Small Model and In-Domain Data are All You Need

Hui Zeng
Independent Researcher
felix_zeng_ai@aliyun.com

Abstract

I participated in the WMT shared news translation task and focus on one high resource language pair: English and Chinese (two directions, Chinese to English and English to Chinese). The submitted systems (ZengHuiMT) focus on data cleaning, data selection, back translation and model ensemble. The techniques I used for data filtering and selection include filtering by rules, language model and word alignment. I used a base translation model trained on initial corpus to obtain the target versions of the WMT21 test sets, then I used language models to find out the monolingual data that is most similar to the target version of test set, such monolingual data was then used to do back translation. On the test set, my best submitted systems achieve 35.9 and 32.2 BLEU for English to Chinese and Chinese to English directions respectively, which are quite high for a small model.

1 Introduction

I participated in the WMT shared news translation task and focus on the English and Chinese language pair. This language pair is challenging due to the plentiful in-domain bitext training data and abundant monolingual data. High resource means fierce competition, many high-tech companies and universities chose this language pair also. My neural machine translation system is developed using base transformer (Vaswani et al., 2017) architecture and the toolkit I used is THUMT (Zhang et al., 2020). Rules and word aligning model are used to clean parallel data. Language model is used to clean monolingual data. I use a base transformer (Vaswani et al., 2017) architecture since I have only one GPU. The following techniques are used on model training: a. Increase the number of encoder layers to 12 to further improve the encoder's representation capability; b. Back translation (Sennrich et al., 2016) are applied to fully utilize the monolingual corpus. c. Shared vocabulary is used for better performance. d. Four different models using diversified data are trained for ensemble decoding.

2 Data Filtering and Selection

The parallel data is mainly from CCMT Corpus[1], and the monolingual data is collected from the internet. I did not use any other datasets since I think they are not highly related to this news translation task. To evaluate my model's performance, I merged the test set from WMT2017 to WMT2020 to build a big development set.

2.1 Monolingual Data Filtering Using Language Model

In terms of monolingual data, I collected more than 20 million Chinese sentences and more than 15 million English sentences from various websites.

The Chinese text are collected from the following websites:

http://www.chinanews.com/
https://cn.reuters.com/
http://news.ifeng.com/
http://people.com.cn/
https://www.sina.com.cn/
http://www.xinhuanet.com/
https://news.cctv.com/
https://www.qq.com/

[1] http://mteval.cipsc.org.cn:81/agreement/description

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 255–259
November 10–11, 2021. ©2021 Association for Computational Linguistics

https://www.sohu.com/

The English text are collected from the following websites:

https://www.bbc.com/news
www.theguardian.com
https://www.telegraph.co.uk/
https://www.washingtonpost.com/
https://www.latimes.com/
https://www.smh.com.au/
https://www.brisbanetimes.com.au/
https://www.dailymail.co.uk/
https://www.rt.com/
https://time.com/
https://www.euronews.com/
https://www.msnbc.com/
https://www.cbsnews.com/
https://www.nytimes.com/
https://www.newsweek.com/
https://www.foxnews.com/
https://www.standard.co.uk/

The following rules are used for a simple cleaning:
•Remove duplicated sentences.
•Remove the sentences containing special characters.
•Remove the sentences containing html addresses or tags.

Afterwards, language models are used to filter the monolingual data. For English sentences, lm-scorer[2] is used to calculate a score for each sentence, which is the mean of tokens' probabilities. The pre-trained model used for English is GPT-2 (Radford et al., 2019).[3] For Chinese sentences, a pre-trained Chinese GPT-2 (Radford et al., 2019)[4] model is used to calculate a score for each sentence. Then, the English and Chinese sentences are filtered by their scores.

GPT-2 (Radford et al., 2019) is a large transformer-based language model with 1.5 billion parameters, trained on a dataset of 8 million web pages. GPT-2 (Radford et al., 2019) is trained with a simple objective: predict the next word, given all of the previous words within some text.

The threshold I used is determined based on my personal evaluation on the text. After calculating the scores for all the sentences, I sampled the sentences by their scores and perform a language quality check. I started from the extremely low scores and the extremely high scores, and then gradually move the scale from the two ends to the middle until I find that the language quality is up to my standard.

There are about 16 million Chinese sentences and 10 million English sentences left after filtering using language model.

2.2 Parallel Data Filtering Using Rules

For CCMT parallel Corpus and synthetic parallel corpus from back translation, I used the following rules to filter data.

a. Remove duplicated sentence pairs.
b. Remove the lines having identical source and target sentences.
c. Remove the sentence pairs containing special characters.
d. Remove the sentence pairs containing html addresses or tags.
e. Remove the sentence pairs with empty source or target side.

