle မည် mai (formal form: မည်) is used to indicate the future tense or an action which is yet to be performed. ကစားမည် ka-sa- mai 'will play' Verbs are negated by the particle □: ma, which is prefixed to the verb. When the corpus contains only imperative verb ကစား ka-sa, we can generally decide Myanmar verb tense by looking verb particles ခဲ့: hkai 'past tense', နေ : nay 'continuous tense', မည်: mai 'future tense'. Verb particle ကြ:kyat can be omitted in the sentence. For example: ကျောင်းသားများကစားနေကြသည်။ 'Students are playing'. And ကျောင်းသားများကစားနေသည်။ 'Students are playing.' In second sentence, verb particle ကြ: kyat is omitted. Some verb phrases are same in main verb category but different in suffixes category. They have the same meaning in translation. The system solved this problem by defining possible verb suffixes groups. Nouns in Myanmar language are pluralized by suffixing the particle များ : mya in formal language. The particle တို့: tou, which indicates a group of persons or things, is also suffixed to the modified noun. Subject pronouns begin sentences, though the subject is generally omitted in the imperative forms and in translation. Subject marker particles က ka: in colloquial, သည်: ti in formal must be attached to the subject pronoun, although they are also generally omitted in translation.

Object pronouns must have an object marker particle: ကို: ko in colloquial, အား:ar in formal attached immediately after the pronoun. Object marker particle cannot be omitted in translation. We combine object pronouns and object marker particles and then we translate objects of sentences.

5. PROPOSED SYSTEM

This system needs segmented words and POS annotation corpus. Preprocessing includes segmenting input sentence, finding verb phrases in sentence and morphological analysis on noun and verb phrases. MWS is used to segment input sentence. Verb phrases detection will be presented in next subsection and morphological analysis on noun and verb phrases are presented in subsection 4.3. Myanmar-English bilingual corpus is used as a main knowledge source. We estimate the phrase translation probability distribution for the collected phrase pairs by relative frequency:

$$\Pr(f \mid e) = \frac{count(\bar{f}, \bar{e})}{\sum_{\bar{f}} count(\bar{f}, \bar{e})} \qquad (4)$$

The number of co-occurrences of a phrase pair (f, e) that are consistent with the phrase alignment is denoted as count (f, e). When the system looks up input phrases in the corpus, it also finds main verb of verb phrases. Suffixes extraction can also be applied to training dataset to decide stem words of verb phrases. An example sentence from the corpus is shown in below.

[0]ကလေးများ/[0]children[NNS]
#[1]သည်/[7]null[-]
#[2]ရုပ်ရှင်/[6]film[NN]
#[3]ကြည့်/[5]see[VB]
#[4]ရန်/[4]to[TO]
#[5]ရုပ်ရှင်ရုံ/[3]cinema[NN]
#[6]သို့/[2]to[TO]
#[7]သွားခဲ့ကြသည်/[1]went[VBD]

Each token has index of Myanmar and English word, English POS from tree tagger. English words and POS tags are extracted from this corpus according to Myanmar phrases.

5.1Verb Phrases Detection

Different languages may differ in their syntactic structure in general: for instance the placement of the verb in sentence or the use of postpositional markers in the sentences. Currently, no mature deep analysis that has been worked done is available for Myanmar language. The proposed system detects verb phrases

in Myanmar sentence by using syntactical structure of sentence. Myanmar language is SOV pattern. Verb suffixes are at the end of Myanmar sentences and Myanmar verb (stem) is very complex to define. Example of Myanmar verbs are shown in Table 3.

Table 3. Examples of Myanmar Verbs

Myanmar Verbs	Main Verbs	Suffixes	English meaning
လုပ်အားပေးနေသည်။	လုပ်အားပေး	နေသည်	contributing
အားပေးနေသည်။	အားပေး	နေသည်	encouraging
ပေးနေသည်။	ပေး	နေသည်	giving
နေသည်။	နေ	သည်	live

Verb suffixes are mined from any Myanmar sentences by using N-gram method in the system. Main verb is in front of suffix. We will first provide some examples to illustrate this concept and conclude this section with a formal definition. The following sentences are collected from Myanmar grammar books. The main verbs are marked by parentheses and suffixes are marked by italics. Example:

(1)မောင်မြသည်သူမအားပန်းမျှားကို(ပေး)သည်။
(Mg Mya gives flowers to her.)
(2)မောင်လှသည်ကျောင်းသို့မြန်မြန်(သွား)ခဲ့သည်။
(Mg Hla went to school quickly.)
(3)မောင်ဘသည်စာဖတ်ခြင်းကိုအလွန်(ကြိုက်နှစ်သက်)သည်။
(Mg Ba likes reading very much.)

