Comparing Chinese-English MT Performance Involving ChatGPT and MT Providers and the Efficacy of AI mediated Post-Editing

Larry P Cady larry.cady@chilin.hk
Chilin (HK), Ltd., Santa Cruz CA, USA
Benjamin K. Tsou btsou99@chilin.hk
Chilin (HK), Ltd. and CityU HK, Hong Kong
John S. Y. Lee jsylee@cityu.edu.hk
City University of Hong Kong, Hong Kong

Abstract

The recent introduction of ChatGPT has caused much stir in the translation industry because of its impressive translation performance against leaders in the industry. We review some major issues based on the BLEU comparisons of Chinese-to-English (C2E) and English-to-Chinese (E2C) machine translation (MT) performance by ChatGPT against a range of leading MT providers in mostly technical domains. Based on sample aligned sentences from a sizable bilingual Chinese-English patent corpus and other sources, we find that while ChatGPT performs better generally, it does not consistently perform better than others in all areas or cases.

We also draw on novice translators as post-editors to explore a major component[1] in MT post-editing: Optimization of terminology. Many new technical words, including MWEs (Multi-Word Expressions), are problematic because they involve terminological developments which must balance between proper encapsulation of technical innovation and conforming to past traditions[2]. Drawing on the above-mentioned reference corpus[3] we have been developing an AI mediated MT post-editing (MTPE) system through the optimization of precedent rendition distribution and semantic association to enhance the work of translators and MTPE practitioners.

1. Introduction

In recent decades we have witnessed spectacular advancements in Machine Translation (MT) technology. More recently, a major turning point has appeared with the introduction of ChatGPT, which is based on a large language model and generative AI.

We pose the following questions: 1. What is the range of variation in the performance of popular MT systems on technical subjects? 2. Are there some clear leaders which perform consistently better than others? 3. What kind of tools can effectively enhance MT results (for

[1] MT post-editing includes several key stages and terminological optimization is the foremost. It entails (1) Identification of canididate constituents for improvement; (2) Appreciation of good available alternatives; and (3) Selection of appropriate alternatives. (See Tsou et al. 2022; Green et al. 2013)

[2] See Tsou et al. 2020.

[3] See also Goto et al. 2013.

Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track, pages 205–216
September 4–8, 2023, Macau SAR, China.

example, post-editing)? 4. How realistic is it for a robotic translator to replace human translator in the foreseeable future?

To answer these questions: 1. We select in this study authoritative and parallel English and Chinese texts with recognized human translations (e.g., parallel bilingual patents) for their technical nature and legal status. 2. We compare translation performance of ChatGPT with the others based on BLEU scores. 3. We focus on terminological deficiency and how to assist human subjects in remedying it. This study draws on novice human translators who would review and select among alternate translations of technical terms with and without reference to their authoritative usage frequencies, and analyze their selections with reference to the gold standards in the filed patents. We conduct C2E and E2C tests with and without access to external resources[4] and analyze the results in the context of questions raised above.

This paper begins with an examination of how ChatGPT 3.5 and 4 compare with some notable MT systems and explores the consequential implications for consumers and providers of MT technology, as well as what might be included in the timely introduction of AI mediated Post-Editing technology. We look at translation between Chinese and English on technical subjects, where there is high demand for quality and where cost is an issue. We first discuss our comparative analysis and some results from a preliminary small-scale study on how lexical improvement in Post-Editing may be achieved.

Our study is based on a set of 3,000 bilingual sentences drawn equally from patent documents involving biotechnology as well as computer science and electronics. We focus on the bidirectional translation between English and Chinese for these sentences among a number of well-known MT systems[5].

2. Comparative Performance of 7 Notable MT Systems

Among the large number of MT systems examined, we report on seven systems: Baidu, ChatGPT3.5, ChatGPT4.0, DeepL, Google, Niutrans and Youdao. Their BLEU scores on the 3,000 sentences in science and technology are taken as the basis for this study. An illustrative sentence and its alternate translations are given in Table 1 based on comparison between their MT ouput and the "Gold standard"[6] of human translation in the filed patents.

