UCSMNLP: Statistical Machine Translation for WAT 2019

Aye Thida, Nway Nway Han, Sheinn Thawtar Oo, Khin Thet Htar
AI Research Lab, University of Computer Studies, Mandalay, Myanmar
ayethida, nwaynwayhan, sheinthawtaroo, khinthethtar@ucsm.edu.mm

Abstract

This paper represents UCSMNLP's submission to the WAT 2019 Translation Tasks focusing on the Myanmar-English translation. Phrase based statistical machine translation (PBSMT) system is built by using other resources: Name Entity Recognition (NER) corpus and bilingual dictionary which is created by Google Translate (GT). This system is also adopted with listwise reranking process in order to improve the quality of translation and tuning is done by changing initial distortion weight. The experimental results show that PBSMT using other resources with initial distortion weight (0.4) and listwise reranking function outperforms the baseline system.

1 Introduction

Machine translation system can be formally defined as the task of translating text given in one natural language to others automatically (Koehn, P., et al., 2003). In Natural Language Processing (NLP), machine translation system is one of the important tasks to communicate one language to another. Developing high quality machine translation systems has been special interest in NLP research area. Many different preprocessing and post-processing tasks have also been studied in order to get high quality. In this work, both tasks are performed by building lexicons and reranking the translations. And translation quality is also observed by changing initial distortion weight.

For the preprocessing task, NER corpus and Bilingual lexicon which support the translation tasks, are built by Standford NER tagger and Google Translate (GT). These two resources are used to combine and retrain with existing ALT corpus for translation task. For the post-processing tasks, reranking is performed with the combination of baseline pointwise reranking and listwise reranking which takes into account the similarity score of each translation to all other translations included in n-best list. And the initial distortion weight that gives better translation result is analyzed by changing various initial distortion weights.

This paper describes phrase based statistical machine translation (PBSMT) by building bilingual lexicons, changing distortion weight and reranking for English-Myanmar translation in both directions. Section 2 describes system description. PBSMT is described in Section 3 followed by building bilingual lexicons in Section 4 and Section 5 describe experimental results. Finally, Section 6 will conclude this report.

2 System Description

This system is built phrase based statistical machine translation (PBSMT) system using other resources: Name Entity Recognition (NER) corpus and bilingual dictionary which is created by Google Translate (GT). These two resources are combined with existing ALT corpus which is used as the training data. This system is also adopted with listwise reranking process in order to improve the quality of translation and tuning is done by changing initial distortion weight.

2.1 Phrase Based Statistical Machine Translation (PBSMT)

A PBSMT translation model strives to produce the best possible translations based on probabilistic models analyzing phrase units, sequences of words, extracted from sentence aligned Myanmar-English parallel corpus. A phrase based translation model typically gives better performance than word-based translation model because one word in one language may not be one word in other languages (Koehn, P., et al., 2003). Changing the initial distortion weights for tuning process and reranking are the

Proceedings of the 6th Workshop on Asian Translation, pages 94–98
Hong Kong, China, November 4, 2019. ©2019 Association for Computational Linguistics

crucial processes to acquire the better translation result.

2.1.1 Distortion

Distortion is one of phrase based models used to justify the placement of words in different orders in the output translation. Before tuning process, initial distortion weight value is needed to assign. This system performs tuning process by changing the initial weight of distortion model from 0.1 to 0.6. Table 1 shows BLEU scores by changing various initial distortion weights in Myanmar-English bidirectional translations.

Data Set	Initial Distortion Weight	BLEU	
		my-en	en-my
ALT	0.1	6.96	24.16
ALT	0.2	6.93	23.86
ALT	0.3	7.02	24.13
ALT	**0.4**	**7.15**	**24.24**
ALT	0.5	7.04	24.22
ALT	0.6	7.02	24.05

Table 1: BLEU scores by changing initial distortion weight in Myanmar-English bidirectional translations.

According to the experiments, the BLEU score result by changing the initial distortion weight (0.4) is better than other initial distortion weights for both Myanmar-English directions. Therefore, we choose the initial distortion weight (0.4) for tuning to get the better translation results.

2.1.2 Reranking

Reranking aims to consider the entire list of best possible translations as a whole through the adoption of a listwise ranking function, which calculates the reranking score by asking each translation to report its similarity to all other translations (Zhang, M. et al., 2016). Reranking is the combination of pointwise and listwise reranking score. Pointwise score is calculated based on 14 baseline features such as 4 translation models, a language model , a word penalty, a phrase penalty and 7 reordering models.

