Exploring the traditional NMT model and Large Language Model for chat translation

Jinlong Yang, Hengchao Shang, Daimeng Wei, Jiaxin Guo,
Zongyao Li, Zhanglin Wu, Zhiqiang Rao, Shaojun Li,
Yuhao Xie, Yuanchang Luo, Jiawei Zheng, Bin Wei, Hao Yang
Huawei Translation Service Center, Beijing, China
{yangjinlong7,shanghengchao,weidaimeng,guojiaxin1,lizongyao,
wuzhanglin2,raozhiqiang,lishaojun18,xieyuhao2,luoyuanchang,
zhengjiawei15,weibin29,yanghao30}@huawei.com

Abstract

This paper describes the submissions of Huawei Translation Services Center(HW-TSC) to WMT24 chat translation shared task on English↔Germany (en-de) bidirection. The experiments involved fine-tuning models using chat data and exploring various strategies, including Minimum Bayesian Risk (MBR) decoding and self-training. The results show significant performance improvements in certain directions, with the MBR self-training method achieving the best results. The Large Language Model also discusses the challenges and potential avenues for further research in the field of chat translation.

1 Introduction

Neural machine translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015; Gehring et al., 2017; Wu et al., 2023) has made substantial progress in recent years, largely due to the adoption of the transformer (Vaswani et al., 2017) architecture. NMT has demonstrated promising translation results across various scenarios. However, research in the field of chat translation remains limited, primarily due to the scarcity of chat data. In prior chat-related tasks, we utilized data from related domains, such as spoken dialogue and subtitles, to augment our translation models, but the outcomes were only mediocre.

Like the preceding two chat shared tasks, the WMT24 chat shard task concentrates on translating conversations between consumers and servers in different languages. We participated in the en-de bidirectional translation task. The en-de bidirectional models we submitted to the WMT22 chat task (Yang et al., 2022) function as our baseline models, leveraging the deep transformer (Dou et al., 2018) architecture. Building on this foundation, we employed the Minimum Bayesian Risk (MBR) strategy to select the optimal translation outcomes, and iterative self-training yielded the best results on the development set.

Beyond traditional NMT models, the emergence of large language model(LLM) has introduced a new paradigm to translation tasks(Wang et al.; Moslem et al., 2023; Guo et al., 2024). Due to its extensive context length and powerful language modeling capabilities, large language models significantly outperform NMT in the translation of lengthy texts and the fluency of translation results. We input the translation output from the NMT model into the LLM as a prompt, allowing the LLM to combine the reference translatio from traditional NMT model to produce an improved translation. However, the comet metric of the LLM's output did not surpass the optimal results of the NMT model.

Recognizing that chat translation is a context-aware task, we conducted a series of context-aware experiments(Wu et al., 2024) using LLMs with WMT and IWSLT document data . We fine-tuned the LLM by constructing streamed translations and contextualized translation data, and translated the development set in the same format. Unfortunately, the results were unsatisfactory.

The structure of this paper is as follows: Section 2 describes our data volume and format for fine-tuning the LLM. The model structure and key methods utilized are presented in Section 3. Section 4 outlines our experiment setting. Results and analysis are presented in Section 5, and we conclude our work in Section 6.

2 Data

2.1 Data Size

All experiments conducted for this task are based on the model developed by our team, as participated in the WMT22 chat shared task. For details on the training data and strategies used for this model, please refer to the system report Yang et al. (2022); Wei et al. (2021). Table 1 and Table 2 list all

Proceedings of the Ninth Conference on Machine Translation, pages 1031–1037
November 15-16, 2024 ©2024 Association for Computational Linguistics

24 train	24 valid	22 valid	22 test
17805	2569	2109	2488

Table 1: Chat shared task en-de bilingual data lines used for training

Dataset	lines	documents
iwslt_2017_ted	209522	1705
news-commentary-v18	449333	11396

Table 2: Document-level data used for LLM related experiments

the data used in this experiment. Based on the prior tasks experience, the contribution of out-domain data to the improvement of translation quality is limited. Therefore, we only further optimize our translation model using the data shown in Table 1, which consists of historical chat tasks. The data in Table 2 is used for fine-tuning the LLM, enabling it to translate context-aware texts and validate the impact of paragraph information on dialogue translation quality.

