The SETU-ADAPT Submissions to WMT 2024 Chat Translation Tasks

Maria Zafar, Antonio Castaldo[a], Prashanth Nayak[b], Rejwanul Haque, Andy Way[c]
South East Technological University, Carlow, Ireland
[a]University of Pisa, Tuscany, Italy
[b]KantanAI, Dublin, Ireland
[c]ADAPT Centre, Dublin City University, Dublin, Ireland
c00304029@setu.ie,antonio.castaldo@phd.unipi.it,pnayak@kantanai.io
rejwanul.haque@setu.ie,andy.way@adaptcentre.ie

Abstract

This paper presents the SETU-ADAPT submissions to the WMT24 Chat Translation Task. Large language models (LLM) currently provides the state-of-the-art solutions in many natural language processing (NLP) problems including machine translation (MT). For the WMT24 Chat Translation Task we leveraged LLMs for their MT capabilities. In order to adapt the LLMs for a specific domain of interest, we explored different fine-tuning and prompting strategies. We also employed efficient data retrieval methods to curate the data used for fine-tuning. We carried out experiments for two language pairs: German-to-English and French-to-English. Our MT models were evaluated using three metrics: BLEU, chrF and COMET. In this paper we describes our experiments including training setups, results and findings.

1 Introduction

There have been drastic transformation in many areas of natural language processing (NLP) in recent times mainly due to the emergence of powerful LLMs. The LLM-based solutions are becoming more powerful and accurate than ever before. Notably, we have seen the unprecedented successes in many MT tasks in recent years, thanks to multilingual LLMs. In sum, the LLMs are the current state-of-the-art in MT research and development.

In our submission for the French-to-English and German-to-English Chat Translation Tasks, we built our MT systems using multilingual LLMs such as NLLB-200-600M (Team et al., 2022),[1] Llama-3-8B (Dubey et al., 2024), [2] and mBART-50 (Tang et al., 2021).[3] We fine-tuned these models using both domain-specific and synthetic back-translated data.

Due to the lack of high-quality domain parallel data, we used a data generation approach. For this, we utilised a freely available monolingual data. We retrieved domain-specific monolingual sentences of target language and translated them back to source language for creating new synthetic data (Sennrich et al., 2016). This synthetic data was then combined with the original data for fine-tuning the LLMs. This approach ensured that our MT models are better adapted to the domain, thereby improving the quality of translations. We retrieved domain-specific monolingual German sentences from OPUS ELRC-4992 Customer Support MT [4] for creating our synthetic data. We also explored the idea of in-context learning by fine-tuning LLMs with a few-shot approach. These techniques helped our MT systems better adapt in translating agent input from the source language to the target language and customer response from the target language to the source.

The rest of the paper is organised as follows. Section 2 describes our related work. Our datasets are explained in Section 3 and Section 4 tells about the models and their fine-tuning. Section 5 discusses the experimental setup describing the parameters tuned in our systems. In Section 6, we discuss our results. Finally, Section 7 presents the conclusion of our work.

[1] NLLB-200: https://ai.meta.com/research/no-language-left-behind/

[2] Llama-3: https://github.com/unslothai/unsloth/

[3] mBART-50: https://huggingface.co/facebook/mbart-large-50

[4] OPUS: https://opus.nlpl.eu/ELRC-4992-Customer_Support_MT/de&en/v1/ELRC-4992-Customer_Support_MT

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 1023–1030
November 15–16, 2024. ©2024 Association for Computational Linguistics

2 Related Works

In this section we discuss the papers that are related to our work. Alves et al. (2022) conducted experiments on fine-tuning `mBART-50` using domain-specific data retrieved through semantic search. For this, they used `LaBSE` (Language-Agnostic BERT Sentence Embedding) (Feng et al., 2020). They demonstrated how this approach leads them to large gains across all language pairs under evaluation. They also performed experiments using this data to further adapt the model using KNN-MT (Khandelwal et al., 2021). Note that this approach involves a nearest neighbor retrieval strategy, through which a set of relevant examples are provided at decoding time. They demonstrate how combining these approach leads to improved translation quality, over regular fine-tuning.

