ISTIC's Neural Machine Translation Systems for CCMT' 2023

Shuao Guo, Ningyuan Deng, and Yanqing He*

Research Center of Information Theory and Methodology, Institute of Scientific and Technical Information of China, Beijing, China, 100038
{guosa2021, dengny2022, heyq}@istic.ac.cn

Abstract. This paper describes the technical details of ISTIC's neural machine translation systems for the 19th China Conference on Machine Translation (CCMT' 2023). ISTIC participated in two evaluation tasks of machine translation (MT) The team participated in two machine translation(MT) evaluation tasks: Low resource MT task(Vietnamese↔Chinese, Czech↔Chinese, Lao↔Chinese, Mongolian↔Chinese) and Chinese-Centric Multilingual MT task(Vietnamese↔Chinese, Thailand↔Chinese, Kazakh↔Chinese, Hindi↔Chinese, Uyghur↔Chinese). Context-aware systems and a multilingual system are built for two tasks respectively. The paper mainly illuminates its system framework based on Transformer, data preprocessing methods and some strategies adopted in this system. In addition, the paper evaluates the system performance under different methods.

Keywords: Low resource languages · Multilingual machine translation· Context-aware

1 Introduction

This paper describes building process and technical details of neural machine translation (NMT) systems developed by the Institute of Scientific and Technical Information of China (ISTIC) for the 19th China Conference on Machine Translation (CCMT'2023). ISTIC participated in two evaluation tasks of machine translation(MT): Low Resource MT Task and Chinese-Centric Multilingual MT Task. For Low Resource MT Task, we built context-aware NMT systems for each translation direction of (Vietnamese↔Chinese, Czech↔Chinese, Lao↔Chinese, Mongolian↔Chinese). Contextual information can be incorporated into NMT systems by additional encoders in context-aware system. For Chinese-Centric Multilingual MT Task, we built a multilingual NMT system involving five language pairs and ten translation directions (Vietnamese↔Chinese, Thailand↔Chinese, Kazakh↔Chinese, Hindi↔Chinese, Uyghur↔Chinese). All systems are built based on Transformer architecture. Some corpus preprocessing methods are introduced in this paper. Experiments proved context-aware system can effectively enhance translation quality than baseline system and multilingual MT system has its translation ability over ten translation directions.

2 Data

2.1 Data Size

There are parallel corpus for 8 languages pairs in our NMT systems. All data comes from CCMT2023 evaluation organizer. All systems we submitted belong to constrained systems. Table 1 presents the data size after pre-processing.

Table 1. Data Size

Task	Language pairs	Data size
Chinese-Centric Multilingual MT task	Thailand-Chinese(thai-zh)	530K
	Vietnamese-Chinese(vi-zh)	530K
	Uyghur-Chinese(ug-zh)	535K
	Hindi-Chinese(hi-zh)	500K
	Kazakh-Chinese(kk-zh)	475K
Low resource MT task	Vietnamese-Chinese(vi-zh)	196K
	Czech-Chinese(cs-zh)	197K
	Lao-Chinese(lo-zh)	197K
	Mongolian-Chinese(mn-zh)	193K

2.2 Data Preprocessing

In data preprocessing, team refines the process of parallel corpus processing into four stages. They are character-level preprocessing, tokenization, sentence-level preprocessing and text-level preprocessing respectively(See Fig.1).

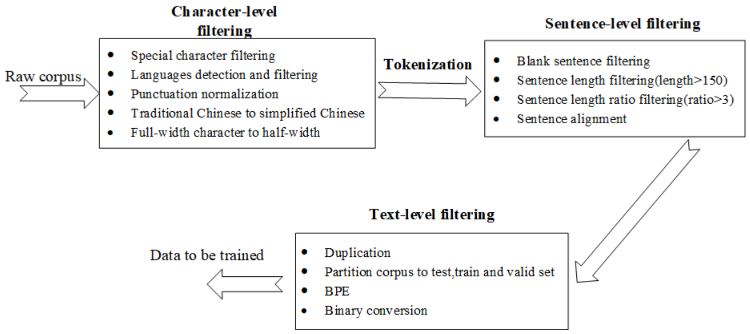

Character-level filtering
Special character filtering
Languages detection and filtering
Punctuation normalization
Traditional Chinese to simplified Chinese
Full-width character to half-width
Raw corpus
Tokenization
Sentence-level filtering
Blank sentence filtering
Sentence length filtering(length>150)
Sentence length ratio filtering(ratio>3)
Sentence alignment
Text-level filtering
Duplication
Partition corpus to test,train and valid set
BPE
Binary conversion
Data to be trained

