Deep Blue Sonics’ Submission to IWSLT 2020 Open Domain Translation Task

Enmin Su
Deep Blue Sonics
enmin.su@pku.edu.cn

Yi Ren
Deep Blue Sonics
yi.ren@dbsonics.net

Abstract

We present in this report our submission to IWSLT 2020 Open Domain Translation Task(Ansari et al., 2020). We built a data preprocessing pipeline to efficiently handle large noisy web-crawled corpora, which boosts the BLEU score of a widely used transformer model in this translation task. To tackle the open-domain nature of this task, back-translation (Sennrich et al., 2016) is applied to further improve the translation performance.

1 Introduction

Neural machine translation (NMT) is a well-studied problem boosted in recent years by powerful transformer models (Vaswani et al., 2017), which find their ways to various sequence-to-sequence tasks, such as automatic speech recognition (ASR) (Dong et al., 2018), speech translation (Gangi et al., 2019), text-to-speech (Shin et al., 2019), to name a few. Nevertheless, many challenges remain in training an efficient transformer NMT model in practice, such as the following ones raised in IWSLT 2020 Open Domain Translation Task:

Handling noisy dataset Hassan et al. pointed out that machine translation models are vulnerable to noise even in small quantity. In practice, manual correction of a massive corpora is prohibitive, thus calling for an automatic data cleaning pipeline.

Leveraging monolingual data Compared to parallel corpora, monolingual data can be acquired at a much lower cost. Common ways of using monolingual data include language modeling (Çaglar Gülçehre et al., 2015), back-translation (Sennrich et al., 2016), and dual learning (He et al., 2016), all exhibiting promising results; furthermore, they could be adopted in a complementary way when carefully designed (Hassan et al., 2018). Last but not least, massive pre-trained language models like BERT perform strongly in NLP tasks like question answering, reading comprehension and text classification (Devlin et al., 2019), motivating our attempts to incorporate them in our NMT system.

Domain mismatch NMT systems trained with data from specific domains may translate poorly in other domains (Freitag and Al-Onaizan, 2016). Training the model with all available corpora, and fine-tuning it on a specific domain generally achieves best results in this domain (Chu et al., 2017). In open domain cases, it’s impractical to keep a dedicated model or to obtain enough training data for every single domain. Hence Multi-Domain NMT, where a single model generalizes to multiple domains, is gaining interest in recent research. For example, Tars and Fishel; Jiang et al.; Zeng et al. injected domain information into model input, leading to convincing and consistent improvements, in which domain information may be derived in both supervised and unsupervised manners.

Our work consists of establishing an efficient data pre-processing pipeline for large web-crawled corpora to train a transformer model for NMT and exploiting large amount of monolingual data with back-translation and language modeling.

This report is organized as follows: Section 2 depicts the different techniques applied to improve the official baseline model, whereas in Section 3 the experiments and results are described in greater details. Finally, we conclude our work and suggest a few future work directions in Section 4.

Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT), pages 140–144
July 9-10, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

2 System Overview

2.1 Noisy Data

According to Hassan et al., the common noises in web-crawled corpora can be categorized into the following groups:

- mis-aligned pairs,
- partially translated pairs,
- inaccurate or low-quality pairs,
- pairs in wrong languages, or as exact duplicates.

Meanwhile, in the context of training Transformer models with large-scale parallel data, Popel and Bojar found out while clean and smaller datasets help the model to converge faster, noisy and larger datasets help in converging to a better result. Our experiments indicate that with a pre-processing pipeline, training larger datasets is of great help in improving translation BLEU score.

2.2 NMT model

Our NMT model is identical to the baseline of IWSLT 2020 Open Domain Translation Task (Ansari et al., 2020), which is a common transformer architecture. The hyper-parameters of the model are listed in Table 1.

Hyper-parameters	
encoder layers	6
decoder layers	6
filter width	4096
attention width	1024
attention heads	16
token type	BPE
source vocabulary size	30k
target vocabulary size	30k
Total Parameters	270M

Table 1: Hyper-parameters for our NMT model.

2.3 Language Modeling

Çaglar Gülçehre et al. proposed language modeling as a way of leveraging monolingual corpora in the context of NMT. Given massive monolingual data, language modeling helps in decoding accuracy, thus ensuring improvements in iterative back-translation training. Among various ways of incorporating language models in an NMT system, we conduct experiments on shallow fusion and deep fusion, following the settings of Çaglar Gülçehre et al..

A rescoring method put forward by Shin et al. is also tested, where the translation candidates from beam search are reranked using a weighed combination of original scores and scores calculated by a pre-trained Japanese BERT model (Takeshi et al., 2019).

2.4 Back-translation

Back-translation (Hoang et al., 2018) has been proven to be an effective and highly applicable way to achieve consistent improvements by increasing both size and diversity of the training corpora (Edunov et al., 2018); we follow their back-translation setting in our experiments.

