Filtering Back-Translated Data in Unsupervised Neural Machine Translation

Jyotsana Khatri
Indian Institute of Technology Bombay
jyotsanak@cse.iitb.ac.in

Pushpak Bhattacharyya
Indian Institute of Technology Bombay
pb@cse.iitb.ac.in

Abstract

Unsupervised neural machine translation (NMT) utilizes only monolingual data for training. The quality of back-translated data plays an important role in the performance of NMT systems. In back-translation, all generated pseudo parallel sentence pairs are not of the same quality. Taking inspiration from domain adaptation where in-domain sentences are given more weight in training, in this paper we propose an approach to filter back-translated data as part of the training process of unsupervised NMT. Our approach gives more weight to good pseudo parallel sentence pairs in the back-translation phase. We calculate the weight of each pseudo parallel sentence pair using sentence-wise round-trip BLEU score which is normalized batch-wise. We compare our approach with the current state of the art approaches for unsupervised NMT.

1 Introduction

Back-translation involves generating a set of pseudo parallel sentence pairs using monolingual data of target language and a target to source machine translation model. Back-translation provides the capability to utilize target-side monolingual data for training.

Unsupervised NMT gained a lot of attention in the last two years. Current state of the art approaches for unsupervised NMT even surpasses supervised baseline for English-French language pair (Song et al., 2019). Unsupervised NMT has three main components: cross-lingual embeddings, denoising, and back-translation, where training involves alternating between denoising and back-translation after a good initialization process (Lample and Conneau, 2019; Song et al., 2019). In this paper our focus is on improving finetuning phase, we are introducing a weight component in unsupervised NMT training based on the quality of pseudo parallel sentence pairs generated for training in back-translation phase. These kinds of techniques have been utilized in domain adaptation to give more weight to in-domain sentences (Wang et al., 2017). Pretraining is also a key component of unsupervised NMT, we utilize existing pretraining approaches proposed in Lample and Conneau (2019) and Song et al. (2019).

2 Related Work

Our work majorly involves the exploration of filtering of back-translated data in unsupervised NMT. We briefly describe some related concepts of back-translation, unsupervised NMT and language model pretraining in this section.

Back-translation utilizes target-side monolingual data to create pseudo parallel sentence pairs using a translation system from target to source which is then utilized to train source to target NMT system (Sennrich et al., 2016). Hoang et al. (2018) show that iteratively generating better synthetic data improves the NMT performance.

Quality of back-translated data plays an important role in performance of NMT systems (Fadaee and Monz, 2018; Poncelas et al.,). Fadaee and Monz (2018) show that the target side words which have high prediction loss gets most benefit from the addition of synthetic data. Filtered pseudo parallel data selected with a threshold on round-trip BLEU score helps in improving the performance of NMT systems for low

Proceedings of the 28th International Conference on Computational Linguistics, pages 4334–4339
Barcelona, Spain (Online), December 8-13, 2020

resource languages (Morishita et al., 2018; Imankulova et al., 2019). In Reinforcement learning based approaches, rewards for pseudo parallel sentence pairs based on language model score and round trip reconstruction error helps NMT models (He et al., 2016). Caswell et al. (2019) provides an identification mark for synthetic sentence pairs while training on a mix set of human generated and synthetic sentence pairs . Junczys-Dowmunt (2018) show that filtering the parallel data based on cross-entropy scores which calculates the agreement between both models of both directions from source to target and target to source helps in selection of good pseudo parallel sentence pairs. Dou et al. (2020) show that for iterative back-translation selecting and weighing sentences based on the quality of sentence pairs improves the performance of NMT systems, they use combination of different scores like round-trip BLEU, tf-idf, language model scores etc. to select top sentences, and then on rest of the sentences they use encoder representation similarities and agreement between forward and backward models to provide weights to back-translated data. Wang et al. (2019a) proposed uncertainty-based confidence measures to select good pseudo parallel sentence pairs. Wang et al. (2019b) proposed to select in-domain and clean data based on co-curricular learning.

In Domain adaptation different techniques have been applied to give more weight to in-domain sentences in the training process of NMT, some of these techniques are: providing a weight component in the loss function (Junczys-Dowmunt, 2018), using curriculum learning (Zhang et al., 2019), and dynamically selecting data with iterations (van der Wees et al., 2017).

Pretraining in unsupervised NMT is generally focused on language model training of both encoder and decoder to make them understand the language properties and to provide a good initialization to the finetuning phase. Finetuning utilizes the approaches proposed in Lample et al. (2018) which involves denoising and back-translation. Artetxe et al. (2019) proposed a good initialization mechanism using statistical machine translation for training unsupervised NMT. Wu et al. (2019) proposed a new architecture for unsupervised NMT which does not use back-translation but try to find the best possible translations from the target corpus and edit them to make pseudo parallel sentence pairs. Yang et al. (2018) proposed to utilize two independent encoders with sharing some partial weights. Lample and Conneau (2019) proposed a pretraining mechanism for unsupervised NMT to pretrain encoder and decoder separately using monolingual data. Song et al. (2019) show that training encoder and decoder simultaneously using monolingual data helps in pretraining of unsupervised NMT.

