UCSYNLP-Lab Machine Translation Systems for WAT 2018

[1]Yi Mon Shwe Sin, [1]Thazin Myint Oo, [1]Hsu Myat Mo, [1]Win Pa Pa, [1]Khin Mar Soe and [2]Ye Kyaw Thu

[1]Natural Language Processing Lab., University of Computer Studies, Yangon, Myanmar
[2]Language and Speech Science Research Lab., Waseda University, Japan
{yimonshwesin, thazinmyintoo, hsumyatmo, winpapa, khinmarsoe}@ucsy.edu.mm,
wasedakuma@gmail.com

Abstract

In this description, we report the experimental results of Machine Translation models conducted by a team from University of Computer Studies, Yangon (UCSY) for the translation tasks of WAT 2018. Generally, our models are based on neural methods and statistical methods for both Myanmar-English and English-Myanmar direction of languages pair. For the neural method experiments, attention-based neural machine translation (NMT) that uses word level segmentation and Transformer that uses sub-word level segmentation have been carried out. In the portion of statistical machine translation (SMT), we used three different statistical approaches: phrase-based, hierarchical phrase-based, and the operation sequence model (OSM). Different Machine Translations are conducted on the ALT and UCSY datasets and the best scores from the experiments are described in this system description.

1 Introduction

Machine Translation (MT) which is also known as Computer Aided Translation is the task of specifically designing to translate both verbal and written texts between natural languages by a computer system. MT uses a machine translation engine to perform substitution of words or phrases or any other in one language for words or phrases or any other in another language. MT is widely used in Natural Language Processing (NLP) tasks such as online translation services applications in information extraction, document retrieval, intelligence analysis, electronic mail, and much more. A few different types of MT are available in the market today, the most widely used are Statistical Machine Translation (SMT), Rule-Based Machine Translation (RBMT), Hybrid Systems, which combine RBMT and SMT and Neural Machine Translation (NMT). However, there are still many challenges for high-quality translations in real-world applications.

To date, there have been very few studies on the MT from Myanmar language to other languages (T.Zin, 2011), (W. Pa, 2016). And Myanmar MT is still in its early stages and researchers are faced with many difficulties such as the lack of resources. Existing research on Myanmar MT has been either rule-based or more recently statistical-based have been tried. There have been some studies on the SMT of Myanmar language. Ye Kyaw Thu et al. (2016) presented the first large-scale study of the translation of the Myanmar language. A total of 40 language pairs were used in the study that included languages both similar and fundamentally different from Myanmar. The results show that the hierarchical phrase-based SMT (HPBSMT) approach gave the highest translation quality in terms of both the BLEU and RIBES scores. Win Pa Pa et al (2016) presented the first comparative study of five major machine translation approaches applied to low-resource languages, PBSMT, HPBSMT, tree-to-string (T2S), string-to-tree (S2T) and OSM translation methods to the translation of limited quantities of travel domain data between English and {Thai , Laos, Myanmar} in both directions. The experimental results indicate that in terms of adequacy (as measured by BLEU score), the PBSMT approach produced highest quality translations. From their RIBES scores, we noticed that OSM approach achieved best machine translation performance for Myanmar to English translation. There was also a study of SMT on word segmented and syllable segmented data for

Myanmar language by (Ye Kyaw Thau et al., 2016) and they proved word information had large effect in MT.

We have prepared 200K of parallel corpus and tries on both statistical machine translation system and neural machine translation system. In the experiments, there are two different NMT models, NMT with attention and NMT with Transformer model and two SMT models, OSM, PBSMT and HPBSMT and we will refer them as two NMT models as NMT1, NMT2 in the rest of the sections.

The toolkits we used for NMT1 is PyTorch OpenNMT[1] for NMT1 and Sockey Sequence-to-sequence Framework[2] for NMT2. NMT1 is a simple NMT model with an attention mechanism. We implement NMT1 with word level segmentation. For word level segmentation of Myanmar language, we use UCSY_NLP lab segmenter[3].

To build SMT models, we used the Moses (P. Koehn, 2007) which is the de facto tool among the numerous MT tools. Language Modeling is trained by using kenLM using 5-grams, with modified Kneser-Ney discounting (smoothing). Alignment with GIZA++[4] implementation of IBM word alignment model 4 with grow-diagonal-final- and heuristic for phrase-extraction. The lexicalized reordering model was trained with the msd-bidirectional-fe option. Minimum error rate training (MERT) was used to tune the decoder parameters and the decoding was done using the Moses decoder (version 2.1.1).