2.2 Data pre-processing

Since the domain-specific data listed in Table 1 is limited, no special treatment was applied to this portion of the data; it was simply tokenized and input into the NMT model. For the document data in Table 2, we constructed the two formats shown in Table 3 by considering the characteristics of chat tasks, and used them to fine-tune the LLM, separately validating the impact of only preceding information and both preceding and context information on chat translation quality.

In the format of streamlined translation, during each translation session, only preceding information is visible. The LLM generates results based on this preceding information and the previews translation output, resulting in a translation that leans more towards the style of the reference.

In the context-aware translation format, during each translation session, preceding and following N sentences are provided along with the output of the NMT model, guiding the LLM to combine context information to produce a more natural translation.

3 System Overview

3.1 Model

The baseline models for WMT24 chat task use the Transformer-Big architecture. Deep transformer is an improvement of Transformer, which increases the number of encoder layers and uses pre-layer-normalization to further improve model performance. Therefore, in this task, we adopt the following model architecture:

- Deep 25-6 large Model: This model features 25-layer encoder, 6-layer decoder, 1024 dimensions of word vector, 4096 domensions of FFN, 16-head self-attention, and pre-layer-normalization.

For experiments related to large language model, we choose llama2-8b as the base.

3.2 MBR Decoding

Minimum Bayesian Risk (MBR) decoding was initially introduced during the era of statistical machine translation(Kumar and Byrne, 2004; Jinnai et al., 2024). This strategy calculates the output with the minimum expected error among multiple candidates, rather than simply selecting the result with the highest probability during the decoding process. In our experimental approach, we utilize the outputs of 10 distinct models as candidates. These candidates are then used to score each other's comet, and the candidate with the highest average comet is chosen as the final output. Algorithm 1 show the detail.

3.3 Regularized Dropout

Regularized Dropout (R-Drop) [1](Liang et al., 2021) presents a simple yet more effective approach to regulate the training inconsistency caused by dropout (Srivastava et al., 2014). Specifically, during each mini-batch training, each data sample is processed twice through the forward pass, with each pass utilizing a distinct sub-model and randomly dropping out some hidden units. R-Drop minimizes the bidirectional Kullback-Leibler (KL) divergence (van Erven and Harremos, 2014) between the two distributions outputted by the two sub-models for the same data sample, thereby regulating the outputs of two sub-models randomly sampled from dropout for each data sample in training. This method effectively alleviates the inconsistency between the training and inference stages.

3.4 Self-Training

Self-Training(ST) (Imamura and Sumita, 2018), also known as forward translation (FT) (Wu et al., 2019), typically involves utilizing a forward NMT

[1] https://github.com/dropreg/R-Drop

Streaming Translation Data Format
Natural English: <src1>, Translated German: <mt1>, Natural German:<ref1>
Natural German: <src2>, Translated English: <mt2>, Natural English:<ref2>
Natural English: <src3>, Translated German: <mt3>, Natural German:<ref3>
Translate the following sentence into German with a style bias towards Natural: Natural English: <src4>, Translated German: <mt4>, Natural German: <ref4>
Context-aware Translation Data Format
Natural English: <src1>, Translated German: <mt1>
Natural German: <src2>, Translated English: <mt2>
Natural English: <src3>, Translated German: <mt3>
Natural German: <src4>, Translated English: <mt4>
Natural English: <src5>, Translated German: <mt5>
Translate the following sentence into German with a style bias towards Natural: Natural English: <src3>, Natural German: <ref3>

