Myanmar Phrase Translation for Myanmar-English Machine Translation System

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

Abstract

In statistical machine translation (SMT), the currently best performing systems are based in some way on phrases or word groups. In this paper, phrase-based translation system is described. Translation model and word-based lexicon model are used in this system. This system is implemented as a separated part of the Myanmar to English machine translation system. Myanmar language does not use space between words. Therefore, the system uses Myanmar Word Segmenter, which is implemented in UCSYNLP lab and is available for research purpose, to segment Myanmar sentence. Myanmar-English bilingual corpus is also used as a main knowledge source. The vast amount of information is needed to guide the translation process. The large scale Myanmar corpus is unavailable at present. This system aims to increase correct translation results with limited bilingual corpus. The experimental results show that the proposed system seems promising for Myanmar to English machine translation system.

1. Introduction

The goal of creating statistical machine translation (SMT) systems incorporating rich, spare, features over syntax and morphology has consumed much recent research attention. In SMT approach, MT is treated as a decision problem: given the source language sentence, we have to decide for the target language sentence that is the best translation. Bayes rule and statistical decision theory are used to address this decision problem. Statistical models estimated parameters based on bilingual text corpora. It differs from example-based MT (EMT) is the ranking between fragments is done with probabilities rather than matching measures. SMT approach relies only on automatically detected word correspondences and alignment patterns. In the source-channel approach, $P(e|f)$ depends on two factors $P(e)$ and reverse translation probability $P(f|e)$. In a direct translation approach, various feature function $h_m(e|f), m = 1,\ldots,M$ is developed. In this case, the free model parameters are λ_1^m. A standard training criterion for the translation model in the source-channel approach is the maximum likelihood criterion and in maximum entropy based translation models. Today's statistical translation models $P(f|e)$ are only rough approximations to the "true" probability distributions $P(e|f)$. Therefore, certain natural language phenomena cannot be handled well. Source language architecture uses language model to get well form of target language sentence. SMT approach has advantages in machine translation. Firstly, In SMT, we have a mathematically well-founded machinery to perform on optimal combination of these knowledge sources. Second, translation knowledge is learned automatically from example data. As a result, MT system based on statistical methods is very fast compared to the rule-based approach. Third, SMT can deal with lexical word ambiguity involving context or

other informational sources. Myanmar and English languages have different word order. However, reordering model is implemented as a separate part of the translation system. The proposed system uses two types of information: translation model and word-based lexicon model. The translation model links the source language sentence to the target language sentence. It uses conditional probability to get translation probability of Myanmar and English phrase pairs. By using relative frequencies to estimate the phrase translation probabilities, most of the longer phrases are see only once in the training corpus. Therefore, relative frequencies overestimate the probability of those phrases. To overcome this problem, word-based lexicon model is used to smooth the phrase translation probabilities.

This paper is structured as followings: Previous works in machine translation for various languages are described in section 2. The proposed system is presented in section 3. Experimental results and conclusion are presented in section 4 and 5.

2. Related Work

In SMT, we are given a source language string, $F = f_1^J = f_1 \ldots f_j \ldots f_J$, which is to be translated into a target language sentence $E = e_1^I = e_1 \ldots e_i \ldots e_I$, where I and J represent the number of words of the sentences in target and source languages. In statistical machine translation, every English string $\boldsymbol{E}$ is a possible translation of $\boldsymbol{F}$. We assign a probability of $P(E \mid F)$ to every pair of strings <F, E> and when translating, we want to find the string $\bar{E}$ along with the highest probability. Searching problem is defined as ***arg max*** operation. In 1993, Brown et al. took the translation process as a noisy-channel model [1]. In terms of modeling Berger et al., 1996 appended context-based information based on the Maximum Entropy principle to enrich the word-based models [2]. Alignment model which is based on phrase structure is firstly proposed by Wang and Waible in 1998, which was automatically acquired from parallel corpus [3]. Och et al., 1999 used beam search algorithm, which could make use of pruning strategies for balancing efficiency and accuracy [4]. In 2002, Och and Ney first introduced the log-linear model into SMT [5]. In 2004 Koehn suggested using features of lexical weighting. In this year, the famous phrase-based decoder, Pharaoh [6], was released to be a free SMT toolkit by Philipp Koehn and further updated to Moses [7] by Koehn et al., 2007. In 2003 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) [8]. Log-linear based statistical machine translation model is proposed by Zens and Ney in 2004. 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 [9].