Liang et al. (2022a) used pre-trained LLMs and fine-tuned them to the domain of interest. For this, they first trained their models on general domain data and then fine-tuned them with chat translation training data. They used strategies such as including the multi-encoder framework, speaker tag prompt-based fine-tuning and boosted Self-COMET-based (Rei et al., 2020a) ensemble models to incorporate the potential context. They found their strategies helpful in improving the quality of translations produced by their MT models.

Yang et al. (2022) participated in the English-to-German task of the WMT22 Chat Translation Task. For this, they utilised the models previously submitted to the WMT21[5] news task (Wei et al., 2021) as their MT baseline systems. These baseline models are built upon a deep Transformer architecture (Vaswani et al., 2017). They used widely adopted optimisation strategies to improve model performance, including domain transfer, data selection, back-translation, self-training, noisy self-training, fine-tuning, and model averaging. Their results showed the effectiveness of their approached in improving the quality of translations.

Zhou et al. (2022) presented a multi-task multi-stage transitional (MMT) training framework, where they trained their model using the bilingual chat translation dataset and additional monolingual dialogues. To incorporate dialogue coherence and speaker characteristics in their model, they designed two auxiliary tasks: utterance discrimination and speaker discrimination. Their training had three stages: sentence-level pre-training on the large-scale parallel corpus, intermediate training with auxiliary tasks using additional monolingual dialogues and context-aware fine-tuning with a gradual transition. They found that the second stage served as a medium to reduce the training discrepancy between the pre-training and fine-tuning stages. They also trained their model using a gradual transition strategy, i.e. gradually transitioning from monolingual to bilingual dialogues, to make their stage transition smoother. Their results demonstrated the effectiveness of their framework, giving them better translations.

Liang et al. (2022b) contributed to the two large-scale in-domain paired bilingual dialogue corpora (28M for English-to-Chinese and English-to-German) through their framework. Their framework consisted of scheduled multi-task learning with three training stages, in which a gradient-based scheduling strategy was designed to take advantage of the auxiliary tasks for their model for higher translation quality. They conducted extensive experiments on four chat translation tasks, and their model achieved new state-of-the-art performance and outperformed the existing chat MT models by a significant margin.

3 Data Statistics

For our experiments we used the data provided by the WMT-24 Chat Translation Task[6] organisers. The dataset consists of authentic bilingual customer support conversations. This includes parallel data of interactions between an agent and a customer within the customer support domain. We detail the data description in Table 1. Note that we removed duplicates from the training data.

4 The LLMs

This section details the configurations of the LLMs that were used for our experiments.

4.1 mBART

mBART (Liu et al., 2020) is a pre-trained encoder-decoder Transformer model that was first trained on an auto-denoising task with monolingual data of twenty five languages. For adapting the mBART to the MT task, Tang et al. (2021) performed multilingual fine-tuning using data from fifty supported languages. For our experiments we used `facebook/mbart-large-50-many-to-many-mmt`

[5] https://www.statmt.org/wmt21/index.html

[6] https://www2.statmt.org/wmt24/chat-task.html

Dataset	EN–to–DE	EN–to–FR
	WMT-24	
Train	10,556	7,856
Validation	2,569	3,007
Blind Test	2,041	2,091
+ Back-translation	1,317	-
	WMT-22	
Train	2,110	2,754
Validation	-	-
	WMT-20	
Train	10,248	-
Validation	1,619	-

Table 1: Overview of datasets.

checkpoint. We used the following hyperparameters setup for our experiments: batch size: 4, number of training epochs: 5, predict_with_generate: True, evaluation strategy: epoch, logging steps: 2,000, and checkpoint save steps: 500. The remaining parameters were set to default values.

