Improving Statistical Myanmar to English Machine Translation through Morphological Analysis

Thet Thet Zin, Khin Mar Soe, Ni Lar Thein
University of Computer Studies, Yangon
thetsmt@gmail.com,nilarthein@gmail.com

Abstract

In this paper, we present a translation model of Myanmar phrases for statistical machine translation. Phrase based SMT models have limitations in mapping from the source to target language without using linguistic information. Morphological analysis is needed especially for morphology rich language and small amount of training data are available. Myanmar language is inflected language and we have very modest parallel resources for machine translation. Therefore, we present Myanmar language morphology analysis in noun and verb phrases. Especially analysis is performing on number category of noun phrase, suffixes and tense particle of verb phrases. We test our system on Myanmar-English bilingual corpus. The experiment results show that the quality of statistical machine translation is improved by applying morphology analysis of Myanmar language.

1. Introduction

Machine translation (MT) is the task of automatically translating a text from one natural language into another. Recent Statistical machine translation systems based on phrase or word groups. They use conditional probability for maximum likelihood of source language and foreign language. Searching is one of the problems of machine translation when training data is very large. They use various decoders and searching methods to solve searching problem. In statistical machine translation, the large amount of information is needed to guide the translation process. However, we have very modest parallel resources available for machine translation. Therefore, data sparseness is an important issue for our statistical machine translation. This paper considers morphological analysis on number category of noun phrases and suffixes and particle of Myanmar verb phrases.

Plural number particles of Myanmar noun are "များ၊ တို့တွေ :*mya, tot, twe*". Myanmar verb has many particles and suffixes. It is not easy to define tense like English. Some suffixes have same meaning. (1)For example: these suffixes: "ကြသည်: *kyti* ၊ ကြပါသည်: *kyparti* ၊ ကြ၏:*kyei*" have same meaning. Some verb behave particle to support previous verb in sentence. (2)For example: "ပြောပေးသည်: *pyawpayti*; talk" in this verb "ပေး: *pay*" behave particle to support previous verb "ပြော: *pyaw*". But "ပြော: *pyaw;talk*" and "ပေး: *pay;give*" can behave individual verb in sentence. More than two individual verbs can include in Myanmar compound verb. (3)For example: "ဝင်ထွက်သွားလာသည်: *win twaet twa lar ti"*- four individual verb "ဝင်+ထွက်+သွား+လာ" includes in this compound verb. It is difficult to translate English language. Some verb particles indicate tense of Myanmar verb. (4)For example: (၏၊သည်:*ei,ti*-present tense, ခဲ့:*khe*-past tense, နေ: *nay*-continuous tense, မည်၊လိမ့်:*mi,leint* - future tense). This paper focuses on singular and plural number of noun phrase, verb suffixes which have same meaning in translation and particles which indicate verb tense. In our baseline system, we use direct modeling of posterior probability by using log linear model for translation probability of Myanmar phrase and English phrase pair. We use N-gram language model. Myanmar language does not place space between words. Therefore, we use Myanmar Word Segmenter implemented from UCSYNLP Lab which is available for research purpose. We also use N-gram method to extract phrases from segmented input sentence. The rest of this paper is organized as follows: In Section 2, previous works in statistical machine translation is presented. Section 3 describes analysis of Myanmar language. Section 4 presents phrase-based translation model. The proposed system is presented in section 5. Finally, Section 6 and 7 discusses our translation results and conclusion.

2. Related Work

In this section, previous works in Statistical machine translation on different languages are reviewed. Recent Statistical machine translation systems based on phrase or word group and use probabilistic model by using source channel approach or direct probability model (log linear model). They solve searching problem by using various heuristic methods and pruning strategies. Philipp Koehn, Franz Josef Och, Daniel Marcu [1] used noisy channel based translation model and beam search decoder. They achieved fast decoding, while ensuring high quality. They compared the performance of the three methods for phrase extraction, using the same decoder and the same trigram language model. Learning all phrases consistent with the word alignment (AP) is superior to the joint model. The performance of IBM model-4 word-based translation system is worse than both AP and Joint. Limiting the length to a maximum of only three words per phrase achieves top performance. Richard Zens and Hermann Ney [3] proposed Phrase-based Statistical Machine Translation based on log-linear model with components and scaling factors. They solve search problem using dynamic programming and beam search with three pruning methods. A comparison with Moses [5] showed that the presented decoder is significantly faster at the same level of 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 2005, Sharon Goldwater and David McClosky [4] 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. In 2006, Thai Phuong Nguyen and Akira Shimazu [6] proposed morphological transformational rules and Bayes' formula based transformational model to translate English to Vietnamese. In 2007, Philipp Koehn [2] presented factored translation models. They use confusion network decoding to deal with ambiguous factors in translation. Their morphological analysis and generation model have three mapping steps. However, the more complex a multi-factored scenario is the worse the results are. In 2008, morphology generation models for machine translation are presented in [5]. They applied their inflection generation models in translating English into two morphologically complex languages, Russian and Arabic and their model improves the quality of SMT over both phrasal and syntax-based SMT systems according to BLEU and human judgements.

