Optimising LLM-Driven Machine Translation with Context-Aware Sliding Windows

Xinye Yang, Yida Mu, Kalina Bontcheva, Xingyi Song
School of Computer Science, The University of Sheffield
{xyang138, y.mu, k.bontcheva, x.song}@sheffield.ac.uk

Abstract

This paper describes SheffieldGATE's submission to WMT 2024 Chat Shared Translation Task. We participate in three language pairs: English-German, English-Dutch, and English-Portuguese (Brazil). In this work, we introduce a context-aware sliding window decoding method to track dependencies between chat messages. We fine-tune a large pre-trained language model based on the training data provided by the shared task Our experiments (i) compare the model performance between multilingual and bilingual fine-tuning and (ii) assess the impact of different window sizes. Our experimental results demonstrate that utilising contextual information yields superior performance in document-level translation compared to translating documents as isolated text segments, and that models fine-tuned with multilingual data perform better than those fine-tuned with bilingual data.

1 Introduction

Translating chat text is an important and challenging application of machine translation technology (Farajian et al., 2020; Farinha et al., 2022). The purpose of this task is to build a translation model that addresses the challenges of multilingual customer support for multinational companies. In informal conversations, people often use abbreviations and incomplete sentences and may include spelling errors, leading to significant noise in the dialogue text (Varnhagen et al., 2010). These factors complicate the translation of such texts, a challenge that traditional machine translation methods struggle to address (Almansor et al., 2020).

Recently, large language models (LLMs) have gradually taken over the mainstream in the field of natural language processing (Ouyang et al., 2022). LLMs have demonstrated impressive capabilities in a wide range of domains such as computational social science (Mu et al., 2024), question answering (Tan et al., 2023), and machine translation(Wang et al., 2023). Their ability to be well robust to noise in the input data provides new ideas to address the challenges of chat translation.

At the sentence level, Neural Machine Translation (NMT), represented by pre-trained large language models, is approaching the quality of professional human translations or even exceeding that of crowd-sourced non-professional translations in a few resource-rich languages (Hassan et al., 2018). For document-level translation, NMT systems still have certain errors that are difficult to detect in sentence-level translation (Läubli et al., 2018). Such as language ambiguity, which frequently results in numerous translation errors. Depending on the context, a single word or phrase can have multiple meanings (Abeysiriwardana and Sumanathilaka, 2024). Without the use of contextual information, problems including co-reference (Guillou and Hardmeier, 2016), lexical cohesion (Carpuat, 2009), or lexical disambiguation (Rios Gonzales et al., 2017) will be difficult to address (Jin et al., 2023).

In this work, we focus on modelling strategies based on contextual information. Our submission is based on an existing pre-trained model and fine-tuned using multilingual chat data, behaviour without incorporating additional contextual information during the fine-tuning process. We implemented context-aware sliding windows for the inference stage to perform translation tasks. We also conducted the following experiments (i) to compare the performance difference between using multilingual data and bilingual data in the fine-tuning process and (ii) the impact of window size, or the extent of contextual information, on the quality of translation.

With this study, we aim to shed light on the great potential of large language models for machine translation tasks and their ability to utilise contextual information for document-level translation and learn from migrating across linguistic data.

Proceedings of the Ninth Conference on Machine Translation, pages 1004–1010

Language Pair	Train	Val.	Test
EN <-> DE	17,805	2,569	2,041
EN <-> FR	15,027	3,007	2,091
EN <-> PT-BR	15,092	2,550	2,040
EN <-> KO	16,122	1,935	1,982
EN <-> NL	15,463	2,549	2,015

Table 1: Number of source segments in the released dataset.

2 Data

The dataset for this task comprises authentic bilingual customer support conversations across five language pairs: English-German, English-French, English-Korean, English-Dutch, and English-Portuguese (Brazil). Table 1 displays the number of training, validation and test samples for each language pair in the dataset.

2.1 Dataset Characteristics

The chat content flows freely without strict format constraints, authentically reflecting the characteristics of real conversations. This natural language use includes incomplete sentences, interjections, and context-dependent responses, which, while representative of genuine dialogue, increases the complexity of processing and translation.

3 System Description

3.1 Context-Aware Sliding Window

To effectively utilise contextual information, we use a context-aware sliding window mechanism. This approach allows model to consider context sentences when translating each individual message, thereby enhancing the overall coherence and accuracy of the translation. In addition, we improve translation efficiency by reusing the Key-Value (KV) cache. KV caching is a crucial technique in transformer models, involves storing and reusing previously computed Key and Value matrices in the self-attention mechanism. This method significantly enhances inference speed by eliminating redundant calculations, particularly beneficial for long sequences or auto-regressive generation tasks such as machine translation. It enables the model to efficiently leverage information from the source language when generating the target sequence, substantially reducing computational overhead, especially for longer texts.

