Statistical Machine Translation Using 5-grams Word Segmentation in Decoding

Aye Thida **Nway Nway Han** **Sheinn Thawtar Oo**
AI Research Lab, University of Computer Studies, Mandalay, Myanmar
{ayethida, nwaynwayhan, sheinthawtaroo}@ucsm.edu.mm

Abstract

This paper presents UCSMNLP's submission to the WAT 2018 Translation Tasks focusing on the Myanmar-English translation for mixed domain tasks. In statistical machine translation (SMT), word segmentation is a necessary step to generate more fluent translation results. Myanmar is one of the low resource languages and many researches are supposed to develop word segmentation for Myanmar language. However, there are no public tools for Myanmar word segmentation. This paper addresses the problem of word segmentation for Myanmar language in SMT by developing syllable, 5-grams and longest matching word segmentation with monolingual lexicon. This paper describes the phrase-based statistical machine translation (PBSMT) with batch version of MIRA tuning using 5-grams word segmentation for English-Myanmar language pairs in both directions. The experimental results showed that the baseline SMT with 5-grams could outperform the baseline system in terms of BLEU, RIBES and AMFM scores.

1 Introduction

In Natural Language Processing (NLP), machine translation system is one of the important tasks to communicate one language to another. In Myanmar script, sentences are clearly delimited by a sentence boundary maker but words are not always delimited by spaces. Words are composed of one or more syllables and that are not usually separated by white space. Sometimes spaces are used to separating the phrases for easier reading, but it is not essential, and these spaces are rarely used in short sentences.

Syllable and word segmentation are the necessary steps for statistical machine translation. This paper describes the phrase-based statistical machine translation (PBSMT) using 5-grams word segmentation for English-Myanmar language pairs in both directions.

In this work, we focus on comparing baseline PBSMT and PBSMT with 5-grams. Section 2 describes our system description. Section 3 describes the experimental setup. Section 4 describes the results and observations of our experiments. Finally, section 5 concludes the report.

2 System Description

In this section, we describe the methodology used in the machine translation experiments for this paper. This system mainly relies on the phrase-based statistical machine translation system and focus on comparing baseline PBSMT and PBSMT with 5-grams. It is implemented using the Moses toolkit (Philipp and Haddow, 2009).

2.1 Syllable 5-grams

Many other western languages use alphabetic writing system like English. However, Myanmar language uses a syllabic writing system and every syllable has a meaning is interestingly. This system firstly introduced how to do the 5-grams word segmentation. One 5-grams word consists of 5 syllables and one longest word includes maximum number of syllables in Myanmar Lexicon of Words (MLW). This system used Myanmar word segmentation by using MLW with a set of heuristics to identify word boundaries in text. In Myanmar lexicon, there are 39854 different words from myPOS corpus (draft released 1.0)[1] and myG2P[2] dictionary (Thu et al., 2016).

2.2 Phrase-based Statistical Machine Translation (PBSMT)

A PBSMT translation model is based on phrasal units. Here, a phrase is simply a contiguous sequence of words and generally, not a linguistically motivated phrase. A phrase-based translation model typically gives better translation performance than word-based models. We describe a simple phrase-based translation model consisting of phrase-pair probabilities extracted from corpus and a basic reordering model, and an algorithm to extract the phrases to build a phrase-table.

2.3 Phrase-based Statistical Machine Translation with 5-grams (PBSMT_5grams)

PBSMT with 5-grams Myanmar word segmentation uses Myanmar lexicon, which includes 39854 different words from myPOS corpus and myG2P dictionary. Table 1 shows the precision, recall and F1 scores for different segmentations. In segmentation experiments, we used the test corpus sentences from WAT-18 Myanmar-English dataset as reference corpus and segmented with 5-grams on these sentences. To evaluate the performance of the segmentation, we against the result of our segmenter with reference data. According to our experiment, the accuracy of longest segmentation is better than 5-grams segmentation. This system firstly applies the phrase-based statistical machine translation (PBSMT) using 5-grams word segmentation for English-Myanmar language pairs in both directions.

Segmentation	Precision	Recall	F1
5-grams	35.8	34.5	35.1
longest	38.9	35.4	37.0

Table 1: Precision, Recall and F1 scores for different segmentations

3 Experiments

To evaluate the translation quality of baseline PBSMT and PBSMT with 5-grams, our analysis looked through the translation tasks of two corpora, the ALT corpus and UCSY corpus.

