Detailed Error Detection in Machine Translation Using Large Language Models

Da-Hyun Jung[1], Jeong-Bae Park[2,*]
[1]Korea University, [2]Human-inspired AI Research
{dhaabb55,insmile}@korea.ac.kr

Abstract

Machine Translation (MT) is a crucial technology that bridges language gaps, enabling seamless global communication. Despite recent advancements in Large Language Models (LLMs), MT systems still face challenges with translation accuracy, especially when critical errors occur that can distort the intended meaning of entire sentences. This study addresses the detection of word-level critical errors in MT. We propose a novel approach leveraging LLMs to detect these critical errors across multiple language pairs. Our method integrates a smaller-scale language model to preemptively flag potential translation issues, subsequently refining the error detection capabilities of LLMs. Through extensive experimentation with LLM, we demonstrate the superiority of our approach in accurately identifying critical errors. This research lays a foundational framework for the development of more robust MT systems, ultimately enhancing global communication.

Keywords: Word-level Critical Error Detection, Large Language Models

1. Introduction

Machine Translation (MT) is a pivotal technology that facilitates seamless communication across diverse languages worldwide, effectively breaking down barriers to global communication [1, 2]. In recent years, the remarkable advancements in Large Language Models (LLMs) have significantly improved the performance of MT systems. However, translation accuracy remains imperfect, with critical errors that can severely distort the intended meaning of entire sentences, sometimes leading to serious misunderstandings [3, 4, 5, 6]. Accurate detection of such errors is crucial for enhancing the reliability and usability of MT systems, especially in high-stakes domains such as legal documents, medical records, and international communications.

Critical errors in machine translation are defined as those that could result in a negative user experience, particularly in contexts involving ethical, economic, or legal issues where distortions of meaning could lead to financial losses or legal liabilities. These errors often arise from subtle details in the translated text, such as word choice or grammatical structure, making error detection at the word level particularly important.

*Corresponding author

This study explores methods for detecting word-level critical errors in MT using LLMs. We evaluate the performance of LLMs in identifying critical errors across various language pairs and propose strategies to enhance this capability effectively. Specifically, we introduce an approach that involves using a smaller-scale language model to preemptively detect potential errors in translated documents, which are then used as input for LLMs to improve their detailed error detection performance.

To this end, we conducted experiments using the state-of-the-art LLMs, GPT-3.5 [7] and GPT-4 [8], across multiple multilingual language pairs. Our findings demonstrate that our methodology offers a superior error detection capability compared to existing LLMs. By clarifying the word-level error detection potential of LLMs, this study provides a critical foundation for developing more accurate and reliable MT systems. Such advancements contribute to smoother international communication and significantly enhance access to information across diverse languages.

2. Related Work

MT has been an area of research for several decades, and recently, Transformer-based approaches have gained signifi-

		En-De			Zh-En			En-Ru		
Method	**Model**	**F1**	**Recall**	**Presicion**	**F1**	**Recall**	**Presicion**	**F1**	**Recall**	**Presicion**
Zero-shot	GPT-3.5	0.0332	0.0276	0.0771	0.0408	0.0388	0.0767	0.0153	0.0108	0.0412
	GPT-4	0.1659	0.2117	0.1867	0.1399	0.1867	0.1276	0.0808	0.0909	0.0758
Few-shot	GPT-3.5	0.0587	0.0687	0.0792	0.0341	0.0525	0.0464	0.0157	0.0138	0.0199
	GPT-4	0.2151	0.2566	0.2292	0.1159	0.1481	0.1204	0.0895	0.0899	0.1047
Ours										
Zero-shot	GPT-3.5	0.1362	0.1206	0.2534	0.0910	0.0677	**0.2042**	**0.1552**	**0.1130**	**0.4379**
	GPT-4	0.1770	0.2318	0.1778	0.1137	0.1572	0.1019	0.0111	0.0167	0.0083
Few-shot	GPT-3.5	0.1899	0.2411	0.2427	0.1099	0.1250	0.1122	0.0646	0.0571	0.1167
	GPT-4	**0.2356**	**0.2898**	**0.2575**	**0.1758**	**0.2446**	0.1622	0.1429	0.1000	0.2500

Table 1. Comparison of LLM performance with our method

cant attention for dramatically improving the performance of MT systems [9, 10]. Previous studies have proposed various evaluation methods for MT outputs to minimize errors and enhance performance. These efforts have primarily focused on building sentence-level and word-level error detection frameworks, which have been further refined by leveraging the strong language understanding capabilities of pretrained language models [11, 12, 13]. Moreover, recent research has particularly highlighted the potential of LLMs to preemptively detect and correct errors in MT systems, which is especially critical in high-risk domains where translation accuracy is paramount [14, 15].

3. Our Method

This study explores a method for detecting critical word-level errors in machine translation using LLMs. Our approach consists of two stages. First, we use a smaller, pretrained language model, XLM-RoBERTa [16], to preliminarily detect potential errors in translated documents. The model takes both the source and translated sentences as input and performs a binary classification to determine whether an error is present.

Second, the preliminary error detection results are used as input to the LLM, leveraging its advanced language understanding capabilities to identify more detailed errors. Specifically, a prompt containing the definition of the error existence task and critical word-level error detection is provided to the LLM. This allows the LLM to recognize the presence of potential errors in advance and perform a more fine-grained analysis, thereby enhancing its ability to analyze sentences at a more detailed level.