Table 1. Alternate Translations of an illustrative sentence[7]

From Chinese Patent	其中 SGLTs 家族中具有葡萄糖转运功能的成员主要分布于肠道和肾脏的近端小管等部位，进而推断其在肠葡萄糖的吸收和肾脏葡萄糖的重摄取等过程中均发挥着关键作用，因而使其成为治疗糖尿病的理想潜在靶点之一。

[4] See Tsou et al. 2022.

[5] The 3,000 test sentences are taken from more than 30 million parallel sentences from the PatentLex corpus developed by Chilin (HK) Ltd in Hong Kong and available from TAUS and LDC (see references). Chilin first cultivated and curated a large corpus of 300,000+ Chinese-English parallel/comparable patents, and from it 30+m bilingually aligned Chinese-English sentences. From that, a large corpus of bilingual multi-word expressions is being cultivated in conjunction within the developments of an AI-mediated Machine Translation post-editing system that makes use of these progressively winnowed databases.

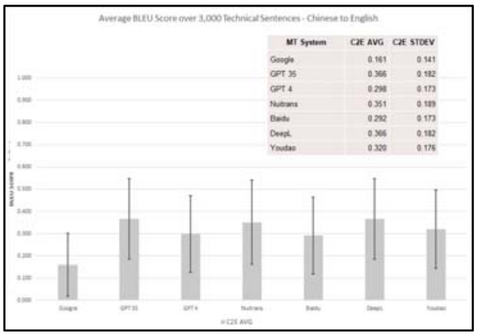
[6] The Gold standard for the 3,000 test sentences is taken to be from the corresponding target language sentences in the filed patents.

[7] From patent WO2004040948A1 (US Priority) "Apparatus and method for controlling registration of print steps in a continuous process for the manufacture of electrochemical sensors"

From English Patent	Members of SGLTs acting as glucose transporters are mainly distributed in the intestine and the proximal tubules of the kidneys, indicating that SGLTs are responsible for the majority of glucose reuptake in the intestine and the kidneys. SGLTs are considered as potential and ideal antidiabetic targets.	
Google	Among them, members of the SGLTs family with glucose transport function are mainly distributed in the proximal tubules of the intestine and kidney, and it is inferred that they play a key role in the process of intestinal glucose absorption and renal glucose reuptake, thus making them It has become one of the ideal potential targets for the treatment of diabetes.	BLEU Score 0.1704
GPT 3.5	Among the SGLT (sodium-glucose co-transporter) family, members with glucose transport function are mainly distributed in the intestine and proximal tubules of the kidneys. It is inferred that they play a crucial role in processes such as intestinal glucose absorption and renal glucose reabsorption, making them one of the ideal potential targets for the treatment of diabetes.	0.2005
GPT 4.0	The members of the SGLTs family with glucose transport functions are mainly distributed in the proximal tubules of the intestine and kidney, thereby hypothesizing that they play a key role in processes such as intestinal glucose absorption and renal glucose reuptake. Therefore, they are considered one of the ideal potential targets for the treatment of diabetes.	0.1876
Niutrans	Among them, the members of SGLTs family with glucose transport function are mainly distributed in the proximal tubules of intestine and kidney, and it is inferred that SGLTs play a key role in the process of intestinal glucose absorption and renal glucose reuptake, which makes them one of the ideal potential targets for the treatment of diabetes.	0.1592
Baidu	Among them, the members of the SGLTs family with glucose transport function are mainly distributed in the intestinal tract and the proximal tubules of the kidney. It is inferred that they play a key role in the absorption of intestinal glucose and the reabsorption of renal glucose, which makes them one of the ideal potential targets for the treatment of diabetes.	0.1718
DeepL	The glucose transporting members of the SGLTs family are mainly located in the proximal tubules of the intestine and the kidney, and are thus hypothesized to play a key role in both intestinal glucose absorption and renal glucose reuptake, thus making them an ideal potential target for the treatment of diabetes.	0.1635
Youdao	Members of the SGLTs family with glucose transport function are mainly distributed in the proximal tubules of the intestine and kidney, and it is inferred that they play a key role in the process of intestinal glucose absorption and renal glucose reuptake, which makes them one of the ideal potential targets for the treatment of diabetes.	0.1911