Table 3: LLM Supervised fine-tuning(SFT) data format

Algorithm 1 MBR decoding algorithm

```
Input:
    The set of translation candidates file, $MT_n$;
    The source text file, $SRC$;
    Comet metric model, $M_{comet}$;
Output: final translation output
initialize output list $out[]$
for each $line \in [MT_1, ..., MT_n, SRC]$ do
    initialize $tmp\_max\_comet = 0$
    initialize $candidate\_mt = ''$
    for each $candidate \in [mt_1, mt_2, ..., mt_n]$ do
        let each $mt_x$ as ref, $candidate$ as mt and calculate the comet score with source text using $M_{comet}$
        $mean\_comet = \frac{\sum_{x=1}^{n} comet_x}{n}$
        if $mean\_comet > tmp\_max\_comet$ then
            $tmp\_max\_comet = mean\_comet$
            $candidate\_mt = candidate$
        end if
    end for
    out.append($candidate\_mt$)
end for
return out
```

model to translate source-side monolingual data into target-side text, thereby generating synthetic bilingual data. The generated data is then employed to train the forward translation model. Typically, beam search (Freitag and Al-Onaizan, 2017) is applied for forward translation. In our experimental approach, we set the beam size to 4. Furthermore, we utilized the MBR selection results as self-training data, which led to the best results on the validation set.

3.5 Back Translation

Back-translation (Edunov et al., 2018; Wei et al., 2023) is acknowledged as a highly effective data augmentation strategy to boost NMT model performance. Unlike forward translation, back-translation converts target-side monolinguals into source-side text, thereby producing synthetic parallel corpora. Numerous back-translation techniques have been explored, with sampling (Graça et al., 2019), noise (Edunov et al., 2018), and tagged back-translation (Caswell et al.) demonstrating superior results. In our experimental setup, we opted for sampling back-translation.

3.6 Model Averaging

Model averaging (Dormann et al., 2018) is a widely utilized technique to enhance translation quality. Typically, models (in our experiment, 5 models) that exhibit the highest performance on the development set are chosen for parameter averaging, which leads to substantial improvements.

3.7 LLM Few-shot Prompting

Although large language models exhibit impressive zero-shot capabilities, they still struggle with more complex tasks in the zero-shot setting. To address this, few-shot prompting can be employed as a technique for in-context learning, where demonstrations are provided in the prompt to guide the model towards enhanced performance. In our approach, we provide 5 reference translations to assist the large language model in producing superior results.

3.8 LLM SFT with LoRA

LLM SFT (Supervised Fine-Tuning) is a technique for fine-tuning large language models using specific datasets, which effectively enhances the performance of large language models on tasks such as text generation, machine translation, or sentiment analysis. LoRA (Low-Rank Adaptation)(Hu et al., 2022) is a technique that reduces the computational burden during large language model training by decreasing the number of model parameters through matrix decomposition. This technique maintains performance while lowering computational and memory requirements. By applying LoRA, large language models can perform better under limited computational resources, reducing training costs and resource consumption.

4 Experiment Setting

During the NMT model training phase, we use Pytorch-based Fairseq[2] (Ott et al., 2019) open-source framework as our benchmark system. Each model is trained using 8 GPUs with a batch size of 2048. The update frequency is 4 and the learning rate is 5e-4. The label smoothing rate is set to 0.1, the warm-up steps to 4000, and the dropout to 0.3. Adam optimizer (Kingma and Ba, 2015) with $\beta 1$=0.9 and $\beta 2$=0.98 is also used. Beyond that, we have configured the hyper parameter reg-alpha of the R-Drop technique to a value of 5. In the evaluation phase, We employ the official automatic evaluation scripts and primarily base our model and result selection on the comet metric(Rei et al., 2022)[3].

In the experiments related to large models, we utilize the open-source model llama2_8b_instruct from Meta and the training scripts from HF to train our models, setting the max_seq_length to 1024. For inference on large models, we employ the vllm tool.

5 Result and Analysis

Table 4 displays the results of the official test set, ranked according to the comet-22 score, where our system achieved the top position in comet-22, chrF, and BLEU metrics.

