Choose the Final Translation from NMT and LLM hypotheses Using MBR Decoding: HW-TSC's Submission to the WMT24 General MT Shared Task

Zhanglin Wu, Daimeng Wei, Zongyao Li, Hengchao Shang, Jiaxin Guo, Shaojun Li, Zhiqiang Rao, Yuanchang Luo, Ning Xie, Hao Yang
Huawei Translation Service Center, Beijing, China
{wuzhanglin2,weidaimeng,lizongyao,shanghengchao,guojiaxin1,
lishaojun18,raozhiqiang,nicolas.xie,yanghao30}@huawei.com

Abstract

This paper presents the submission of Huawei Translate Services Center (HW-TSC) to the WMT24 general machine translation (MT) shared task, where we participate in the English to Chinese (en→zh) language pair. Similar to previous years' work, we use training strategies such as regularized dropout, bidirectional training, data diversification, forward translation, back translation, alternated training, curriculum learning, and transductive ensemble learning to train the neural machine translation (NMT) model based on the deep Transformer-big architecture. The difference is that we also use continue pre-training, supervised fine-tuning, and contrastive preference optimization to train the large language model (LLM) based MT model. By using Minimum Bayesian risk (MBR) decoding to select the final translation from multiple hypotheses for NMT and LLM-based MT models, our submission receives competitive results in the final evaluation.

1 Introduction

Machine translation (MT) (Brown et al., 1990) predominantly utilizes transformer encoder-decoder architectures (Vaswani et al., 2017), which is evident in prominent models such as NLLB-200 (Costa-jussà et al., 2022), M2M100 (Fan et al., 2021), and MT5 (Xue et al., 2021). Significant research effort has been devoted to task-specific neural machine translation (NMT) models (Wei et al., 2022; Wu et al., 2023b) trained in a fully supervised manner with large volumes of parallel data. Their performance has been enhanced through techniques such as regularized dropout (Wu et al., 2021), bidirectional training (Ding et al., 2021), data diversification (Nguyen et al., 2020), forward translation (Abdulmumin, 2021), back translation (Sennrich et al., 2016), alternated training (Jiao et al., 2021), curriculum learning (Zhang et al., 2019), and transductive ensemble learning (Wang et al., 2020b).

The emergence of decoder-only large language models (LLMs) such as the GPT series (Wu et al., 2023a; Achiam et al., 2023), Mistral (Jiang et al., 2023), and LLaMA (Touvron et al., 2023a,b) shows remarkable efficacy in various NLP tasks, providing a fresh perspective on the MT task. Recent studies (Hendy et al., 2023; Jiao et al., 2023) indicate that larger LLMs such as GPT-3.5 (175B) and GPT-4 exhibit strong translation abilities. However, the performance of smaller-sized LLMs (7B or 13B) still falls short when compared to conventional NMT models (Zhu et al., 2024). Therefore, there are studies (Yang et al., 2023; Zeng et al., 2024) intend to enhance the translation performance for these smaller LLMs, but their improvements are relatively modest, primarily due to the predominant pre-training of LLMs on English-centric datasets, resulting in limited linguistic diversity. Addressing this limitation, Xu et al. (Xu et al., 2023) initially continue pre-training (CPT) LLaMA-2 (Touvron et al., 2023b) with extensive non-English monolingual data to enhance their multilingual abilities, and then perform supervised fine-tuning (SFT) with high-quality parallel data to instruct the model to generate translations. Nonetheless, the performance still lags behind leading translation models such as GPT-4 and WMT competition winners. Subsequently, Xu et al. (Xu et al., 2024) bridged this gap by further fine-tuning the LLM-based MT model using contrast preference optimization (CPO).

Ensembling (Zhou et al., 2002) has a long history in machine learning, being well known for leveraging multiple complementary systems to improve performance on a given task and provide good/robust generalization. Minimum Bayesian risk (MBR) (Finkelstein and Freitag, 2023; Farinhas et al., 2023) decoding has successfully improved translation quality using task-specific NMT models, and subsequently it has also been shown to be suitable for LLM-based MT models.