4. Experiment

The experimental results of this study aim to evaluate the performance changes when applying our proposed methodology to LLMs. The experiments were conducted under zero-shot and few-shot settings across various language pairs (En-De, Zh-En, En-Ru). We use datasets from the WMT23 QE [17], based on the MQM dataset [4].

Overall, the addition of our method improved the performance of LLMs in most experimental settings. Our approach led to performance gains for both GPT-3.5 and GPT-4 across all language pairs in the few-shot setting. Notably, GPT-4 showed a significant increase in F1 score, rising to 0.2356 for the En-De language pair and achieving an F1 score of 0.1758 for the Zh-En language pair, demonstrating the highest performance improvements. Moreover, our methodology exhibited substantial improvements in the zero-shot setting with GPT-3.5 for the En-Ru language pair, demonstrating that it enhances error detection capabilities even for LLMs that typically show weaker performance in this area.

5. Conclusion

This study proposes a method for effectively detecting critical word-level errors in machine translation by leveraging LLMs. Experimental results demonstrate that our method offers a superior error detection capability compared to existing LLM-based systems. Future research will focus on eval-

uating the generalizability of the proposed approach across various language pairs and translation document types and exploring additional optimization techniques to further enhance the performance of LLMs. Such advancements are expected to facilitate smoother international communication and maximize information accessibility.

Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425). This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2023S1A5C2A07095987)

Reference

[1] L. Wang, C. Lyu, T. Ji, Z. Zhang, D. Yu, S. Shi, and Z. Tu, "Document-level machine translation with large language models," *arXiv preprint arXiv:2304.02210*, 2023.

[2] H. Xu, Y. J. Kim, A. Sharaf, and H. H. Awadalla, "A paradigm shift in machine translation: Boosting translation performance of large language models," *arXiv preprint arXiv:2309.11674*, 2023.

[3] K. Sudoh, K. Takahashi, and S. Nakamura, "Is this translation error critical?: Classification-based human and automatic machine translation evaluation focusing on critical errors," *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pp. 46–55, 2021.

[4] M. Freitag, G. Foster, D. Grangier, V. Ratnakar, Q. Tan, and W. Macherey, "Experts, errors, and context: A large-scale study of human evaluation for machine translation," *Transactions of the Association for Computational Linguistics*, Vol. 9, pp. 1460–1474, 2021.

[5] K. Al Sharou and L. Specia, "A taxonomy and study of critical errors in machine translation," *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pp. 171–180, 2022.

[6] C. Zerva, F. Blain, R. Rei, P. Lertvittayakumjorn, J. G. De Souza, S. Eger, D. Kanojia, D. Alves, C. Orăsan, M. Fomicheva *et al.*, "Findings of the wmt 2022 shared task on quality estimation," *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 69–99, 2022.

[7] OpenAI-Blog, "Chatgpt: Optimizing language models for dialogue," 2022. [Online]. Available: https://openai.com/blog/chatgpt/

[8] OpenAI, "Gpt-4 technical report," 2023.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., pp. 4171–4186, Jun. 2019. [Online]. Available: https://aclanthology.org/N19-1423

[11] J. Ive, F. Blain, and L. Specia, "deepQuest: A framework for neural-based quality estimation," *Proceedings of the 27th International Conference on Computational Linguistics*, E. M. Bender, L. Derczynski, and P. Isabelle, Eds., pp. 3146–3157, Aug. 2018. [Online]. Available: https://aclanthology.org/C18-1266

[12] F. Kepler, J. Trénous, M. Treviso, M. Vera, and A. F. T. Martins, "OpenKiwi: An open source framework for quality estimation," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, M. R. Costa-jussà and E. Alfonseca, Eds., pp. 117–122, Jul. 2019. [Online]. Available: https://aclanthology.org/P19-3020

[13] T. Ranasinghe, C. Orasan, and R. Mitkov, "TransQuest: Translation quality estimation with cross-lingual transformers," *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds., pp. 5070–5081, Dec. 2020. [Online]. Available: https://aclanthology.org/2020.coling-main.445

[14] L. Specia, F. Blain, M. Fomicheva, C. Zerva, Z. Li, V. Chaudhary, and A. F. T. Martins, "Findings of the WMT 2021 shared task on quality

estimation," *Proceedings of the Sixth Conference on Machine Translation*, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussa, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, T. Kocmi, A. Martins, M. Morishita, and C. Monz, Eds., pp. 684–725, Nov. 2021. [Online]. Available: https://aclanthology.org/2021.wmt-1.71

[15] C. Zerva, F. Blain, R. Rei, P. Lertvittayakumjorn, J. G. C. de Souza, S. Eger, D. Kanojia, D. Alves, C. Orăsan, M. Fomicheva, A. F. T. Martins, and L. Specia, "Findings of the WMT 2022 shared task on quality estimation," *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 69–99, Dec. 2022. [Online]. Available: https://aclanthology.org/2022.wmt-1.3

[16] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2020.

[17] F. Blain, C. Zerva, R. Ribeiro, N. M. Guerreiro, D. Kanojia, J. G. de Souza, B. Silva, T. Vaz, Y. Jingxuan, F. Azadi *et al.*, "Findings of the wmt 2023 shared task on quality estimation," *Proceedings of the Eighth Conference on Machine Translation*, pp. 629–653, 2023.