The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Application of Unsupervised NMT Technique to Japanese–Chinese Machine Translation

Yuting Zhao*1 Longtu Zhang*2 Mamoru Komachi*3

Tokyo Metropolitan University

Neural machine translation (NMT) often suffers in low-resource scenarios where sufficiently large-scale parallel corpora cannot be obtained. Therefore, a recent line of unsupervised NMT models based on monolingual corpus is emerging. In this work, we perform three sets of experiments that analyze the application of unsupervised NMT model in Japanese–Chinese machine translation. We report 30.13 BLEU points for ZH–JA and 23.42 BLEU points for JA–ZH.

1. Introduction

Neural machine translation (NMT) has recently shown impressive results thanks to the availability of large-scale parallel corpora [Bahdanau 14]. NMT models typically fit hundreds of millions of parameters to learn distributed representations which may generalize better when data is redundant. Unfortunately, finding massive amounts of parallel data remains challenging for vast majority of language pairs, especially for low-resource languages, as it may be too costly to manually produce or nonexistent. Conversely, monolingual data is much easier to find, and many languages with limited parallel data still possess significant amounts of monolingual data.

Recently, remarkable results have been shown in training NMT systems relying solely on monolingual data in the source and target languages by using an unsupervised approach [Artetxe 18, Lample 18a]. They proposed unsupervised NMT models that are effective on English–French and English–German. Following their practice, we try to apply unsupervised NMT model to Japanese–Chinese translation.

In this work, we perform experiments from two data domains. They are divided into two types of monolingual corpus and quasi-monolingual corpus. Among them, the best BLEU score can reach 30.13 of ZH–JA and 23.42 of JA–ZH with using ASPEC-JC (Japanese Chinese language pairs) parallel corpus [Nakazawa 16] in the quasi-monolingual setting.

2. System Architecture

The unsupervised NMT model [Lample 18b] we used is composed of two encoder-decoder models for source and target languages and in series with back-translation models. The encoders will encode monolingual sentences into latent representations for respective decoders. One decoder is used as a translator to decode the latent representations, and the other decoder perform the denoising effect of a language model on the target side that refines the latent representation of the source sentence. Then, it jointly train two back-translation models together with the two encoder-decoder language models. In the forward translation, the model generates data which will be trained to the backward translation and in the backward translation, the model trained from the generated target to the source generates translations. The generated sentences from back-translation are added to the regular training set in order to regularize the model.

Contact: Yuting Zhao, zhao-yuting@ed.tmu.ac.jp

3. Experiment

3.1 Datasets

We prepare three data sets from ASPEC-JC (Japanese Chinese language pairs) parallel corpus [Nakazawa 16] and Wikipedia dump *1.

For quasi-monolingual data, the Japanese–Chinese portion of ASPEC-JC was used. Note that although this is a parallel corpus, we shuffled it and used it monolingually. In this paper, we call it ASPEC-Quasi. Official training/development/testing split contains totally 670,000 Chinese and Japanese sentences for training and 2,000+ sentences for evaluating and testing.

For monolingual data, the Japanese–Chinese portion of ASPEC-JC was also used. Note that we shuffled it monolingually and divided the monolingual Chinese and Japanese data into the first half and the second half. Then, they were staggered and combined to form two groups of monolingual data sets with a size of 335,000, and one group was randomly selected for experiment. In this paper, we call it ASPEC-Mono. In addition, we created a Japanese–Chinese monolingual corpus with a training size of 10 million from Wikipedia articles. As above, evaluation and test data are all official data from ASPEC-JC.

3.2 Preprocessing

Firstly, we tokenize Japanese and Chinese datasets separately. We use MeCab *2 with dictionary IPADic for Japanese and Jieba *3 with its default dictionary for Chinese. Secondly, we join the source and target monolingual corpora to learn fastBPE tokens with the vocabulary size of 30,000. Finally, we apply fastText [Bojanowski 17] on the

∗1 https://dumps.wikimedia.org/
∗2 http://taku910.github.io/mecab/
∗3 https://github.com/fxsjy/jieba

Corpora	Amount	JA-ZH	ZH-JA
ASPEC-Mono	335,000	8.9	10.37
Wikipedia	10,000,000	9.74	12.51
ASPEC-Quasi	670,000	**23.42** *(31.19)*	**30.13** *(39.18)*

Table 1: BLEU scores of 3 datasets. (The BLEU score of OpenNMT model is presented in parentheses)

BPE tokens. This way, we obtain cross-lingual BPE embeddings for Chinese and Japanese language pairs to initialize lookup tables. More specifically, we use the skip-gram model with ten negative samples, a context window of 5 words, and 512 dimensions.