4.2 NLLB

NLLB is a cutting-edge multilingual translation model developed to support many languages, mainly low-resource languages. Initially, the model was trained using diverse, multilingual data that includes various underrepresented languages. This comprehensive pretraining allows NLLB to effectively handle translation tasks across many languages that typically lack sufficient data. For our experiments we used facebook/nllb-200-distilled-600M checkpoint for building our MT systems. Our training configuration is as follows: batch size: 4, 8; max sequence length: 128 tokens; training steps: 10,000, 20,000, 40,000; learning rate: 0.0001; optimiser: Adafactor; weight decay and gradient clipping applied; and model saved every 1000 steps.

4.3 Llama

Llama is an auto-regressive language model that pretrained and fine-tuned in different sizes of data. We used unsloth/llama-3-8b-bnb-4bit checkpoint for building our MT systems. Our training parameters we set are as follows: max seq length: 2048 tokens, batch size: 2 per device, gradient accumulation steps: 4, learning rate: 2e-4, mixed precision training enabled: (fp16 or bf16), learning rate scheduler: linear with 5-step warmup, maximum training steps: 500, optimiser: adamw-8bit, logging steps: 1, seed: 3407.

5 Methodology

In this section, we discuss our methodologies.

5.1 mBART50

We used mBART for building three different MT systems for German-to-English. More specifically, mBART was fine-tuned on three distinct datasets: WMT-20,[7] WMT-22,[8] and WMT-24[9]. For French-to-English we used two datasets from WMT: WMT-22 and WMT-24. we detailed the datasets and hyperparameters setups in Section 3 and 4, respectively.

We performed data preprocessing the original data such as normalisation by removing special characters, removing duplicates and performing lowercase conversions. The source and target sentences were then tokenized using a predefined tokenizer.

In order to handle data during training and evaluation, a collator named as DataCollatorForSeq2Seq[10] is instantiated with the tokenizer and pretrained model checkpoint from Transformers library. This collator is designed to dynamically pad inputs to the maximum length within a batch, ensuring efficient processing. The Seq2SeqTrainer is then instantiated with the pretrained model checkpoint, training arguments, tokenized datasets, evaluation function, data collator, and tokenizer. This setup ensures a structured and efficient fine-tuning process, evaluating the model's performance at each epoch.

Fine-tuning is performed using the Seq2SeqTrainer class from the Transformers library. The training arguments are specified through Seq2SeqTrainingArguments, where parameters such as the output directory, batch sizes for training and evaluation and the number of training epochs were defined in Section 4.

[7] https://www.statmt.org/wmt20/chat-task.html

[8] https://wmt-chat-task.github.io

[9] https://www2.statmt.org/wmt24/chat-task.html

[10] https://huggingface.co/docs/transformers/en/main_classes/data_collator

5.2 NLLB

We built four MT systems for English-to-German and two MT systems for English-to-French considering NLLB as the baselines. We evaluated our MT systems on the development and blind test sets. The MT training setups are detailed below. Our first MT systems involves fine-tuning the baseline NLLB model on the original data. For our second MT system we used normalised data (i.e. removing special characters and duplicates, and lowercasing) for fine tuning to observe any impact of data cleaning on performance.

We build two additional MT systems for for English–German. We back-translated monolingual data in order to create a synthetic bilingual data. For this, we mined monolingual data from OPUS [11]. The domain of the monolingual data is customer support. We combined the generated synthetic data with the original data for fine-tuning. The first and second MT systems were fine-tuned on the combined data, and this gave us the third and fourth MT systems, respectively.