Different authors have different objectives that they attempt to achieve high translation precision on many languages. Our translation model aims is to get correct translation phrases with very limited bilingual corpus for Myanmar to English machine translation. 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.

3. Proposed System

In the proposed system, translation model, and word-based lexical model are used which ensures fluent output. The combination of these models is mathematically justified by the noisy channel model. In this approach, the input sentence is broken up into a sequence of phrases; these phrases are mapped one-to-one to output phrases, which may be reordered. Commonly, phrase models are estimated from parallel corpora that were annotated with word alignments. All phrase pairs that are consistent with the word alignment are extracted.

Probabilistic scores are assigned based on relative counts or by backing off to lexical translation probabilities.

3.1. System Architecture of the Proposed Myanmar Phrase Translation System

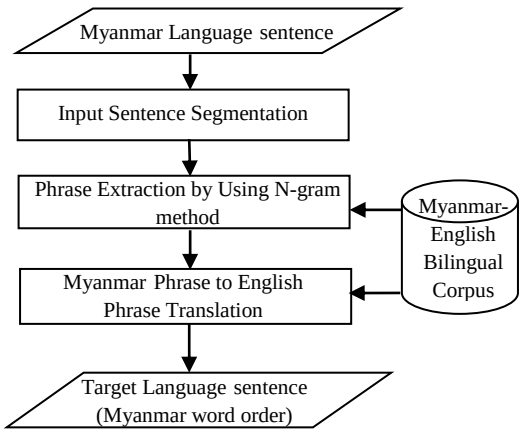

Figure 1: Overview Design of the Proposed System

Myanmar language does not use space between words. Therefore, Myanmar Word Segmenter which is implemented from UCSYNLP Lab is used to segment input Myanmar sentence. The system extracts possible phrases for segmented input sentence by using N-gram method. And then searches translation options for these extracted phrases by using Myanmar-English bilingual corpus. Translation probabilities of these options are calculated by using conditional probability. Some Myanmar phrases have more than one translation options. For example; "အများအပြား" has four translation options "a lot of, many, much, plenty of". The system selects the best options by using bigram combination Myanmar phrases in sentence. Myanmar and English languages have different word order. However, the system does not focus on Myanmar phrases rearrangement. It mainly considers getting the best translation options for Myanmar phrases. Reordering of Myanmar language is implemented as a separate part of the Myanmar to English machine translation system. The system is based on statistical approach.

3.2. Statistical Machine Translation (SMT)

Statistical Machine Translation is the most widely technique in machine translation system. The key feature of Statistical Machine Translation is that it is based on the principle of translating a source sentence into target language sentence using statistical information (probabilities) drawn from the parallel training corpus. First SMT systems were based on the so called source channel (noisy-channel) approach. Due to this approach equation 1 can be decomposed according to the Bayes rule as follows:

$$\hat{E} = \arg\max_e \{P(e).P(f \mid e)\} \qquad (1)$$

Thus, the problem of finding conditional probability grows into the *argmax* operation over the product of two models:

- $P(f \mid e)$ refers to translation model probability
- $P(e)$ refers to target language model probability

At training time, the system uses a parallel corpus to estimate the translation model probabilities and the target language model probabilities. An alternative to the classical source-channel approach is the direct modeling of the posterior probability $P(e \mid f)$ as follow:

$$\hat{E} = \arg\max_e \{\sum_{m=1}^{M} \lambda_m h_m(e \mid f)\} \qquad (2)$$

h_m refer to the system models. λ_m refers to the weights corresponding to these models. The architecture of the translation approach based on source channel model is described in figure 3. Post processing includes grammar checking for translated sentence. The proposed system uses translation probability and lexicon probability.

