Enhancing Translation Quality: A Comparative Study of Fine-Tuning and Prompt Engineering in Dialog-Oriented Machine Translation Systems. Insights from the MULTITAN-GML Team

Lichao Zhu[1], Maria Zimina-Poirot[1], Behnoosh Namdarzadeh[1], Nicolas Ballier[1,3] and Jean-Baptiste Yunès[2]
[1]CLILLAC-ARP, [2]IRIF, [3]LLF
Université Paris Cité, F-75013 Paris, France
Contact: lichao.zhu@u-paris.fr

Abstract

For this shared task, we have used several machine translation engines to produce translations (en ⇔ fr) by fine-tuning a dialog-oriented NMT engine and having NMT baseline translations post-edited through prompt engineering. Our objectives are to assess the effectiveness of a fine-tuning strategy with a robust NMT model, to advance towards a comprehensive pipeline that covers the entire translation process (from fine-tuning and machine translation to automatic post-editing (APE)), and to evaluate the strengths and weaknesses of NMT systems.

1 Introduction

We had three research objectives in carrying out our experiments. The first objective was to assess the feasibility of fine-tuning an in-domain neural machine translation (NMT) baseline model using minimal unlabelled data. The second objective involved utilising large language models (LLMs) and prompt engineering techniques to post-edit translations within the same domain. The third objective was to examine the linguistic features of various models' erroneous translations, particularly in bilingual customer service conversations. For example, in their description of the data of the first edition of the Chat Task, (Farajian et al., 2020) noted the excessive use of pronouns in the dataset.

The remaining sections of the paper are organised as follows: section 2 mentions previous research, section 3 outlines our methods and describes our NMT systems, section 4 delves into our results[1], section 5 provides a discussion of these results, and section 6 outlines future work.

[1]https://github.com/lichaozhu/team_MULTITAN-GML_WMT24_Chat_Shared_Task

2 Previous Research

2.1 Fine-tuning Strategies for NMT and Domain Adaptation

Fine-tuning a pre-trained LLM baseline model with low-resource NMT has been the subject of previous MT empirical studies (Galiano-Jiménez et al., 2023) and the back-translation approach is often used to improve the accuracy of models (Hoang et al., 2018). Open source toolkits are available for building pipelines, such as `fairseq`[2]. However, some models require a higher level of expertise in pipeline construction and rely on cutting-edge hardware for optimal performance[3]. In terms of domain adaptation, filtering back-translations is considered one of the most frugal and efficient techniques (Kumari et al., 2021). In addition, more and more domain adaptations rely on prompt engineering.

Based on what was reported in the findings of the Chat Task 2022 (Farinha et al., 2022), MT systems handle source-related issues more or less similarly. Analysing the distribution of error types presented in the task indicates that "mistranslation" is the most frequent error across all systems. Furthermore, prompt-based machine translation has shown a significant impact in medical domains. For example, Ramachandran et al. (2023) demonstrated that using GPT-4 for extracting Social Determinants of Health (SDOH) from electronic health records achieved a 0.652 F1 score, which is comparable to the 7th best system among traditional supervised approaches.

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

[3]For example, NLLB-200-3.3.B requires Hydra (Yadan, 2019) and very high GPU resources. We were unable to load and train the model using a dual A100 40GB setup due to persistent memory overflow problems.

Proceedings of the Ninth Conference on Machine Translation, pages 1016–1022
November 15-16, 2024

2.2 Automatic Post-editing of MT and Prompt Engineering

Automatic post-editing (APE) systems are designed to enhance the quality of machine translation (MT) by *leveraging* data (Raunak et al., 2023; Gao et al., 2023). These systems work by taking both the source text and the initial MT output as inputs, then applying learned post-editing patterns to refine the translation, and the final output is an improved translation (Chollampatt et al., 2020; Sharma et al., 2021; Bhattacharyya et al., 2023). To further improve performance, APE systems often employ domain adaptation and fine-tuning on in-domain data (Moslem et al., 2023). Based on previous studies, prompting for machine translation still suffers from issues such as copying, mistranslation of entities, and hallucinations (Zhang et al., 2023). Furthermore, previous comprehensive evaluations of GPT models for machine translation across various language pairs indicate that GPT models perform competitively for high-resource languages, but face limitations with low-resource languages (Hendy et al., 2023; Jiao et al., 2023; Peng et al., 2023).