Fig. 1. Preprocessing operations

Character-level preprocessing. To improve data quality, we filtered special characters such as emoji character, illegal character, the same schedule character before the language pairs, etc[1]. To achieve format uniformity, we used Moses[1] to make punctuation normalization and performed full to half width operations on all characters. We converted traditional Chinese to simplified Chinese with python toolkit Hanziconv[2].

Tokenization. We used different tokenization tools according to different languages. We use Jieba[3] for Chinese tokenization, Underthesea[4] for Vietnam tokenization, Pythainlp[5] for Thailand and Lao tokenization, Kaznlp[6] for Kazakh tokenization, Nltk[7] for Czech tokenization and Monparser[8] for Mongolian tokenization and Asianlp[9] for Hindi tokenization. Since words in Uyghur sentences are connected by spaces and there are no appropriate Uyghur tokenization toolkits, so we directly consider words connected by spaces in Uyghur sentences as its tokens.

Sentence-level preprocessing. We delete language pairs which have at least one blank sentence and use language detection toolkit Py3langid[10] to detect language pairs which don't meet language requirements and delete them. After that, we delete the language pairs whose sentence length is greater than 150 and sentence length ratio is greater than 3.

Text-level preprocessing. Firstly we conduct duplication for each bilingual data and patition the data into validation set, test set and train set. Then we learn Byte-Pair Encoding(BPE)[2] for each language in bilingual data from Low resource MT task and learn a joint BPE over all languages involved in Chinese-Centric Multilingual MT task. BPE merge operations in two tasks are both 32K. At last , we converse data to binary fomat with fairseq-preprocess[11].

3 System

All systems we built for two tasks are all based on standard Transformer[3]. Standard Transformer is an Encoder-Decoder structure(see Figure2), which has 12 blocks containing 6 layers stacked encoders and 6 layers stacked decoders. Model dimension is 521, the number of attention head in every encoder and every decoder is 8, the dimension of feed forward network in every encoder and every decoder is 2048.

[1] https://github.com/moses-smt/mosesdecoder
[2] https://github.com/berniey/hanziconv
[3] https://github.com/fxsjy/jieba
[4] https://github.com/undertheseanlp/underthesea
[5] https://github.com/PyThaiNLP/pythainlp
[6] https://github.com/nlacslab/kaznlp
[7] https://www.nltk.org/
[8] https://github.com/realzoberg/Mon-Parser
[9] https://github.com/sheoguo/Asianlp
[10] https://github.com/adbar/py3langid
[11] https://github.com/facebookresearch/fairseq

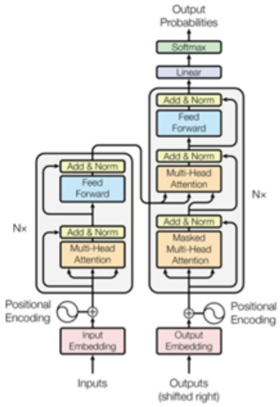


Fig. 2. Transformer architecture

3.1 Systems for Low resource MT task

We built 8 NMT systems according to 8 translation direction specified by Low resource MT task. They are Vietnamese -to-Chinese NMT system, Lao-to-Chinese NMT system, Mongolian-to-Chinese NMT system, Czech-to-Chinese NMT system, Chinese-to-Vietnamese NMT system, Chinese-to-Lao NMT system, Chinese-to-Mongolian NMT system and Chinese-to-Czech NMT system respectively. All systems in this task are context-aware NMT systems with multi encoders based on Transformer-base architecture. Context-aware NMT is a model to incorporate contextual information into NMT[4]. In this model multi-encoder can take the surrounding sentences as the context and encode them by an additional neural networks. There are two methods of integrating the context into NMT, they are outside integration[5] and inside integration[6]. For outside integration, as Figure3 shows, the representations of the context