3 Experiments and Results

3.1 Data Acquisition and Pre-processing

All the datasets used in our experiments are listed in Table 2. While larger datasets boost model performance in general, we observe considerable amount of noises in all the datasets in Table 2 apart from the ”clean parallel” set. As mentioned in Section 2.1, these noises are of various nature, and show negative impact in our primary experiments. To deal with them, several pre-processing steps have been applied as follows.

First, the noisy datasets turn out to contain a lot of rare or meaningless characters. In order to remove them, we define a valid Unicode range, consisting of basic Latin, Greek alphabet, Japanese alphabet and CJK symbols and punctuations. Then we discard sentence pairs including more than 20% of invalid characters, and delete the invalid symbols in the remaining pairs.

Second, we normalize these sentences with neologdn[1] to handle encoding issues and special punctuations.

Third, a naive de-duplicate algorithm is applied to get rid of redundancy in training data, which also eliminates invalid text containing only error messages.

Finally, the sentences in wrong languages in the datasets are filtered by a pre-trained Fasttext language classification model (Joulin et al., 2017), where sentences with wrong language labels or low confidence are removed.

The use of pre-processed noisy data results later in a notable increase of BLEU score (see Table 4).

[1] https://github.com/ikegami-yukino/neologdn

Dataset	**Size**	**Source**
training set		
clean parallel	2M	`existing_parallel`
noisy parallel	17M	pre-processed `web_crawled_parallel_filtered`
monolingual(Ja)	10M	`unaligned_documents`
monolingual(Zh)	10M	Large Scale Chinese Corpus for NLP (Xu, 2019)
validation set		
basic expressions	5304	JEC Basic Sentence Data (Kurohashi-Kawahara Lab.)

Table 2: Datasets used in our experiments. The size is in number of sentence pairs for parallel datasets, and number of sentences for monolingual ones.

3.2 Baseline Model

We train the transformer model in Section 2.2 on clean data as baseline. We use Jieba[2] and Mecab[3] to tokenize the Chinese and Japanese text respectively, and use subword-nmt[4] to perform BPE encoding/decoding (Gage, 1994), with vocabulary size approximately to 30k for each language. We use Tensor2Tensor (Vaswani et al., 2018) implementation of Transformer, with 4 GPU and accumulates gradient for 4 steps, resulting in an equivalent batch-size of 32768.

3.3 Language Modeling

Here we attempt to acquire some improvements utilizing unpaired data by means of language models (LM). The methods tested are:

- shallow fusion with language model (Çaglar Gülçehre et al., 2015)
- deep fusion with language model(Çaglar Gülçehre et al., 2015)
- BERT rescoring (Shin et al., 2019)

As summarized in Table 3, none of the LM-based methods leads to gain in BLEU score just yet, and further research needs to be conducted to beat the baseline with language models.

Methods	**Zh2Ja**
baseline model	27.48
shallow fusion	26.79
deep fusion	21.84
BERT rescoring	24.80

Table 3: BLEU scores after incorporating with language models.

3.4 Back-translation

To generate a back-translation dataset, we first augment clean target sentences using the exact 'beam + noise' setting in (Edunov et al., 2018), with $p(deletion) = 0.1$, $p(substitution) = 0.1$ for each token in the sentence; for substitution, we randomly pick the i^{th} token and draw a random number n from uniform distribution of {-3, 2, -1, 1, 2, 3}, and replace this token with the $(i+n)^{th}$ token. We generate noisy source sentences using a target-to-source NMT model trained from previous steps, and construct a dataset using noisy source sentences with their clean target counterparts. During training, parallel data and back-translated data are sampled at 1:1 ratio.

3.5 Final Results

As is shown in previous sections, using large normalized corpora and back-translation both improve the baseline system in two translation directions. The overall result on validation set is depicted in Table 4. The final result on test dataset is depicted in (Ansari et al., 2020).

To further confirm the effectiveness of our back-translation approach acorss different domains, we classify the validation set into 14 different topics using a validated pre-trained bag-of-words model, and compute the validation BLEU scores of each topic before and after back-translation. In Figure 1, an overall improvement is observed in all categories with a few exceptions, which is expected.

Methods	**Ja2Zh**	**Zh2Ja**
official baseline	20.28	26.57
clean parallel	20.37	27.48
+ noisy parallel	25.48	30.32
+ back-translation	27.79	35.87

Table 4: Overall BLEU Scores on Validation Set.