In this report, section 2 will describe our MT systems. In section 3, the experimental setup will be described. In section 4, the results of our experiments will be reported followed by the conclusion in section 5.

2 System Description

NMT systems and SMT systems are used for Myanmar-English translations in both directions. To reduce the vocabulary size, we apply byte pair encoding (BPE; Sennrich et al., 2016) which breaks all words into sub-word units in Transformer model and SMTs, with different number of BPE segmentations.

The NMT1 model is based on the standard encoder-decoder architecture with attention as proposed by (Bahdanau et al., 2015). The encoder is a bidirectional recurrent neural network (BiRNN) using Gated Recurrent Units. In each step, it takes an embedded token from the input sequence and its previous output and outputs a representation of the token. The encoder works in both directions; the resulting vector representations at corresponding positions are concatenated. Additionally, the final outputs of both the forward and backward run are concatenated and used as the initial state of the decoder. At each decoding step, it takes its previous hidden state and the embedding of the token produced in the previous step as the input and produces the output vector. This vector is used to compute the attention distribution vector over the encoder outputs. The RNN output and the attention distribution vector are then used as the input of a linear layer to produce the distribution over the target vocabulary. During training, the previously generated token is replaced by the token present in the reference translation.

For building NMT2, we applied Transformer NMT that based on self-attention mechanism. The architecture is single layer encoder and decoder. The model is trained with a sub-word vocabulary and we apply it to all the training and evaluation data.

In the experiment description of SMT, we trained PBSMT and OSM models for English to Myanmar translation and HPBSMT and OSM for Myanmar to English. In this system description, we propose 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 phrased to build a phrase table. We model it using 5-gram language model under the PBMT paradigm. The hierarchical phrase-based SMT approach is a model based on synchronous context free grammar and the model is able to learn from corpus of unannotated parallel text. The benefit of this technique is that the hierarchical structure is able to represent the word reordering process. As a

[1] http://github.com/OpenNMT/OpenNMT-py
[2] https://awslabs.github.io/sockeye
[3] http://nlpresearch-ucsy.edu.mm/NLP_UCSY/wordsegmentation.html
[4] http://www.statmt.org/moses/giza/GIZA++.html

consequence of this advantage, this makes particularly applicable to language pairs that requires long distance reordering in the case of Myanmar-English translation process. The OSM combines the benefits of phrase-based and N-gram based SMT. It is based on minimal translation units, capture source and target context across phrasal boundaries and simultaneously generate source and target units. OSM motivates better reordering mechanism that uniformly handles local and non-local reordering and strong coupling of lexical generation and reordering. It means that OSM can handle both short and long distance reordering. The list of operations can be divided into two groups and there are five translation operations Generate(X, Y), Continue Source Cept, Generate Identical, Generate Source Only (X) and Generate Target Only (Y) and three reordering operations such as Insert Gap, Jump Back (N) and Jump Forward.

3 Experimental Setup

3.1 Datasets and preprocessing

The parallel data for Myanmar-English and English-Myanmar translation tasks at WAT2018 consists of two corpora: the ALT corpus and the UCSY corpus. The ALT corpus is one part from the Asian Language Treebank (ALT) Project, consisting of twenty thousand Myanmar-English parallel sentences from Wiki news articles. The UCSY corpus is constructed by the NLP Lab, University of Computer Studies, Yangon (UCSY), Myanmar, aiming to promote machine translation research on Myanmar language. This corpus consists of 200K Myanmar-English parallel sentences collected from different domains, including local news articles and textbooks (Yi Mon et.al, 2018). The UCSY corpus and a portion of the ALT corpus are used as training data, which are around 220,000 lines of sentences and phrases. The development and test data are from the ALT corpus. Therefore, the training data for Myanmar-English and English-Myanmar translation tasks is a mix domain data collected from different sources. Table 1 shows data statistics used for the experiments.

Data Type	File Name	Number of Sentences
TRAIN	train.ucsy.[my|en]	208,638
	train.alt.[my|en]	17,965
DEV	dev.alt.[my|en]	993
TEST	test.alt.[my|en]	1,007

Table 1: Statistics of Datasets

Due to Myanmar Language being an unsegmented language with no clear definition of word boundaries, proper text segmentation is essential. Although the Myanmar textual data given form the WAT2018 have been segmented into writing units and Romanized, the data provided was segmented into word level. Moses tokenizer is used for English side of parallel data in NMT1.