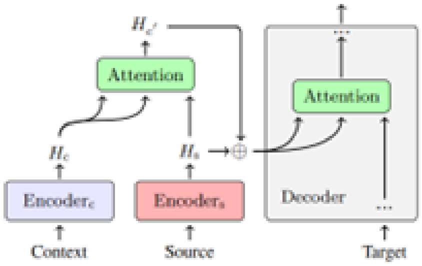


Fig. 3. Outside integration

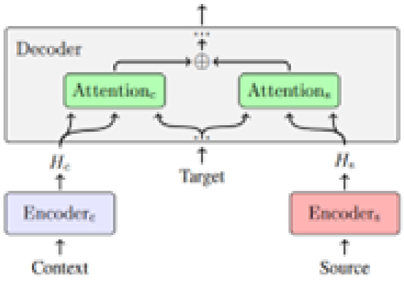


Fig. 4. Inside Integration

and the current sentence are firstly transformed into a new representation by an attention network. Then the attention output and the source sentence representation are fused by a gated sum. For inside integration, as Figure4 shows, decoder can attend to two encoders respectively. Then, the gating mechanism inside the decoder is employed to obtain the fusion vector. In our experments, the context we use to integrate is source language sentences from train set for each translation direction.

3.2 System for Chinese-Centric Multilingual MT task

We built a multilingual NMT system for Chinese-Centric Multilingual MT task. The multilingual NMT model uses a shared encoder and a shared decoder for Vietnamese, Thailand, Hindi, Kazakh, Uyghur and Chinese. The whole multilingual system is based on multilingual Transformer (mTransformer)[7]. mTransformer has the same encoder-decoder architecture as standard Tranformer but instead introduces an lanuage identifying token at the beginning of the input sentence(See Figure5). The language identifying token is a label used to represent the language of the sentences in train set.

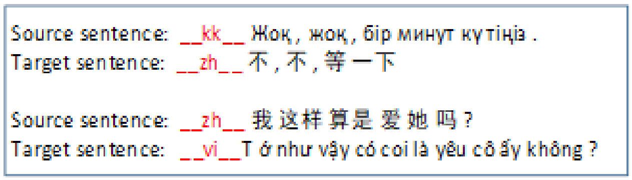
Define our system's multilingual dataset[8]:

$D_{multi} = \{D_{src \to zh}, D_{zh \to tgt}\}, src, tgt \in \{vi, ug, hi, thai, kk\}$

We train our multilingual MT model with the following loss:

$\ell = \sum_{d \in D_{multi}} \sum_{<x,y> \in d} - \log P_\theta(y|x)$

where d is dataset for each language pair in D_{multi}, $< x, y >$ is a sentence pair from s_i to t_i in dataset d, and θ is the model parameter.



Source sentence: __kk__ Жоқ , жоқ , бір минут күтіңіз .
Target sentence: __zh__ 不 , 不 , 等 一 下

Source sentence: __zh__ 我 这 样 算 是 爱 她 吗 ?
Target sentence: __vi__T ớ như vậy có coi là yêu cô ấy không ?

Fig. 5. Sentence examples for lanuage identifying token

4 Experiments

4.1 System environment

Context-aware NMT system and multilingual NMT system are trained in different training environments. Tables 3 shows systems's environment settings.

Table 2. Environment settings

	Context-aware NMT system	Multilingual NMT system
DL framework	Pytorch 1.5.0	Pytorch 1.8.0
NMT framework	Fairseq 0.6.0	Fairseq 0.10.0
Number of GPU	4	8
OS	CentOS Linux release 7.6.1810 (Core)	
GPU	NVIDIA TITAN Xp(12GB)	

4.2 Model and Train

A baseline system and a context-aware system are trained for every translation direction in Low resourece MT task. Standard Transformer architecture is used to train baseline system and context-aware system is trained by both inside and outside integration methods. The trained data of source language in every translation direction is copied as contextual information incorporated into corresponding context-aware system. Every model was optimized with Adam[9] with an initial learning rate of 0.0001, which was multiplied by 0.7 whenever perplexity on the validation set was not improved for three checkpoints. When it was not improved for eight checkpoints, we stopped the training. Dropout probabilities is set to 0.3, the loss function is set to "label smoothed cross entropy" and warm-up steps are set to 4000. Beam search[10] is adopted in decoding stage.

