An Empirical Study of Machine Translation for the Shared Task of WMT18

Chao Bei, Hao Zong, Yiming Wang, Baoyong Fan, Shiqi Li, Conghu Yuan
Global Tone Communication Technology Co., Ltd.
{beichao,zonghao, wangyiming, fanbaoyong, lishqi, yuanconghu}@gtcom.com.cn

Abstract

This paper describes the Global Tone Communication Co., Ltd.'s submission of the WMT18 shared news translation task. We participated in the English-to-Chinese direction and get the best BLEU (43.8) scores among all the participants. The submitted system focus on data clearing and techniques to build a competitive model for this task. Unlike other participants, the submitted system are mainly relied on the data filtering to obtain the best BLEU score. We do data filtering not only for provided sentences but also for the back translated sentences. The techniques we apply for data filtering include filtering by rules, language models and translation models. We also conduct several experiments to validate the effectiveness of training techniques. According to our experiments, the Annealing Adam optimizing function and ensemble decoding are the most effective techniques for the model training.

1 Introduction

We participated in the WMT shared news translation task and focus on the English-to-Chinese direction. Our neural machine translation system is developed as transformer (Vaswani et al., 2017a) architecture and the toolkit we used is Marian (Junczys-Dowmunt et al., 2018). Since BLEU (Papineni et al., 2002) is the main ranking index for all submitted systems, we apply BLEU as the evaluation matrix for our translation system. We aim to verify whether the techniques we applied in the Encoder Decoder architecture of recurrent neural network(RNN) and attention mechanism (Bahdanau et al., 2014) are also positive for transformer architecture (Vaswani et al., 2017b) and the effectiveness of the data filtering.

For data preprocessing, the basic methods include Chinese word segmentation, tokenization, byte pair encoding(BPE) (Sennrich et al., 2015b). Besides, human rules and translation model are also involved for cleaning parallel data, as well as using language model for cleaning monolingual data. As to the techniques on model training, Annealing Adam (Denkowski and Neubig, 2017), back-translation (Sennrich et al., 2015a) and right-to-left reranking (Sennrich et al., 2016) which have proven to be effective in the Encoder Decoder model with RNN layer and attention mechanism are applied to verify whether these techniques in transformer architecture are also effective.

When comparing our baseline model, we show the increase in 5.57 BLEU scores of English to Chinese direction for news. And comparing the best score in last year, transformer architecture is more powerful than RNN with attention mechanism with 3.65 BLEU score improvement. However, not all the techniques we applied to RNN with attention mechanism are equally effective against transformer architecture, especially reranking by right-to-left model.

This paper is arranged as follows. We firstly describe the task and provided data information, then introduce the method of data filtering, including rules, language model and translation model. After that, we describe the techniques on transformer architecture and show the conducted experiments in detail, including data preprocessing, postprocessing and model architecture. At last, we analyse the results of experiments and draw the conclusion.

2 Task Description

The task focuses on bilingual text translation in news domain and the provided data is show in Table 1, including parallel data and monolingual data. The parallel data is mainly from News Commentary v13 (Tiedemann, 2012), UN Parallel Corpus V1.0 (Ziemski et al., 2016) and CWMT Cor-

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 344–348

https://doi.org/10.18653/v1/W18-6404

Direction	parallel data	monolingual data
en-zh	22,587,593	9,061,023

Table 1: The number of provided data including parallel data and monolingual data.

pus, and monolingual we used is XMU corpus from CWMT Corpus. To compare with others in last year, WMT17 test set in English to Chinese direction is used as the development set to compare with the best score in last year.

3 Data Filtering

This section introduces the methods we used for data filtering in the news task. For this task, because we found that it is very difficult to make a significant improvement for training technique in a short time. Therefore, we pay more attention on the data filtering than exploring different training techniques. In this task we do the data filtering for both of the provided parallel sentences and the generated sentence from back translation.

3.1 Data Filtering through Rules

According to our observations the provided data has two types of noise: misalignment and translation error. One of the misalignment noise we found in the parallel corpus is that the translation only translates half or even a very small part of the source text. The translation error behaves like one punctuation repeated many times. Obviously language model cannot solve the problem of alignment or translation error from parallel sentences. It only evaluates the quality of the monolingual sentences. Thus, we clean up sentences with these problems with calculating the number of punctuation in both source sentence and target sentence. The parallel sentences where the difference between the number of punctuation of source and target sentences that exceeds the threshold A are removed. Besides, the sentences which contain punctuation more than threshold B will be removed because these sentences may appear as the table of contents or other sentences with some punctuation error. Here threshold A is named relative punctuation frequency threshold and threshold B is named absolute punctuation frequency threshold.