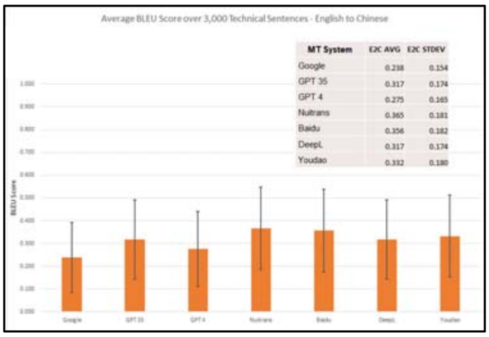
Even though the range of BLEU scores varies from 0.16 to 0.2 for this example sentence, human evaluation has found these translations to be useful for reference but not as final products.

It will be useful to look at the overall performance of Chinese-English translations of seven among the many MT systems we have evaluated. Figure 1A and 1B show the scores in the bidirectional English-Chinese translations of the seven systems. The error bars are based on one standard deviation.



MT System	C2E AVG	C2E STDEV
Google	0.161	0.141
GPT 35	0.366	0.182
GPT 4	0.298	0.173
Nuitrans	0.351	0.189
Baidu	0.292	0.173
DeepL	0.366	0.182
Youdao	0.320	0.176

Figure 1A. Average BLEU Scores Chinese to English



MT System	E2C AVG	E2C STDEV
Google	0.238	0.154
GPT 35	0.317	0.174
GPT 4	0.275	0.165
Nuitrans	0.365	0.181
Baidu	0.356	0.182
DeepL	0.317	0.174
Youdao	0.332	0.180

Figure 1B. Average BLEU Scores English to Chinese

We note there is a wide range of overall BLEU scores (0.16 to 0.36) in the translation of this set of 3,000 technical sentences, which are low in general even though the results provide useful references.

3. Pairwise Comparison between common MT Systems and ChatGPT

To make the comparison more meaningful, we did pairwise comparison among the translation systems by plotting the corresponding BLEU scores on a grid. Each axis ranges from 0.0 to 1.0, where 1.0 corresponds to a perfect match versus the reference sentence. Figure 2A shows a comparison between Google and ChatGPT-4 for the Chinese to English direction. Figure 2B shows the English to Chinese direction.

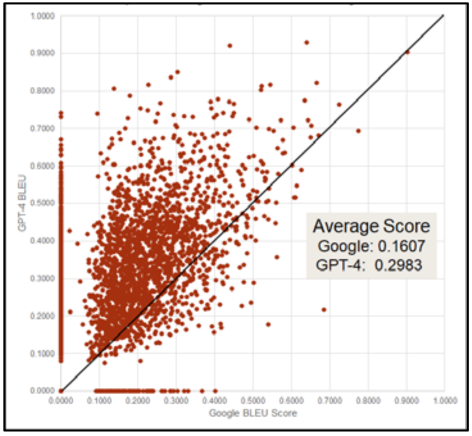


Figure 2A. Google vs GPT-4 : Chinese to English

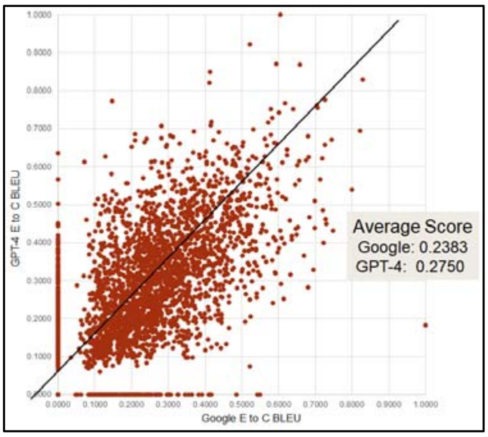


Figure 2B. Google vs GPT-4 : English to Chinese

Note that in the Chinese to English direction, ChatGPT4 outperforms Google. In the English to Chinese direction, they are very close. Note also that the data is widely scattered with many data points on the X and Y axis indicating 0.0 BLEU scores for one of the systems.

Figures 3A and 3B below compare the output of GPT-4 with Baidu.