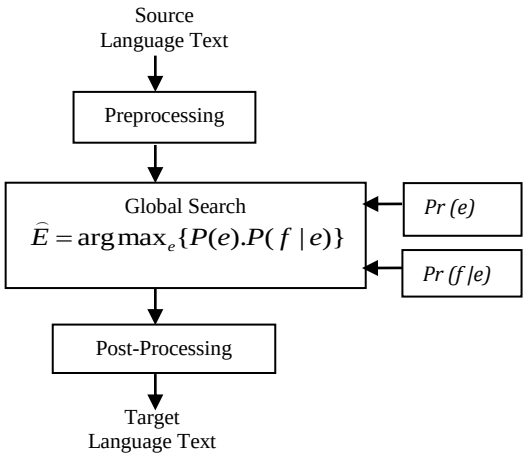

Figure 2 Architecture of the translation approach based on source-channel model

3.3. Phrase-Based Translation Model

The basic idea of phrase-based translation is to segment the given source sentence into phrases, then translate each phrase and compose the target sentence from these phrase translations. The phrase pairs for the source language sentence (Myanmar sentence) "ကျွန်တော်တို့၏ကျောင်းသည်နှစ်ထပ်အဆောက်အဦးဖြစ်သည် " are listed in the table1.

Table1. The list of used phrase pairs for above sentence

Source Phrases	Target Phrases
ကျွန်တော်တို့၏	Our
ကျောင်းသည်	school
နှစ်ထပ်အဆောက်အဦး	a two-storey building
ဖြစ်သည်	is

We define a segmentation of a given sentence pair (f_1^J, e_1^I) into K phrase pairs:

$$k \to s_k := (i_k; b_k, j_k), \; for k = 1...K. \quad (3)$$

i_k denotes the end position of the K^{th} target phrase. The pair (b_k, j_k) denotes the start and end positions of the source phrase that is aligned to the target phrase. Phrases are defined as nonempty contiguous sequence of words. We constrain the segmentations such that all words in the source and target sentence are covered by exactly one phrase. Thus, there are no gaps and there is no overlap.

The number of co-occurrences of a phrase pair (e, f) that are consistent with the phrase alignment is denoted as count (e, f). If one occurrence of a target phrase e has count>1 possible translations and they have sense meaning, WSD (Word Sense Disambiguation) model is used to handle ambiguous words. WSD model is implemented as a separated part of the translation model. A phrase translation table of English translations for the Myanmar phrase "အိမ်အနီးတွင်" may look like the following:

Table2. Phrase translation table

Translation	Probability $P(e \mid f)$
near to the house	0.5
near my home	0.3
nearby the home	0.1
nearby the house	0.1

The phrase translation probability for collected phrase pairs is estimated by using relative frequency:

$$P(e_1^I \mid f_1^J) = \frac{count(\overline{f}, \overline{e})}{\sum_{\overline{f}} count(\overline{f}, \overline{e})} \quad (4)$$

$P(e_1^I \mid f_1^J)$ is the frequencies of English phrase e_1^I translation of Myanmar phrase f_1^J that occur in the bilingual corpus. For example, the translation options and their probabilities for Myanmar sentence "သူသည်အိမ်တွင်ရှိသည်" are shown in table3. By using Myanmar-English bilingual corpus, the total frequency counts of Myanmar phrases "သူ၊ သည်၊ အိမ်၊ တွင်၊ ရှိသည်" are 479, 365, 78,100 and 63 respectively.

