Preliminary evaluation of ChatGPT as a machine translation engine and as an automatic post-editor of raw machine translation output from other machine translation engines

Michael Farrell[1][0000-0002-7138-6639]

[1] IULM University, Milan, Italy
michael.farrell@iulm.it

Abstract. This preliminary study consisted of two experiments. The first aimed to gauge the translation quality obtained from the free-plan version of ChatGPT in comparison with the free versions of DeepL Translator and Google Translate through human evaluation, and the second consisted of using the free-plan version of ChatGPT as an automatic post-editor of raw output from the pay-for version of DeepL Translator (both monolingual and bilingual full machine translation post-editing). The experiments were limited to a single language pair (from English to Italian) and only one text genre (Wikipedia articles).
In the first experiment, DeepL Translator was judged to have performed best, Google Translate came second, and ChatGPT, last.
In the second experiment, the free-plan version of ChatGPT equalled average human translation (HT) levels of lexical variety in automatic monolingual machine translation post-editing (MTPE) and exceeded average HT lexical variety levels in automatic bilingual MTPE. However, only one MT marker was considered, and the results of the post-editing were not quality-assessed for other features of MTPE that distinguish it from HT. It would therefore be unadvisable to generalize these findings at present.
The author intends to carry out new translation experiments during the next academic year with ChatGPT Plus, instead of the free-plan version, both as an MT engine and as an automatic post-editor. The plan is to continue to evaluate the results manually and not automatically.

Keywords: machine translation post-editing, human machine translation output evaluation, DeepL Translator, Google Translate, ChatGPT, automatic post-editing

1 Introduction

Although ChatGPT has only been available to the public since the end of November last year, some evaluation studies have already been carried out on the chatbot's ability to translate between natural languages, including Turkish, Romanian, Chinese, English and German [5, 6 and 7]. However, to the best of the author's knowledge, all of these

Proceedings of the International Conference HiT-IT 2023, pages 108–113,
7-9 July 2023, Naples, Italy
https://doi.org/10.26615/issn.2683-0078.2023_009

have so far used automatic metrics for raw output quality evaluation. Moreover, the author is unaware of any attempts to use ChatGPT as an automatic post-editor of machine translation (MT) output from other MT engines.

This preliminary study consisted of two experiments. The first aimed to gauge the translation quality obtained from ChatGPT in comparison with DeepL Translator and Google Translate, and the second attempted to use ChatGPT as an automatic post-editor of raw output from the pay-for version of DeepL Translator, examining both monolingual and bilingual full MT post-editing (MTPE).

The results were analysed to assess how to proceed with a new series of further-reaching experiments.

2 Design and methods

2.1 First experiment

Seven post-graduate students of translation (IULM University, Milan) comparatively assessed the raw output from the free-plan version of ChatGPT, based on the GPT-3.5 architecture, and the free versions of DeepL Translator and Google Translate. Three short extracts from the biographies of heterogeneous celebrities (Yehoshua Bar-Hillel [313 words], G. H. MacDermott [342 words] and Michael Jackson [358 words]) were taken from the English language version of Wikipedia and machine-translated into Italian on 6 April 2023. The three outputs were then segmented in Raw Output Evaluator[1] [3] and presented to the students, who simply assessed the three translations as *best*, *second best* and *worst* on a segment-by-segment basis (ties were allowed). A score was then calculated by assigning three points for each segment regarded as *best*, two points for *second best* and one point for *worst*. This simple ranking technique was chosen both for its speed and because the students had not yet received any training on the use of analytic metrics.

Wikipedia articles were used since they are likely to be less challenging for a machine translation system than classic works of literature but more problematic than the boilerplate-style texts that are normally considered to lend themselves best to machine translation.

The simple prompt used to generate the translation in ChatGPT was “Please translate the following text into Italian”, followed by a line break and then the source text.

2.2 Second experiment

A short extract from an English-language Wikipedia entry on Slovakia was taken for the second experiment (262 words). This text was chosen since it contained the bigram *there are* four times. This was the first short extract that contained at least four examples of the chosen n-gram in the space of approximately 250 words that the author came across while searching randomly through Wikipedia. Again, a Wikipedia entry was chosen in order to pose a medium-level challenge.