3 Methods and Tools

3.1 Fine-tuning via NMT Engine

For our primary submission, we have used a neural machine translation (NMT) engine, its in-domain baseline model, and in-domain training data to fine-tune the model. To create our fine-tuning dataset, we used the Chat Task 2022's valid and test sets (en ⇔ fr) as well as the Chat Task 2024's `train` and `valid` sets and compiled 13,622 aligned segments (122,905 words in English and 127,335 words in French). We used this dataset to fine-tune the *Dialog* in-domain model on the training server Model Studio Lite of Systran®[4] since we did not manage to fine-tune Facebook's NLLB-200-3.3B model, which was our first choice.

3.2 Translation and Post-editing with LLMs

For our two contrastive submissions, we have used NLLB-200-3.3B (NLLB Team et al.) baseline model and deep-translator[5] which was used by ChatGPT (GPT-4-turbo) to generate translations. All translations are then post-edited using prompt engineering via ChatGPT-4o.

[4] https://modelstudio-lite.systran.net/

[5] https://github.com/nidhaloff/deep-translator

4 Results

4.1 Qualitative Assessment

We have then compared three models in Systran Model Studio Lite to verify whether the in-domain Dialog model is adapted or not to the custom service conversation domain, by using the test set and reference translations published by the organisers of the Chat Task 2024. Table 1 compares the performance of three different models for language translation tasks: a fine-tuned model, an in-domain baseline model, and a generic baseline model. The performance is measured for two translation directions: English to French (en → fr) and French to English (fr → en).

	Fine-tuned model	In-domain baseline model	Generic baseline model
en → fr	**57.19**	48.05	50.47
fr → en	**55.02**	48.28	48.19

Table 1: Comparison of generic baseline, in-domain baseline and fine-tuned models of Systran®

The fine-tuned model shows a significant improvement over both baseline models in both translation directions. This highlights the effectiveness of fine-tuning in enhancing model performance for specific tasks. The in-domain baseline model performs slightly worse than the generic baseline model for en → fr but slightly better for fr → en. This suggests that the in-domain data may not always provide a consistent advantage over generic data without further fine-tuning. The results indicate the importance of model fine-tuning in achieving superior translation quality and accuracy, especially in specialised domains. They seem to support our approach and the effectiveness of our fine-tuning dataset.

To compare translations, we used quantitative methods such as *vocabulary growth*, *characteristic elements computation*, and *correspondence analysis* (Lebart et al., 1997; Fleury and Zimina, 2014; Zimina-Poirot et al., 2020) implemented in *iTrameur*[6] and *Voyant Tools*[7]. In Figure 1, generated with *iTrameur*, the vocabulary growth curves of three predictions, fine-tuned Systran (*systran_ft*), NLLB-200-3.3B (*nllb*), and Deep translator (*deep-translator*) can be compared

[6] https://itrameur.clillac-arp.univ-paris-diderot.fr

[7] https://voyant-tools.org

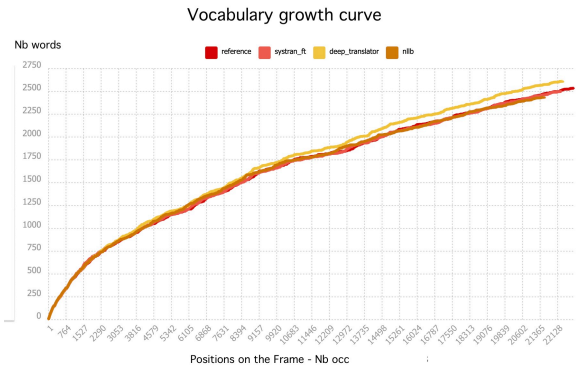

Figure 1: Vocabulary growth curve of reference translation and predictions of fine-tuned Systran, NLLB-200-3.3B and Deep translator.