3.2 Data Filtering through Language Model

It has been proved that back translation (Sennrich et al., 2015a) is an effective way to improve the translation quality, especially in low-resource condition. In this task we firstly train an initial translation model(from Chinese to English) using transformer architecture, then we use this model to translate the provided monolingual Chinese data onto English and then get the generated synthetic data. To filter the generated synthetic data, we organize the filtering procedure as follows:

- Train two language models with Chinese and English monolingual data extracted from provided parallel corpus. To train the models we utilized the Marian toolkit, the model type of Marian is lm-transformer whose architecture is based on transformer.
- Calculate the cross entropy of each sentence with the trained language model in Chinese.
- Analyse the cross entropy, according to our observation, we removed the sentences with cross entropy higher than -30.971481 or lower than -299.529816. After this operation the number of remaining parallel sentences is 6,280,000 out of 9,061,023.
- Remove the duplicated sentences in the remaining 6,280,000. This operation further reduced the remaining sentences to 5,891,328.
- Remove the sentences that contain HTML tag such as "$\langle p \rangle \langle /p \rangle$","$\langle strong \rangle \langle /strong \rangle$", the remaining sentences then reduced to 4,981,288.
- Calculate the cross entropy of each translated English sentence with the trained English language model.
- Remove the sentences with cross entropy lower than -396.643829, the remaining parallel sentences further reduced from 4,981,288 to 4,975,094.

The reason why our filtering procedure is more complicated is that we believe the quality of the data can heavily affect the translation performance. We trained two language models to filter the synthetic data from both source text and target text. Through the above filtering procedure the synthetic data is reduced from 9,061,023 to 4,975,094.

3.3 Data Filtering Through Translation Model

Beside the generated synthetic data, we also suppose the provided parallel corpus is not clean enough to directly put into the training procedure. Since the language model cannot evaluate the quality of translation for parallel sentences which means that tow irrelevant or bad-translated sentences can't be distinguished through language model. Therefore, we use the rescorer tool of Marian to evaluate the parallel sentences in loss. In this case, we trained a translation model with the provided paralleled data, then we assume the translation model is generally correct and fix all the parameters in the model to calculate the cross entropy loss of each pair of provided parallel sentences. We remove the provided parallel sentences with cross entropy loss lower than -165.529449. This operation accompanies by the filter rules make the number of parallel sentences reduces from 22,587,593 to 17,969,826.

4 Optimizing transformer

The intuition for optimizing transformer is to try those optimizing methods which have proven to be effective in RNN architecture. According to our previous experiments right-to-left reranking, back translation synthetic data, Annealing Adam and ensemble decoding are the most effective approaches to improve the translation performance.

Right-to-left reranking means training a right-to-left model in target side. It can rerank the n-best translations and the expected averaged probabilities will be more robust for general evaluation. In this task, we reverse the target sentences and train the rights-to-left model.

Back translation is trying to improve the translation quality through data aspect. It is a simple but effective approach especially in low-resource condition. In this task, we have nearly 20 million parallel sentences from English to Chinese, but we are still trying to translate the Chinese monolingual data to construct the back translation data.

Annealing Adam is an optimizing function which is significantly faster than stochastic gradient descent with Annealing. Besides, it can also obtain a better performance in most cases. In this task we set the baseline with Annealing Adam optimizing function.

Ensemble decoding is trying to combine different models together to explore a better translation balance between different translation preference. The most common solution is to average the parameters of the latest server saved models during the training procedure. We can also combine models with different parameter initialization or even models with different hyper parameters. Normally to do ensemble decoding requires many different trained models. Therefore, it needs a lot of time and hardware resources which is the main reason that we only participate in one direction of the whole evaluation task. Unlike some other participants, we take a greedy ensemble strategy to combine our trained models instead of directly ensemble decoding them all. The greedy ensemble strategy firstly choose one model with the best single model BLEU score as the base model, and choose one model from the rest models again as the ensemble result to get a better BLEU score, then repeatedly choose one of the rest model to obtain a better BLEU until the BLEU doesn't increase.

5 Experiment

This section describes the all experiments we conducted and illustrate how we get the evaluation step by step.

5.1 Data pre-processing

In the news translation task we only focus on English to Chinese direction. Both of the parallel data and monolingual data are fully filtered at first. After that, we normalized the punctuation of English texts by normalize-punctuation.perl in Moses toolkit(Koehn et al., 2007) and normalized the punctuation of Chinese texts by converting the double byte character(DBC) to single byte character(SBC). We applied Jieba(Sun, 2012) as our Chinese word segmentation tool for segment Chinese text in both parallel data and monolingual data. For English text, tokenizer and truecase in Moses toolkit are applied. Finally, we applied BPE on both tokenized Chinese and English text.