For training we handled out-of-memory errors by dynamically creating training batches, i.e. the Adafactor (Shazeer and Stern, 2018) optimizer is employed instead of AdamW (Loshchilov and Hutter, 2017) to save GPU memory. Weight decay and gradient clipping (Loshchilov and Hutter, 2017) were applied to stabilize the training. Training batches were created by randomly choosing the translation direction (source to target or reverse) and sampling sentence pairs. To enhance robustness against memory issues, a function was implemented to release memory, with parameters set to different batch-sizes, maximum sentence length, and different training-steps. For the German-to-English and French-to-English tasks the best performing models were found to be the ones with 40,000 and 10,000 training steps, respectively. The model is saved every 1,000 steps, allowing for interruptions to adjust parameters or evaluate translations. Training typically runs for a short period of time, which is sufficient for a language similar to those already known by NLLB.

Post-training evaluation involves testing translation quality using parameters like $num - beams = 4$, which affects accuracy, speed, and memory consumption, and parameters a and b, which control the maximum length of a generated text. The number of beams (or beam size) controls how many alternative sequences are kept during the search. This means that the model keeps the top 4 translations at each step during decoding.

5.3 Llama

We also used LLaMA for English-to-German and English-to-French. We built two MT systems for each of the translation tasks. This time, we focused on a specific learning technique, i.e. few-shot in-context learning. For this, we constructed a sentence retrieval system based on dense vector embeddings. Initially, sentence embeddings were generated using SentenceTransformer. More specifically, we used `all-MiniLM-L6-v2` [12] for our task. This model was applied to the source sentences of the dataset, transforming them into high-dimensional vector representations. These embeddings were then indexed using FAISS (Facebook AI Similarity Search) (Douze et al., 2024)[13], creating a searchable database of vectors. In other words, in order to create in-context learning examples, we encode the source test sentence using the pre-trained SentenceTransformer model. The resulting embedding is then used to query the FAISS index, which retrieves the most semantically similar sentences from the training dataset (note that we set $k = 3$). For constructing prompts, we retrieve the source sentences, their corresponding target sentences, and the associated language labels from the training dataset using the indices returned by FAISS. These sentences are then iteratively combined to construct three-shot prompts. Figure 1 shows the structure of a prompt. As can be seen from Figure 1, the prompt consists of initial instruction followed by three components: an instruction, an input, and a response. The instruction guides the task of translating from English-to-German/French or German/French-to-English. The language in an instruction is set dynamically based on labels provided with the sentence in our dataset. The input provides context, and the response is the desired output.

For the fine-tuning process tokenizer was instantiated using the `FastLanguageModel` class with parameters tailored to support efficient training on large sequences. The model is loaded from the `unsloth/llama-3-8b-bnb-4bit` pre-trained

[11] https://opus.nlpl.eu/ELRC-4992-Customer_Support_MT/de&en/v1/ELRC-4992-Customer_Support_MT

[12] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[13] https://github.com/facebookresearch/faiss

checkpoint. Subsequently, the model was further configured with $get-peft-model$, which applies Parameter-Efficient Fine-Tuning (PEFT) (Xu et al., 2023) techniques.

The fine tuning process was managed by `SFTTrainer`, which was integrated into `TrainingArguments` from the Transformers library. The training configurations were discussed in Section 4. Throughout the training, logging was performed at every step and the training process was seeded with a fixed value to ensure reproducibility. This training approach leveraged the state-of-the-art techniques to enable fine-tuning LLMs on extensive datasets while minimizing the computational overhead. The same parameters were tuned for both language-pairs.

The construction of a prompt for inference is identical to those constructed for training. The prompt instructs the model to translate a chat abstract from one language (source) to another (target). The instruction is specifically tailored to the languages involved in the translation, which are specified in the input-row. The source text is included in the prompt, while the output field is left blank, allowing the model to generate the translation. The prompt is tokenized using the pre-trained tokenizer, and the inputs are formatted as tensors compatible with PyTorch, with the computation offloaded to a GPU. The tokenized inputs are then passed to the model's $generate$ method, which performs the translation. The generated output is then decoded from the tokenized format back into a human-readable string using the batch-decode method.

6 Results

This section describes the results obtained. Table 2 shows the performance of our MT systems on the validation sets. Tables 3 and 4 show the results obtained on the blind test set provided by the task organisers.