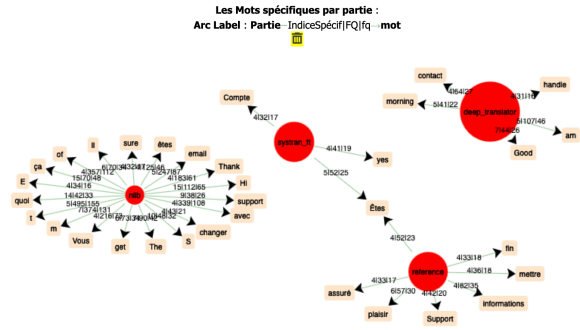

Figure 2: *Characteristic elements* computation for comparison of specific lexical features of reference translation and predictions of fine-tuned Systran (`systran_ft`), NLLB-200-3.3B (`nllb`) and Deep translator (`deep_translator`).

with the (*reference*) translation. While the reference translation is the longest (Nb occurrences: 22,834), it is followed by fine-tuned Systran (Nb occurrences: 22,291), which is the closest to the reference in terms of vocabulary growth.

In Figure 2 generated with *iTrameur*, we used *characteristic elements* computation to compare three predictions with the reference translation. The results show that many translation errors (including the occurrences of *E, S, t, Thank*, etc.) are over-represented in NLLB-200-3.3B prediction, while the reference translation and fine-tuned Systran prediction share common lexical features, such as identical translations *Are you still there?* ⇒ *Êtes-vous toujours là ?* attested by the over-representation of *Êtes*.

In Figure 3, we used correspondence analysis in *Voyant Tools* to compare our three predictions with the reference translation. The results suggest that the reference translation was carried out with human intervention, as it is clearly opposed

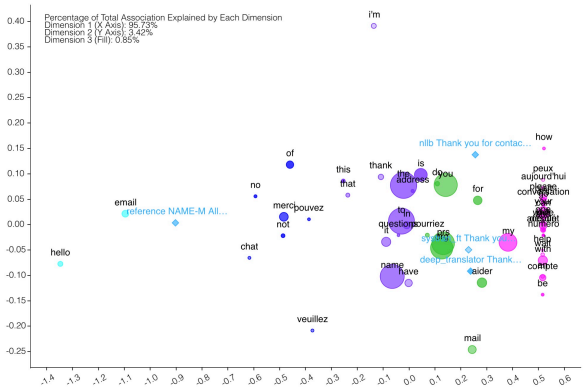

Figure 3: Correspondence analysis of the *reference* translation and tree predictions: fine-tuned Systran (*systran_ft*), NLLB-200-3.3B (*nllb*), and Deep translator *deep-translator*.

to three predictions (Zimina-Poirot et al. (2020) provides a discussion on this phenomenon). Although fined-tuned Systran is closer to `reference`, it is also very close to Deep translator, with NLLB-200-3.3B having a distinct profile.

Table 2 presents examples of segments that were incorrectly translated in our primary submission. It includes a comparison between the original source text, the reference translation, and our system's primary output, along with corresponding sentence-level BLEU and TER scores.

4.2 Comparisons of Primary and Contrastive Translations

In Table 3, we compared sentenceBLEU and TER scores of our *Primary* predicted by fine-tuned Systran model and two *Contrastives* predicted respectively by NLLB-200-3.3 baseline and Deep Translator. Except NLLB-200-3.3's predictions which have noticeably lower score, Deep Translator and fine-tuned Systran model have higher similar scores, which confirms our analysis of Figure 3. Deep Translator gets a slightly higher mean sentenceBLEU score, but its TER score is also higher. We noticed however that Deep Translator provided more literal or inaccurate translations of pragmatic expressions. It has translated *Bonjour* (greetings in French used in the daytime) by *Good morning*, and wrongly translated *You're welcome* by *Vous êtes les bienvenus*, which means "You are most welcome" in French.

