GTCOM Neural Machine Translation Systems for WMT19

Chao Bei, Hao Zong, Conghu Yuan, Qingming Liu, Baoyong Fan
Global Tone Communication Technology Co., Ltd.
{beichao, zonghao, yuanconghu, liuqingming, fanbaoyong}@gtcom.com.cn

Abstract

This paper describes the Global Tone Communication Co., Ltd.'s submission of the WMT19 shared news translation task. We participate in six directions: English to (Gujarati, Lithuanian and Finnish) and (Gujarati, Lithuanian and Finnish) to English. Further, we get the best BLEU scores in the directions of English to Gujarati and Lithuanian to English (28.2 and 36.3 respectively) among all the participants. The submitted systems mainly focus on back-translation, knowledge distillation and reranking to build a competitive model for this task. Also, we apply language model to filter monolingual data, back-translated data and parallel data. The techniques we apply for data filtering include filtering by rules, language models. Besides, We conduct several experiments to validate different knowledge distillation techniques and right-to-left (R2L) reranking.

1 Introduction

We participated in the WMT shared news translation task and focus on the bidirections: English and Gujarati, English and Lithuanian, as well as English and Finnish. 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. In addition to data filtering, which is basically the same as the techniques we applied in WMT 2018 last year, we verify different knowledge distillation and reranking techniques to improve the performance of all our systems.

For data preprocessing, the basic methods include punctuation normalization, tokenization, truecase and byte pair encoding(BPE) (Sennrich et al., 2015b). Besides, human rules and language model are also involved to clean English parallel data, monolingual data and synthetic data. Regard to the techniques on model training, back-translation (Sennrich et al., 2015a), knowledge distillation and R2L reranking (Sennrich et al., 2016) are applied to verify whether these techniques could improve the performance of our systems.

In order to explore the application of knowledge distillation technology in the field of neural machine translation, we conduct a number of experiments for sequence-level knowledge distillation and sequence-level interpolation (Kim and Rush, 2016). Another, R2L reranking didn't get the better performance in last year experiment. In order to improve the performance of R2L reranking, we increase the beam size step by step, and explore the effect of any combination for R2L models with every step.

This paper is arranged as follows. We firstly describe the task and provided data information, then introduce the method of data filtering, mainly in the application of language model. After that, we describe the techniques on transformer architecture and show the conducted experiments in detail of all directions, including data preprocessing, model architecture, back-translation and knowledge distillation. At last, we analyze 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. For the direction between English and Lithuanian, the parallel data is mainly from Europarl v9, ParaCrawl v3, Wiki Titles v1 and Rapid corpus of EU press releases (Rozis and Skadiņš, 2017). For the direction between English

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 116–121

direction	number of sentence
en-lt parallel data	4.21M
en-gu parallel data	155K
en-fi parallel data	9.17M
en monolingual data	18M
lt monolingual data	3.09M
gu monolingual data	4.35M
fi monolingual data	18M
en-gu unconstrained data	4.63M

Table 1: Task Description.

direction	number of cleaned data
en-lt parallel data	4.08M
en-gu parallel data	77K
en-fi parallel data	9M
en monolingual data	17.6M
lt monolingual data	2.92M
gu monolingual data	4.28M
fi monolingual data	15M
en-gu unconstrained data	4.55M

Table 2: Number of cleaned data.

and Gujarati the parallel data is from Wiki Titles v1, Bible Corpus, OPUS (Tiedemann, 2012) and govin crawled corpus, as well as our own parallel data. Thus, this direction is unconstrained. The Corpus, from Europarl v9, ParaCrawl v3, Wiki Titles v1 and Rapid corpus of EU press releases, are used to the directions between English and Finnish. Another, monolingual data we used are News crawl, Europarl and Europarl v9. All directions we participated are new for this year, we use newsdev2019 as our development set.

3 Data Filtering

The methods of data filtering by human rules are mainly the same as we did in English to Chinese(Bei et al., 2018) last year, but language models are used to clean all data, including monolingual data, parallel data and synthetic data. We use Marian to train the transformer language model for each language (i.e. English, Gujarati, Lithuanian and Finnish). We introduce this section in two condition:

- For monolingual data and synthetic data (i.e. back-translate data from target side and knowledge distillation from source side), Every sentence are scored by language model, and the score for sentence is calculated as follows:

$$Score_{sentence} = \frac{Score_{lm}}{\sqrt{L_{sentence}}}$$

  Here $Score_{lm}$ is score of language model for sentence, and $L_{sentence}$ is length of sentence in token level.

- For parallel data, considering scores of two sides, we combine the two side score of parallel data with liner:

$$Score_{combine} = \lambda * Score_{src} + (1-\lambda) Score_{tgt}$$

.