The listwise reranking process contains the two main functions, tuning and similarity calculation. In the similarity calculation, the translation scores of candidates correspond to the current candidate is also considered to get higher similarity between translations. In this system, two evaluation metrics, Bilingual Evaluation UnderStudy (BLEU) (Papineni et al., 2002) and Metric for Evaluation of Translation with Explicit Ordering (METEOR) (Denkowski, M. and Lavie, A., 2014), are used as two feature functions for reranking to measure the similarity between translations in n-best list. And then the weights of these two feature functions are tuned on development set using z-mert tuning (Zaidan, O., 2009). This system chooses the 100 translation candidates (N=100) which impact on reranking model because of consideration of similarity between translations in n-best list.

2.2 Building Bilingual Resources

In machine translation, bilingual resources are essential language resources to get the influent translations. Moreover, the areas concerned with NER are also needed to be developed for translation tasks from Myanmar language to other languages.

2.2.1 Name Entity Recognition (NER) Corpus

This system uses Stanford NER tagger[1] to make the tagging process for every English token e (in the parallel data). If e has any tag in tagging process, this system extracts the translation of e by using the Myanmar ALT Treebank. In order to decide whether the two tokens are correctly translated in extracting NER corpus, we manually checked if the two tokens have translation of each other. Finally, we added the translation pairs to the bilingual NER corpus one at a time. The data statistics of NER corpus is shown in Table 2.

Corpus	Translation pairs
ALT	230,240
NER (Raw)	14313
NER (clear)	14310

Table 2: Data statistics of the NER corpus.

[1] https://nlp.stanford.edu/software/CRF-NER.html#Download

2.2.2 Bilingual Lexicon

For bi-directional translation tasks of Myanmar-English, the system built bilingual lexicon to retrain the data with existing corpus to get the fluent translations. This bilingual corpus is built by using Google Translate (GT)[2]. When building the bilingual lexicon, distinct English and Myanmar tokens from ALT my-en corpus is used as input words for GT to get Myanmar-English translation pairs and then add these translations pairs to the bilingual lexicon. The data statistics of Bilingual lexicon is shown in Table 3.

Bilingual Lexicon	Translation pairs
English	54674
Myanmar	35532
Total	90206

Table 3: Data statistics of the Bilingual Lexicon.

3 Experiments

To evaluate the translation quality of baseline PBSMT and PBSMT with reranking, our analysis looked through the translation tasks of ALT corpus by adding bilingual lexicons. All experiments are trained on Dell PowerEdge R720.

3.1 Corpus Statistics

This system used ALT corpus for Myanmar-English translation tasks at WAT 2019. The training data was the combination of ALT corpus and two new resources, NER corpus and bilingual lexicon, which are built using ALT TreeBank, Standford NER tagger and Google Translate (GT).

Data Set				**#sentences**
TRAIN	**ALT**	**NER**	**Bilingual**	112570
	18088	14310	80172	
DEV				1000
TEST				1018

Table 4: Statistics of data sets.

ALT corpus is one part from the Asian Language Treebank (ALT) Project, consists of twenty thousand Myanmar-English parallel sentences from news articles. In this experiment, the

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. The 5-grams language model was trained by KENLM (Heafield, 2011) with modified Knerser-Ney discounting (smoothing). The alignment process is implemented using GIZA++ (Och, F.J. and Ney, H., 2000). This system used grow-diag-final-and heuristic for symmetrized alignment and msd-bidirectional-fe (Koehn et al., 2003) 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. To tweak the parameters of decoding, Minimum Error Rate Training (MERT) (Och, F.J., 2003) is used by changing various initial distortion weights. Reranking is performed based on 100 (default) best possible target translations generated by Moses decoder.

4 Results and Discussion

This system reports the translation quality of those methods in terms of Bilingual Evaluation Understudy (BLEU), Rank-based Intuitive Bilingual Evaluation Measure (RIBIES) (Isozaki et al., 2010) and Adequacy-Fluency Metrics (AMFM) (Banchs et al., 2015) in Table. 5.