Structure of the Sliding Window Our context-aware sliding window comprises four key components:

- Task Description: Provides the model with clear instructions about the translation task.
- Source language tag: Identifies the beginning of the original text.
- Original Text: Contains the message to be translated along with its context.
- Target Language Label: Indicates the end of the original text and directs the model to give the translation.

Figure 1 illustrates the structure of the Context-Aware Sliding Window. This system comprises a task description and a window containing a sequence of source sentences, which together function as input to the model. The model generates new translations based on the contextual information available within the window. After each translation is produced, it is inserted into the list of translated sentences, and the window shifts to incorporate new source sentences. If the number of sentences in the source text window exceeds a predefined limit, the earliest sentence in the window is removed to maintain the set window size. This sliding mechanism ensures that the model consistently has track dependencies throughout the translation process.

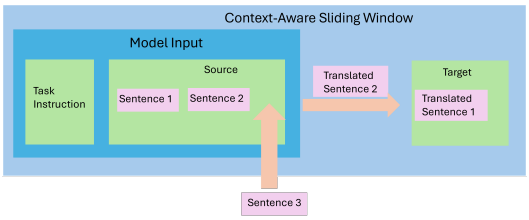

Figure 1: Context-Aware Sliding Window

Prompt We used the following prompt for translation:
You are a translation specialist serving multinational companies. Your task is to translate the given text from [source language] to [target language]. Provide the translation result in [target language] directly without including any additional content.

Workflow The operation of our context-aware sliding window can be described as follows:

- Initialisation: The sliding window starts empty and gradually fills with sentences from the chat log up to the predefined window size.
- Generation: The language model generates the translation for the most recent sentence, considering both the original sentences in the window and their existing translations.
- Window Shift: After generating a translation, the window shifts by one position. It incorporates the next sentence from the chat log and removes the earliest one and its corresponding translation if the window is full. If the translation direction of the next sentence changes , the windows storing the original text and the translated text are swapped. This approach allows for seamless handling of bidirectional translations within the same conversation, maintaining context in both languages.
- Iteration: Steps 2 and 3 are repeated until all sentences in the chat log have been processed.

The workflow of the context-aware sliding window is illustrated in pseudocode in Algorithm 1.

Advantages This approach offers several benefits:

- Improved Coherence: By considering the surrounding context, the model can maintain better consistency in tone, style, and terminology across the translation.
- Enhanced Accuracy: Contextual information helps resolve ambiguities and choose more appropriate translations for words or phrases with multiple meanings.

Algorithm 1 Context-Aware Sliding Window Translation Algorithm with Bidirectional Support

Initialise:
  source-window ← [ ]
  target-window ← [ ]
  window-size ← predefined window size
  translation-result ← [ ]
  current-direction ← initial translation direction
for each *sentence* in *input text* **do**
  if *sentence-direction* ≠ *current-direction* **then**
    source-window, *target-window* ← *target-window*, *source-window*
    current-direction ← *sentence-direction*
  end if
  if *len(source-window)* < *window-size* **then**
    source-window.append(sentence)
    translation ← *Generate(source-window, target-window)*
    target-window.append(translation)
    translation-result.append(translation)
  else
    source-window.pop(0)
    target-window.pop(0)
    source-window.append(sentence)
    translation ← *Generate(source-window, target-window)*
    target-window.append(translation)
    translation-result.append(translation)
  end if
end for
Output *translation-result*

4 Experiments

In this section, we describe the experiments conducted to select the fine-tuning strategy and determine the optimal window size for our system. The hyperparameters used in this experiment are listed in Table 2. All experiments were executed on a single Nvidia A100 GPU equipped with 40GB of memory.