Considering the diversity of dataset volume, transformer_iswlt_de_en architecture is used to train the multilingual NMT system. This architecture belongs to variants of Transformer architecture, the number of attention head in every encoder and every decoder is 4, the dimension of feed forward network in every encoder and every decoder is 1024. The method of temperature sampling is used in model training and sampling_temperature is 4. Other model parameter settings and training process are the same as systems in Low resourece MT task.

4.3 Experiments Results

We use character BLEU[12] to evaluate translation quality with fairseq-score[13].Table 3 shows the NMT systems' BLEU in Low resource MT task. Table 4 shows multilingual system's BLEU in Chinese-Centric Multilingual MT task.

As shown in table 3, no matter inside integration or outside integration, context-aware system's performance is prior to baseline system in 8 translation directions. So we believe context-aware system can enhance the model's performance effectively. The best performance system for every translation direction is choosed to submit to evaluation organizer. As shown in table 4, our multilingual MT system demonstrates its translation ability over ten translation directions. We submitted this multilingual MT system to evaluation organizer.

Table 3. BLEU for systems in Low resource MT task

	Baseline	Inside integration	Outside integration
cs → zh	33.93	34.03	34.11
lo → zh	26.70	26.97	26.83
mn → zh	21.34	21.78	22.09
vi → zh	28.59	28.56	28.86
zh → cs	26.00	27.01	26.83
zh → lo	11.85	12.76	13.01
zh → mn	29.44	30.55	30.42
zh → vi	26.94	27.40	27.53

Table 4. BLEU for the multilingual system

Translation direction	BLEU
kk → zh	33.93
thai → zh	26.70
vi → zh	21.34
ug → zh	28.59
hi → zh	26.00
zh → kk	11.85
zh → thai	29.44
zh → vi	26.94
zh → ug	26.94
zh → hi	26.94

5 Conclusion

In this paper, we describe building process and technical details of translation systems for Low Resource MT Task and Chinese-Centric Multilingual MT Task. In Low Resource MT Task, we construct a NMT system for eight translation directions and our experiments proved context-aware system can effectively enhance translation quality. In Chinese-Centric Multilingual MT Task, we trained a multilingual NMT system with ten translation directions. Experiments proved this multilingual MT system has its translation ability over ten translation directions, but there are imbalance in translation abilities between different language pairs.

Due to the time constraint, we didn't attempt to use LLM pre-training models approaches to enhance NMT model performance. In the future we will further explore such approaches for these two tasks.

6 Acknowledgements

This work was supported by ISTIC Fund (ZD2022-20).

References

1. Guo, S., Guo, H., He, Y., Lan, T. (2022). ISTIC's Thai-to-Chinese Neural Machine Translation System for CCMT' 2022. In: Xiao, T., Pino, J. (eds) Machine Translation. CCMT 2022. Communications in Computer and Information Science, vol 1671. Springer, Singapore.
2. 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.
3. Ashish Vaswani et al.: Attention is all you need. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 30 (NIPS 2017), pp. 5998–6008. Neural Information Processing Systems, Online. (2017)
4. Bei Li et al.: Does multi-encoder help? a case study on context-aware neural machine translation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3512—3518. Association for Computational Linguistics, Online. (2020)
5. Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the Transformer Translation Model with Document-Level Context. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.
6. Elena Voita et al.: Context-aware neural machine translation learns anaphora resolution. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 1264—1274. Association for Computational Linguistics, Online. (2018)
7. 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.
8. Zehui Lin, Liwei Wu, Mingxuan Wang, and Lei Li. 2021. Learning Language Specific Sub-network for Multilingual Machine Translation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 293–305, Online. Association for Computational Linguistics.
9. Kingma, Diederik P. and Jimmy Ba. "Adam: A Method for Stochastic Optimization." CoRR abs/1412.6980 (2014): n. pag.
10. Vijayakumar, Ashwin K , et al. "Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models." (2016).
11. 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
12. 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

13. Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML'17). JMLR.org, 1243–1252.