Table3. Translation Probability

f	**e**	**P(e|f)**
သူ	he	1
သည်	null	1

အိမ်	home	0.36
	house	0.64
တွင်	null	0.12
	in	0.67
	on	0.03
	at	0.18
ရှိသည်	am	0.03
	is	0.45
	are	0.2
	there is	0.32

3.4. Word-Based Lexicon Model

The proposed system uses relative frequencies to estimate the phrase translation probabilities. Most of the longer phrases are seen only once in the training corpus. Therefore, pure relative frequencies overestimate the probability of those phrases. To overcome this problem, a word-based lexicon model is used to smooth the phrase translation probabilities. This function helps to decide correct translation result. There are two symmetrization methods for the lexical model. The motivation is that a combination of the two training directions should result in better and more reliable parameter estimates for the lexical model than each of the individual directions. However, the proposed system uses lexicon probability $P(e, f)$ and determined the conditional probabilities for the source to target direction $P(e, f)$. The function should return a high value, if an English candidate word e is a common translation. It returns a low value, if an English candidate word e is a rare translation. It returns 0, if the English translation e is impossible. If scoring with the empty word is desired, the constant value of 0.05 is used in the proposed system. We combine the lexical weight P_w with the phrase pair probabilities to get new phrase translation scores:

$$P(E \mid F) = P(e_1^I \mid f_1^J)\ P_w(e_1^I \mid f_1^J) \quad (5)$$

4. Experimental Result

4.1. Corpus Statistics

System is developed from 4 different sizes of training corpora, 500 sentences, 1000 sentences, 1500 sentences and 2000 sentences, as in table 4. The number in each cell indicates the number of sentence pairs in each source (Myanmar Middle School text book, grammar book and other). Corpus statistics are shown in table 5. Zawgyi-One Myanmar font is used for Myanmar Language. The system separately takes 200 sentences as testing dataset, and the remaining (2000 sentences) is used as the training dataset. There is no overlap of sentences between training and testing datasets. There are two types of sentence: simple sentence and complex sentence. Simple sentence contains one subject (ကတ္တားပုဒ်), one verb (ကြိယာပုဒ်) and postpositional markers (PPM) (ဝိဘတ်). For example: "မောင်မောင်သည် (sub) မိဘကို (obj) (ကို=PPM) ကူညီသည် (verb) ။" Complex sentence contains one conjunction clause with two simple sentence. For example: "မောင်မောင်သည်ကြိုးစားသည်။ထို့ကြောင့်ထူးချွန်သည်။" "ထို့ကြောင့်" is conjunction clause. Sentence Type is shown in table 6. Average sentence length of Myanmar and English sentences are 12 and 10 respectively.

Table4. Training Corpus Specifications

Source (Sec)	500	1000	1500	2000
Text Book	300	600	950	1250
Grammar Book	100	250	350	500
other	100	150	200	250

Table 5: Statistics of the Training and test datasets

Sentence Pairs of Datasets		Total Words		Vocabulary Size(words)	
		Myanmar	English	Myanmar	English
Train	2000	9481	8628	338	306
Test	200	948	863	507	409

Table6: Sentence Type

Datasets	Training Dataset		Testing Dataset	
Sentence Type	Simple	Complex	Simple	Complex
sentence	1500	500	120	80

4.2. Evaluation Criteria

In this paper, evaluation of the system is measured in terms of the standard measure of BLEU (Bilingual Evaluation Understudy). The primary task for a BLEU implementer is to compute n-grams of the candidate with the n-gram of the reference translation and count the number of matches. Only single manual translated reference is used in this system. To compute precision, one counts up the number of candidate translation words (unigram) which occur in reference translation and then divides by the total number of words in the candidate translation. N-gram precision in BLEU is computed as follows:

$$p_n = \frac{\sum_{c\in\{Candidates\}} \sum_{n-gram\in C} Count_{clip}(n-gram)}{\sum_{C\in\{Candidates\}} \sum_{n-gram\in C} Count(n-gram)} \quad (6)$$

Where $Count_{clip}(n\text{-}gram)$ is the maximum number of n-grams co-occurring in a candidate translation and a reference translation, and $Count(n\text{-}gram)$ is the number of n-grams in the candidate translation. To prevent very short translations that try to maximize their precision scores, BLEU adds a brevity penalty, *BP*, to the formula:

$$BP = \begin{cases} 1 & if\ |c| > |r| \\ e^{(1-|r|/|c|)} & if\ |c| \le |r| \end{cases}$$

|c| =the length of the candidate translation
|r|=the length of the reference translation

The BLEU formula is written as follows:

$$BLEU = BP * \exp\left(\sum_{n=1}^{N} w_n \log P_n\right) \quad (7)$$

In the proposed system, N=1 and uniform weights w_n=1/N=1.