Following the release of human evaluations, we have focused on mistranslations which scored 0 points, e.g. *I hope you have an excellent day* (source) is translated to *Merci pour l'information*

	Source	Reference	Primary	sentenceBLEU	TER
1	Is there anything else I can assist you with today?	Avez-vous besoin d'aide pour autre chose aujourd'hui ?	Y a-t-il autre chose que je puisse faire pour vous aider aujourd'hui ?	0.25	1.125
2	I am so sorry to hear that.	Je regrette sincèrement d'apprendre cela.	Je suis vraiment désolé de l'apprendre.	0.00	1.0
3	You are welcome!	Avec plaisir !	Je vous en prie.	0.00	1.33
4	You are welcome!	Ce fut un plaisir de vous parler.	C'était agréable de parler avec vous.	0.00	1.0
5	ok merci	Ok, thanks	Ok, thank you	0.00	1.0

Table 2: Mistranslated segments in our primary submission

("Thank you for the information"). The presence of these translation segments probably reflects misalignments in the fine-tuning data, as Systran Model Studio Lite does not necessarily filter out mismatching segments during the training process. These segments of the translation memory can be deemed correct as part of the normalisation process.

5 Discussion

5.1 Automatic Post-editing vs. Prompt Engineering

Pipelines for translation and post-editing using LLM engines were proposed with LLM engines (Vidal et al., 2022). The primary submission and the two contrasting submissions were subsequently post-edited by ChatGPT-4o using instructions such as:

```
"Post-edit the translations in file XX according to the source texts in file YY where English sentences are translated into French, and French sentences translated into English. Send me back in one single file",
```

where two raw text files are given: `XX` is line-separated `source` file and `YY` `translation` file. We noticed that when we asked ChatGPT-4o to post-edit by performing domain adaptation considering our dataset as a reference or knowledge base, it did not work.

The default instructions are ineffective when used with Anthropic Claude. To detect the language accurately, it is necessary to use language columns. In this context, using tags enhances the precision of the translation (without them, the translation will default to a single language). Adhering to the token limit is crucial, as failure to do so may lead to overlooking the total number of tokens in the input. Although the tag has been modified to "tear", it still functions as the translated message.

Another hallucination occurred when the instructions themselves were translated. Figure 4 illustrates the interface and the applied prompt. The French text contained several misspellings, homophonic confusions, such as *est* versus *ait*, participle versus infinitive confusions, and various conjugation errors. We also attempted to prompt LLMs to translate from the initial CSV file, but this strategy has limitations. The LLMs may suggest Python code to extract sentences in both languages, translate only one language, or perform the task for a limited number of sentences.

Using Anthropic Claude for translation also highlights the variability in LLM translations. For example, for the sentence *Pardonnez-moi je n'ai pas du bien formuler ma question.* Three translations were obtained: 1) "I'm sorry I must not have formulated my question well." 2) "I apologize I must not have phrased my question well." 3) "I apologize, I may not have phrased my question well."

6 Further Research

6.1 Retrieval-Augmented Generation (RAG)

The database serves as a vital resource for addressing the challenges posed by rare or complex structures that may not be well-represented in translation models (Gao et al., 2024). Retrieval-augmented generation (RAG) is a technique for enhancing the accuracy and reliability of generative AI models with facts fetched from external sources. Future improvements could involve aug-

Primary (Systran® fine-tuned)		Contrastive 1 (NLLB-200-3.3)		Contrastive 2 (Deep translator)	
sentenceBLEU	TER	sentenceBLEU	TER	sentenceBLEU	TER
0.70	0.25	0.57	0.50	0.71	0.28

Table 3: Primary and Contrastives metrics comparison (arithmetic mean)