Proceedings of the Ninth Conference on Machine Translation, pages 155–164

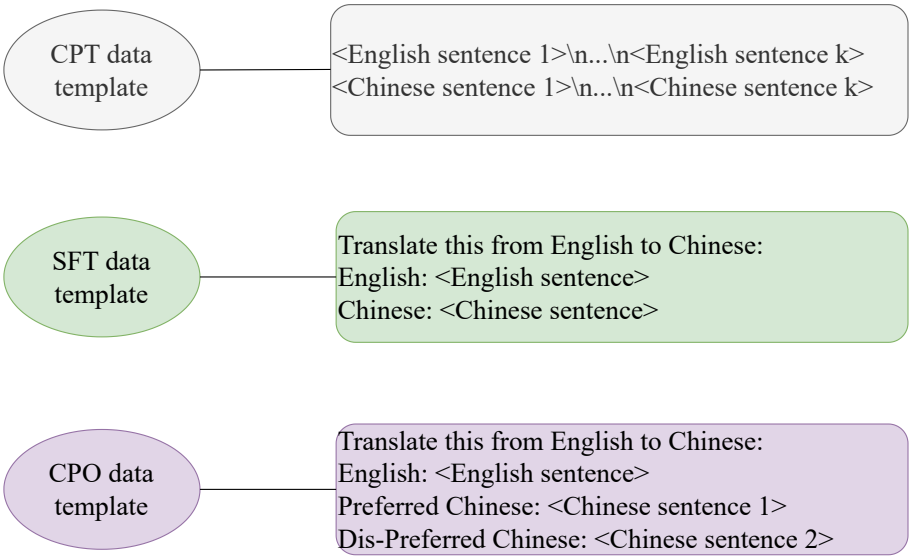

Figure 1: CPT, SFT and CPO data templates used for LLM-based MT training.

For the WMT24 general MT shared task, we participate in the en→zh language pair. Similar to previous years' work (Wei et al., 2021, 2022; Wu et al., 2023b), we use training strategies such as regularized dropout (Wu et al., 2021), bidirectional training (Ding et al., 2021), data diversification (Nguyen et al., 2020), forward translation (Abdulmumin, 2021), back translation (Sennrich et al., 2016), alternated training (Jiao et al., 2021), curriculum learning (Zhang et al., 2019), and transductive ensemble learning (Wang et al., 2020b) to train NMT models based on the deep transformer-big architecture. In addition, we use CPT, SFT and CPO methods to train LLM-based MT models. Finally, we use MBR decoding to select the final translation from multiple hypotheses of NMT and LLM-based MT models.

2 Data

2.1 Data Source

We obtain bilingual and monolingual data from ParaCrawl v9, News Commentary v18.1, Wiki Titles v3, UN Parallel Corpus V1.0, CCMT Corpus, WikiMatrix, News Crawl and Common Crawl data sources. The amount of data we used for training NMT and LLM-based MT models is shown in Table 1. It should be noted that in order to obtain better translation performance in the general domain, we mix the monolingual data from Common Crawl and News Crawl.

language pairs	bitext data	monolingual data
en→zh	25M	en: 50M, zh: 50M

Table 1: Bilingual and monolingual used for training NMT and LLM-based MT models.

2.2 NMT Data Pre-processing

Our data pre-processing methods for NMT include:

- Remove duplicate sentences or sentence pairs.
- Convert full-width symbols to half-width.
- Use fasttext[1] (Joulin et al., 2016) to filter other language sentences.
- Use jieba[2] to pre-segment Chinese sentences.
- Use mosesdecoder[3] (Koehn et al., 2007) to normalize English punctuation.
- Filter out sentences with more than 150 words.
- Use fast-align (Dyer et al., 2013) to filter sentence pairs with poor alignment.
- Sentencepiece[4] (SPM) (Kudo and Richardson, 2018) is used to perform subword segmentation, and the vocabulary size is set to 32K.