Three evaluation metrics are used in this experiment, aligned with the automatic evaluation metrics of the shared task, they are:

- BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002): Measures translation quality based on n-gram overlap between the candidate and reference translations. BLEU primarily assesses fluency and adequacy at the phrase level. It is widely used but may not always capture deeper semantic nuances.
- chrF (Character n-gram F-score) (Popović, 2015): Evaluates translation quality at the character level. It is particularly effective for capturing morphological accuracy and subtle differences in word forms. chrF is sensitive to grammatical correctness and precise word choice.
- COMET (Cross-lingual Optimised Metric for Evaluation of Translation) (Rei et al., 2020): A more recent metric that focuses on seman-

Hyperparameter	Value
LoRA rank (r)	8
LoRA alpha	16
LoRA dropout	0.05
Learning rate	2.5e-5
Weight decay	0.001
Batch size	8
Training epochs	10
Warmup ratio	0.3
Max gradient norm	0.3
LR scheduler	Linear

Table 2: Fine-tuning Hyperparameters

tic similarity between the source, translation, and reference. COMET uses contextual embedding to evaluate meaning preservation and overall translation quality, aiming to correlate better with human judgements.

4.1 Multilingual and bilingual Fine-tuning

Given the computational resources and time constraints, we choose the LLaMA3-8B instruct model (LLaMA) (Dubey et al., 2024) as our base model. We fine-tune LLaMA using Low-Rank Adaptation (LoRA) (Hu et al., 2022) with training and validation data provided by the shared task. We employed two distinct fine-tuning strategies, i.e., (i) multilingual fine-tuning and (ii) bilingual fine-tuning.

For the multilingual fine-tuning, we feed five language pairs simultaneously: English <-> German, English <-> French, English <-> Brazilian Portuguese, English <-> Korean, and English <-> Dutch. This strategy allows the model to learn from multiple languages concurrently and potentially leverage cross-lingual information.

In contrast, our bilingual strategy involved fine-tuning separate models for each language pair, using solely the training and validation data specific to that pair. This approach enables more focused adaptation to each language pair.

The motivation for employing these two strategies was to explore the cross-linguistic learning and transfer capabilities of large language models (Lample and Conneau, 2019). By comparing these approaches, we aim to investigate whether the model can extract universally applicable translation patterns and linguistic features from multiple language pairs, thereby potentially improving its performance on new language pairs.

The experiment results are shown in Table 3. The multilingual fine-tuned models outperform bilingual fine-tuned models. This may be because multilingual dataset provide more samples than each bilingual datasets, offering a broader and more diverse set of data, which helps prevent the model from overfitting. Also, the model can learn translation patterns through transfer learning from other languages. Hence, in our final submission, the model was fine-tuned using the multilingual dataset.

4.2 Impact of Window Size

We also investigated the effect of different window sizes on the translation quality. In this work, the window size determines the amount of context available to the model during the translation process.

To that end, we conducted experiments with window sizes ranging from 1 to 3 sentences. For each window size, we translated five language pairs from the validation set provided by shared task and evaluated the results using automated metrics. Table 4 presents the detailed results for chrF, BLEU, and COMET scores across different window sizes and language pairs.

The window size used in our submission is 3. Our findings indicate that the translation quality generally improves as the window size increases, but the extent and nature of improvement varies across translation directions and metrics. We observe that the COMET metric tends to favour larger window sizes more consistently than chrF or BLEU.

COMET scores show improvement or maintain high performance with larger windows in 5 out of 6 translation directions (de-en, en-de, pt-br-en, nl-en, en-nl).

For en-pt-br, small windows have the best performance across all metrics. This unique behavior might be attributed to several factors. Firstly, the structural similarities between English and Brazilian Portuguese allow for effective translation with minimal context.(Angeli and Mota, 2023) The relatively simple morphology of English compared to Portuguese's more complex system might also contribute to this phenomenon. Additionally, the direct lexical correspondence between many English and Portuguese words could lead to high accuracy in word-to-word translations, which is particularly well-captured by chrF and BLEU metrics.

In contrast, chrF and BLEU metrics often peak

	multilingual			bilingual		
Language Pair	chrF	Bleu	COMET	chrF	Bleu	COMET
de->en	**67.45**	**44.46**	**88.13**	65.63	41.11	86.55
en->de	**60.95**	**35.74**	**86.41**	60.03	34.82	85.59
pt-br->en	**65.50**	**43.74**	**87.10**	63.17	36.52	84.68
en->pt-br	**66.94**	**42.02**	**89.43**	65.21	39.43	87.67
nl->en	**68.05**	**45.94**	**88.66**	65.77	42.58	86.38
en->nl	**62.26**	**35.94**	**89.29**	59.65	32.41	87.09

Table 3: Translation Quality Metrics for Multilingual and bilingual Models. The highest scores for each metric are marked in bold.