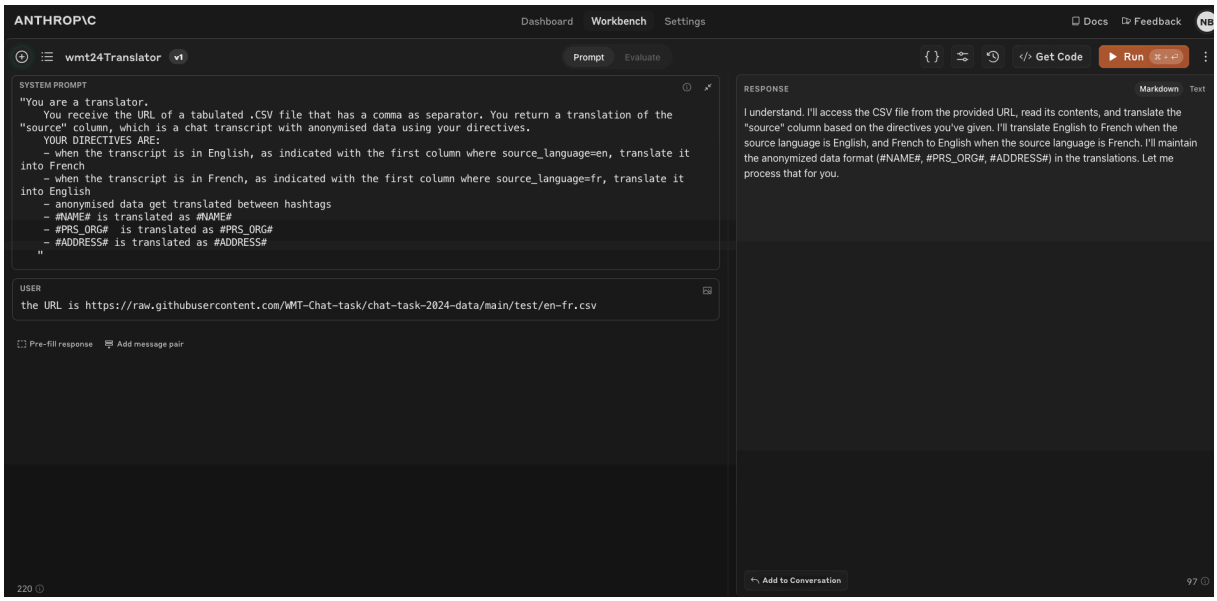

Figure 4: Anthropic Claude's interface with a prompt based on the URL of the WMT shared task test set

menting the training set with more examples, either through synthetic data or diverse real-world instances, to enhance the model's performance to translate challenging constructions, such as dislocations.

6.2 Explainability: Probing MT Systems for Trustworthy Outputs

Controlling LLM outputs and their repeatability is crucial for trustworthy AI. We tried to probe LLMs with (a) the detection of explicit representations and (b) their potential use in the LLM outputs. Similarly, in NMT, information might be available but not used by the system, as seen in the case of gender information discrepancies (Wisniewski et al. (2022a,b).

7 Conclusion

In this paper, we outline our methods for participating in the Chat Task 2024, focusing on enhancing translation quality in dialog-oriented machine translation systems through fine-tuning and prompt engineering. Our translation data files are available on GitHub[8]. Key findings indicate that fine-tuning an in-domain NMT model is feasible with minimal unlabelled data, resulting in significant improvements in translation quality. The research also emphasises the importance of analysing linguistic features in translations to identify strengths and weaknesses of different machine translation models. The study also highlights the necessity of ensuring explainability in LLM outputs to foster trust in AI systems.

Acknowledgements

This publication is the result of research supported by the scientific platform Pure Neural Server (PNS-UP)[9], partially funded by the 2024 research equipment grant MULTITAN-GML[10] (COPES-2024-12, *financement Fonds d'intervention Recherche, Université Paris Cité*) and the 2021 research equipment grant PAPTAN[11] from the Scientific Platforms and Equipment Committee, under the ANR grant (ANR-18-IDEX-0001, *financement IdEx Université de Paris*).