2.3 Parallel Data Filtering Using Word Alignment

In order to get word alignment results, fast_align (Dyer et al., 2013) is used on the CCMT Corpus filtered by rules, then extract-lex[5] is used to generate bilingual phrase tables. The phrase tables are then pruned according to probabilities. Afterwards, I use the pruned phrase table to measure the confidence of the sentence pairs being mutual translations. The confidence score is calculated like this: check each token of the target sentence to find if it has a counterpart in the source side, then perform this operation in the reverse direction, the final confidence score is calculated by summing up the two percentages from two respective directions and then getting the average.

Then the confidence score is used to remove bad sentence pairs. The sentence pairs with confidence scores below 0.6 are discarded. In this way, I finally got a high quality parallel CCMT Corpus.

3 System Description

This section illustrate how I train the model step by step.

[2] https://github.com/simonepri/lm-scorer
[3] https://openai.com/blog/better-language-models/
[4] https://huggingface.co/uer/gpt2-chinese-cluecorpussmall
[5] https://github.com/marian-nmt/extract-lex

3.1 Data pre-processing

For data preprocessing, I use the tokenizer developed on my own to process both Chinese and English. Chinese text (including punctuations and numbers) is split to single character level. I keep the upper and lower case letters of English as they are, since I believe they are also important features for the model. Numbers in English text are also split into single digits. I use byte pair encoding (BPE) (Sennrich et al., 2016) to create a shared vocabulary, so that the vocabulary size is reduced to 45467. I also wrote a post-processor to restore the Chinese and English text to normal form.

3.2 Normal Model Training

To evaluate my model's performance, I merged the test set from WMT2017 to WMT2020 into a big development set. First, I use the CCMT parallel Corpus filtered by rules and word aligning model to train base transformer (Vaswani et al., 2017) English to Chinese and Chinese to English translation models. Two sets of training parameters were used with only one difference: the number of encoder layers. The detailed parameters are as follows:

batch_size=15000,
max_length=384,
hidden_size=512,
filter_size=2048,
num_heads=8,
num_encoder_layers=6 or 12,
num_decoder_layers=6
max_relative_dis=16,
layer_preprocess="layer_norm",
eval_steps=2000,
warmup_steps=4000

Validation is performed every 2000 steps, the training is terminated if there is no gain in BLEU for 20 consecutive validations.

As shown in Table 1, using the same filtered CCMT Corpus, the BLEU scores of models with deeper encoder (12-layer-encoder, 6-layer-decoder) are slightly higher than that of the base version.

Back translation (Sennrich et al., 2016) is a useful data augmentation technique to boost model performance with target side monolingual data. The technique starts from training a target to source translation model using initial bilingual corpus, which is later used to translate the monolingual data in the target language back to source language. Then the synthetic back-translated corpus is concatenated with the original bilingual corpus to train the source to target translation model. After the source to target model is enhanced, the same method can be applied

Model + Corpus	BLEU EN2 ZH	BLEU ZH to EN
filtered CCMT Corpus base transformer 6-layer-encoder, 6-layer-decoder, base transformer	32.7	21.0
filtered CCMT Corpus base transformer 12-layer-encoder, 6-layer-decoder, base transformer	32.9	21.1
filtered CCMT Corpus half of the filtered monolingual data multiple rounds of back translations 12-layer-encoder, 6-layer-decoder, base transformer	35.3	24.5
filtered CCMT Corpus in-domain monolingual data extracted using test set multiple rounds of back translations 12-layer-encoder, 6-layer-decoder, base transformer best single model	38.3	28.0
filtered CCMT Corpus in-domain monolingual data extracted using test set multiple rounds of back translations 12-layer-encoder, 6-layer-decoder, base transformer ensemble of four models	39.5	29.1

Table 1: Different models and their BLEU scores
I merged the test set from WMT2017 to WMT2020 into a big development set.
The BLEU scores are calculated on this big development set.

again to train the back-translation system in the reversed direction.

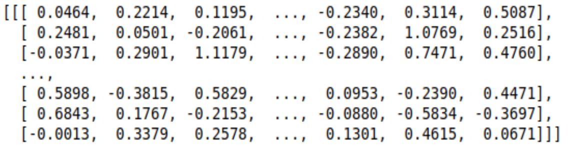
I repeat this process using **half** of the filtered monolingual data for several iterations until the BLEU is not increasing.

3.3 Training on In-domain Data

BERT (Devlin et al., 2019) is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. Before feeding word sequences into BERT (Devlin et al., 2019), 15% of the words in each sequence are replaced with a [MASK] token. The model then attempts to predict the original value of the masked words, based on the context provided by the other, non-masked, words in the sequence.