In our experiments, firstly the initial distortion weights are changed from 0.1 to 0.6 as shown in Table 1, we found that the translation result did not improve significantly compared with baseline. Second, we analyze with new reranking method (listwise) which is combined with pointwise. The translation quality is not good enough. Finally, two bilingual lexicons are added to the existing parallel data to reuse as the training data. In PBSMT without reranking model, the translation result is significantly improved from 7.15 to 10.70 in Myanmar-English and 24.24 to 28.20 in English to Myanmar translation. On the other hand, from Myanmar to English translation, the PBSMT with reranking is better than baseline PBSMT in

[2] https://github.com/ssut/py-googletrans

Source-Target	BLEU			RIBIES			AMFM		
	PBSMT	PBSMT (Without Reranking)	PBSMT (Reranking)	PBSMT	PBSMT (Without Reranking)	PBSMT (Reranking)	PBSMT	PBSMT (Without Reranking)	PBSMT (Reranking)
en-my	24.24	28.20	-	58.15	59.68	-	61.67	69.34	-
my-en	7.15	10.70	10.84	53.29	57.08	57.11	53.01	53.82	54.04

Table 5: BLEU, RIBES and AMFM scores for PBSMT, PBSMT with reranking.

Source	ပြည်ထဲရေး ဝန်ကြီး ဌာန နှင့် ထောက်လှမ်း ရေး အဖွဲ့အစည်း ၏ တာဝန် ရှိ သူ များ က သူမ ဆီသို့ ကောင်လေး ကို ပေး ခဲ့ ပြီးနောက် ဒေါက်တာ ဖာဇီယ က သတင်းထောက် များ ကို ပြောကြား ခဲ့ သည် ။
Reference	Dr. Fauzia told journalists after the boy had been given to her by officials of the interior ministry and intelligence agencies .
Baseline	Interior Ministry and intelligence officials of her towards the boy after the Dr. ဖာဇီယ told reporters .
Baseline with Reranking	Interior Ministry and Intelligence of officials said she was given to the boy after the Dr. ဖာဇီယ told reporters .
Baseline+NER+GT	Interior Ministry and intelligence agency in charge of the boy to her after Dr. **Fauzia** told reporters .
Baseline+NER+GT with Reranking	Interior Ministry and intelligence **agency 's** officials to her **after the boy** Dr. **Fauzia** told reporters .

Table 6: Comparison between my-en translation results.

terms of BLEU (7.15 to 10.84), RIBES (53.29 to 57.11) and AMFM (53.01 to 54.04) scores.

In table 6, the comparison between translation results of my-en is described. In this table, “Source” and “Reference” sentences are shown in the first two rows. The translation of “baseline” and the translation of baseline with reranking cannot translate the name “ဖာဇီယ”. After using NER and GT, this name can translate as “**Fauzia**”. The translation result is a slightly smooth after reranking. The result “agency in charge of the boy to her” to “agency 's officials to her” and “the boy to her after” has been changed to “after the boy”. Even though the translation result is not definitely perfect, using resources with reranking can change to better translation is one of the worthy evidences.

According to our experiments, using resources with PBSMT model get better translation result significantly. Even though the translation result is better than the baseline, the current resources that we used in this system is not still covered for fluent translation, we need to extend the current resources and build new resources in future.

5 Conclusion

In this paper, we have described our submissions to WAT 2019. To improve the translation result, two bilingual resources were added to the training data and the result of our system was comparable to baseline PBSMT model. The reranking result of my-en is better than baseline system, however, our team can not submit PBSMT with reranking results of en-my because of time constraint. This is the initial learning of PBSMT model and still need to explore with other models to get the adequate and fulfilled translation results. In future, we would like to extend the existing Myanmar resources and investigate the better models for Myanmar to other language machine translation system.

References

Banchs, R.E., D'Haro, L.F. and Li, H., 2015. Adequacyfluency 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.

Denkowski, M. and Lavie, A., 2014, June. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation* (pp. 376-380).

Finkel, J.R., Grenager, T. and Manning, C., 2005, June. Incorporating non-local information into information extraction systems by gibbs sampling. *In Proceedings of the 43rd annual meeting on association for computational linguistics 2005 Jun 25 (pp. 363-370). 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 Linguistic (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. 944952). 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., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R. and Dyer, C., 2007, June. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions* (pp. 177-180).

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.*

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.*

Och, F.J., 2003, July. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1 (pp. 160-167). 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.*

Thida, A., Han, N.N. and Oo, S.T., 2018. Statistical Machine Translation Using 5-grams Word Segmentation in Decoding. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation*.

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.*

Zaidan, O., 2009. Z-MERT: A fully configurable open source tool for minimum error rate training of machine translation systems. *The Prague Bulletin of Mathematical Linguistics*, *91*, pp.79-88.

Zhang, M., Liu, Y., Luan, H. and Sun, M., 2016. Listwise ranking functions for statistical machine translation. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, *24*(8), pp.1464-1472.