Since there may be some semantically dissimilar sentence pairs in bilingual data, we use LaBSE[5]

[1] https://github.com/facebookresearch/fastText
[2] https://github.com/fxsjy/jieba
[3] https://github.com/moses-smt/mosesdecoder
[4] https://github.com/google/sentencepiece
[5] https://huggingface.co/sentence-transformers/LaBSE

(Feng et al., 2022) to calculate the semantic similarity of each bilingual sentence pair, and exclude bilingual sentence pairs with a similarity score lower than 0.7 from our training corpus.

2.3 LLM-based MT Data Pre-processing

The training of the LLM-based MT model requires three stages: CPT, SFT and CPO. As shown in Figure 1, the training data templates of the LLM-based MT model in these three stages are different.

In the CPT stage, considering that most LLMs are trained on English-dominated data, we using Chinese and English monolinguals for CPT to improve LLM's proficiency in Chinese. To preserve the long-context modeling capability of LLM, we concatenate multiple sentences into a long text with no more than 4096 words, and preferentially concatenate sentences from the same document.

In the SFT stage, drawing inspiration from the recognized significance of data quality in other applications (Zhou et al., 2024; Maillard et al., 2023), we fine-tune the model with high-quality parallel data. In order to obtain high-quality parallel data, we use cometkiwi model [6] (Rei et al., 2022) to calculate the score of bilingual data on the en→zh language pair, and then retain bilingual data with a cometkiwi score greater than 0.8.

In the CPO stage, to learn an objective that fosters superior translations and rejects inferior ones, access to labeled preference data is essential, yet such data is scarce in machine translation. The following describes our process of constructing the triplet preference data required for CPO training. First, we randomly sample 50,000 data from high-quality bilingual data. Then, we use the NMT model to obtain N-best (N=10) hypotheses based on beam search decoding, and then use the cometda model[7] (Rei et al., 2020) to calculate the score of each hypothesis, select the hypothesis with the highest score as the preferred translation, and select the hypothesis with the lowest score as the dis-preferred translation.

3 NMT System

3.1 System Overview

Transformer is the state-of-the-art model structure in recent NMT evaluations. There are two parts of research to improve this kind: the first part uses wide networks (eg: Transformer-Big (Vaswani et al., 2017)), and the other part uses deeper language representations (eg: Deep Transformer (Wang et al., 2019)). For the WMT24 general MT shared task, we combine these two improvements, adopting the Deep Transformer-Big (Wei et al., 2022; Wu et al., 2023b) model structure to train the NMT system. Deep Transformer-Big uses pre-layer normalization, features 25-layer encoder, 6-layer decoder, 16-heads self-attention, 1024-dimensional word embedding and 4096-dimensional FFN embedding.

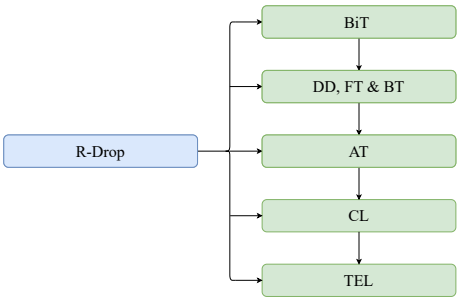

Figure 2: The overall training flow of NMT system.

Fig. 2 shows the overall training flow of NMT system. We use training strategies such as regularized dropout (R-Drop) (Wu et al., 2021), bidirectional training (BiT) (Ding et al., 2021), data diversification (DD) (Nguyen et al., 2020), forward translation FT) (Abdulmumin, 2021), back translation (BT) (Sennrich et al., 2016), alternated training (AT) (Jiao et al., 2021), curriculum learning (CL) (Zhang et al., 2019), and transductive ensemble learning (TEL) (Wang et al., 2020b) for training.