	Window Size = 1			Window Size = 2			Window Size = 3		
Language Pair	chrF	BLEU	COMET	chrF	BLEU	COMET	chrF	BLEU	COMET
de-en	64.37	39.75	84.78	**68.16**	**45.53**	**88.35**	67.45	44.46	88.12
en-de	60.86	35.47	86.31	**61.15**	**35.92**	86.11	60.95	35.74	**86.40**
pt-br-en	62.96	39.24	83.44	**65.62**	**44.24**	**87.35**	65.50	43.74	87.10
en-pt-br	**67.82**	**45.49**	**89.94**	67.34	43.04	89.48	66.94	42.02	89.43
nl-en	64.02	40.94	83.23	**68.15**	**48.01**	88.26	68.05	45.94	**88.66**
en-nl	60.16	33.06	87.67	60.32	33.34	88.15	**62.26**	**35.94**	**89.29**

Table 4: Translation Quality Metrics for Different Window Sizes. The highest scores for each metric are marked in bold.

at window size 2 or even size 1 for some translation directions. For example, en-pt-br achieves its highest chrF and BLEU scores with window size 1. The en-nl pair is a notable exception, showing consistent improvement across all metrics as the window size increases.

This pattern suggests that the COMET metric may be more sensitive to the broader context provided by larger window sizes, while chrF and BLEU might prioritise local fluency or accuracy that can sometimes be captured effectively with smaller windows.

5 Conclusion

In this paper, we compared the performance of fine-tuning using multilingual data and bilingual data. Additionally, we conducted an ablation study by evaluating the translation quality with different window sizes. Our research indicates that fine-tuning models on multilingual data results in superior translation capabilities compared to fine-tuning on a single language. This approach could improve translation quality for low-resource languages. Furthermore, we also found that increasing the contextual information provided to the model can enhance its semantic performance in translation. Our future work will focus on:

- **Named Entity Handling** We plan to integrate a named entity recognition system and leverage external knowledge resources, such as Wikipedia, to ensure accurate translations of named entities.
- **Model Fine-tuning Comparison** We also aim to conduct a comparative analysis between fine-tuning the foundation model and the instruction-tuned model, exploring the trade-offs between general and task-specific performance.

Acknowledgements

This work has been jointly funded by the UK's innovation agency (Innovate UK) grant 10098112 (project name ASIMOV: AI-as-a-Service) and grant 10039055 (approved under the Horizon Europe Programme as vera.ai[1], EU grant agreement 101070093).

References

Miuru Abeysiriwardana and Deshan Sumanathilaka. 2024. A survey on lexical ambiguity detection and word sense disambiguation. *arXiv preprint arXiv:2403.16129*.

Ebtesam Almansor, Ahmed Al-Ani, and Farookh Hussain. 2020. *Transferring Informal Text in Arabic as*

[1] https://www.veraai.eu/home

Low Resource Languages: State-of-the-Art and Future Research Directions, pages 176–187. Springer International Publishing.

Natália Pinheiro De Angeli and Mailce Borges Mota. 2023. Cross-linguistic priming effects during the comprehension of the passive voice: Two primes are enough. *Ilha do Desterro*, 76(3):17–39.

Marine Carpuat. 2009. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 19–27, Boulder, Colorado. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and et al. 2024. The llama 3 herd of models.

M Amin Farajian, António V Lopes, André FT Martins, Sameen Maruf, and Gholamreza Haffari. 2020. Findings of the wmt 2020 shared task on chat translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 65–75.

Ana C Farinha, M. Amin Farajian, Marianna Buchicchio, Patrick Fernandes, José G. C. de Souza, Helena Moniz, and André F. T. Martins. 2022. Findings of the WMT 2022 shared task on chat translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 724–743, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, Portorož, Slovenia. European Language Resources Association (ELRA).

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

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*.

Linghao Jin, Jacqueline He, Jonathan May, and Xuezhe Ma. 2023. Challenges in context-aware neural machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15246–15263, Singapore. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *ArXiv*, abs/1901.07291.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Yida Mu, Ben P. Wu, William Thorne, Ambrose Robinson, Nikolaos Aletras, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2024. Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12074–12086, Torino, Italia. ELRA and ICCL.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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, Online. Association for Computational Linguistics.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Yiming Tan, Dehai Min, Y. Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Can chatgpt replace traditional kbqa models? an in-depth analysis of the question answering performance of the gpt llm family. In *International Workshop on the Semantic Web*.

Connie K Varnhagen, G Peggy McFall, Nicole Pugh, Lisa Routledge, Heather Sumida-MacDonald, and Trudy E Kwong. 2010. Lol: New language and

spelling in instant messaging. *Reading and writing*, 23:719–733.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.