5.2 Experiments setup

We describe all the experiment setups for this task in detail. The transformer baseline is trained with only parallel data, including CWMT corpus, UN Parallel Corpus V1.0 and News Commentary v13, after data preprocessing. We trained the baseline system not only in English to Chinese direction, but also in Chinese to English direction in order to translate the filtered monolingual data and do

configuration	value
architecture	transformer
English vocabulary size	40500
Chinese vocabulary size	50000
word embedding	512
Encoder depth	6
Decoder depth	6
transformer heads	8
size of FFN	2048

Table 2: The main model configuration. FFN means feed forward network.

parameter	value
maximum sentence length	100
batch fit	true
learning rate	0.0003
label-smoothing	0.1
optimizer	Adam
learning rate warmup	16000
clip gradient	5

Table 3: The training and decoding parameter.

data	number
original data	22,587,593
cleaning by rules and TM	17,969,826
original synthetic data	9,061,023
synthetic sentences cleaning by LM	4,981,288

Table 4: Cleaning parallel data and synthetic data. TM means translation model and LM means language model.

the parallel data filtering. During the training procedure the number of BPE merge operation is set to 40,000 for both English and Chinese. The hyperparameter of our baseline model configuration is shown in Table2 and the training parameter is in Table 3. After the baseline, we filter parallel data through rules and translation model. The relative punctuation frequency threshold and absolute punctuation frequency threshold we mentioned in section 3 is 5 and 15 respectively. We construct the synthetic data with back translation baseline model from Chinese to English. The synthetic data is firstly filtered by Chinese language model and then filtered by English language model. Table 4 shows the detail information about the data filtering.

In general, we trained 3 models to explore the effect of data filtering, which are: 1. baseline model with provided parallel sentences; 2. baseline model with parallel sentences filtered by rules and translation model; 3. baseline model with sentences mixed parallel sentences filtered by rules and translation model and synthetic sentences filtered by language model. Beside the baseline models, we trained four groups of translation model with fully filtered parallel data and synthetic data. Each model in the four groups is trained with different random seed and also apply Annealing Adam which get better performance compared with Adam. Therefore, we got 8 different translation models with the filtered data. We applied the greedy ensemble strategy to combine the 8 models and finally obtain the best translation performance on the development set with 3 models. Another, the right-to-left model in target side is also trained to rerank n-best translation of three best translation performance models.

6 Result and analysis

Table 5 shows the BLEU score we evaluated on development set. For data filtering, we observed that the methods improve the quality of sentences and get a better BLEU score. The methods can solve some problems of corpus quality. For model training techniques, back-translation is still the most effective method of improvement on 3.83-3.93 BLEU score. Annealing Adam has an improvement of BLEU score ranging from 0.04 to 0.36. The evaluation table shows that the higher BLEU score we get from the neural machine translation model, the smaller improvement can we get from Annealing Adam. When ensemble decoding, the greedy ensemble decoding strategy get the improvement on 0.56 BLEU score. However, when trying to decode our models ensemble with right-to-left reranking it did not improve the BLEU score as we expected.

Regard to the official evaluation we add one more post-processing step which is to convert all the SBC punctuation to DBC punctuation and it consequently further improved the BLEU score form 43.2 to 43.8.

7 Summary

We explored how to optimize the quality of machine translation in two different ways:1. through the data; 2 through the training and decoding approaches. In data aspect, we illustrated how we filter the provided parallel corpus through the trained

model	BLEU
baseline with PS	34.38
+ Annealing Adam	34.74
clean PS by rules and TM	35.42
+ Annealing Adam	35.56
mix cleaned PS and SS cleaned by LM	39.35
+ Annealing Adam	39.39
greedy ensemble decoding	39.95
r2l reranking	39.91

Table 5: The BLEU score in character level for development set of English-to-Chinese direction. SS means synthetic sentences, TM means translation model, LM means language model and PS means parallel sentences. The greedy ensemble decoding means decoding the 8 models and finally obtain the best translation performance on development set with 3 models.

language model and trained translation model and showed the improvement of the data filtering, as well as constructing the synthetic through the back translation approach. In the training and decoding aspect, we applied transformer architecture as our main machine translation framework. To optimize it we utilized Annealing Adam optimize function and ensemble decoding. We also found that right to left reranking is not working according to our experiments.

Acknowledgments

This work is supported by 2020 Cognitive Intelligence Research Institute[1] of Global Tone Communication Technology Co., Ltd.[2]

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Michael J. Denkowski and Graham Neubig. 2017. Stronger baselines for trustable results in neural machine translation. *CoRR*, abs/1706.09733.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, Andr F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*.

[1]http://www.2020nlp.com/

[2]http://www.gtcom.com.cn/

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.

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. 2015a. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for WMT 16. *CoRR*, abs/1606.02891.

J Sun. 2012. jiebachinese word segmentation tool.

Jrg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. *CoRR*, abs/1706.03762.

Michal Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *LREC*.