3.2 Regularized Dropout

Regularized Dropout (R-Drop)[8] (Wu et al., 2021) is a simple yet more effective alternative to regularize the training inconsistency induced by dropout (Srivastava et al., 2014). Concretely, in each mini-batch training, each data sample goes through the forward pass twice, and each pass is processed by a different sub model by randomly dropping out some hidden units. R-Drop forces the two distributions for the same data sample outputted by the two sub models to be consistent with each other, through minimizing the bidirectional Kullback-Leibler (KL) divergence (Van Erven and Harremos, 2014) between the two distributions. That is, R-Drop regularizes the outputs of two sub models ran-

[6] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

[7] https://huggingface.co/Unbabel/wmt20-comet-da

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

domly sampled from dropout for each data sample in training. In this way, the inconsistency between the training and inference stage can be alleviated.

3.3 Bidirectional Training

Many studies have shown that pre-training can transfer the knowledge and data distribution, hence improving the model generalization. Bidirectional training (BiT) (Ding et al., 2021) is a simple and effective pre-training method for NMT. Bidirectional training is divided into two stages: (1) bidirectionally updates model parameters, and (2) tune the model. To achieve bidirectional updating, we only need to reconstruct the training samples from "src→tgt" to "src→tgt & tgt→src" without any complicated model modifications. Notably, BiT does not require additional parameters or training steps and only uses parallel data.

3.4 Data Diversification

Data Diversification (DD) (Nguyen et al., 2020) is a data augmentation method to boost NMT performance. It diversifies the training data by using the predictions of multiple forward and backward models and then merging them with the original dataset which the final NMT model is trained on. DD is applicable to all NMT models. It does not require extra monolingual data, nor does it add more parameters. To conserve training resources, we only use one forward model and one backward model to diversify the training data.

3.5 Forward Translation

Forward translation (FT) (Abdulmumin, 2021), also known as self-training, is one of the most commonly used data augmentation methods. FT has proven effective for improving NMT performance by augmenting model training with synthetic parallel data. Generally, FT is performed in three steps: (1) randomly sample a subset from the large-scale source monolingual data; (2) use a "teacher" NMT model to translate the subset data into the target language to construct the synthetic parallel data; (3) combine the synthetic and authentic parallel data to train a "student" NMT model.

3.6 Back Translation

An effective method to improve NMT with target monolingual data is to augment the parallel training data with back translation (BT) (Sennrich et al., 2016; Wei et al., 2023). There are many works expand the understanding of BT and investigates a number of methods to generate synthetic source sentences. Edunov et al. (2018) find that back translations obtained via sampling or noised beam outputs are more effective than back translations generated by beam or greedy search in most scenarios. Caswell et al. (2019) show that the main role of such noised beam outputs is not to diversify the source side, but simply to tell the model that the given source is synthetic. Therefore, they propose a simpler alternative strategy: Tagged BT. This method uses an extra token to mark back translated source sentences, which generally outperforms noised BT (Edunov et al., 2018). For better joint use with FT, we use sampling back translation (ST) (Edunov et al., 2018).

3.7 Alternated Training

While synthetic bilingual data have demonstrated their effectiveness in NMT, adding more synthetic data often deteriorates translation performance since the synthetic data inevitably contains noise and erroneous translations. Alternated training (AT) (Jiao et al., 2021) introduce authentic data as guidance to prevent the training of NMT models from being disturbed by noisy synthetic data. AT describes the synthetic and authentic data as two types of different approximations for the distribution of infinite authentic data, and its basic idea is to alternate synthetic and authentic data iteratively during training until the model converges.