[2] https://github.com/fxsjy/jieba

[3] https://taku910.github.io/mecab/

[4] https://github.com/rsennrich/subword-nmt

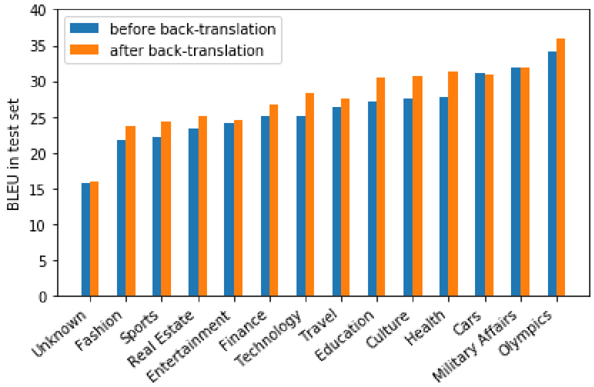

Figure 1: BLEU score in different domains in the validation set

4 Conclusion and Future Work

In this paper we described our submission to IWSLT 2020 open domain translation task. We improved the baseline model with a large amount of cleaned-up web-crawled data and the back-translation technique. Our final system achieved 27.79 and 35.87 BLEU scores on Ja2Zh and Zh2Ja tasks respectively, out running the official baseline by about 35%.

For future work, we first plan to improve the baseline model architecture, since it is left unchanged in our current experiments (e.g. by following (Sun et al., 2019)). Furthermore, loss masking (Rusiecki, 2019) would also be appealing, which ignores the samples of highest losses in each batch during training. Proven to be effective for noisy-label classification, loss masking may also be helpful to our NMT model trained with noisy sentence pairs. Another possibility is to filter noisy data with a learned representation in both languages (Hassan et al., 2018), which can further eliminate incomplete or mismatched translation pairs and help with model accuracy.

Initializing NMT decoder with a pre-trained BERT model is also stated to be useful; this technique is named ’cold fusion’ in the context of ASR(Sriram et al., 2017), and we expect to see similar effects in the case of NMT. An alternative way of incorporating pre-trained BERT into NMT models is to merge hidden activations of these models together(Zhu et al., 2020). The results show that such a fusion is an effective way to utilize monolingual data as complementary to back-translation.

Finally, to tackle the multi-domain translation scenario, specific loss functions and model structures exhibit promising results (Zeng et al., 2018; Jiang et al., 2019); meanwhile, adding special domain tokens to source text may also achieve comparable results (Tars and Fishel, 2018).

References

Ebrahim Ansari, Amittai Axelroad, Nguyen Bach, Ondrej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, and Changhan Wang. 2020. Findings of the IWSLT 2020 Evaluation Campaign. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT 2020)*, Seattle, USA.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of simple domain adaptation methods for neural machine translation. *ArXiv*, abs/1701.03214.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *EMNLP*.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *ArXiv*, abs/1612.06897.

Philip J. Gage. 1994. A new algorithm for data compression. *The C Users Journal archive*, 12:23–38.

Mattia Antonino Di Gangi, Matteo Negri, Roldano Cattoni, Roberto Dessì, and Marco Turchi. 2019. Enhancing transformer for end-to-end speech-to-text translation. In *MTSummit*.

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *ArXiv*, abs/1503.03535.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mengnan Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *ArXiv*, abs/1803.05567.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *NIPS*.

Cong Duy Vu Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *NMT@ACL*.

Haoming Jiang, Chen Liang, Chonggang Wang, and Tuo Zhao. 2019. Multi-domain neural machine translation with word-level adaptive layer-wise domain mixing. *ArXiv*, abs/1911.02692.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Martin Popel and Ondrej Bojar. 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110:43 – 70.

Andrzej Rusiecki. 2019. Trimmed robust loss function for training deep neural networks with label noise. In *ICAISC*.

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, Berlin, Germany. Association for Computational Linguistics.

Joongbo Shin, Yoonhyung Lee, and Kyomin Jung. 2019. Effective sentence scoring method using bidirectional language model for speech recognition. *ArXiv*, abs/1905.06655.

Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates. 2017. Cold fusion: Training seq2seq models together with language models. In *INTERSPEECH*.

Meng Sun, Bojian Jiang, Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Baidu neural machine translation systems for WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 374–381, Florence, Italy. Association for Computational Linguistics.

Sakaki Takeshi, Sakae Mizuki, and Naoyuki Gunji. 2019. Bert pre-trained model trained on large-scale japanese social media corpus.

Sander Tars and Mark Fishel. 2018. Multi-domain neural machine translation. *ArXiv*, abs/1805.02282.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation. *CoRR*, abs/1803.07416.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Bright Xu. 2019. Nlp chinese corpus: Large scale chinese corpus for nlp.

Jiali Zeng, Jinsong Su, Huating Wen, Yang Liu, Jun Xie, Yongjing Yin, and Jianqiang Zhao. 2018. Multi-domain neural machine translation with word-level domain context discrimination. In *EMNLP*.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating bert into neural machine translation. *ArXiv*, abs/2002.06823.