3.1 Corpus Statistics

This system used parallel data for Myanmar-English translation tasks at WAT2018 which consists of two corpora, the ALT corpus and UCSY corpus. The ALT corpus is one part from the Asian Language Treebank (ALT) Project, consists of nearly twenty thousand Myanmar-English parallel sentences from news articles. The UCSY corpus is constructed by the NLP Lab, University of Computer Studies, Yangon (UCSY), and Myanmar. This corpus consists of over 200 thousand Myanmar-English parallel sentences collected from different domains, including news articles and textbooks.

In this experiment, the training data was combined with ALT and UCSY corpus. And then, removing mismatch sentences from training corpus. After cleaning sentences from training corpus, 226601 parallel sentences were remained. Table-2 shows data statistics used for the experiments.

Set	#Sentences
TRAIN	226,601
DEV	993
TEST	1,007

Table 2: Statistics of data sets

For the baseline, Moses tokenizer is used for both English and Myanmar language of the parallel data. Table-3 shows data statistics used for baseline PBSMT.

Set	#Sentences	#Tokens	
		En	Myan
TRAIN	226,601	743,028	170,216
DEV	993	25,360	54,268
TEST	1,007	25,903	55,022

Table 3: Statistics of data sets for BaseLine

For 5-grams PBSMT, Moses tokenizer is used for English language and 5-grams word segmenter

[1] https://github.com/ye-kyaw-thu/myPOS

[2] https://github.com/ye-kyaw-thu/myG2P

Source-Target	BLEU Scores		RIBES Scores		AMFM Scores	
	PBSMT	PBSMT (5-grams)	PBSMT	PBSMT (5-grams)	PBSMT	PBSMT (5-grams)
en-my	6.79	**19.17**	53.07	**55.40**	45.88	**61.52**
my-en	3.51	**6.01**	52.61	**53.63**	47.05	**51.90**

Table 5: BLEU, RIBES and AMFM scores for PBSMT, PBSMT with 5-grams (with Batch-Mira tuning)

is used for Myanmar language. Table-4 shows data statistics used for 5-grams PBSMT.

Set	#Sentences	#Tokens	
		Eng	Myan
TRAIN	226,601	743,028	170,216
DEV	993	25,360	65,028
TEST	1,007	25,903	66,040

Table 4: Statistics of data sets for 5-grams

3.2 Moses SMT system

We used the PBSMT system provided by the Moses toolkit (Philipp and Haddow, 2009) for training PBSMT statistical machine translation systems. According to our knowledge, there is no publicly available word segmenter for Myanmar language. Thus, we used 5-grams word segmenter only for Myanmar language which is developed by the NLP Lab, University of Computer Studies, Mandalay (UCSM), Myanmar, aiming to promote machine translation research on Myanmar language. The segmented word of source language was aligned with the segmented word of target language using MGIZA++ (Gao and Vogel, 2008). This system used grow-diag-final-and heuristic (Koehn et al., 2003) for the symmetrized alignment and msd-bidirectional-fe (Tillmann, 2004) option for the lexicalized reordering model was trained with. Although the sentences in test data are long, this system used (default 6) distortion limit in MOSES. The language model was trained by KENLM (Heafield, 2011) with interpolated modified Kneser-Ney discounting (Heafield et al., 2013). To do the tuning for decoder parameters, this system used Batch (J=60) with MIRA tuning (Cherry and Foster, 2012) and Moses decoder (version 2.1.1) for decoding.

3.3 Evaluation

This system reports the translation quality of those methods in terms of Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002), Rank-based Intuitive Bilingual Evaluation Measure (RIBES) (Isozaki et al., 2010) and Adequacy-Fluency Metrics (AMFM) (Banchs, et al., 2015).

In PBSMT without tuning method, this system gets -96.5 pairwise human evaluation score in English to Myanmar translation and -99.5 in Myanmar to English translation.

To get the more influent translation results, this system used the Batch-Mira tuning method. Table 5 shows the BLEU, RIBES and AMFM scores for PBSMT and PBSMT with 5-grams by using batch size 60 with mira tuning. Due to time constraint, the results of PBSMT with 5-grams by using batch-mira tuning method could not be submitted for manual evaluation.

4 Results and Discussion

The BLEU, AMFM and RIBES score results for machine translation experiments with PBSMT and PBSMT with 5-grams segmentation are shown in Table 5. The highest scores of the different approaches are indicated as bold numbers. Comparing to existing baselines of PBSMT, PBSMT with 5-grams approach achieved higher scores in our experiments.