Here, λ is 0.5. According the sorted score for each sentence or sentence pair, we clean the sentences that is obviously not influence. Table 2 shows the number of cleaned data.

4 Back-translation

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. Same as we did in last year, we firstly train models from target to source, then we use these model to translate the provided monolingual data in target side onto source side. Besides, the target parallel data is also translated to source side. It should be noticed that the ratio of parallel data and synthetic data is 1:1.

Joint-training (Zhang et al., 2018) is another method which has been proved that it can improve the performance of back-translation. In another perspective, back-translation is the first step of joint-training. When getting the best model from back-translation, we consecutively translate the monolingual data from the target side of parallel data and mix parallel data and synthetic data with the ratio of 1:1. Then the new training set is used to train a new model until there is no improvement. We only repeated this procedure twice due to the time limitation.

5 Knowledge Distillation

5.1 Sequence-level Knowledge Distillation

Sequence-level Knowledge distillation describes the method of training a smaller student network to perform better by learning from a teacher network. Knowledge distillation suggests training by matching the student's predictions to the teacher's

predictions. We consider two different kinds of methods to improve the performance for NMT:

- **Ensemble Teacher** As according (Freitag et al., 2017), we translate the source side sentences of parallel data with ensemble models and get the synthetic target side sentences. The synthetic data is applied to training.
- **R2L Teacher** Inspired by (Wang et al., 2018) (Hassan et al., 2018), we translate the source side sentences of parallel data to target side with R2L model to improve L2R model.

To avoid bad translation, we filter the synthetic data with BLEU score lower than 30.

5.2 Sequence-level Interpolation

After sequence-level Knowledge distillation, the trained models are fine-tuned with n-best knowledge distillation data. The n-best knowledge distillation data is from the n-best translation from sequence-level knowledge distillation with different kinds of teachers. For every translation with the same source side sentence in an n-best translation, we extract the highest BLEU score and get the n-best knowledge distillation data.

6 R2L Reranking

Last year we didn't get better result with applying R2L reranking technique from English to Chinese. And we found out that the reason is we didn't increase the beam size step by step and didn't use all combination of R2L models. Therefore, to increase search space and get better translation, we applied the above procedure this time.

7 Experiment

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

7.1 Model Architecture

We use transformer big model to train our model with Marian according (Vaswani et al., 2017b). The model configuration and the training parameters are show in Table 3 and Table 4 respectively.

configuration	value
architecture	transformer
word embedding	1024
Encoder depth	6
Decoder depth	6
transformer heads	16
size of FFN	4096
transformer dropout attention	0.1
transformer dropout FFN	0.1

Table 3: The main model configuration.

parameters	value
maximum sentence length	100
learning rate	0.0003
label-smoothing	0.1
optimizer	Adam
learning rate warmup	16000
clip gradient	5

Table 4: The main training parameters.

7.2 Date preprocessing

Both of parallel data and monolingual data are fully filtered. After that, we normalize the punctuation of all sentences by normalize-punctuation.perl in Moses toolkit (Koehn et al., 2007). We apply tokenizer and truecaser in Moses toolkit for English, Lithuanian and Finnish sentences and use polyglot [1] to tokenize Gujarati sentences. Finally, BPE is applied on tokenized English, Lithuanian, Finnish and Gujarati sentences respectively. Here, the BPE merge operation is set to 30000, and the vocabulary size is 30500.

7.3 Training Step

Here we introduce the training step in detail.

- **Baseline model** We use transformer big model to train our baseline model with only parallel data cleaned by human rules and language model. Besides, R2L models are trained with the same data with 4 different seeds.
- **Back-translation** When getting the baseline model, we decode monolingual data in target side to source side with ensemble models trained from source side to target side. For example, if we want to train an English to Gujarati model with synthetic data, using Gujarati-to-English baseline model to translate Gujarati sentences to English. Then, the translated English sentences are filtered by

[1] https://github.com/aboSamoor/polyglot

language model. The synthetic data and parallel data, which are mixed with ratio of 1:1, are applied to train back-translation model.

- **Joint Training** When getting the back-translation model, repeat back-translation step until there is no improvement. We repeated this step twice.

- **Sequence-level Knowledge Distillation** Different from back-translation, we use different teachers of source-to-target model to translate the source sentence of parallel data to target side. For example, we use English-to-Gujarati model to translate English sentences to Gujarati. Compared with golden reference, each translation with the BLEU score lower than 30 will be removed. Considering the low-resource condition, we mix parallel data, synthetic data and knowledge distillation data with ratio of 1:1:1 to train the new model.

- **Sequence-level Interpolation** After sequence-level knowledge distillation, the best models are fine-tuned with the n-best knowledge distillation data.

