GTCOM Neural Machine Translation Systems for WMT22

Hao Zong, Chao Bei, Conghu Yuan
Global Tone Communication Technology Co., Ltd
{zonghao,yuanconghu}@gtcom.com.cn
chaobei001@gmail.com

Abstract

This paper describes the Global Tone Communication Co., Ltd.'s submission of the WMT22 shared general MT task. We participate in six directions: English to/from Ukrainian, Ukrainian to/from Czech, English to Croatian and English to Chinese. Our submitted systems are unconstrained and focus on backtranslation, multilingual translation model and finetuning. Multilingual translation model focus on X to one and one to X. 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 by using sacreBLEU[1].

For data preprocessing, punctuation normalization, tokenization and BPE(byte pair encoding) are applied for all languages. Further, we apply truecase model for English, Ukrainian, Czech and Croatian according to the character of each language. Regarding to the tokenization, we use polyglot as the tokenizer for Ukrainian and Croatian, and mosese tokenizer.perl for English and Czech. Besides, knowledge based rules and language model are also involved to clean parallel data, monolingual data and synthetic data.

This paper is arranged as follows. We firstly describe the task and show the data information, then introduce our baseline and multilingual translation model. After that, we describe the conducted experiments in detail in all directions, including data preprocessing, model architecture, back-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 and the provided data is shown in Table 1, including parallel data and monolingual data. For the directions between English and Ukrainian, the parallel data is mainly from ParaCrawl v9, WikiMatrix, Tilde MODEL corpus and OPUS, as well as the directions English to Croatian. For the directions between Ukrainian and Czech, the parallel data is mainly from WikiMatrix and OPUS. The monolingual data we used includes: News Crawl in English, Ukrainian, Croatian and Czech; Leipzig Corpora in Croatian, Ukrainian and Czech; News discussions in English. All language directions we participated in are new tasks this year, therefore we only use the provided development set from FLoRes101 dataset for all directions.

Usually, the news translation task will take the human evaluation result as the final ranking index. And this requires each participated team contribute 8 hours of human evaluation for each participating translation direction. For some low resource language directions, it is not very easy for the organizer to employ human translators from the participating team or translation agency. Besides, due to the number of sentences in the test set and the quantity of participating teams, it is not possible to employ human evaluation for all the test sets. Besides, with recent improvements of MT quality, the organizer decided to move away from testing only in the news domain and we are shifting the WMT focus on testing the general capabilities of MT systems.

3 Billingual Baseline Model and Multilingual Translation Model

To set a strong baseline for our multilingual model as a comparison. Our Billingual base-

[1] https://github.com/mjpost/sacrebleu

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 428–431
Abu Dhabi, December 7–8, 2022.

language	number of sentences
en-hr parallel data	318M
en-uk parallel data	13M
uk-cs parallel data	4M
en monolingual data	40M
uk monolingual data	15M
cs monolingual data	40M
hr monolingual data	13M
en-uk development set	997
en-hr development set	997
uk-cs development set	997

Table 1: Task Description

line model is different from the transformer base model transformer_wmt_en_de with 6 encoding layers and 6 decoding layers. Instead, we set our bilingual baseline model by using transformer_vaswani_wmt_en_de_big architecture with 12 encoding layers and 4 decoding layers.

The multilingual translation model is almost the same as GTCOM2021(Bei and Zong, 2021), but focuses on one to X and X to one this year. To obtain a better translation quality, we include Russian as the main auxiliary language since Russian and Ukrainian are very similar. We train four multilingual models: 1. ru-en, uk-en and hr-en to translate uk-en; 2. en-ru, en-uk and en-hr to translate en-uk and en-hr; 3. cs-uk, en-uk and ru-uk to translate cs-uk 4. en-uk; uk-cs and en-cs to translate uk-cs and en-cs. We use joint BPE for all languages in the multilingual model separately.

For English to Chinese direction, we just test our online system as a comparison with other participating systems. Therefore we did not conduct data augmentation, finetuning, or any other adaption experiments.

4 Experiment

4.1 Training Step

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

- **Date Filtering** The methods of data filtering are mainly the same as we did last year, including human rules, language models, and repeat cleaning.

- **Baseline.** We use big transformer architecture with 24 layers of encoder and 4 layers of decoder to construct our baseline.

model	en2uk	uk2en
baseline	32.43	40.08
back translation	32.58	40.84
joint training	32.97	42.33
deep multilingual translation model	33.72	43.27

Table 2: The BLEU score between English and Ukrainian.

model	uk2cs	cs2uk
baseline	22.52	22.00
back translation	25.51	23.59
joint training	25.72	24.09
deep multilingual translation model	26.14	24.89

Table 3: The BLEU score between Czech and Ukrainian.

