GTCOM Neural Machine Translation Systems for WMT21

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 WMT21 shared news translation task. We participate in six directions: English to/from Hausa, Hindi to/from Bengali and Zulu to/from Xhosa. Our submitted systems are unconstrained and focus on multilingual translation model, back-translation and forward-translation. We also apply rules and language model to filter monolingual, parallel sentences and synthetic sentences.

1 Introduction

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 data preprocessing, punctuation normalization, tokenization and BPE(byte pair encoding) (Sennrich et al., 2015) are applied for all language. Further, we apply truecase model for English, Hausa, Zulu and Xhosa according to the character of each language. Regarding to the tokenization, we use polyglot [1] as the tokenizer for Hausa, Hindi, Bengali, Zulu and Xhosa. Besides, knowledge based rules and language model are also involved to clean parallel data, monolingual data and synthetic data.

Due to the quantity limitation of parallel corpus in low-resource language pair, we use forward-translation with monolingual data to generate more synthetic data instead of knowledge distillation (Kim and Rush, 2016). Here forward-translation refers to translate the source language sentences to the target language, and then clean this synthetic data with the above described method. In order to enrich the low-resource language corpus, we add English to X corpus to construct a multilingual translation model. This multilingual model is expected to obtain the inner deep information among all languages and give us a better translation.

This paper is arranged as follows. We firstly describe the task and show the data information, then introduce our multilingual translation 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, forward-translation and multilingual translation model. 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 directions between Hindi and Bengali, the parallel data is mainly from CC-Aligned, as well as the directions between Zulu and Xhosa. For the directions between English and Hausa, the parallel data is mainly from English-Hausa Opus corpus, Khamenei corpus, ParaCrawl v8. The monolingual data we used includes: News Crawl in English, Hindi and Bengali; extended Common Crawl in Hausa, Xhosa and Zulu; Common Crawl in Hausa. All language directions we participated are new tasks in this year, therefore we only use the provided newsdev2021 as our development set for the directions of English to/from Hausa, flores-dev for the directions of Hindi to/from Bengali and Zulu to/from Xhosa.

3 Multilingual Translation Model

In low-resource condition, data augmentation and pretrained model are the most effective approaches to improve translation quality. According Google's Multilingual Neural Machine Translation System(Johnson et al., 2017), we use other language

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

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 100–103

language	number of sentences
bn-hi parallel data	3.3M
en-bn parallel data	2.2M
en-hi parallel data	2.2M
en-ha parallel data	750K
xh-zu parallel data	60K
en-xh parallel data	41K
en-zu parallel data	45K
en monolingual data	93.4M
bn monolingual data	59.7M
hi monolingual data	46.1M
ha monolingual data	46.1M
xh monolingual data	1.6M
zu monolingual data	2M
en-ha development set	2000
bn-hi development set	997
xh-zu development set	997

Table 1: Task Description

pairs parallel data along with the provided bilingual data to training a multilingual translation model, the low-resource language pair is expected to get the benefits from other language pair's parallel data, especially in similar language. For the multilingual model preprocessing, we add a language tag at beginning of each source sentence, and use joint BPE for all languages in one multilingual translation model.

4 Experiment

4.1 Model architecture

- **Baseline** Table 2 shows the baseline model architecture.

- **Big transformer** We use fairseq to train our model with transformer big architecture. The model configuration and training parameters is almost same as GTCOM2020(Bei et al., 2020).

4.2 Training Step

This section introduces all the experiments we set step by step and Figure1 shows the full improvement status.

- **Date Filtering** The methods of data filtering are mainly the same as GTCOM2020, including knowledge based rules, language model and repeat cleaning.

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.

- **Baseline** We use FLoRes (Guzmán et al., 2019) architecture to construct our baseline in low-resource condition.

- **Multilingual translation model.** Due to the language distinction, We construct two multilingual translation models with the corpus organized as: 1. English-Bengali parallel data, English-Hindi parallel data and Bengali-Hindi parallel data; 2. English-Hausa parallel data, English-Xhosa parallel data, English-Zulu parallel data and Xhosa-Zulu parallel data. Each multilingual translation model has a shared vocabulary.

- **Back-translation** We use multilingual translation model to translate the target language sentence to source language, and clean synthetic data with language model. Here, we translate all language pairs we have added into this multilingual translation model. Then we combine the cleaned back-translation data and provided parallel sentences to train a new multilingual translation model.