3.3 Model

In this work, our models use transformer cells as basic units in the encoders and decoders with PyTorch toolkit version 0.5. We set the number of layers of both the encoders and decoders to 4, and the hidden layers is set to 512. Adam optimizer is used with a learning rate of 0.0001 and a batch size of 25. We set a maximum length of 175 tokens per sentence for each type of dataset and a dropout rate of 0.1. We also set random blank-out rate to 0.1 and word shuffle of 3.

BLEU score is used to evaluate translation in both directions with every iteration, and training will stop when the scores from the last 3 iteration did not improve any more.

4. Results and Discussions

The BLEU scores obtained by all the tested datasets are reported in Table 1.

Amount of data. Firstly, we see the results obtained from the complete monolingual datasets ASPEC-Mono and Wikipedia. As our baseline, the results of ASPEC-Mono obtained 8.9 BLEU points for JA–ZH and 10.37 BLEU points for ZH–JA. As the amount of sentences increases from 335,000 to 10,000,000, the results of Wikipedia obtained 9.74 BLEU points for JA–ZH and 12.51 BLEU points for ZH–JA. Comparing with ASPEC-Mono, scores have gone up in both directions despite of domain difference.

Quasi-monolinguality. Secondly, we see the results in the last row, which is from ASPEC-Quasi corpus. It gets 23.42 BLEU points for JA–ZH and 30.13 BLEU points for ZH–JA. This result exceeds all the previous two results. Moreover, the OpenNMT model using ASPEC-JC parallel corpus reports 31.19 BLEU points for JA–ZH and 39.18 BLEU points for ZH–JA. In contrast, the BLEU score of unsupervised NMT is lower than that of supervised NMT, but the gap is not big.

5. Related Work

From the work of Sennrich et al. [Sennrich 16], they proposed a straightforward approach to create synthetic parallel training data by pairing monolingual training data with an automatic back-translation.

Recently, Artetxe et al. [Artetxe 18] and Lample et al. [Lample 18a] have achieved substantial improvement for fully unsupervised machine translation. They leverage strong language models through training the sequence-to-sequence system as a denoising autoencoder.

6. Conclusion

Based on the above analysis, it can be inferred as follows:

- For monolingual data, the larger the data, the better the translation results.
- For quasi-monolingual data, the effectiveness of unsupervised NMT model on Japanese–Chinese is quite promising, even if it uses smaller training dataset.

From the experiment, we can see unsupervised NMT is effective in Japanese–Chinese machine translation. However, it is worth considering that why there is a huge gap between the results of using monolingual corpus and quasi-monolingual corpus on Japanese–Chinese unsupervised NMT. Even though the amount of monolingual Wikipedia corpus is 15 times more than that of ASPEC-Quasi corpus, the result is much worse. We hope to start from this significant gap and continue to study the factors affecting unsupervised NMT in Japanese–Chinese machine translation.

References

[Artetxe 18] Artetxe, M., Labaka, G., Agirre, E., and Cho, K.: Unsupervised Neural Machine Translation, in *Proceedings of ICLR* (2018)

[Bahdanau 14] Bahdanau, D., Cho, K., and Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate, in *Proceedings of ICLR* (2014)

[Bojanowski 17] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T.: Enriching word vectors with subword information, in *Proceedings of TACL*, pp. 135–146 (2017)

[Lample 18a] Lample, G., Conneau, A., Denoyer, L., and Ranzato, M.: Unsupervised Machine Translation Using Monolingual Corpora Only, in *Proceedings of ICLR* (2018)

[Lample 18b] Lample, G., Ott, M., Conneau, A., Denoyer, L., and Ranzato, M.: Phrase-Based & Neural Unsupervised Machine Translation, in *Proceedings of EMNLP*, pp. 5039–5049 (2018)

[Nakazawa 16] Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S., and Isahara, H.: ASPEC: Asian Scientific Paper Excerpt Corpus, in *Proceedings of LREC*, pp. 2204–2208 (2016)

[Sennrich 16] Sennrich, R., Haddow, B., and Birch, A.: Improving Neural Machine Translation Models with Monolingual Data, in *Proceedings of ACL*, pp. 86–96 (2016)