4.3. Translation Results

We evaluate the effects of lexicon model on translation quality by considering lexicon probability within a phrase-based SMT system. Lexicon model is used to smooth the phrase translation probabilities. The training corpus (Train) is used to train phrase-based translation model and the word-based lexicon. By using word-based lexicon model, the system not only can reduce OOV words but also improve translation accuracy. Translation results and OOV rate are shown in Figure3 and4.

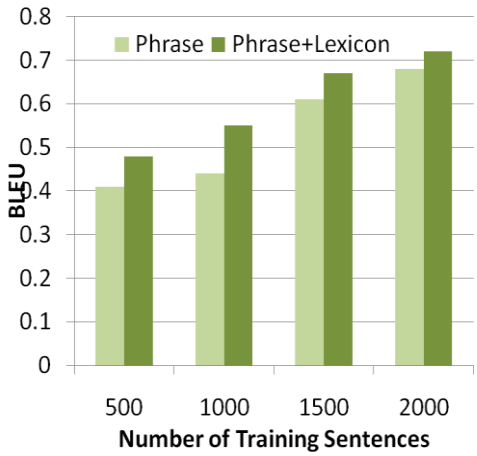

Figure3. Translation Results

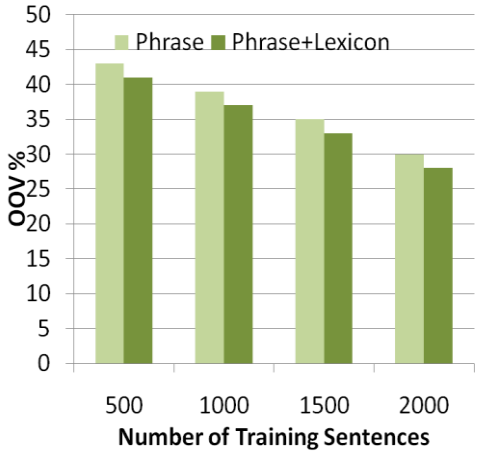

Figure4. OOV (unknown words) Rate

4.4. Error Analysis

The system treats phrases as sequences of fully-inflected words and do not incorporate any additional linguistic information. It is unable to learn translations of words that do not occur in the data, because the system is unable to generalize. This problem is severe for language

which are highly inflective and in cases where only small amounts of training data are available. Second problem is that the system is unable to distinguish between different linguistic contexts. When current models have learned multiple possible translation options for a particular word or phrase, the choice of which translation to use is guided by frequency information rather than by linguistic information. Third problem is Out-Of-Vocabulary (OOV) words in translation process. The OOV rate exceeds 40% for tested dataset with 500 training sentences, which means that near 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. OOV words can be reduced by using large training data or morphology of source and/or target language. Compound verbs and proper nouns pose problems to the robustness of a translation method. For example: "အားပေးသည်" 'encourage', "ချီးမြှင်သည်" 'award' include in compound verb: "အားပေးချီးမြှင့်သည်" 'encourage and award'. Although the corpus contains "အားပေးသည်" 'encourage, and "ချီးမြှင်သည်", we have difficult to translate these words "အားပေးချီးမြှင့်သည်" 'encourage and award' to get correct translation. Some errors occurred in adjective. Myanmar adjectives vary according to sentence patterns. Table7 shows error analysis for tested dataset. There are 63 errors in tested sentences (exclude OOV words).