3.8 Curriculum Learning

A practical curriculum learning (CL) (Zhang et al., 2019) method should address two main questions: how to rank the training examples, and how to modify the sampling procedure based on this ranking. For ranking, we choose to estimate the difficulty of training samples according to their domain feature (Wang et al., 2020a). The calculation formula of domain feature is as follows, where θ_{in} represents an in-domain NMT model, and θ_{out} represents a out-of-domain NMT model. One thing to note is that we treat domains including news, user-generated (social), conversational, and e-commerce domains as in-domain, and others as out-of-domain. Specifically, we use the WMT22 test set to fine-tune a baseline model, and then use the baseline model and the fine-tuned model as the out-of-domain model and the in-domain model respectively.

$$q(x, y) = \frac{\log P(y|x; \theta_{in}) - \log P(y|x; \theta_{out})}{|y|} \quad (1)$$

For sampling, we adopt a probabilistic CL strategy that leverages the concept of CL in a nondeterministic fashion without discarding the original standard training practice, such as bucketing and mini-batching.

3.9 Transductive Ensemble Learning

Ensemble learning (Garmash and Monz, 2016), which aggregates multiple diverse models for inference, is a common practice to improve the performance of machine learning models. However, it has been observed that the conventional ensemble methods only bring marginal improvement for NMT when individual models are strong or there are a large number of individual models. Transductive Ensemble Learning (TEL) (Zhang et al., 2019) studies how to effectively aggregate multiple NMT models under the transductive setting where the source sentences of the test set are known. TEL uses all individual models to translate the source test set into the target language space and then fine-tune a strong model on the translated synthetic data, which significantly boosts strong individual models and benefits a lot from more individual models.

4 LLM-based MT System

4.1 System Overview

There is recently a surge in research interests in Transformer-based LLMs, such as ChatGPT (Wu et al., 2023a), GPT-4 (Achiam et al., 2023), and LLaMA (Touvron et al., 2023a,b). Benefiting from the giant model size and oceans of training data, LLMs can understand better the language structures and semantic meanings behind raw text, thereby showing excellent performance in a wide range of natural language processing (NLP) tasks. Although the training methodology of LLMs is simple, high computational requirements have limited the development of LLMs to a few players. In order to avoid training LLM from scratch, we chose to conduct research work on the open source Llama2-13b[9] (Touvron et al., 2023b) model. Llama2-13b is an autoregressive language model using an optimized transformer architecture that is pre-trained on 2 trillion tokens of data from publicly available

[9] https://huggingface.co/meta-llama/Llama-2-13b-hf

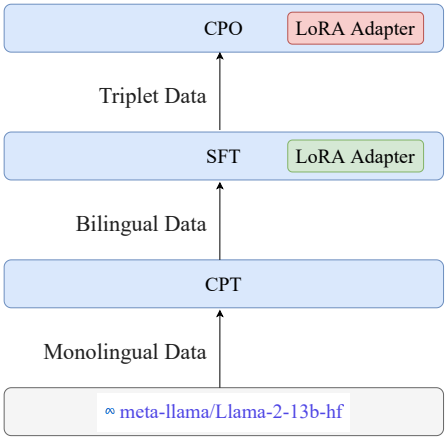

Figure 3: The training flow of LLM-based MT system.

sources. As shown in Figure 3, we train Llama2-13b into a powerful LLM-based MT model through three-stage training of CPT, SFT and CPO.

4.2 Continue Pre-training

LLMs like LLaMA are pre-trained on English-dominated corpora. This potentially explains their inadequate translation performance which necessitates cross-lingual capabilities. To ameliorate this, our first stage is to perform continue pre-training (CPT) on LLM with Chinese and English monolingual data to improve proficiency in Chinese and prevent forgetting of English knowledge. Previous studies also offer some clues that monolingual data help in translation. For instance, guo et al. (Guo et al., 2024) proposed a three-stage training method, which proved that using CPT can improve the performance of MT task in the SFT stage. Note that we use full fine-tuning at this stage.

