GTCOM Neural Machine Translation Systems for WMT20

Chao Bei, Hao Zong, Qinming Liu and Conghu Yuan
Global Tone Communication Technology Co., Ltd.
{beichao, zonghao, liuqingmin and yuanconghu}@gtcom.com.cn

Abstract

This paper describes the Global Tone Communication Co., Ltd.'s submission of the WMT20 shared news translation task. We participate in four directions: English to (Khmer and Pashto) and (Khmer and Pashto) to English. Further, we get the best BLEU scores in the directions of English to Pashto, Pashto to English and Khmer to English (13.1, 23.1 and 25.5 respectively) among all the participants. Our submitted systems are unconstrained and focus on mBART (Multilingual Bidirectional and Auto-Regressive Transformers), back-translation and forward-translation. Also, we apply rules, language model and RoBERTa model to filter monolingual, parallel sentences and synthetic sentences. Besides, we validate the difference of the vocabulary built from monolingual data and parallel data.

1 Introduction

We participated in the WMT shared news translation task and focus on the bidirections: English and Khmer, English and Pashto. We applied fairseq(Ott et al., 2019) as our develop tool and use transformer(Vaswani et al., 2017) as the main architecture. The primary ranking index for submitted systems is BLEU (Papineni et al., 2002), therefore we apply BLEU as the evaluation matrix for our translation system. For Khmer, we use polyglot [1] as the tokenizer before evaluation.

For data preprocessing, the basic method includes punctuation normalization for all language. Further, according to the different language characteristics. Tokenization, truecase and byte pair encoding (BPE) (Sennrich et al., 2015b) are applied for English, and sentencepiece (Kudo and Richardson, 2018) is applied for Khmer and Pashto. Besides, human rules, language model and RoBERTa model (Liu et al., 2019) are also involved to clean parallel data, monolingual data and synthetic data. Regard to the techniques on model training, back-translation (Sennrich et al., 2015a) and forward-translation are applied to verify whether these techniques could improve the translation performance especially in low-resource condition.

We all know that it is more difficult to train a model in low-resource condition, because it suffers from data sparsity and out-of-vocabulary problem. Normally knowledge distillation (Kim and Rush, 2016) is a good way to generate synthetic data. But in this task we suppose that knowledge distillation can only generate 100 thousand to 1 million parallel sentences due to the size of provided data. Therefore, we use forward-translation with monolingual data to generate more synthetic data. Here forword-translation refers to translate the source sentences to target language, and clean synthetic data.

This paper is arranged as follows. We firstly describe the task and show the data information, then introduce how we do data filtering, including human rules, language model and RoBERTa model. After that, we describe the techniques on low-resource condition and show the conducted experiments in detail of all directions, including data preprocessing, model architecture, back-translation and forwor-translation. 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 Khmer, the parallel data is mainly from ParaCrawl v5.1 and shared task on parallel corpus filtering (mostly from OPUS (Tiedemann, 2012)), as well as the direction between English and Pashto. Another, monolin-

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

Proceedings of the 5th Conference on Machine Translation (WMT), pages 100–104

language	number of sentences
en-ps parallel data	1M
en-km parallel data	4.17M
en monolingual data	16.9M
ps monolingual data	4.2M
km monolingual data	12.7M
en-ps development set	3162
en-km development set	2378
en-ps devtest set	2698
en-km devtest set	2309

Table 1: Task Description.

configuration	value
architecture	transformer
word embedding	512
Encoder depth	5
Decoder depth	5
transformer heads	2
size of FFN	2048
attention dropout	0.2
dropout	0.4
relu dropout	0.2

Table 2: The FLoRes model architecture.

configuration	value
architecture	transformer
word embedding	768
Encoder depth	6
Decoder depth	6
transformer heads	12
size of FFN	3072
attention dropout	0.1
dropout	0.1
relu dropout	0

Table 3: The mBART model architecture.

gual data we used are News crawl both for English, Common Crawl and Wiki dumps both for Khmer and Pashto. All directions we participated are new for this year, we use wikipediadev as our development set and wikipediadevtest as our test set.

3 Data Filtering

The methods of data filtering are mainly the same as we did in last year (Bei et al., 2019), including human rules and language model. Further, another methods we used this year are as follows:

Clean repeated translation sentences in synthetic data. For example, we often see the translation like: I want to eat an apple apple apple apple apple, when translating a source language sentence with repeated words until the end of the sentence. In this task, we made a simple clean strategy which is to remove the sentences that repeat one word four times, two words three times or three words two times.

Clean synthetic data by RoBERTa model. In order to clean synthetic data, especially from forword-translation, we represent the source and target sentences by RoBERTa model and calculate the cosine distance. Remove the sentences with low score or without translation (the source sentence and target sentence are same).

4 Forward-translation

In low-resource condition, out-of-vocabulary is a problem. There is a difference between the test scenario and training scenario, which means the words appear in test set may be not existed in training vocabulary. Back-translation is a common way to extend the word vocabulary. However, with the generated synthetic data from back-translation, only target vocabulary can be enriched. To extend the source side vocabulary, we use source-to-target model to translate the source monolingual data to target side. Further, it is necessary to clean the forward-translation sentences to avoid cascading error for the next training. We use RoBERTa to represent the source and target sentence and calculate the cosine distance. Remove the sentences with low score or without translation (the source sentence and target sentence are same).

5 Experiment

5.1 Model architecture

- **Baseline** Table 2 shows the baseline model architecture.
- **mBART** We fine-tune on mBart model to get better translation. Table 3 shows the model architecture.
- **Big transformer** We use transformer big model to train our model with fairseq. The model configuration and training parameters is almost same as last year we use. In order to training more stable in low-resource condition, we add layer normalize before encoder and decoder.