3 Approach

Current state of the art approaches for unsupervised NMT (Lample and Conneau, 2019; Song et al., 2019) do not consider the quality of each generated pseudo parallel sentence pair in the process of training, all generated pseudo parallel sentence pairs have the same weight. There exist different methods to filter bad pseudo parallel sentence pairs with a threshold, which is explained in the previous section. In case of iterative back-translation it is difficult to select a threshold for each batch separately as the training progresses. In the initial iterations of back-translation phase the quality of generated pseudo parallel sentence pairs is very poor. To decrease the weights of bad pseudo parallel sentence pairs we propose to modify back-translation training to include weight of each pseudo parallel sentence pair based on sentence wise round-trip BLEU score which is normalized batch-wise. We perform batch-wise normalization because it helps in maintaining a steady progress in training and also helps maintaining equal weightage of denoising and back-translation. Round-trip BLEU score is the BLEU score between the source sentence and the translation of source sentence to target then back to source language. As the systems in both directions (source to target and target to source) are trained simultaneously we can calculate the round-trip BLEU score based on the current trained systems in both directions. The sentence wise round-trip BLEU score is added in the cross-entropy loss function as weight of each pseudo parallel sentence pair. In general, the cross-entropy loss function is given by:

$$-\sum_{b=1}^{BS}\sum_{l=1}^{|L|}\sum_{n=1}^{|N|} y_{l,n} log(\hat{y}_{l,n}) \tag{1}$$

where BS is batch-size, $\mid N \mid$ is length of the vocabulary, $\mid L \mid$ is length of the sentence, $\hat{y}_{l,n}$ is predicted probability of word n from vocabulary on word l in sentence, and $y_{l,n}$ is 1 when l from vocab is correct word otherwise 0. We utilize the weighted cross-entropy loss function:

$$-\sum_{b=1}^{BS} w_b * (\sum_{l=1}^{|L|} \sum_{n=1}^{|N|} y_{l,n} log(\hat{y}_{l,n})) \tag{2}$$

w_b is the batch-wise normalized round-trip BLEU score between source and round-trip translation of source. The normalization function is given by:

$$w_b = \frac{w_{un_b}}{\sum_{b=1}^{BS} w_{un_b}} \tag{3}$$

where w_b is normalized round trip bleu score for bth sentence in the batch. w_{un_b} is un-normalized round trip BLEU score for bth sentence. BLEU score is bilingual evaluation understudy which is commonly utilized to evaluate machine translation systems (Papineni et al., 2002). There exist various other methods to evaluate machine translation systems but for start we are considering sentence-wise BLEU score which is the most popular one.

We provide results for the same approaches shown in Song et al. (2019), Lample and Conneau (2019) with and without filtering in the back-translation phase. In Song et al. (2019) finetuning is only done using iterative back-translation and in Lample and Conneau (2019) finetuning is done using denoising and back-translation similar to (Lample et al., 2018).

4 Experiment and Results

In this section we show the impact of inclusion of filtering of back-translated data using above approach with two state of the art unsupervised NMT benchmarks for three language pairs.

4.1 Data

We utilize the same BPE codes and vocab as utilized in (Lample and Conneau, 2019) and (Song et al., 2019) for en-fr, en-fe and en-ro. We utilize the same data with mentioned number of sentences in table 1 for our experiments. All this data is from WMT[1]. We perform all pre-processing (normalization, tokenization and byte pair encoding) similar to Song et al. (2019). The validation data is newstest2013 and test data is newstest2014 of WMT.

Lang-pair	# sentences in source language	# sentences in target language	Dataset
en-fr	5M	5M	WMT
en-de	5M	5M	WMT
en-ro	5M	2.28M	WMT

Table 1: Dataset details

4.2 Model configuration

We utilize the pretrained models from (Lample and Conneau, 2019) and (Song et al., 2019) for language model pretraining. For XLM we utilize masked language model pretraining [2]. We utilize transformer architecture with 6 layers, 8 heads, 1024 hidden units, GELU activation units, attention drop-out of 0.1, learning rate starts from 10^{-4} and batch size of 32 sentences. We perform decoding using beam search. BPE codes are learnt using FastBPE[3] using 60000 BPE codes over the combined data of both languages. The epoch size is set to 200000 sentences. We use adam (Kingma and Ba, 2002) optimizer. We perform tokenization using moses(Koehn et al., 2007). For calculating sentence-wise BLEU scores

[1] http://www.statmt.org/wmt16/translation-task.html
[2] https://github.com/facebookresearch/XLM
[3] https://github.com/glample/fastBPE

using tensors we utilize allennlp[4] (Gardner et al., 2018) toolkit. We choose best model from different iterations according to BLEU score on validation set. We evaluate our results using tokenized BLEU scores calculated using multi-bleu.pl[5].