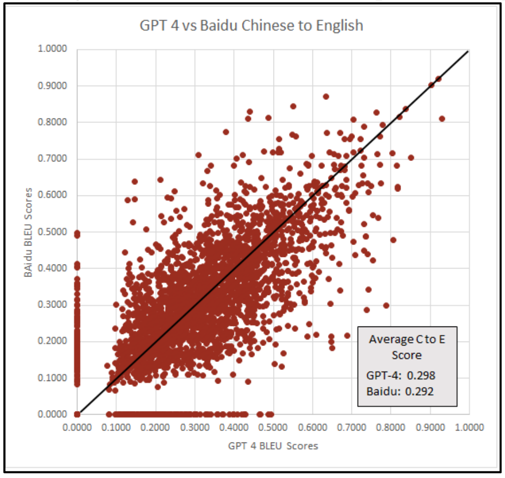


Figure 3A. GPT-4 vs Baidu:
Chinese to English

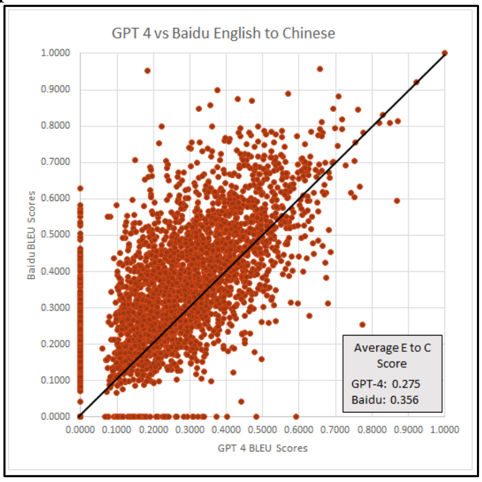


Figure 3B. GPT-4 vs Baidu:
English to Chinese

Baidu and GPT-4 are very close for Chinese to English but Baidu outperforms for English to Chinese.

Figures 4A and 4B compare GPT-4 with DeepL.

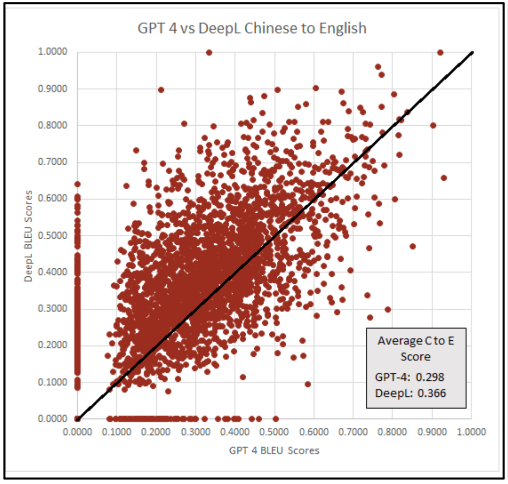


Figure 4A. GPT-4 vs DeepL:
Chinese to English

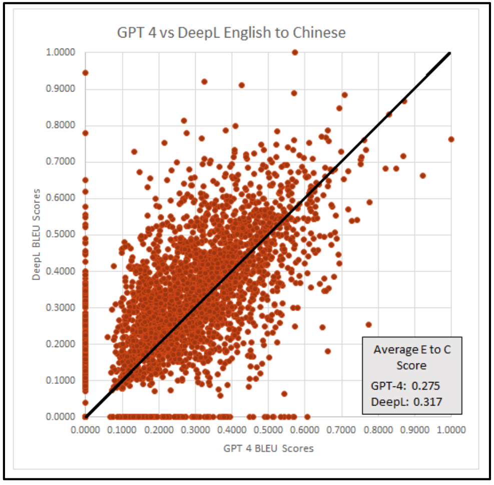


Figure 4B. GPT-4 vs DeepL:
English to Chinese

Note that DeepL outperforms GPT-4 in both directions.

Figures 5A and 5B below compare the output of GPT-4 and Niutrans.