3. Analysis of Myanmar Language

The Myanmar language is the official language of Myanmar. It is 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 word order. The language uses the Myanmar script, derived from the Old on script and ultimately from the Brāhmī script.

3.1. 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 *mranma sa* (written Myanmar language), or မြန်မာစကား *mranma sa.ka:* (spoken Myanmar language). This paper focuses on written Myanmar language. Much of the differences between formal and colloquial Myanmar occurs in grammatical particles and lexical items.

4. Phrase-based Translation Model

In statistical machine translation, we are given a source language sentence $f_1^J = f_1 \cdots f_j \cdots f_J$ which is to be translated into a target language sentence $e_1^I = e_1 \cdots e_i \cdots e_I$. Among all possible target language sentences, we will choose the sentence with the highest probability by using log linear model:

$$E^* = \arg\max_E \{ \sum_{m=1}^{M} \lambda m hm(E,F) \} \qquad (1)$$

We use the following two feature functions shown in equation (2) and (3).

$$h_1(e^I_1, f^J_1) = \log P(f^J_1, e_1^I) \qquad (2)$$

$$h_2(^I_1, f^J_1) = \log P(e_1^I) \qquad (3)$$

We use relative frequency to get translation probabilities in equation (4) and language model probabilities in equation (5). N is the total number of word or the size of the training dataset.

$$P(f^J_1, e_1^I) = \frac{count\ (f, e)}{\sum_f count\ (f, e)} \qquad (4)$$

$$P(e_1^I) = \frac{count\ (e)}{N} \qquad (5)$$

We set $\lambda_1 = \lambda_2 = 1$. This approach is a generalization of the source-channel approach. It has the advantage that additional model can be easily integrated into the overall system. We calculate translation probabilities for any pair of Myanmar and English and then select translation options which have maximum translation probabilities. We also use N-gram language model. The language model determines the well-formed of target sentence. We define log linear model based translation model as the baseline model to compare proposed system.

5. The Proposed System

The proposed system is to translate Myanmar phrases to English phrases for Statistical Machine translation. We implement this system as a subsystem of Myanmar to English machine translation. In proposed system, translation model consider morphology analysis on noun and verb phrases in preprocessing step. Processing procedure of the system is shown in Figure 1. The system needs segmented and tagged of Myanmar input sentence. Therefore, we used Myanmar POS tagger and Segmenter which are available for research purpose. The main knowledge source is Myanmar-English bilingual corpus.

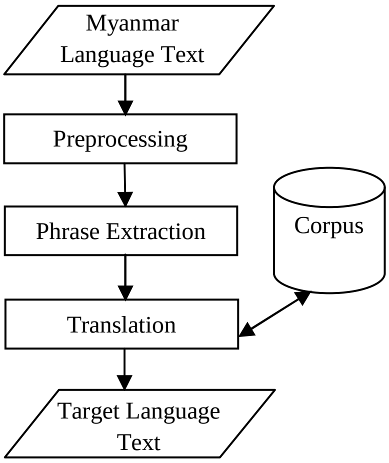

Figure 1.System Architecture of Translation Model

5.1. Preprocessing

Preprocessing step includes segmenting, POS tagging and morphological analysis of Myanmar sentence.

Table1. Preprocessing procedure

Input Sentence	သူမသည်ဈေးမှပန်းများဝယ်လာသည်
1.Segmenter Output	သူမ_သည်_ဈေး_မှ_ပန်းများ_ဝယ်လာသည်_။_
2.POS tagging	သူမ/PRN$_1$.person သည်/PPM$_1$.Subj ဈေး/NN$_1$.Location မှ/PPM$_2$.Place ပန်းများ/NNR$_1$.Objects ဝယ်လာသည်/VB.compound
3.Morphological Analysis	ပန်း/NN.stem များ/NN.Plu ဝယ်လာ/VB.main သည်/VB.part

This paper focuses on step 3 of preprocessing procedure. We find noun phrase and verb phrase from step2 by using Myanmar POS tagging. And then we define stem word and number particle for noun phrase and suffixes particle and main verb for verb phrase. We add stem word of noun phrase and main verb of verb phrase into the searching list. By searching not only surface word but also stem word, we can solve the problem of unknown word related to singular and plural number of noun phrase and different verb suffixes with same meaning which have not appeared in the training corpus, but for which other inflectional forms related to the given unknown word can be found in the corpus. Examples for the generation of inflectional forms of verbs and nouns are given in Table 2 and Table 3, respectively.