Table7. Error analysis with 200 tested sentences

Error Types	Words (%)
Segmentation Error	18(28.6%)
Phrase Left	17(27%)
Phrases Selection Error	16(25.4%)
Other (PPM.etc,)	12(19%)

4.5. Efficiency

In this section, the translation speech of the phrase-based translation system is analyzed. Experiment was carried out on Intel Core i3 with 2.40 GHz. Note that the systems were not optimized for speed. We used the best performing system to measure the translation times. The translation time can be attributed to the large training corpus and sentence length. The system is evaluated on 2000 training sentences and 200 test sentences. The average sentence length of test set is 12 segmented words. Fig.5 shows the average translation time per sentence according to sentence length.

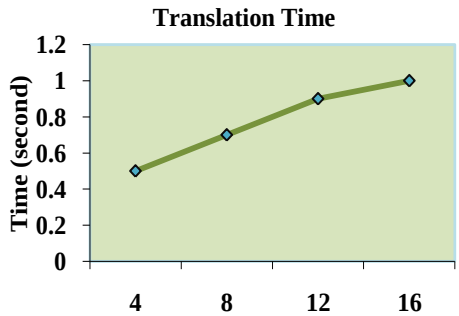

Figure5: Average Translation Time per second according to Sentence Length

4.6. Discussion

The maximum likelihood phrase translation probability is estimated by using the relative frequency count. Sometimes, the probability associated with a phrase pair computed in this way suffers from training data sparsity. To improve the quality of these scores, the proposed system uses lexical translation probabilities for individual words in the phrase pair. A comparison of BLEU scores when translating into English and training size is shown in figure3. We see that improvements can be obtained by using lexical scores. With 500 training data, translation with lexical score gains 0.08 BLEU points over translation with only phrase translation probability. In the large-corpus setting, lexical score is not too affecting on translation accuracy. To get high score in BLEU, the system also needs to use more than one reference translation.

5. Conclusion

In this paper, Myanmar phrase translation system is described. Phrase translation model

combine with lexical model to get better translation probability and to reduce unknown words in translation. The experimental results show that the system seems promising for Myanmar to English machine translation system. The system based on statistical approach. A major drawback with the statistical model is that it presupposes the existence of parallel corpus. For the translation model to work well, the corpus has to be large enough that the model can derive reliable probabilities from it, and representative enough of the domain or sub-domain it is intended to work for. The large scale Myanmar-English bilingual corpus is not available at present. Therefore, to reduce unknown words and to increase translation accuracy, morphological analysis on source or/and target language is needed. In the future, Myanmar morphological features could be considered in more training data in domain specific corpus.

References

[1] P. F. Brown, V.J. Della Pietra, S. Della. Pietra, and R. Mercer. 1993, "The mathematics of statistical machine translation: parameter estimation" Computational Linguistics, 19(2), 263-311.

[2] Berger, S. A. Della Pietra, and V. J. Della Pietra, "A maximum entropy approach to natural language processing," Computational Linguistics 22(1), 39-72, 1996.

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

[4] F. J. Och, 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.

[5]F. J.Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," In proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL),Philadelphia, July 2002, pp. 295-302.

[6] P. Koehn, "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.

[7] P. Koehn, Marcello, B. Cowan, 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.

[8]P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," In proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL), Edomonton, Canada,2003.

[9] R. Zens and H. Ney, "Improvements in Phrase-Based Statistical Machine Translation," 2004.

[10]W.P.Pa and N.L.Thein, "Disambiguation in Myanmar Word Segmentation," In proceedings of the seventh international conference on computer application, February,26-27,2009,PP 1-4.

[11] Ministry of Education, Union of Myanmar, Department of the Myanmar Language Commission, 1998, "Myanmar-English Dictionary".

[12] Ministry of Education, Union of Myanmar, Department of the Myanmar Language Commission, 2005, "Myanmar Grammar".

[13]H.H.Htay, G.B.Kumar, K.N.Murthy, "Statistical Analyses of Myanmar Corpora", Technical report, Department of Computer and Information Science, University of Hyderabad, 26 march 2007.