- **Forward-translation** Source language sentences are translated to target language, and then cleaned by language model. Again we add this forward translation data with cleaned back-translation data and provided parallel sentences to train another multilingual translation model.

- **Joint training** Repeat generating back-translation data and forward-translation data by currently trained best multilingual model until there is no improvement.

- **Transformer big** Using bilingual parallel data and synthetic data generated by cur-

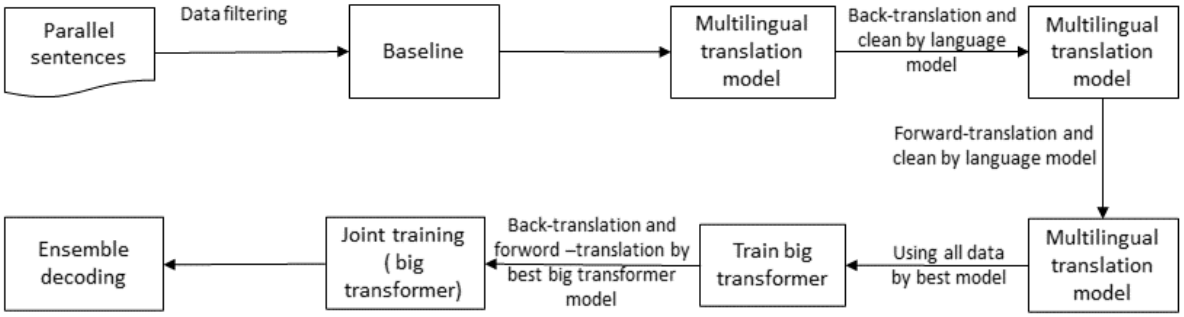

Figure 1: The whole work flow.

model	bn2hi	hi2bn
baseline	19.00	11.20
multilingual translation model	19.33	11.22
+ back-translation	23.63	14.80
+ forward-translation	23.95	14.95
+ joint training	24.05	15.02
big transformer	24.11	15.14
+ Ensemble Decoding	25.13	15.86

Table 3: The BLEU score between Hindi and Bengali.

model	en2ha	ha2en
baseline	11.04	12.02
multilingual translation model	12.20	13.09
+ back-translation	18.27	17.56
+ forward-translation	18.74	18.21
+ joint training	18.95	18.59
big transformer	19.32	18.91
+ Ensemble Decoding	21.09	21.58

Table 4: The BLEU score between English and Hausa after truecase.

rently best multilingual model to train a bilingual model with transformer big architecture and repeat back-translation step and forward-translation step, until there is no improvement.

- **Ensemble Decoding** We use GMSE Algorithm (Deng et al., 2018) to select models to obtain the best performance.

5 Result and analysis

Table 3, Table 4 and Table 5 show the BLEU score we evaluated on development set for Hind to/from Bengali, English to/from Hausa and Xhosa to Zulu

model	xh2zu	zu2xh
baseline	10.58	10.60
multilingual translation model	11.66	10.73
+ back-translation	12.48	10.76
+ forward-translation	12.70	10.86
+ joint training	12.74	10.92
big transformer	12.77	10.95
+ Ensemble Decoding	12.95	11.02

Table 5: The BLEU score between Xhosa and Zulu after truecase.

respectively. Back-translation is still the most effective method with improvement ranging from 0.03 to 6.07 BLEU score in low-resource condition. And multilingual translation model gets the improvement ranging from 0.02 to 1.16 BLEU score. Forward translation enrich the information in low-resource condition, with improvement of 0.1 to 0.65 BLEU score. Further, ensemble decoding increase the performance with 0.07 to 2.67 BLEU score.

6 Summary

This work mainly focus data augmentation and pay less attention on modeling. Because optimizing translation by data augmentation is the most elegant way for a commercial system. It can avoid many unexpected translation result generated by a newly proposed model which may give our customers worse translating experience.

This paper describes GTCOM's neural machine translation systems for the WMT21 shared news translation task. For all translation directions, we build systems mainly base on multilingual translation model and enrich information by back-

translation and forward-translation. The effect of increasing information is also dependent on data filtering.

Acknowledgments

The authors gratefully acknowledge the financial support provided by the National Key Research and Development Program of China (2020AAA0108005). And this work is supported by Global Institute of Intelligent Language Technology[2] of Global Tone Communication Technology Co., Ltd.[3]

References

Chao Bei, Hao Zong, Qingmin Liu, and Conghu Yuan. 2020. GTCOM neural machine translation systems for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 100–104, Online. Association for Computational Linguistics.

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*.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

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

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. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

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.

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

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