Table2.Verb Inflections with singular noun

Category	"သွား" (go)
Present Tense	သွား**သည်**၊သွား**၏**၊သွားပါသည်
Past Tense	သွား**ခဲ့**သည်၊ သွားခဲ့ပါသည်
Future Tense	သွား**လိမ့်**မည်၊ သွား**မည်**
Continuous Tense	သွား**နေ**သည် ၊သွားနေပါသည်

If noun is plural,we use "ကြ" partical for verb suffixes. e.g; သွားကြသည်. Bold letters are tense marker. In noun inflection word, bold letters are plural marker. We define verb suffixes and particle for Myanmar verb. We have 20 verb particles and 8 same suffixes pairs. These verb suffixes (နေခဲ့ကြပါသည်၊နေခဲ့ကြသည်၊နေခဲ့ကြ၏) have same meaning by combining with main verb.

Table3. Noun Inflection

Number	"ကလေး" (child)
Singular	ကလေး
Plural	ကလေး**များ**၊ကလေး**တို့**၊ကလေး**တွေ**

In this example we define noun phrase of input Myanmar sentence ကလေးများ*(ka lay myar)*/ NNR$_1$.Objects with ကလေး/ NNR$_1$.stem and များ/ NNR$_{1.}$plu.

5.2. Phrase Extraction from corpus

The system used Myanmar-English bilingual corpus for translation. An example sentence from the corpus is shown in below.

[0]ကလေးများ[NNR$_1$.person]/[0]children[NNS]
#[1]သည်[PPM$_1$.Subj]/[7]null[-]
#[2]ရုပ်ရှင်[NN$_1$.objects]/[6]film[NN]
#[3]ကြည့်[VB$_1$.common]/[5]see[VB]
#[4]ရန်[PPM$_2$.cause]/[4]to[TO]
#[5]ရုပ်ရှင်ရုံ[NN$_2$.location]/[3]cinema[NN]
#[6]သို့[PPM$_3$.direction]/[2]to[TO]
#[7]သွားခဲ့ကြသည်[VB$_2$.common]/[1]went[VBD]

Each token has index of Myanmar word and English word in the sentence and English POS (Part-Of-Speech) from tree tagger and Myanmar POS from Myanmar tagger. We extract English word and Myanmar and English POS tagging from this corpus according to Myanmar phrase. To extract this information from corpus, we need to create Myanmar phrase from segmented input sentence. For example:

Input Myanmar sentence:
သူတို့သည်လိမ္မာသောကျောင်းသားများဖြစ်သည်။
Segmenter Output:
သူတို့_သည်_လိမ္မာ_သော_ကျောင်းသားများ_ဖြစ်သည်_။_

To create Myanmar phrase from Segmenter output, we use N-gram method. In this case, we assume one segmented word is one word. We use left-to-right trigrams on segmented input sentence to create phrases for translation. We find these phrases in the corpus. If all trigram phrases have not been observed in the corpus, we use bigrams and unigram phrases. If some or all of the phrases have the same meaning, we select longer n-grams. Therefore, we generally get less and less number of phrases.

Table4. Possible Phrases of input sentence

Unigram (one segmented word)	Bigram (two segmented word)	Trigram (three segmented word)
သူတို့	သူတို့သည်	သူတို့သည်လိမ္မာ
သည်	သည်လိမ္မာ	သည်လိမ္မာသော
လိမ္မာ	လိမ္မာသော	လိမ္မာသောကျောင်းသားများ
သော	သောကျောင်းသားများ	သောကျောင်းသားများဖြစ်သည်
ကျောင်းသားများ	ကျောင်းသားများ ဖြစ်သည်	
ဖြစ်သည်		

Phrases for input sentence according to the longest N-gram method

သူတို့၊ သည် ၊လိမ္မာသော ၊ ကျောင်းသားများ ၊ ဖြစ်သည်

We calculate translation probabilities and language model probabilities of these phrases by using relative frequency count. If there are more than one translation options, we select phrase with highest translation probability.

5.3. Generation Process

Surface words are not appearing in the training corpus, we use stem word and particles to generate surface word. To generate plural or singular forms of English word, we use English grammar rules. Singular words which end is s, z, sh, ch or x, we add es to become plural words. Singular words which end is consonant with "y" changes the "y" to "i" and add es. All other singular words add "s". But some nouns have irregular form e.g; man (plural men). We cannot handle this irregular noun. We also generate verb tense by using verb stem word and suffixes particles. Stem of verb add "ed" to become past tense. We use English grammar rule to change verb tense but some verb has irregular form e.g; past tense of "read" is also "read". We handle irregular verb by using irregular verb list defined by Oxford Dictionary.