According the results from Table 5, the PBSMT with 5-grams outperforms the baseline PBSMT in terms of BLEU, RIBES and AMFM score from English to Myanmar translation. On the other hand, from Myanmar to English translation, the PBSMT with 5-grams better than the baseline PBSMT in terms of BLEU, RIBES and AMFM score.

Further experimentation is required to see the PBSMT with n-grams segmentation and analyze which one performs better on English-Myanmar in both directions. One way to test this is by increasing the n-grams (longest) segmentation even further.

Our experimental results indicate that in terms of adequacy (as measured by BLEU score), the PBSMT with 5-grams approach produced the higher quality translations than baseline. However in some sentences, PBSMT with 5-grams approach cannot translate in name entities. Likewise, the PBSMT with 5-grams is better word order (as measured by the RIBES score) than the baseline but this system requires to get correct word order in long sentences for both directions. In terms of both semantic and syntactic components of the translation process, the AMFM scores of PBSMT with 5-grams give higher scores than baseline that provide a more balanced view on Myanmar-English bidirectional translation quality.

For Myanmar Language, PBSMT with 5-grams segmentations approach is not suitable to get the adequate and fluent translation results. As an example, the Myanmar word “တပ်မတော်ဒုတိယကာကွယ်ရေးဦးစီးချုပ်” is translated to “Deputy Chief of Defense” in English. Therefore, we need to test n-grams (longest) segmentation to get better translation results. Moreover, according to our experiments, automatic evaluation metrics can sometimes be misleading. In Myanmar-English bilingual language translation, the word order difference of two languages is one of the difficulties in machine translation to get influent results. Therefore, human evaluation with bilingual judges (Vilar et al., 2007) would be required to get better qualities of the machine translation approaches for Myanmar-English language pair.

5 Conclusion

This paper compared two approaches of PBSMT: baseline PBSMT and PBSMT with 5-grams segmentation approach for Myanmar-English language pairs in both directions. Although the evaluation results of PBSMT with 5-grams with batch-mira tuning are better than baseline PBSMT, this approach is not suitable to get the adequate and fluent translation results. In future, we would like to investigate the better model by modifying the PBSMT with 5-grams approach to improve the translation quality of statistical machine translation for Myanmar-English in both directions.

References

Banchs, R.E., D'Haro, L.F. and Li, H., 2015. Adequacy-fluency metrics: Evaluating MT in the continuous space model framework. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, *23*(3), pp.472-482.

Chen, S.F. and Goodman, J., 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, *13*(4), pp.359-394.

Cherry, C. and Foster, G., 2012, June. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 427-436). Association for Computational Linguistics.

Gao, Q. and Vogel, S., 2008, June. Parallel implementations of word alignment tool. In *Software engineering, testing, and quality assurance for natural language processing* (pp. 49-57). Association for Computational Linguistics.

Heafield, K., 2011, July. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation* (pp. 187-197). Association for Computational Linguistics.

Heafield, K., Pouzyrevsky, I., Clark, J.H. and Koehn, P., 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Vol. 2, pp. 690-696).

Isozaki, H., Hirao, T., Duh, K., Sudoh, K. and Tsukada, H., 2010, October. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 944-952). Association for Computational Linguistics.

Koehn, P. and Haddow, B., 2009, March. Edinburgh's submission to all tracks of the WMT2009 shared task with reordering and speed improvements to Moses. In *Proceedings of the Fourth Workshop on Statistical Machine Translation* (pp. 160-164). Association for Computational Linguistics.

Koehn, P., Och, F.J. and Marcu, D., 2003, May. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (pp. 48-54). Association for Computational Linguistics.

Nakazawa, T., Higashiyama, S., Ding, C., Dabre, R., Kunchukuttan, A., Pa, W.P., Goto, I., Mino, H., Sudoh, K. and Kurohashi, S., 2018, December. Overview of the 5th Workshop on Asian Translation.

In *Proceedings of the 5th Workshop on Asian Translation (WAT2018)*.

Och, F.J. and Ney, H., 2000, October. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics* (pp. 440-447). Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.J., 2002, July. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311-318). Association for Computational Linguistics.

Thu, Y.K., Pa, W.P., Sagisaka, Y. and Iwahashi, N., 2016. Comparison of grapheme-to-phoneme conversion methods on a myanmar pronunciation dictionary. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)* (pp. 11-22).

Tillmann, C., 2004, May. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers* (pp. 101-104). Association for Computational Linguistics.

Vilar, D., Leusch, G., Ney, H. and Banchs, R.E., 2007, June. Human evaluation of machine translation through binary system comparisons. In *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 96-103). Association for Computational Linguistics.