5.2 Training Step

This section introduces all the experiments we set step by step and Figure 1 shows the whole flow.

- **Date Filtering** Following the task of Parallel Corpus Filtering and Alignment for Low-Resource Conditions, we use the LASER-based scores to filter the raw parallel sentences and extract 5 million words England tokens.
- **Baseline.** We use FLoRes (Guzmán et al., 2019) architecture to construct our baseline in low-resource condition.
- **Fine-tuning on mBART.** In such low-resource condition, we fine-tune on mBART model with filtered sentences.
- **Back-translation.** We use fine-tuned model to translate the target sentence to source side, and clean synthetic data with language model and RoBERTa model. Mix cleaned back-translation data and parallel sentences and fine-tune on mBART model.
- **Forward-translation.** Source side sentences are translated to target side, and cleaned by language model and RoBERTa model. Mixed with cleaned back-translation data, forward-translation data and parallel sentences, fine-tune on mBART model.
- **Monolingual vocabulary.** To enrich the vocabulary further, we preprocess the monolingual data and build the vocabulary as model vocabulary. Here, we normalize the punctuation of all data by nomalize-puncuation.perl in Moses toolkit (Koehn et al., 2007). We apply tokenizer and truecaser in Moses toolkit for English. Finally, BPE (Byte Pair Encoding) (Sennrich et al., 2016) is applied on tokenized English and sentencepiece is applied on Pashto and Khmer. The BPE and sentencepiece merge operation are both 32000. Therefore, the vocabulary of monolingual data is set to 32500. We use these vocabularies as model vocabulary and train big transformer model.
- **Joint training.** Repeat back-translation step and forward-translation step by best model, until there is no improvement.
- **Ensemble Decoding.** We use GMSE Algorithm (Deng et al., 2018) to select models to obtain the best performance.

6 Result and analysis

Table 4 and Table 5 show the BLEU score we evaluated on development set for English to Pashto, Pashto to English, English to Khmer and Khmer to English respectively.

For fine-tuning on mBART model, we find that it is the most effective method with an improvement from 0.56 to 3.62 BLEU score in low-resource condition. And back-translation gets the improvement from 0.17 to 3.04 BLEU score. Forward translation and monolingual vocabulary enrich the information in low-resource condition, with improvement of 0.16 to 0.74 BLEU score and 0.69 to 0.94 BLEU score respectively. Further, joint training and ensemble decoding slightly increase the performance with 0.31 to 0.4 BLEU score and 0.15 to 0.4 BLEU score.

7 Summary

This paper describes GTCOM's neural machine translation systems for the WMT20 shared news translation task. For all translation directions, we build systems mainly base on mBART model and enrich information by back-translation, forward-translation and using monolingual vocabulary with data filtering, including calculating cosin distance by RoBERTa model, language model and so on. The effect of increasing information is also dependent on data filtering. Finally, we submit the online system including English to Pashto, Pashto to English, Khmer to English and English to Khmer with almost same methods in this paper. Another, we also submit our online system from English to Tamil and Tamil to English.

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, Conghu Yuan, Qingming Liu, and Baoyong Fan. 2019. GTCOM neural machine translation systems for WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 116–121, Florence, Italy. Association for Computational Linguistics.

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

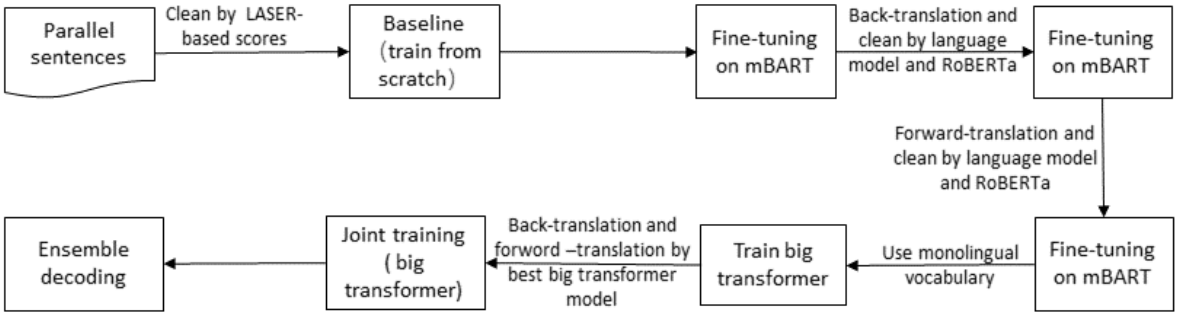

Figure 1: The whole work flow.

model	en2ps	ps2en
baseline	5.73	8.37
fine-tuning on mBART	9.35	11.98
+ back-translation	9.97	12.15
+ forward-translation	10.15	12.31
+ monolingual vocabulary	10.88	13.25
+ joint training	11.19	13.69
+ Ensemble Decoding	11.34	14.09

Table 4: The case-sensitive BLEU score between English and Pashto.

model	en2km	km2en
baseline	8.68	7.47
fine-tuning on mBART	9.24	9.91
+ back-translation	12.28	12.23
+ forward-translation	12.85	12.97
+ monolingual vocabulary	13.54	13.73
+ joint training	13.87	14.04
+ Ensemble Decoding	14.13	14.29

Table 5: The case-sensitive BLEU score between English and Khmer.

Yongchao Deng, Shanbo Cheng, Jun Lu, Kai Song, Jingang Wang, Shenglan Wu, Liang Yao, Guchun Zhang, Haibo Zhang, Pei Zhang, et al. 2018. Alibaba's neural machine translation systems for wmt18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 368–376.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The flores evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. *arXiv preprint arXiv:1902.01382*.

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.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible

toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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. 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, Berlin, Germany. Association for Computational Linguistics.

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. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.