6. Translation Results

6.1. Corpus Statistics

For experiments, we used general domain corpus as shown in Table 1. The corpus contains sentences from Myanmar text books, Myanmar grammar books and websites.

Table5. Corpus statistics

Corpus	Sentence Pairs	Average Sentence Length(word)	
		Myanmar	English
General	13042	18	14

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, and the F-measure. The F-measure has significantly higher correlation with human judgments than recently proposed alternatives. In this paper, we measure evaluation of our translation system in term of the standard measure of precision, recall and F-measure in equation 6, 7 and 8. We test our system in general domain. Sentence types in corpora are simple and compound. The lengths of source sentences are between 5 and 15. Only single references are used in our measure. These reference sentences are manually translated. Our system does not consider word order of Myanmar and English language. Therefore, we ignore the word order of candidate and reference sentences.

$$\text{Precision } (C \mid R) = \frac{|C \cap R|}{C} \qquad (6)$$

$$\text{Recall } (C \mid R) = \frac{|C \cap R|}{R} \qquad (7)$$

$$F - measure = \frac{2*(precision * recall)}{(precision + recall)} \quad (8)$$

C=set of candidate sentences

R=set of reference sentences

6.3. Results

In baseline system, translation model does not consider morphological analysis of Myanmar sentence. In proposed system, we consider noun and verb morphology in preprocessing phrase. But mention above we have limitation in analysis process. We have very modest parallel resources available. Therefore, we have unknown words in translation. We tested with 215 sentences which lengths are between 5 and 15. We use zawgyi-One Myanmar font. In proposed system, the precision get 68.7%, recall get 76.7% and F-measure get 72.4%. In baseline system, the precision get 60.2%, recall get 69.5% and F-measure get 64.5%.

6.4. Errors analysis

In baseline system, when we search input phrase in the corpus, we need exact match phrase for our translation. Although we use smoothing method to overcome zero probability for input phrases, when a form of a word does not occur in the training data, systems are unable to translate it. In proposed system, to reduce unknown words related to verb suffixes and particle of verb phrases and singular and plural number of noun phrase, we perform analysis on these unknown words and then generate surface word form. Errors in proposed system are ambiguous in noun and verb phrase (“သွား;go and သွား;teeth”). Compound verb (သွားစားသည်:twe sar ti) has two meaning. (သွားသည်:go) and (စားသည်:eat) and meaning of (သွားစားသည်:twe sar ti) is (go and eat). Althoug the corpus contain (သွားသည်:go) and (စားသည်:eat), we have difficult to translate (သွားစားသည်:twe sar ti: go and eat) to get correct translation. Some verb support to previous verb (“ပြောပေးသည်,give”), correct translation is “talk”. Errors in generation are irregular noun (child, children) in singular and plural form and ambiguous in postposition marker (“တွင်; null and တွင်;at”). English particles are missing between noun and postpositional marker.

7. Conclusion

We have shown that Myanmar-English phrase-based SMT can improve by combining the syntactic structure and morphology of Myanmar Language. The contribution of this work includes syntactic structure and morphological analysis of Myanmar language to improve translation. The use of small corpora was a limitation in our work. . If we get larger corpus size, we can get the best translation result. 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.

References

[1]P. Koehn, F. J. Och, D. Marcu, “Statistical Phrase-Based Translation”, *Presentation at DARPA IAO Machine Translation Workshop*, July 22-23,2002, Santa Monica, CA.

[2]P. Koehn and H. Hoang, “Factor Translation Models ”,*Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, PP.868-876, Pragun, June

[3]R. Zens and H. Ney, “Improvements in Phrase-Based Statistical Machine Translation”.

[4] S. Goldwater and D. McClosky, “Improving Statistical MT through Morphological Analysis”, *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP),* pages 676-683,Vancouver,October 2005.

[5] Toutanova, K., Suzuki, H., and Ruopp, “Applying morphology generation models to machine translation”, *In Proceedings of ACL-08: HLT.* Columbus, Ohio, 514-522, 2008.

[6]T. P. Nguyen and A. 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.

[7]Department of the Myanmar Language Commission,Ministry of Education, Union of Myanmar, “Myanmar-English Dictionary.”

[8]Department of the Myanmar Language Commission,Ministry of Education, Union of Myanmar, “Myanmar Grammar”, 2005