In experiment of SMT, byte pair encoding (BPE) is trained using the source and target side of the data. A technique is to segment words into smaller sub-word unit. BPE word segmenter conceptually proceeds by first splitting all words in the whole corpus into individual characters. The most frequent adjacent pairs of symbols are then consecutively merged, until a specified limit of merge operations has been reached. The merge operations learned on a training corpus and that is purely frequency-based. The frequent sequence of characters will be joined through the merge operations, resulting the common words not being segmented. Words containing rare combination of characters will not be fully merged from the characters splitting all the way back to their original form. They will remain split into two or more sub-word units in the BPE segmented data.

3.2 Training

Table 2 shows the settings of network hyper-parameters for NMT models, and Table 3 for SMT models. The experiments were run on Tesla K80 GPU. Based on different parameter settings, the training time is different.

Hyper-parameter	NMT1 Settings	NMT2 Setting
Source Vocabulary size	25,087	10,000
Target vocabulary size	50,004	10,000
Number of hidden units	500	512
Encoder layer	2	1
Decoder layer	2	1
Learning rate	1.0	0.002
Dropout rate	0.3	0.2
Mini-batch size	64	100

Table 2: Hyper-parameter of NMT models

Alignment model	Grow–diag-final and heuristic
Lexicalized reordering model	Msd-bidirectional-fe
Language Model	kenLM (5-gram)
Smoothing	Modified Knerser-Ney discounting
Decoding	Moses decoder
Tuning	Minimum Error Rate Tuning (MERT)

Table 3: Moses settings

3.3 Experimental Results

Table 4 and the Table 5 show the different evaluation metrics such as Bilingual Evaluation Understudy (BLEU), Rank-based Intuitive Bilingual Evaluation Score (RIBES) and Adequacy-Fluency Metrics (AMFM) (Banchs et al., 2015) for Myanmar- English and English-Myanmar translation pairs. We also investigated how segmentation level affects the MT performance in all experiments. The experimental results reveal that word level segmentation can give better performance for attention-based NMT while sub-word level works better with Transformer. Moreover, experiments are conducted by tuning different parameter settings for all NMT1, NMT2 and SMT. Best scores among those of the experimental results are submitted in this description.

Method	BLEU	RIBES	AMFM
NMT1	19.19	0.671,461	0.717,480
NMT2	21.19	**0.679,800**	**0.756,710**
OSM	**22.78**	0.549,883	0.751,180
PBSMT	22.40	0.544,395	0.749,080

Table 4: English to Myanmar Translation

Method	BLUE	RIBES	AMFM
NMT1	**9.56**	**0.642,309**	0.518,990
HPBSMT	8.91	0.583,956	0.560,800
OSM	8.84	0.553,786	**0.594,800**

Table 5: Myanmar to English Translation

In the direction of Myanmar to English, Table 5 show only 3 system results. Experimental result of NMT2 was not able managed to submit in time. In Myanmar to English translation, NMT1 with outperforms HPBSMT and OSM models in terms of BLEU score and the RIBES score. However, OSM gets highest score in AMFM. In English to Myanmar translation, the OSM model performs better than the other models in terms of BLEU score but NMT2 model is better than the others in RIBES and AMFM score. We used Byte Pair Encoding (BPE) segmentation for SMT experiments. Generally says that OSM is the best method for bi-directional translations. In the results of English to Myanmar translation as shown in Table 4, we got the highest BLEU score in the method of OSM and the RIBES and AMFM scores is nearly the same with PBSMT. Interestingly, we got highest AMFM in the method of OSM in Myanmar to English translation and there is little difference in scores of BLEU and RIBES with comparison of HPBSMT.

Conclusion

In this system description for WAT2018, we submitted our NMT systems, which are NMT with attention and NMT with sockey. And we also submitted SMT systems, which are PBSMT, HPBSMT and OSM. We evaluated our systems on Myanmar-English and English-Myanmar translations at WAT 2018. Our team is the first time of competition in WAT and there are so many weaknesses to fulfillment of our destination. In the future, we will collect the more parallel sentences to get a large-sized MT corpus. And we will remove the noise to clean the existing corpus because it contained a lot of parallel sentences with different content. Moreover, we also intend to do more and more experiments with more recent evolutions of the translation models.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *In Proceedings of ACL – IJCNLP 2015*, Volume 1: Long Papers (2015). arXiv preprint arXiv:1409.0473.

Fabien Cromieres, Fabien Cromieres, Toshiaki Nakazawa and Toshiaki Nakazawa. Kyoto University Participation to WAT 2017, Proceedings of the 4th Workshop on Asian Translation, pages 146–153, Taipei, Taiwan, November 27, 2017. © 2017AFNLP.