- **Ensemble Decoding** To get the best performance over all models efficiently, we use GMSE Algorithm (Deng et al., 2018) to select models.

- **R2L Reranking** To enlarge search space, we increase the beam size step by step and rescore it with all combination of R2L models for each step. Here, the step size is 10 and maximum beam size is 200.

8 Result and analysis

Table 5, Table 6, Table 7, Table 8, Table 9 and Table 10 show the BLEU score we evaluated on development set for English to Lithuanian, Lithuanian to English, English to Gujarati, Gujarati to English, English to Finnish and Finnish to English respectively.

For back-translation, we observe that it is the most effective method with an improvement from 1.54 to 4.87 BLEU score, especially in low-resource condition. And joint training can improve the BLEU score slightly from 0.12 to 0.29. For knowledge distillation, sequence-level knowledge distillation gets an improvement of BLEU score ranging from 0.09 to 1.03, and sequence-level interpolation has 0.12 to 0.21 BLEU score improvement. When ensemble decoding, GMSE algorithm gets the improvement ranging from 0.22 to 0.55. After increasing search space and combining the R2L models, reranking can still improve the result by 0.1 to 0.17 BLEU score.

model	BLEU score
baseline	22.56
back-translation	27.43
joint training	27.72
sequence-level KD	27. 83
sequence-level interpolation	27.97
ensemble decoding	28.22
R2L reranking	28.37

Table 5: The case-insensitive BLEU score of English to Lithuanian.

9 Summary

This paper describes GTCOM's neural machine translation systems for the WMT19 shared news translation task. For all translation directions, we build systems mainly from data aspect, including acquiring more quantities and higher quality data. Besides, decoding strategies such as GSME algorithm and R2L reranking give us more robust and high quality translation. Finally, the directions of English to Gujarati (unconstrained) and Lithuanian to English get the best case-sensitive BLEU score of all systems.

Acknowledgments

This work is supported by 2020 Cognitive Intelligence Research Institute[2] of Global Tone Communication Technology Co., Ltd.[3]

References

Chao Bei, Hao Zong, Yiming Wang, Baoyong Fan, Shiqi Li, and Conghu Yuan. 2018. An empirical study of machine translation for the shared task of wmt18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 340–344.

Yongchao Deng, Shanbo Cheng, Jun Lu, Kai Song, Jingang Wang, Shenglan Wu, Liang Yao, Guchun Zhang, Haibo Zhang, Pei Zhang, et al. 2018.

[2]http://www.2020nlp.com/
[3]http://www.gtcom.com.cn/

model	BLEU score
baseline	29.76
back-translation	32.53
joint training	32.7
sequence-level KD	33.73
sequence-level interpolation	33.94
ensemble decoding	34.59
R2L reranking	34.69

Table 6: The case-insensitive BLEU score of Lithuanian to English.

model	BLEU score
baseline	23.93
back-translation	25.62
joint training	25.74
sequence-level KD	26.17
sequence-level interpolation	26.39
ensemble decoding	27.28
R2L reranking	27.44

Table 7: The case-insensitive BLEU score of English to Gujarati.

model	BLEU score
baseline	24.42
back-translation	27.58
joint training	27.79
sequence-level KD	28.05
sequence-level interpolation	28.21
ensemble decoding	28.54
R2L reranking	28.71

Table 8: The case-insensitive BLEU score of Gujarati to English.

model	BLEU score
baseline	18.04
back-translation	21.29
joint training	21.49
sequence-level KD	21.58
sequence-level interpolation	21.79
ensemble decoding	22.01
R2L reranking	22.12

Table 9: The case-insensitive BLEU score of English to Finnish.

model	BLEU score
baseline	25.55
back-translation	27.09
joint training	27.38
sequence-level KD	27.67
sequence-level interpolation	27.79
ensemble decoding	28.22
R2L reranking	28.34

Table 10: The case-insensitive BLEU score of Finnish to English.

Alibaba's neural machine translation systems for wmt18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 368–376.

Markus Freitag, Yaser Al-Onaizan, and Baskaran Sankaran. 2017. Ensemble distillation for neural machine translation. *arXiv preprint arXiv:1702.01802*.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

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*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.

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.

Roberts Rozis and Raivis Skadiņš. 2017. Tilde model-multilingual open data for eu languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265.

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.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

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.

Mingxuan Wang, Li Gong, Wenhuan Zhu, Jun Xie, and Chao Bian. 2018. Tencent neural machine translation systems for wmt18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 522–527.

Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018. Joint training for neural machine translation models with monolingual data. In *Thirty-Second AAAI Conference on Artificial Intelligence*.