The primary results we submit are obtained by translating the source text of the test set with multiple NMT models, selecting the optimal output using MBR strategy, then training on the best models from the validation set using self-training method. The models are averaged over 5 epochs before being used to translate the test set to yield the final results.

5.1 Sentence-level NMT

In the previous chat tasks, we have tried various strategies to optimize the model, and the results from the validation set indicate that the baseline model from 2022 was already sufficiently powerful. On this basis, we combined this year's training set, the 2022 validation and test sets, and conducted BT and ST reinforcement strategies, only in the direction of translation from English to German has there been a noticeable improvement. The results shown in Table 5.

To further improve the results, we attempted the MBR decoding strategy, generating 10 alternative outputs for the validation set using different NMT models in previous steps. These outputs were scored using comet, and the output with the lowest Bayesian risk was selected as the final result. The results in Table 5 indicate that improvement was only seen in the en→de direction. Further, we utilized the MBR results to perform another ST on each direction, ultimately achieving the best results in both directions in the validation set. The reason for the improvement we observed is that the MBR algorithm can integrate the capabilities of multiple models. When performing self training, it essentially utilizes the optimal results of multiple models for a round of knowledge distillation.

5.2 Document-level MT with LLM

According to the test results shown in Table 6, on the chat task valid set, the results of LLM (Large Language Model) are significantly worse than sentence-level under both comet or doc-comet metrics. The few-shot capabilities of LLM is in-

[2] https://github.com/facebookresearch/fairseq

[3] https://github.com/Unbabel/COMET

team	comet↑	chrf↑	bleu↑	context-comet-qe↑
HW-TSC	**93.4**	**83.2**	**69.8**	0.221
unbabel+it	92.9	78.2	62	**0.253**
clteam	91.3	71.9	53	0.204
ADAPT	90.8	72.1	55	0.168
DCUGenNLP	90.8	71.2	53	0.188
baseline	89.8	70.8	51.1	0.173
SheffieldGate	89.4	67.5	45.2	0.177

Table 4: The official automatic evaluation results of the test set, ranked based on the COMET-22 score

System	en→de	de→en
baseline	86.76	85.88
22_denoise	90.06	91.42
+ ST	91.23	91.40
+ ST&BT	91.23	91.53
+ MBR ST	**91.91**	**91.86**
MBR	91.75	90.87

Table 5: Sentence-level NMT results.

deed far better than zero-shot, but it still falls short of sentence-level results. After using the document-level data for LLM SFT, the results became even worse. We analyzed that the reason is the large domain shift, as the IWSLT and WMT datasets we used are far from the domain of the chat task.

To validate the capability of LLM in translating document-level content, we tested the results on the iwslt2017 en-de document-level test set. The results in the right half of Table 6 demonstrate that LLM's few-shot capability surpassed that of the chat task's sentence-level model on this test set. Further, by fine-tuning the large model with document-level data, we obtained better results.

Comparing the results of stream translation and context-aware translation, we originally expected context-aware format data to yield better results because the model could refer to contextual information during translation. However, we analyzed that stream translation sees the previous step's translation result each time, which is more consistent with the translation style of large model. On the contrary, context-aware requires input of the reference MT result from sentence-level model in one go, which is less consistent with the style of large model, causing the model to fail to effectively utilize these information.

6 Conclusion

This paper presents the submissions of HW-TSC to the WMT 2024 Chat Translation Shared Task. For both direction in en↔de translation task, we perform experiments with a series of training strategies. The results show that MBR self-training achieves the best results. In the future, we will continue to explore the applicability of MBR strategy mentioned in this paper.

Beyond that, due to time constraints, further fine-tuning of large language models using chat task data was not conducted to assess its performance. Additionally, there is room for continued exploration of the translation capabilities of large language models.

References

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Isaac Caswell, Ciprian Chelba, and David Grangier. Tagged back-translation. *WMT 2019*, page 53.