According to syntactic structure of Myanmar language, generally very phrases are always end of the sentence. Firstly verb suffixes are extracted and then define main verb. According to analysis, post positional markers or adverb phrases are in front of main verbs. In above examples post positional marker ကို:ko , adverb phrases မြန်မြန် myan-myan 'quickly', အလွန်:ar-lwan 'very' are in front of main verbs. The system defined five types of adverb, seventeen types of postpositional markers and thirteen types of verb particles according to Myanmar grammar rules to detect verb phrases in the sentences. The system does not consider complex sentences structure with the conjunction words. Some verb phrases are same in main verb category but different in suffixes category. But they have the same meaning in translation. The system solved this problem by defining possible verb suffixes groups. Example of possible verb suffixes groups are shown in Table 4.

Table 4. Example of Possible Verb suffixes Groups

English tense	Myanmar Verb Suffixes which convey the same meaning
Present Tense	ပါသည်:par-ti၊သည်:ti၊၏:ei၊ကြသည်:ky-ti၊ ကြပါသည်:ky-par-ti□
Past Tense	ခဲ့ပါသည်:hkai-par-ti၊ ခဲ့သည်:hkai-ti၊ ခဲ့ကြပါသည်: hkai-ky-par-ti၊
Present Continuous	နေကြပါသည်:nay-ky-par-ti၊ နေကြသည် : nay-ky-ti၊ နေကြ၏:nay-ky-ei
Past Continuous	နေခဲ့ကြပါသည်:nay-hkai-ky-par-ti၊ နေခဲ့ကြသည်: nay-hkai-ky-ti၊
Future Tense	မည်:mai၊ကြမည်:ky-mai၊လိမ့်မည်:leint-mai၊ ကြလိမ့်မည်: ky-leint-mai၊

6. TRANSLATION RESULTS

6.1 Corpus Statistics

For experiments, the corpus contains sentences from Myanmar text books, grammar books and websites. The sentence in corpus is more diverse in sentence form than specific domain corpus. Corpus statistics are shown in table 5. Zawgyi-One Myanmar font is used for Myanmar Language.

Table 5. Corpus Statistic

Myanmar-English		
Sentences Pairs	**13042**	
Language	Myanmar	English
Total Word	61824	56263
Vocubary Size	2713	2405
Average Sentence Length (Word)	18	10

The system separately take 215 parallel sentences as testing datasets, and the remaining is used as the training dataset. There is no overlap of parallel sentences between training and testing datasets.

Table7. Statistics of the Myanmar-English datasets

Sentence Pairs of Datasets		Total Words		Vocabulary Size	
		Myanmar	***English***	***Myanmar***	***English***
Train	12827	60805	55335	2168	1965
Test	215	1019	928	545	440

6.2 Evaluation Criteria

MT evaluation measures are limited by inconsistent human judgment data. Nonetheless, machine translation can be evaluated using the well-known measures precision, recall. In this paper, evaluation of this system is measured in term of the standard measure of BLEU (Bilingual Evaluation Understudy). Manually translated sentences are used in this measure. BLEU is the geometric mean of n-gram precision by the system output with respect to manually correct sentences. Only single manual reference is used in this system.

6.3 Results

Figure 3 shows the results for Myanmar-English translation with varying sizes of training sentences. According to the figure, the proposed method begins to get some improvements over the corresponding baseline. When the size of training data sentences is less than 10000 sentences, morphology analysis method has good compared with the corresponding baselines. In proposed system, most errors occur in postpositional markers.

Postpositional markers have ambiguous meaning in translation. One way of helping the disambiguation of ambiguous words is use syntactic structure of language and to annotate words with their part-of-speech (POS). Therefore, we annotated Myanmar POS tags manually. We appended the Myanmar and English POS tags in training and test corpus to compare with baseline system. By using POS tags, the system reduced ambiguous in postpositional markers. Especially, ambiguous in Subject PPM မှာhmr 'has, have, had' and Place PPM မှာ hmr 'at', Subject PPM က ka; 'null' and Leave PPM က ka 'from', Used PPM နှင့် hnin; 'with'and Compare PPM နှင့် hnint 'and', Used PPM ဖြင့် phyint 'with' and Cause PPM ဖြင့် phyint; 'because of' and Place PPM တွင် twin 'at' and Extract PPM တွင် twin 'among'. The best

results got by adding morphology and POS of Myanmar language to baseline system. We also analyzed OOV words in proposed system. The system reduced OOV words in noun and verb phrases. Compound verbs and proper nouns pose problems to the robustness of a translation method and increased unknown words rate in translation. OOV words reduction is shown in table 7.