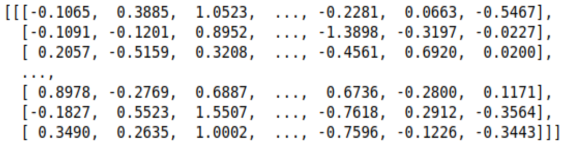
After the WMT2021 test set was released, I first translated the Chinese and English test sentences to target versions using the above models, then I generated feature representations for the target versions of the test sentences using pre-trained English BERT (Devlin et al., 2019)[6] and Chinese BERT (Devlin et al., 2019)[7] models.

The example representations are shown in Figure 1 and Figure 2.



```
[[[ 0.0464,  0.2214,  0.1195,  ..., -0.2340,  0.3114,  0.5087],
  [ 0.2481,  0.0501, -0.2061,  ..., -0.2382,  1.0769,  0.2516],
  [-0.0371,  0.2901,  1.1179,  ..., -0.2890,  0.7471,  0.4760],
  ...,
  [ 0.5898, -0.3815,  0.5829,  ...,  0.0953, -0.2390,  0.4471],
  [ 0.6843,  0.1767, -0.2153,  ..., -0.0880, -0.5834, -0.3697],
  [-0.0013,  0.3379,  0.2578,  ...,  0.1301,  0.4615,  0.0671]]]
```

Figure 1: The BERT representation of "I like this competition very much", the tensor shape is [1, 9 , 768]



```
[[[-0.1065,  0.3885,  1.0523,  ..., -0.2281,  0.0663, -0.5467],
  [-0.1091, -0.1201,  0.8952,  ..., -1.3898, -0.3197, -0.0227],
  [ 0.2057, -0.5159,  0.3208,  ..., -0.4561,  0.6920,  0.0200],
  ...,
  [ 0.8978, -0.2769,  0.6887,  ...,  0.6736, -0.2800,  0.1171],
  [-0.1827,  0.5523,  1.5507,  ..., -0.7618,  0.2912, -0.3564],
  [ 0.3490,  0.2635,  1.0002,  ..., -0.7596, -0.1226, -0.3443]]]
```

Figure 2: The BERT representation of "我非常喜欢这个竞赛。", the tensor shape is [1, 12 , 768]

I also generated feature representation for each sentence in the **other half** of the filtered monolingual data. These features are then used to calculate the cosine similarity scores between the target versions of test sentences generated by the previous trained models and the monolingual sentences that are not used in previous training.

Then, the similarity scores are used to find out monolingual sentences that are most similar to the WMT2021 test set.

For each test set sentence, hundreds of monolingual sentences are extracted. In order to determine a threshold score, I randomly sampled 100 test set sentences and their extracted counterparts. Then I checked their similarities and scores using my personal linguistic competences in these two languages. The determined threshold score was then used to automatically extract in-domain data.

Finally, I extracted around 550 thousand Chinese sentences and 420 thousand English sentences as in-domain monolingual data. These sentences are then divided into four equal portions. On the basis of the best models using back translation and the first half of monolingual data, I use four portions of in-domain English data and four portions of in-domain Chinese data to do back translation until the BLEU stops increasing. Therefore, I get four in-domain English to Chinese and four in-domain Chinese to English translation models. These models are then ensembled to build two most powerful models for each direction.

3.4 Results

The BLEU scores on the aforesaid big development set (I merged the test set from WMT2017 to WMT2020 to build a big development set) for each corpus plus model combination are shown in Table 1.

On the WMT 2021 test set, my best submitted systems achieve 35.9 and 32.2 BLEU for English to Chinese and Chinese to English directions respectively, which are even higher than most of the systems from famous high-tech companies.

4 Conclusion

This paper describes Hui Zeng's translation systems (ZengHuiMT) for the WMT2021 news translation shared task. The potential of small model plus in-domain data is explored. I am pleased to argue that, with high quality in-domain data, small model could achieve BLEU scores comparable to that of huge models.

[6] https://huggingface.co/distilbert-base-uncased

[7] https://huggingface.co/bert-base-chinese

Acknowledgments

Thanks to my wife who spend most of her time to take care of our two kids, so that I am able to participate in the contest and complete this paper.

References

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *In Proceedings of the 31st International Conference on Neural Information Processing Systems,* pages 6000–6010.

Jiacheng Zhang, Yanzhuo Ding, Shiqi Shen, Yong Cheng, Maosong Sun, Huanbo Luan, Yang Liu. 2020. THUMT: An Open Source Toolkit for Neural Machine Translation. *In Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track),* pages 116–122. Association for Machine Translation in the Americas.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. (2013). A Simple, Fast, and Effective Reparameterization of IBM Model 2. *In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *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 4171–4186. Association for Computational Linguistics.

Radford, Alec and Wu, Jeff and Child, Rewon and Luan, David and Amodei, Dario and Sutskever, Ilya. 2019. Language Models are Unsupervised Multitask Learners.