As mentioned in Section 5, we normalised the original data by removing special characters and duplicate sentences and lowercasing to see the impact of data cleaning on performance. We see from Table 2 that the MT models fine-tuned on normalised data are better than those fine-tuned on original data for the German-to-English translation task. This clearly shows us the effectiveness of data normalisation. Our primary submission for German-to-English was based on fine-tuned Llama with few-shot prompting. For this setup, we obtained 42.08 BLEU, 66.84 chrF and 85.25 COMET points on the validation set (cf. row 12 of Table 2). We also submitted two contrastive systems, i.e. NLLB and mBART50 fine-tuned on augmented data. The performance fine-tuned NLLB and mBART50 are shown in rows 9 and 3 of Table 2, respectively. Our constrastive submission 1 was based on NLLB. For this setup, we obtained 48.21 BLEU, 70.31 chrF and 84.60 COMET points on the validation set (cf. row 9 of Table 2). Our constrastive submission 2 was based on mBART50. For this setup, we obtained 47.73 BLEU, 69.17 chrF and 84.09 COMET points on the validation set (cf. row 3 of Table 2).

```
Below is an instruction that
   describes a task, paired with
   an input that provides further
    context. Write a response
   that appropriately completes
   the request.
Instruction:
Translate this chat from German
   to English:
Input:
German 1: <German sentence 1>
English 1: <English sentence 1>
German 2: <German sentence 2>
English 2: <English sentence 2>
German 3: <German sentence 3>
English 3: <English sentence 3>
German 4: <German sentence 4>
English 4:
Response:
<English sentence 4>
```

Figure 1: The structure of a generated prompt.

As for primary submission of the French-to-English task, we considered NLLB as our baseline and its performance is reported in row 19 of Table 2). We see from the table that this setup provided us 54.79 BLEU, 73.88 chrF, and 85.5 COMET points on the validation set. As in the German-to-English-task, we also submitted two contrastive systems for the French-to-English task. We fine-tuned Llama-3-8B and mBART50 following the few-shots prompt generation strategies described in Section 5. Our constrastive submission 1 was based on fine-tuned Llama-3-8B. For this setup, we obtained 38.23 BLEU, 66.54 chrF and 89.08 COMET points on the validation set (cf. row

Model	**BLEU**	**chrF**	**COMET**
	German-English		
mBART50 WMT20	53.27	72.43	86.042
mBART50 WMT22	32.50	55.66	76.94
mBART50 WMT24	**47.73**	**69.17**	**84.096**
NLLB WMT24 SrcNorm → TgtNorm	35.54	61.37	82.93
NLLB WMT24 TgtNorm → SrcNorm	46.08	68.75	83.73
NLLB WMT24 Src → Tgt	20.43	50.81	77.69
NLLB WMT24 Tgt → Src	16.10	51.44	76.16
NLLB WMT24 + BT SrcNorm → TgtNorm	39.62	64.02	84.30
NLLB WMT24 + BT TgtNorm → SrcNorm	**48.21**	**70.31**	**84.60**
NLLB WMT24 + BT Src → Tgt	23.32	53.10	79.05
NLLB WMT24 + BT Tgt → Src	17.33	52.50	77.10
LLaMA WMT24 FS SrcNorm → TgtNorm	**42.08**	**66.84**	**85.25**
LLaMA WMT24 FS TgtNorm → SrcNorm	20.05	52.90	83.03
LLaMA WMT24 FS Src → Tgt	20.07	57.60	85.71
LLaMA WMT24 FS Tgt → Src	35.54	59.79	87.66
	French-English		
mBART50 WMT22	43.51	64.64	80.27
mBART50 WMT24	**53.15**	**72.68**	**84.55**
NLLB WMT24 SrcNorm → TgtNorm	46.24	68.78	85.42
NLLB WMT24 TgtNorm → SrcNorm	**54.79**	**73.88**	**85.50**
NLLB WMT24 Src → Tgt	31.45	59.69	80.65
NLLB WMT24 Tgt → Src	34.07	63.20	79.80
LLaMA WMT24 FS SrcNorm → TgtNorm	31.11	60.02	85.90
LLaMA WMT24 FS TgtNorm → SrcNorm	5.64	30.87	80.42
LLaMA WMT24 FS Src → Tgt	**38.23**	**66.54**	**89.08**
LLaMA WMT24 FS Tgt → Src	24.71	58.38	82.91