4.3 Supervised Fine-tuning

LLMs have shown remarkable performance on a wide range of NLP tasks by leveraging in-context learning (Brown et al., 2020). However, this approach exhibits several drawbacks: performance is highly dependent on the quality of examples (Vilar et al., 2023), outputs are plagued by overgeneration (Bawden and Yvon, 2023), and inference cost are greatly increased by processing all input pairs. When parallel data is available, LLMs can perform supervised fine-tuning (SFT) on translation instructions (Li et al., 2024). Drawing inspiration from the recognized significance of data quality in other applications (Zhou et al., 2024),we use the cometkiwi model (Rei et al., 2022) to filter out large amounts of high-quality parallel data. Here, we use efficient lightweight low-rank adaptation (LoRA) fine-

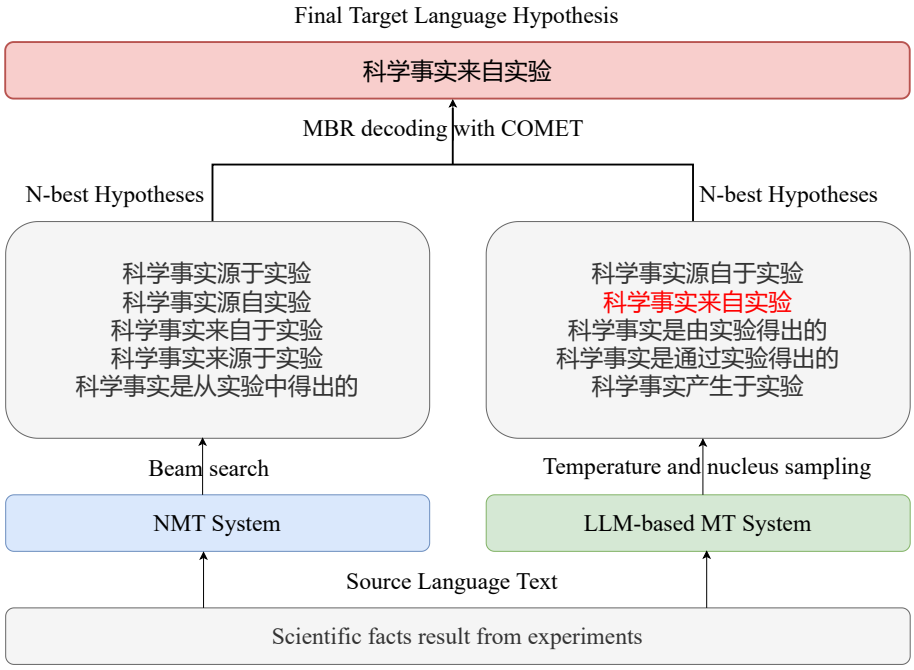

Figure 4: Choose the Final Translation from NMT and LLM hypotheses Using MBR Decoding.

tuning, where we apply LoRA to all modules of feed-forward network.

4.4 Contrastive Preference Optimization

Contrastive Preference Optimization (CPO) (Xu et al., 2024) aims to mitigate two fundamental shortcomings of SFT. First, SFT's methodology of minimizing the discrepancy between predicted outputs and gold-standard references inherently caps model performance at the quality level of the training data. This limitation is significant, as even human-written data, traditionally considered high-quality, is not immune to quality issues. Secondly, SFT lacks a mechanism to prevent the model from rejecting mistakes in translations. While strong translation models can produce high-quality translations, they occasionally exhibit minor errors, such as omitting parts of the translation. Preventing the production of these near-perfect but ultimately flawed translation is essential. To overcome these issues, we introduce CPO to train the LLM-based MT model using specially curated triplet preference data. Here, we construct a high-quality preference data for the WMT24 general MT task, and like the SFT stage, only update the weights of the added LoRA parameters.