4.3 Results

We utilize MASS[6] (Song et al., 2019) as our base implementation and update the back-translation phase to include weight of the pseudo parallel sentence pairs in the loss function.

Method	en-fr	fr-en	en-ro	ro-en	en-de	de-en
Song et al., 2019	26.59	25.42	25.53	24.8	17.62	24.78
Song et al., 2019 + Filtering	26.37	25.5	25.29	24.64	17.51	24.87
Lample and Conneau, 2019	27.95	27.02	26.26	25.8	18.26	25.22
Lample and Conneau, 2019 + Filtering	28.4*	27.69*	26.96*	26.24*	17.35	25.91*

Table 2: Results for filtering of back-translated data in unsupervised NMT (* indicates statistically significant improvement)

It is clear from Table 2 that our approach to filter back-translated data (providing weight as per pseudo sentence-pair quality) gives better results for Lample and Conneau (2019) approach. We also performed an experiment to examine the impact of denoising in the finetuning phase with filtering of back-translated data for en-fr language pair using masked sequence to sequence pretraining, which gave BLEU score of 27.02 for en-fr and 26.01 for fr-en, which is an improvement over the baseline. We perform paired bootstrap re-sampling (Koehn, 2004) for a p-value less than 0.05 for statistical significance test.

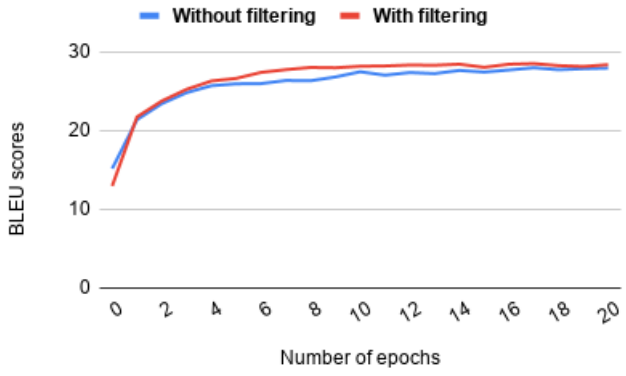

Figure 1: BLEU scores for each epoch for en-fr

Figure 1 represents the BLEU scores of test data for each epoch while training for Lample and Conneau (2019) with and without filtering. In initial iterations model with filtering is not performing good but as training progresses it starts performing better than model with no filtering. This happens because we start the filtering process from the beginning of the finetuning phase when the quality of generated back-translated data is poorer than later iterations. We also observe that the model with filtering tend to converge a little earlier than the model without filtering. As we are giving less weights to poor pseudo parallel sentence pairs, it makes the system learn more from good data which helps in improving the performance of unsupervised NMT.

[4] https://github.com/allenai/allennlp

[5] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

[6] https://github.com/microsoft/MASS

5 Conclusion

In this paper, we show that giving weights to pseudo parallel sentence pairs based on its quality calculated using round trip BLEU score in the back-translation phase helps in improving the performance of unsupervised NMT. In future work, we plan to explore different weighing scores to evaluate quality of back-translated data together with different measures of the quality of individual sentences to improve translation performance and training time.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July. Association for Computational Linguistics.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63.

Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. 2020. Dynamic data selection and weighting for iterative back-translation. *arXiv preprint arXiv:2004.03672*.

Marzieh Fadaee and Christof Monz. 2018. Back-translation sampling by targeting difficult words in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 436–446.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in neural information processing systems*, pages 820–828.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24.

Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. 2019. Filtered pseudo-parallel corpus improves low-resource neural machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(2):1–16.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, Belgium, Brussels, October. Association for Computational Linguistics.

D. P. Kingma and J. Adam Ba. 2002. A method for stochastic optimization. In *In Proceedings of ICLR 2002*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, et al. 2018. Unsupervised machine translation using monolingual corpora only. April.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2018. Ntt's neural machine translation systems for wmt 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 461–466.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

A Poncelas, D Shterionov, A Way, GM de Buy Wenniger, and P Passban. Investigating backtranslation in neural machine translation. arxiv 2018. *arXiv preprint arXiv:1804.06189*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1410.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488.

Shuo Wang, Yang Liu, Chao Wang, Huanbo Luan, and Maosong Sun. 2019a. Improving back-translation with uncertainty-based confidence estimation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 791–802.

Wei Wang, Isaac Caswell, and Ciprian Chelba. 2019b. Dynamically composing domain-data selection with clean-data selection by "co-curricular learning" for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1282–1292.

Jiawei Wu, Xin Wang, and William Yang Wang. 2019. Extract and edit: An alternative to back-translation for unsupervised neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1173–1183.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Unsupervised neural machine translation with weight sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–55.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019. Curriculum learning for domain adaptation in neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1903–1915.