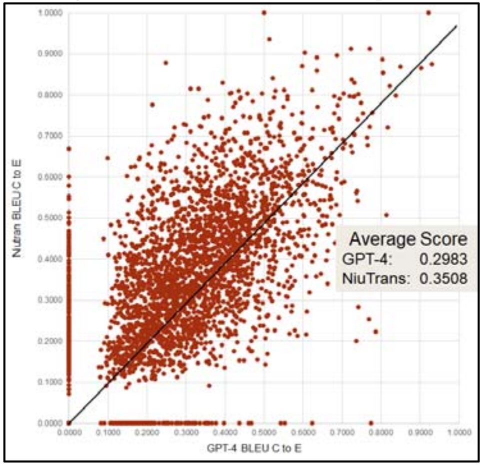


Figure 5A. GPT-4 vs Niutrans: Chinese to English

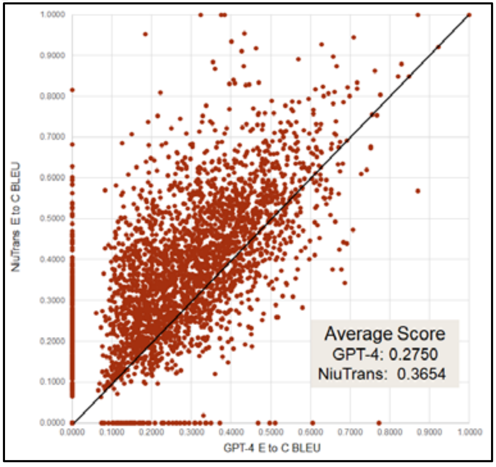


Figure 5B. GPT-4 vs Niutrans: English to Chinese

Note that Niutrans outperforms GPT-4 in both directions.

Figures 6A and 6B compare GPT-4 with Youdao.

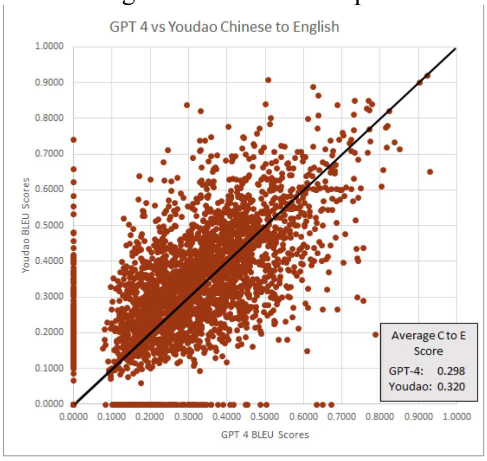


Figure 6A. GPT-4 vs Youdao: Chinese to English

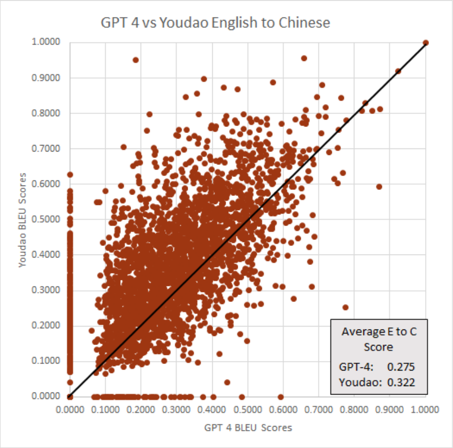


Figure 6B. GPT-4 vs Youdao: English to Chinese

Youdao performs well in both directions.

Furthermore, the scattered nature of the data points on all the comparison graphs show that there is no single system is consistently above all others.

4. Issues with Automated Scoring

Table 2 shows a second detailed example. Compared with Niutrans, the performances of DeepL and ChatGPT3.5 are at opposite ends of the performance range (0.23 to 0.33). We also note that a recent evaluation by Intento[8] ranked Google highest for Chinese to English and English to Chinese. Intento used a different methodology than we have used. Moreover, ChatGPT-4 performed better than ChatGPT3.5 in this example, as can be seen in the Appendix.