4.5 Minimum Bayes Risk Decoding

Minimum Bayesian Risk (MBR) (Kumar and Byrne, 2004; Eikema and Aziz, 2020) decoding aims to find the output that maximizes the expected utility function, which measures the similarity between the hypothesis and the reference. For MT, this could be an automated evaluation metric such as COMET (Rei et al., 2020). Garcia et al. (Garcia et al., 2023) train their own language models, sample multiple hypotheses and choose a final translation using MBR decoding, which has been shown to improve the translation capabilities of task-specific models (Fernandes et al., 2022). Subsequently, Farinhas et al. (Farinhas et al., 2023) find that MBR is also suitable for LLM-based MT. They provide a comprehensive study on ensembling translation hypotheses, proving that MBR decoding is a very effective method and can improve translation quality using a small number of samples. As shown in Figure 4, we simultaneously collect the N-best translations generated by the NMT system based on beam search and the N-best translations generated by the LLM-based MT system based on temperature and nucleus sampling (with t=0.8 and p=0.95), and then use MBR Decoding selects the final translation.

5 Experiment

5.1 Setup

We use the open-source fairseq (Ott et al., 2019) to train NMT models, and then use SacreBLEU

(Post, 2018)[10] and wmt20-comet-da model (Rei et al., 2020) to measure system performance. The main parameters are as follows: each model is trained using 8 GPUs, batch size is 6144, parameter update frequency is 2, and learning rate is 5e-4. The number of warmup steps is 4000, and model is saved every 1000 steps. The architecture we used is described in section 3.1. We adopt dropout, and the rate varies across different training phases. R-Drop is used in model training, and we set λ to 5.

We use Llama2-13B as the backbone model of our LLM-based MT system. In our three-stage training process, the first stage uses full fine-tuning, and the last two stages use LoRA fine-tuning. If LoRA is used, lora_rank is 32, lora_alpha is 64, lora_dropout is 0.05, and lora_modules are "q_proj", "v_proj", "k_proj", "o_proj", "gate_proj", "down_proj", "up_proj". Furthermore, in the first and third stages, we use open-source ALMA [11] for training, while in the second stage, we use open-source llama-recipes [12] for training. The parameters during training are the default configurations of the corresponding codes.

5.2 Results

Tables 2 shows the evaluation results of en→zh NMT systems and LLM-based MT systems on WMT23 general test sets. On NMT systems, we use BiT and R-Drop to build a strong baseline, then use DD, FT and ST for data enhancement, and use AT and CL for more efficient training, and finally use TEL to ensemble multiple models ability. On LLM-based MT systems, we use CPT and SFT to build a strong baseline, and use CPO for further optimization. To ensemble two different types of translation systems, we use MBR decoding to select the final translation, which has been shown to be better than MBR decoding of a single translation system in terms of COMET scores.

5.3 Pre-processing and Post-processing

On the WMT24 general test set, we observe that there are some emoticons and URLs in the source text. To prevent the model from translating them incorrectly, we replace the emoticons and URLs with "Do Not Translate" (DNT) labels in pre-processing, and then restore the DNT labels back in post-processing. By doing so, we can reduce some translation errors for emoticons and URLs.

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

[11] https://github.com/fe1ixxu/ALMA

[12] https://github.com/meta-llama/llama-recipes

WMT23 general test set	BLEU	COMET
NMT baseline (BiT & R-Drop)	54.24	0.6289
+ DD, FT & ST	56.33	0.6580
+ AT	57.03	0.6648
+ CL	58.58	0.6830
+ TEL	**59.34**	0.6928
+ NMT MBR	58.88	0.7178
LLM-based MT baseline (CPT & SFT)	52.18	0.6553
+ CPO	53.09	0.6907
+ LLM-based MT MBR	52.16	0.7102
+ NMT & LLM-based MT MBR	56.41	**0.7234**

Table 2: BLEU and COMET scores of en→zh NMT systems and LLM-based MT systems.

6 Conclusion

This paper presents the submission of HW-TSC to the WMT24 general MT Task. On the one hand, we use training strategies such as R-Drop, BiT, DD, FT, BT, AT, CL, and TEL to train the NMT system based on the deep Transformer-big architecture. On the other hand, we use CPT, SFT, and CPO to train the LLM-based MT system. Finally, we use MBR decoding to select the final translation result from the hypotheses generated by these two systems. By using these enhancement strategies, our submission achieved a competitive result in the final evaluation. Relevant experimental results also demonstrate the effectiveness of our strategies.