[1] www.intelliwebsearch.com/raw-output-evaluator

The bigram *there are* had been identified in a previous experiment caried out between 2017 and 2018 as among the best MT markers, that is n-grams that were translated with a highly statistically significantly greater number of correct translation solutions in human translation (HT) than in MTPE [2]. In the 2017/18 experiment, the frequency with which *there are* was translated into Italian with *ci sono* was successfully used to distinguish MTPE from human translation.

The aim of the second experiment was therefore to see if ChatGPT was able to post-edit MT output automatically and reduce the lexical impoverishment that has been observed to ensue from human post-editing [2 and 8]. Lexical uniformity is not a positive feature in texts that need to be engaging and intellectually stimulating, such as in the fields of marketing, advertising, literature, journalism, education, entertainment, and creative writing in general.

First, the author checked if the free-plan version of ChatGPT could be prompted to produce raw output in which the chosen MT marker occurred with a frequency that was in keeping with its frequency in HT. The prompt used was "Please translate the text below into Italian, keeping in mind that lexical variety is required for a human-quality final text." This was followed by a line break and then "Here is the text in original language:". After that, there was another line break, followed by the English language version of the Slovakia text.

Next the author asked ChatGPT to do automatic bilingual post-editing of raw output obtained from DeepL Translator. This MT engine was chosen because, in a recent survey among professional translators [4], the MT system most used by the respondents who declared that they use MT at some point in their workflow turned out to be DeepL Translator (183 users). Its nearest rival Google Translate was only chosen by just over half that number (93 users). Moreover, the majority of DeepL Translator users surveyed stated that they use the pay-for version (102).

The prompt used was "Please post-edit the text below, which was machine-translated into Italian, keeping in mind that lexical variety is required for a human-quality final text." This was followed by a line break and then "Here is the text in original language:". After that, there was another line break, followed by the English language version of the Slovakia text. This was then followed by "Here is the text to post-edit:", a line break and then the raw output from the pay-for version of DeepL Translator.

Lastly the author asked ChatGPT to do automatic monolingual MTPE. The prompt used was "Please post-edit the text below, which was machine-translated into Italian, keeping in mind that lexical variety is required for a human-quality final text." This was followed by a line break and then the same raw output as used before.

To establish the normal average frequency of the most chosen translation solution in HT for the MT marker *there are*, the same text on Slovakia was translated into Italian by 18 post-graduate students of translation [1]. The most frequent translation solution, *ci sono*, occurred 50% of the time. This solution occurred in the DeepL machine translated text three times out of four (75%). So, a human post-editor would tend to be primed to use this solution with a higher than natural frequency.

3 Results

3.1 First experiment

DeepL Translator was assessed to have performed best with a total score of 279 points, Google Translate came second with 239 points, and ChatGPT, last with 186 points. This preliminary result cannot however be generalized since it concerns only one language pair (English to Italian) and one text genre.

3.2 Second experiment

In the first part of the second experiment, ChatGPT was prompted to produce raw output in which lexical variety was required for *a human-quality final text*. However, ChatGPT failed to deliver: the MT marker *there are* was translated with the bigram *ci sono* four times out of four (100%), which is twice the previously measured average natural HT frequency in this text (50%).

ChatGPT was then asked to do automatic bilingual post-editing of raw output obtained from the pay-for version of DeepL Translator. This time, ChatGPT left the bigram *ci sono* as the translation of the MT marker *there are* only once despite being primed by the raw output with three occurrences. In other words, in the case of this specific MT marker and this specific text, ChatGPT produced greater lexical variety than the students did on average.

Lastly, ChatGPT was asked to do automatic monolingual post-editing of the same raw output. The result was two occurrences of the bigram *ci sono*. Therefore, ChatGPT reached human parity as far as the chosen MT marker is concerned in this particular monolingual post-editing.