Table 2. Second Example Sentence for MT comparison

From Chinese Patent	合成/对抗疗法药物雷尼替丁的有效率为 82.20%，与制剂 F5 几乎相似，但长期使用会阻止胃液的正常分泌。	
From English Patent	The synthetic/allopathic drug ranitidine showed 82.20% which is almost similar to that of formulation F5 but long use block the normal secretions in the stomach.	
Google	The effective rate of synthetic/confrontation drugs Rennitine is 82.20 %, which is almost similar to the preparation F5, but long -term use will prevent the normal secretion of gastric juice.	BLEU Score 0.2641
GPT 3.5	The efficacy of the combination/antagonist therapy drug Ranitidine is 82.20%, which is almost similar to the formulation F5, but long-term use will prevent the normal secretion of gastric juice.	0.2326
GPT 4.0	The efficacy of the synthetic/antagonistic therapy drug Rennitidine is 82.20%, which is almost similar to Formulation F5, but long-term use may block the normal secretion of gastric acid.	0.2773
Niutrans	The effective rate of synthetic/allopathic drug ranitidine is 82.20%, which is almost similar to preparation F5, but long-term use will prevent the normal secretion of gastric juice.	0.3321
Baidu	The effective rate of the synthetic/antagonistic therapy drug ranitidine is 82.20%, which is almost similar to the formulation F5, but long-term use can prevent the normal secretion of gastric juice.	0.2696
DeepL	The synthetic/allopathic drug ranitidine is 82.20% effective, almost similar to preparation F5, but its long-term use prevents the normal secretion of gastric juice.	0.2326

The major differences involve the appropriate selection of translation for terms such as *synthetic*/*combination* and *confrontation*/antagonistic/*allopathic*. There are also issues of stylistic variations and anaphora, which affect the BLEU evaluation. However, these features may be important in high-value, high-demand translations in the legal or technical sector. Moreover, it may be noted that the "Gold standard" is not infallible, as it is ultimately human based[9].

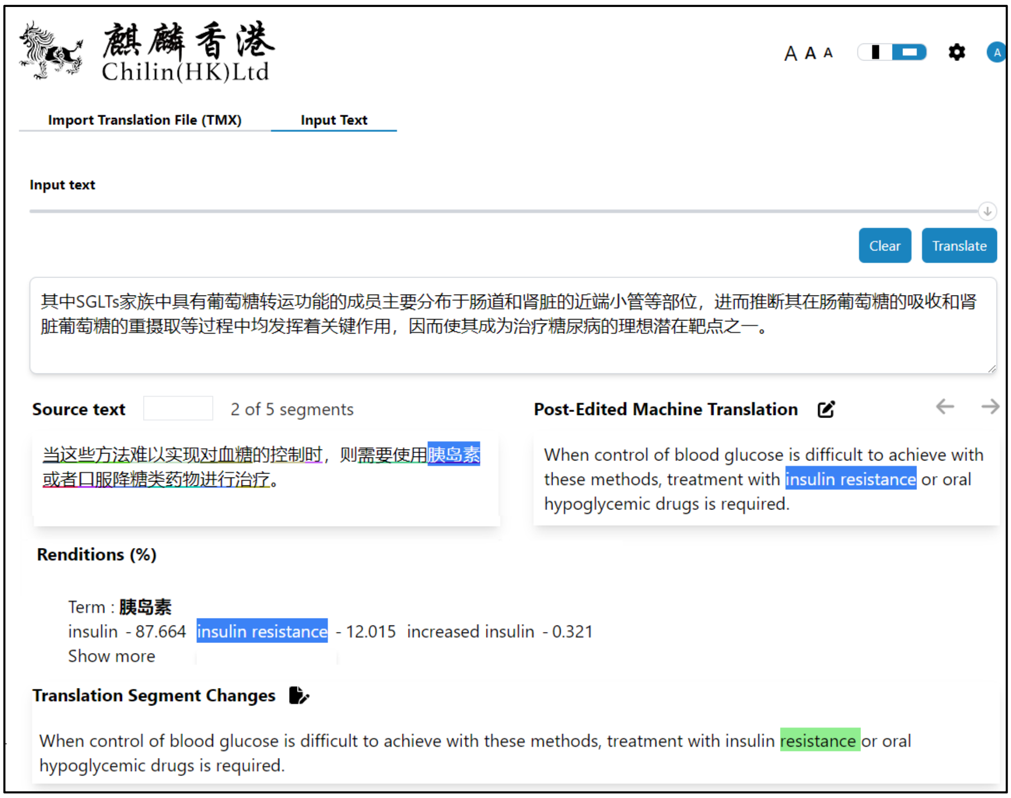
[8] See Intento 2023 State of Machine Translation Report. Intento rates Google and DeepL highest for Chinese to English; Google highest for English to Chinese. GPT is rated as competitive.