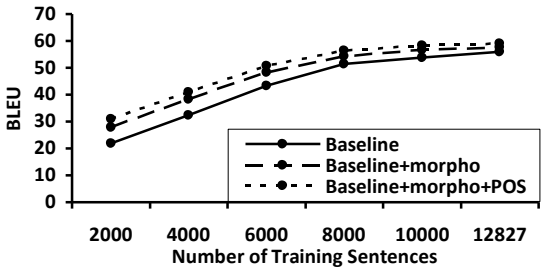

Figure3. Translation Results

Table 7. oov reduction rate

Category of OOV words	OOV%	OOV words
Nouns	22.8	89
Verbs	22.3	87

6.4 Error Analysis

Compound verbs and proper nouns pose problems to the robustness of a translation method. For example: compound verb သွားစားသည် twe-sar-ti ‘go and eat’ has two meaning. သွားသည် thwr-ti; ‘go’, စားသည် sar- ti ‘eat’ and meaning of သွားစားသည် twe sar ti is ‘go and eat’. Althoug the corpus contain သွားသည် thwr-ti ‘go’ and စားသည် sar-ti ‘eat’, we have difficult to translate these words သွားစားသည် twe-sar-ti ‘go and eat’ to get correct translation. Some errors occurred in adjective. Myanmar adjectives vary according to sentence patterns. Results There are 95 errors in tested sentences. The causes in detail are:

- Unknown words: The foreign word did not occur in the training corpus, so translation was not possible at all.
- Unknown translation: The word occurred in the training corpus, but fails to translate: fail to align the word to its correct translation, which often happens for rare words.
- Segmentation Error: Word Segmenter output is not suitable for correct translation result.
- Detecting verb phrases Error: Errors in finding verb phrases in the input sentence especially when input sentence is too long and include conjunction words.
- Untranslatable: Some phrases are not translatable into English phrase correctly.
- Others: missing English particle in noun phrases and so on.

7. CONCLUSION

We have shown that Myanmar-English phrase-based SMT can be improved by combining the syntactic structure, POS and morphological analysis of Myanmar Language. By adding these three features the system can achieve a better result than can be obtained with each individually. The system improved the translation quality with 0.031 BLEU scores over surface based baseline system. This improvement was primarily due to a reduction of the sparse data problem caused by the highly inflected nature of Myanmar language. An alternative method for reducing this problem is to use a larger parallel corpus. However, the large scale Myanmar Corpus is unavailable at present. For that reason, we believe that the approach presented in this paper is a promising one. In the future, we would like to apply other Myanmar morphological features in translation model and to test in more training data and domain specific corpus.

8. REFERENCES

[1] Brown, P.E., J.Cocke, S. A. Della Pietra, V.J. Della Pietra, F.Jelinek, J.D. Lafferty, R. L. Mercer, and P. S. Roossin, “A statistical approach to machine translation”, Computational Linguistics 16(2), 79-85, 1990.

[2] Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra.1996. *A maximum entropy approach to natural language processing.* Computational Linguistics 22(1), 39-72.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society, 39, 1-38.

[4] Wang, Y.-Y. And A. Waibel, “Modeling with structures in statistical machine translation.”, In proceedings of COLING/ACL 1998, Montreal, Quebec, Canada, pp. 1357-1363.

[5] Och, F. J., C. Tillmann, and H. Ney, “Improved alignment models for statistical machine translation”, In proceedings of the Conference on Empirical Method in Natural Language Processing and Very Large Corpora, University of Maryland, College Park, MD, pp. 20-28.

[6] Franz. J. Och and Hermann Ney. *Discriminative training and maximum entropy models for statistical machine translation*. 2002. in proceedings of ACL, pp. 295-302.

[7] Koehn, P., “Pharaoh: a beam search decoder for phrase-based statistical machine translation models”, in proceedings of the Sixth Conference of the Association for Machine Translation in the Americas, pp. 115-124, (2004).

[8] Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, “Moses: Open source toolkit for statistical machine translation”, In proceedings of the ACL 2007 Demo and Poster Sessions, pp, 177-180, (2007).

[9] Philipp Koehn, Fran Josef Och, Daniel Marcu, “Statistical Phrase-Based Translation”, Presentation at DARPA IAO Machine Translation Workshop, July 22-23, Santa Monica, CA, (2002).

[10] Goldwater, S. and McClosky, D. 2005. “ Improving statistical MT through morphological analysis”, In HLT'05: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. 676-683.

[11] Thai Phuong Nguyen and Akira Shimazu, "Improving Phrase-Based SMT with Morpho-Syntactic Analysis and Transformation", Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pages 138-147, Cambridge, August 2006.

[12] Ashin Thumana. 2004. *New Method English Grammar*. Second Edition. Yangon, Myanmar.

[13] Department of the Myanmar Language Commission, Ministry of Education Yangon, 1998. *Myanmar-English Dictionary*.

[14] Department of the Myanamar Languge Commission, Ministry of Education, Union of Myanmar. 2005. Myanmar Grammar.