4 Conclusion

Wenxiang Jiao et al. [6] report that the ChatGPT Plus version, based on GPT-4 architecture, scores higher than the free-plan version in automatic MT raw output evaluation metrics. Consequently, the planned future experiments will be carried out using ChatGPT Plus, and not the free-plan version.

Another limitation of the first experiment was that the evaluators knew which MT engine had been used to produce the raw output they were evaluating. Although it is unlikely that they expressed biased opinions on the basis of this knowledge, the future experiments will be carried out blind.

The prompt used to ask ChatGPT to translate the text in the first experiment does not take advantage of ChatGPT's ability to emulate different styles [7]. Better results may have been achieved with a prompt like "Please translate the following text into Italian in the style of a Wikipedia entry" or by providing information about the source text and purpose of the translation.

Seven human evaluators is a small number, which will be increased in the planned future experiments to reduce subjective biases. However, the author will in any case be limited by the size of the class for all experiments, which is unlikely to be much in

excess of forty students. Another limitation that cannot be overcome is the language pair (English to Italian). Academic time constraints will also limit the length and number of texts that may be analysed and the complexity of the analysis metrics.

In the first part of the second experiment, ChatGPT was prompted to produce raw output with human-like lexical variety. However, it failed to do so, at least in the case of the test MT marker chosen. Again, the result may have been different if ChatGPT Plus had been used. The same experiment should also be repeated on more than one text.

The most remarkable results were seen when ChatGPT was asked to post-edit raw output from the pay-for version of DeepL Translator. In the case of the specific MT marker considered and with the particular text chosen, ChatGPT reached average human-level lexical variety in monolingual MTPE and exceeded it in bilingual MTPE.

It is a little unfair to tell ChatGPT to consider lexical variety and not give the same instruction to the human post-editors. In future experiments, it might be interesting to divide the human post-editors into two groups and ask half of them to bear lexical variety in mind.

The automatic post-editing output produced by ChatGPT also needs to be evaluated to see to what extent a further stage of human post-editing is required. Again, future experiments will be carried out with ChatGPT Plus.

Interestingly, the author has recently received an offer from a language service provider, based in Hong Kong, that specifically offers human-post-edited ChatGPT MT output as a service.

The author intends to carry out the new translation experiments with ChatGPT Plus as an MT engine and as an automatic post-editor during the next academic year. The plan is to continue to evaluate the results manually and not automatically.

References

1. Farrell, M.: Current evidence of post-editese: differences between post-edited neural machine translation output and human translation revealed through human evaluation. Proposed for: International Conference HiT-IT 2023 - Human-informed Translation and Interpreting Technology. Publication pending (2023).
2. Farrell, M.: Machine Translation Markers in Post-Edited Machine Translation Output. In: Proceedings of the 40th Conference Translating and the Computer, pp. 50–59. AsLing: The International Association for Advancement in Language Technology (2018).
3. Farrell, M.: Raw Output Evaluator, a Freeware Tool for Manually Assessing Raw Outputs from Different Machine Translation Engines. In: Proceedings of the 40th Conference Translating and the Computer, pp. 38–49. AsLing: The International Association for Advancement in Language Technology (2018).
4. Farrell, M: Do translators use machine translation and if so, how? Results of a survey held among professional translators. Presented at the 44th Conference Translating and the Computer. Preprint pending publication, DOI:10.13140/RG.2.2.33996.69768, (2022).
5. Işım, C., Balcıoğlu, Y. S.: ChatGPT: performance of translate. In Proceedings of 3rd International ACHARAKA Congress on Humanities and Social Sciences (2023).
6. Jiao, W., Wang, W., Huang, J., Wang, X., Tu, Z.: Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. Preprint (2023).

7. Peng, K., Ding, L., Zhong. Q., Shen, L., Liu, X., Zhang, M., Ouyang, Y., Tao. D.: Towards Making the Most of ChatGPT for Machine Translation. Preprint (2023).
8. Toral, A.: Post-editese: an Exacerbated Translationese. In: Proceedings of Machine Translation Summit XVII: Research Track, pp. 273–281. European Association for Machine Translation, Dublin, Ireland (2019).