- **Back-translation.** We use a multilingual translation model to translate the target sentence to the source side, and clean synthetic data with language model. Here, we translate each language pairs we have added into the multilingual translation model. Mix cleaned back-translation data and parallel sentences and train multilingual translation model.

- **Joint training.** Repeat the back-translation step by the best model, until there is no improvement.

- **Multilingual translation model.** We focus on one to X and X to one model, and each multilingual model has joint BPE and a shared vocabulary. The multilingual translation model setting follows Google's Multilingual Neural Machine Translation System(Johnson et al., 2017).

- **Deep multilingual translation model.** Using bilingual parallel data and synthetic data by the best model, train the multilingual transformer model with 12 encoding layers and 4 decoding layers, then repeat the 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.

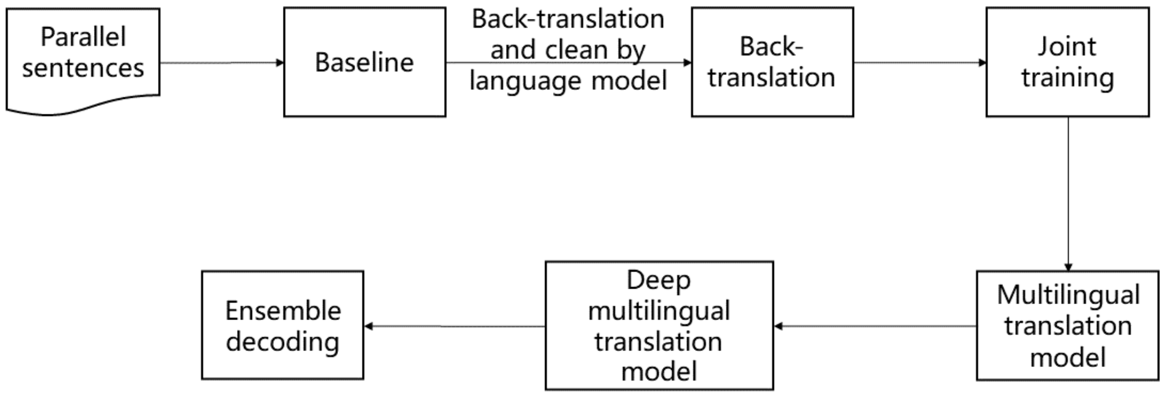

Figure 1: The work flow of GTCOM machine translation competition systems

model	en2hr
baseline	30.15
back translation	32.90
joint training	33.80
deep multilingual translation model	34.93

Table 4: The BLEU score for English to Croatian.

Direction	BLEU	COMET Rank
en2uk	30.8	1
uk2en	43.9	2
cs-uk	36.8	2
uk-cs	31.3	7
en-hr	17.6	2
en-zh	47.7	1

Table 5: The final online automatic evaluation result.

5 Result and Analysis

Table 2, Table 3 and Table 4 show the BLEU score we evaluated on development set for English to/from Ukrainian, Czech to/from Ukrainian and English to Croatian respectively. As shown in the above table, back-translation is still the best data augmentation measure to improve translation quality from the data aspect. Joint training and deep multilingual translation model also show solid improvement in all five directions.

We notice that when adding Russian (a very similar language to Ukrainian) into the multilingual corpus, we did not obtain as much improvement as we expect. This is probably because the original English to Ukrainian data is rich enough and decreased the positive impact of adding Russian data into the multilingual model.

6 Conclusion

This paper describes GTCOM's neural machine translation systems for the WMT22 shared general MT task. We applied 3 major techniques to improve the translation quality: back-translation, joint training, and deep multilingual translation model. With these 3 techniques, the final automatic evaluation matrix is shown in Table 5. Besides BLEU, this year the organizer introduce a new evaluation matrix COMET(Rei et al., 2020) to inspect the translation quality. Our system is ranking 1st place in English->Ukrainian and English->Chinese, 2nd place in Ukrainian-English, Czech ->Ukrainian and English->Croatian, 7th place in Ukrainian->Czech with COMET index.

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 2030 Aritificial Intellegence Research Institute of Global Tone Communication Technology Co., Ltd.[2]

References

Chao Bei and Hao Zong. 2021. GTCOM neural machine translation systems for WMT21. In *Proceedings of the Sixth Conference on Machine Translation*, pages 100–103, Online. Association for Computational Linguistics.

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

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.

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

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 of the Association for Computational Linguistics*, pages 311–318.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

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.