[8] https://github.com/lichaozhu/team_MULTITAN-GML_WMT24_Chat_Shared_Task

[9] https://plateformes.u-paris.fr/category/plateformes/traitement-automatique

[10] https://u-paris.fr/eila/actualites-projet-multitan-gml

[11] https://u-paris.fr/plateforme-paptan

References

Pushpak Bhattacharyya, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri, and Marco Turchi. 2023. Findings of the WMT 2023 shared task on automatic post-editing. In *Proceedings of the Eighth Conference on Machine Translation*, pages 672–681, Singapore. Association for Computational Linguistics.

Shamil Chollampatt, Raymond Hendy Susanto, Liling Tan, and Ewa Szymanska. 2020. Can automatic post-editing improve NMT? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2736–2746, Online. Association for Computational Linguistics.

M. Amin Farajian, António V. Lopes, André F. T. 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, Online. Association for Computational Linguistics.

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.

Serge Fleury and Maria Zimina. 2014. Trameur: A framework for annotated text corpora exploration. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 57–61, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Aarón Galiano-Jiménez, Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, and Juan Antonio Pérez-Ortiz. 2023. Exploiting large pre-trained models for low-resource neural machine translation. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 59–68, Tampere, Finland. European Association for Machine Translation.

Yuan Gao, Ruili Wang, and Feng Hou. 2023. How to design translation prompts for chatgpt: An empirical study. *arXiv preprint arXiv:2304.02182*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

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

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen-Tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is ChatGPT a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Surabhi Kumari, Nikhil Jaiswal, Mayur Patidar, Manasi Patwardhan, Shirish Karande, Puneet Agarwal, and Lovekesh Vig. 2021. Domain adaptation for NMT via filtered iterative back-translation. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 263–271, Kyiv, Ukraine. Association for Computational Linguistics.

Ludovic. Lebart, André Salem, and Lisette Berry. 1997. *Exploring Textual Data*. Text, Speech and Language Technology. Springer Netherlands.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Domain terminology integration into machine translation: Leveraging large language models. *ArXiv*, abs/2310.14451.

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. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore. Association for Computational Linguistics.

Giridhar Kaushik Ramachandran, Yujuan Fu, Bin Han, Kevin Lybarger, Nic Dobbins, Ozlem Uzuner, and Meliha Yetisgen. 2023. Prompt-based extraction of social determinants of health using few-shot learning. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 385–393, Toronto, Canada. Association for Computational Linguistics.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023. Leveraging GPT-4 for automatic translation post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024, Singapore. Association for Computational Linguistics.

Abhishek Sharma, Prabhakar Gupta, and Anil Nelakanti. 2021. Adapting neural machine translation for automatic post-editing. In *EMNLP 2021 Sixth Conference on Machine Translation (WMT21)*, pages 315–319.

Blanca Vidal, Albert Llorens, and Juan Alonso. 2022. Automatic post-editing of MT output using large language models. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 84–106, Orlando, USA. Association for Machine Translation in the Americas.

Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier, and François Yvon. 2022a. Analyzing gender translation errors to identify information flows between the encoder and decoder of a NMT system. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 153–163, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier, and François Yvon. 2022b. Biais de genre dans un système de traduction automatique neuronale : une étude des mécanismes de transfert cross-langue [gender bias in a neural machine translation system: a study of crosslingual transfer mechanisms]. In *Traitement Automatique des Langues, Volume 63, Numéro 1 : Varia [Varia]*, pages 37–61, France. ATALA (Association pour le Traitement Automatique des Langues).

Omry Yadan. 2019. Hydra - a framework for elegantly configuring complex applications. *Github https://github.com/facebookresearch/hydra*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: a case study. pages 41092–41110.

Maria Zimina-Poirot, Nicolas Ballier, and Jean-Baptiste Yunès. 2020. Approches quantitatives de l'analyse des prédictions en traduction automatique neuronale (TAN). In *JADT 2020 : 15èmes Journées Internationales d'Analyse statistique des Données Textuelles*, Toulouse, France. Université de Toulouse.