Guillaume Klein, Yoon Kim, Yoon Kim, Jean Senellart, Alexander M. Rush, SYSTRAN and Harvard SEAS. OpenNMT: Open-Source Toolkit for Neural Machine Translation. (2017). Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, pages 67–72 Vancouver, Canada, July 30- August 4, 2017. ©2017 Association for Computationsl Linguistics http://doi.org/10.18653/v1/P17-4012

Makoto Morishita, Jun Suzuki and Masaaki Nagata. NTT Neural Machine Translation Systems at WAT 2017. Proceedings of the 4th Workshop on Asian Translation, pages 89–94, Taipei, Taiwan, November 27, 2017. © 2017 AFNLP.

Minh-Thang Luong, Hieu Pham and Christopher D. Maiining. Effective Approaches to Attention-based Neural Machine Translation. Proceedings of the 2015 Conference on Empirical Methods in Natu-ral Language Processing, pages 1412-1421(2015).

Rafael E Banchs, Luis F D'Haro, and Haizhou Li. 2015. Adequacy-fluency metrics: Evaluating mt in the continuous space model framework. IEEE Transactions on Audio, Speech, and Language Processing, 23(3):472-482.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In Proceedings of the An-nual Meeting of the Association for Computational Linguistics, pages 86–96.

Rico Sennrich, Barry Haddow and Alexandra Birch (2016): Neural Machine Translation of Rare Words with Subword Units Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin, Germany

Sandhya Singh, Ritesh Panjwani, Anoop Kunchukuttan and Pushpak Bhattacharyya. Comparing Recurrent and Convolutional Architectures for English-Hind Neural Machine Translation. Proceedings of the 4th Workshop on Asian Translation, pages 167–170, Taipei, Taiwan, November 27, 2017. ©2017 AFNLP.

Thet Thet Zin, Khin Mar Soe and Nilar Thein. Myanmar Phrases Translation Model with Morphological Analysis for Statistical Myanmar to English Translation System. 25th Pacific Asia Conference on Language, Information and Computation, pages 130-139(2011).

Ye Kyaw Thu, Andrew Finch, Win Pa Pa, and Eiichiro Sumita, " A large scale study of Statistical Machine Translation Methods for Myanmar Language ", in Proc. Of SNLP2016, February 10-12, 2016

Win Pa Pa, Ye Kyaw Thu, Andrew Finch and Eiichiro Sumita. A Study of Statistical Machine Translation Methods for Under Resourced Languages. 29th Pacific Asia Conference on Language, Information and Computation pages 259-269(2016).

P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation." in Proc. of HTL-NAACL, 2003, pp. 48–54.

Chiang, D., "Hierarchical phrase-based translation", Computational Linguistics 33(2), 2007, pp. 201-228.

Nadir Durrani, Helmut Schmid, Alexander M. Fraser, Philipp Koehn and Hinrich Schutze, "The Operation Sequence Model - Combining N-Gram-Based and Phrase-Based Statistical Machine Translation", Computational Linguistics, Volume 41, No. 2, 2015, pp. 185-214.

Prachya, Boonkwan and Thepchai, Supnithi, "Technical Report for The Network-based ASEAN Language Translation Public Service Project", Online Materials of Network-based ASEAN Languages Translation Public Service for Members, NECTEC, 2013

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

Och Franz Josef and Ney Hermann, "Improved Statistical Alignment Models", in Proc. of the 38th Annual Meeting on Association for Computational Linguistics, Hong Kong, China, 2000, pp. 440-447.

Tillmann Christoph, "A Unigram Orientation Model for Statistical Machine Translation", in Proc. of HLT-NAACL 2004: Short Papers, Stroudsburg, PA, USA, 2004, pp. 101-104.

Heafield, Kenneth, "KenLM: Faster and Smaller Language Model Queries", in Proc. of the Sixth Workshop on Statistical Machine Translation, WMT '11, Edinburgh, Scotland, 2011, pp. 187-197.

Chen Stanley F and Goodman Joshua, "An empirical study of smoothing techniques for language modeling", in Proc. of the 34th annual meeting on Association for Computational Linguistics, 1996, pp. 310-318.

Och Franz J., "Minimum error rate training in statistical machine translation", in Proc. of the 41st Annual Meeting n Association for Computational Linguistics – Volume 1,Association for Computer Linguistics, Sapporo, Japan, July, 2003, pp.160-167.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton and Matt Post. 2017. Sockeye: A Toolkit for Neural Machine Translation. ArXiv e-prints.

Yi Mon Shwe Sin and Khin Mar Soe, "Large Scale Myanmar to English Neural Machine Translation System". Proceeding of the IEEE 7th Global Conference on Consumer Electronic (GCCE 2018).