References

Idris Abdulmumin. 2021. Enhanced back-translation for low resource neural machine translation using self-training. In *Information and Communication Technology and Applications: Third International Conference, ICTA 2020, Minna, Nigeria, November 24–27, 2020, Revised Selected Papers*, volume 1350, page 355. Springer Nature.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of bloom. *arXiv preprint arXiv:2303.01911*.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind

Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Liang Ding, Di Wu, and Dacheng Tao. 2021. Improving neural machine translation by bidirectional training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3278–3284.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

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*, page 489. Association for Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2020. Is map decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

António Farinhas, José de Souza, and André FT Martins. 2023. An empirical study of translation hypothesis ensembling with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11956–11970.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Patrick Fernandes, António Farinhas, Ricardo Rei, José GC de Souza, Perez Ogayo, Graham Neubig, and André FT Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412.

Mara Finkelstein and Markus Freitag. 2023. Mbr and qe finetuning: Training-time distillation of the best and most expensive decoding methods. *arXiv preprint arXiv:2309.10966*.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. In *International Conference on Machine Learning*, pages 10867–10878. PMLR.

Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418.

Jiaxin Guo, Hao Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, and Xiaoyu Chen. 2024. A novel paradigm boosting translation capabilities of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 639–649.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Rui Jiao, Zonghan Yang, Maosong Sun, and Yang Liu. 2021. Alternated training with synthetic and authentic data for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1828–1834.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Parrot: Translating during chat using large language models tuned with human translation and feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15009–15020.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open

source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *EMNLP 2018*, page 66.

Shankar Kumar and Bill 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.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Transactions of the Association for Computational Linguistics*, 11:576–592.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756.

Xuan-Phi Nguyen, Shafiq Joty, Wu Kui, and Ai Ti Aw. 2020. Data diversification: a simple strategy for neural machine translation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 10018–10029.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC de Souza, Taisiya Glushkova, Duarte M Alves, Alon Lavie, et al. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. *WMT 2022*, page 634.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96. Association for Computational Linguistics (ACL).

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting palm for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. 2019. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Wei Wang, Ye Tian, Jiquan Ngiam, Yinfei Yang, Isaac Caswell, and Zarana Parekh. 2020a. Learning a multi-domain curriculum for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7711–7723.

Yiren Wang, Lijun Wu, Yingce Xia, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. 2020b. Transductive ensemble learning for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6291–6298.

Daimeng Wei, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Minghan Wang, Lizhi Lei, Min Zhang, et al. 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.

Daimeng Wei, Zhiqiang Rao, Zhanglin Wu, Shaojun Li, Yuanchang Luo, Yuhao Xie, Xiaoyu Chen, Hengchao Shang, Zongyao Li, Zhengzhe Yu, et al. 2022. Hw-tsc's submissions to the wmt 2022 general machine translation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 403–410.

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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 7944–7959, Toronto, Canada.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905.

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023a. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.

Zhanglin Wu, Daimeng Wei, Zongyao Li, Zhengzhe Yu, Shaojun Li, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Yuhao Xie, Lizhi Lei, et al. 2023b. The path to continuous domain adaptation improvements by hw-tsc for the wmt23 biomedical translation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 271–274.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages. *arXiv preprint arXiv:2305.18098*.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2024. Teaching large language models to translate with comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 17, pages 19488–19496.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019. Curriculum learning for domain adaptation in neural machine translation. In *Proceedings of NAACL-HLT*, pages 1903–1915.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Zhi-Hua Zhou, Jianxin Wu, and Wei Tang. 2002. Ensembling neural networks: many could be better than all. *Artificial intelligence*, 137(1-2):239–263.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781.