Carsten F. Dormann, Justin M. Calabrese, Gurutzeta Guillera-Arroita, Eleni Matechou, Volker Bahn, Kamil Bartoń, Colin M. Beale, Simone Ciuti, Jane Elith, Katharina Gerstner, Jérôme Guelat, Petr Keil, José J. Lahoz-Monfort, Laura J. Pollock, Björn Reineking, David R. Roberts, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Simon N. Wood, Rafael O. Wüest, and Florian Hartig. 2018. Model averaging in ecology: a review of bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs*, 88(4):485–504.

Zi-Yi Dou, Zhaopeng Tu, Xing Wang, Shuming Shi, and Tong Zhang. 2018. Exploiting deep representations for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4253–4262.

	chat en→de		chat de→en		iwslt en→de		iwslt de→en	
System	comet	d-comet	comet	d-comet	comet	d-comet	comet	d-comet
Baseline	86.76	79.40	85.88	79.77	-	-	-	-
MBR ST	**91.91**	**85.41**	**91.86**	**86.21**	84.70	77.55	87.05	80.81
llama2_8b_instruce	87.56	79.99	86.96	80.89	82.53	75.07	86.21	79.74
+ 5 best	90.05	83.34	88.72	83.11	85.10	77.98	87.20	81.04
stream	85.47	78.50	83.98	78.80	**85.69**	**78.91**	**87.45**	**81.73**
context-aware	81.82	73.81	83.89	77.37	84.80	77.51	86.65	80.43

Table 6: The results of LLM MT

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60.

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*, pages 1243–1252.

Miguel Graça, Yunsu Kim, Julian Schamper, Shahram Khadivi, and Hermann Ney. 2019. Generalizing back-translation in neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 45–52.

Jiaxin Guo, Hao Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, and Xiaoyu Chen. 2024. A novel paradigm boosting translation capabilities of large language models.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Kenji Imamura and Eiichiro Sumita. 2018. Nict self-training approach to neural machine translation at nmt-2018. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 110–115.

Yuu Jinnai, Tetsuro Morimura, Ukyo Honda, Kaito Ariu, and Kenshi Abe. 2024. Model-based minimum Bayes risk decoding for text generation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 22326–22347. PMLR.

Diederik P Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-Drop: Regularized Dropout for Neural Networks. *arXiv e-prints*, page arXiv:2106.14448.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. Adaptive machine translation with large language models.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2*, pages 3104–3112.

Tim van Erven and Peter Harremos. 2014. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. Document-level machine translation with large language models.

Daimeng Wei, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Minghan Wang, Lizhi Lei, Min Zhang, Hao Yang, and Ying Qin. 2021. HW-TSC's participation in the WMT 2021 news translation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 225–231, Online. Association for Computational Linguistics.

Daimeng Wei, Zhanglin Wu, Hengchao Shang, Zongyao Li, Minghan Wang, Jiaxin Guo, Xiaoyu Chen, Zhengzhe Yu, and Hao Yang. 2023. Text style transfer back-translation.

Lijun Wu, Yiren Wang, Yingce Xia, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019. Exploiting monolingual data at scale for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4207–4216.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *arXiv preprint arXiv:2401.06468*.

Zhanglin Wu, Zongyao Li, Daimeng Wei, Hengchao Shang, Jiaxin Guo, Xiaoyu Chen, Zhiqiang Rao, Zhengzhe Yu, Jinlong Yang, Shaojun Li, Yuhao Xie, Bin Wei, Jiawei Zheng, Ming Zhu, Lizhi Lei, Hao Yang, and Yanfei Jiang. 2023. Improving neural machine translation formality control with domain adaptation and reranking-based transductive learning. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 180–186, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Jinlong Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, Xiaoyu Chen, Zhengzhe Yu, Zhiqiang Rao, Shaojun Li, Zhanglin Wu, Yuhao Xie, Yuanchang Luo, Ting Zhu, Yanqing Zhao, Lizhi Lei, Hao Yang, and Ying Qin. 2022. HW-TSC translation systems for the WMT22 chat translation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 962–968, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.