Table 2: Performance of our MT systems on the validation set. SrcNorm and TgtNorm stand for Source and Target normalised, respectively. BT stands for back-translation and FS stands for Few-Shot.

Tag	Precision	Recall	F1
	French-English		
formality	90.2	78.8	84.1
lexical cohesion	46.4	42.5	44.3
pronouns	90.8	72	80.3
verb form	62.9	56.8	59.7

Table 3: Official results for the French-English translation task (blind set).

24 of Table 2). Our constrastive submission 2 was based on mBART50. For this setup, we obtained 53.15 BLEU, 72.67 chrF and 84.55 COMET points on the validation set (cf. row 17 of Table 2).

Our primary submission of the German-to-English task was based on Llama. As can be seen from Table 4, we obtained 55.0 BLEU, 72.1 chrF, 90.8 COMET and 0.167827 CONTEXT-COMET-QE (Rei et al., 2020b) points on the WMT 2024 blind test sets. Our primary submission of French-to-English was based on NLLB. For this setup we obtained 31.3 BLEU, 60.9 chrF, 82.4 COMET and -0.23095 CONTEXT-COMET-QE points. Our best-performing system of the German-to-English task is Llama with few-shot learning. We secured the top place for German-to-English in this competition. Table 3 shows our precision for pronouns in the French-to-English system. Our submission for the French-to-English translation task is in fact the best-performing system in terms of pronoun translatiosn.

Model	COMET	ChrF	BLEU	COMET-QE
German-English				
LlaMa WMT24 FS	90.8	72.1	55.0	0.16
French-English				
NLLB WMT24	82.4	60.9	31.3	-0.23

Table 4: Official results. Performance of our MT systems on the blind set (primary submissions).

7 Conclusion

This paper described our submissions to the WMT 2024 Chat Translation Task for German-to-English and French-to-English language pairs. We applied several training and fine-tuning strategies such as standard fine-tuning and fine-tuning with few-shot prompting. We investigated our approaches using three different LLMs: NLLB, Llama and mBART. This allowed us to make a comparative analysis between different architectures and strategies. One of the key findings of our investigation is that the performance of the MT systems on translating conversational messages can be improved with knowledge transfer. We also found that our MT systems exhibit robustness on this *difficult-to-translate* domain.

For future investigations, given the shortage of conversational data, we plan to focus on exploring the use of advanced data augmentation techniques. We also intend to further investigate to what extent synthetic data can be beneficial in chat translation scenarios.

References

João Alves, Pedro Henrique Martins, José G. C. de Souza, M. Amin Farajian, and André F. T. Martins. 2022. Unbabel-IST at the WMT chat translation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 943–948, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. ArXiv:2401.08281 [cs].

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation.

Yunlong Liang, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. 2022a. BJTU-WeChat's systems for the WMT22 chat translation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 955–961, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yunlong Liang, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. 2022b. Scheduled multi-task learning for neural chat translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4375–4388, Dublin, Ireland. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. 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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. *CoRR*, abs/1804.04235.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual Translation from Denoising Pre-Training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

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.

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.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment.

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.

Chulun Zhou, Yunlong Liang, Fandong Meng, Jie Zhou, Jinan Xu, Hongji Wang, Min Zhang, and Jinsong Su. 2022. A multi-task multi-stage transitional training framework for neural chat translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):7970–7985.