[9] In the Chilin database, it has sentence ID WO2005063271-100314. The WIPO document is WO2005063271A1 and https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2005063271. The Chinese document is https://patentscope.wipo.int/search/en/detail.jsf?docId=CN83099623. The original application is from India and appears to be in English. In addition to incorrectly using ***USE*** instead of ***USES*** in the English document, the drug name "Ranitidine" is sometimes capitalized and sometimes not capitalized. These are examples of imperfection in the "gold standard" text. This application was filed in several countries (including the US and China) but was only granted in the UK.

The average scores and plots in Figure 1 to Figure 7 are informative, but the widely scattered data and large number of very low scores suggests that MT performance is highly unpredictable.

Given the widely scattered nature of the results, we conclude that the seven systems are all generally competitive with each other, without any system consistently outperforming all others. It should be noted that all MT systems benefit from large data based on past practices and are not sensitive to innovations which are natural in any evolving system as may be found in natural language. In selecting a system for production use, the user should run a suitable test with data that is representative of what he or she will encounter operationally, and examine the results in the light of appropriate criteria and requirements.

Given that MT performance has become very effective, the likelihood of manual preliminary draft translations will diminish quickly. This is especially so for high value documents, such as legal contracts and patent filings, where a careful review and post-editing of the machine translation outputs is inevitable in the foreseeable future. With support from the Hong Kong Innovation and Technology Commission, Chilin (HK) Ltd is developing the PatentLex Translation Assistant (PaTTA) to perform AI-mediated Post-Editing. It utilizes Chilin's large corpus of bilingual technical terms and parallel sentences to help translators review and edit machine translations. Figure 7 shows how PaTTA can identify technical terms and suggest translation options (multiple renditions with frequencies) in post-editing.



其中SGLTs家族中具有葡萄糖转运功能的成员主要分布于肠道和肾脏的近端小管等部位，进而推断其在肠葡萄糖的吸收和肾脏葡萄糖的重摄取等过程中均发挥着关键作用，因而使其成为治疗糖尿病的理想潜在靶点之一。

Source text 2 of 5 segments

当这些方法难以实现对血糖的控制时，则需要使用胰岛素或者口服降糖类药物进行治疗。

Post-Edited Machine Translation

When control of blood glucose is difficult to achieve with these methods, treatment with insulin resistance or oral hypoglycemic drugs is required.

Renditions (%)

Term : **胰岛素**
insulin - 87.664 insulin resistance - 12.015 increased insulin - 0.321
Show more

Translation Segment Changes

When control of blood glucose is difficult to achieve with these methods, treatment with insulin resistance or oral hypoglycemic drugs is required.

5. Towards Terminological Enhancement in Post-Editing

As a consequence of the findings reported above, a new generation of MT post-editors will be needed who would be able to perform the following tasks:

1. Review different MT outputs and identify the most suitable ones for necessary subsequent post-editing before incorporation in the final translation.
2. Improve on the selected output through both light post-editing and heavy post-editing as necessary. A major task or challenge will be to select the best amoung alternate translations of unfamiliar terms to conform to user or client requirements.

In view of these requirements, we have been curating our big database to provide a timely and useful means to improve post-editing results. On this basis, our PatentLex bilingual terminology database provides access to comprehensive alternate renditions of technical terms which should be beyond the usual repertoire of the translator. This is especially true for relatively novice translators who have had insufficient training or exposure relating to technical subjects. It also provides access to additional information such as the authoritative usage frequencies of the alternate renditions in authentic context.

Table 3 summarizes the results of a recent study focused on the post-editing process involving terminological improvement: Translators were asked to look at two sets of uncommon terms and for each given term, two alternate translations taken from the output of two separate MT systems. One of these being the same as that found in the referenced "Gold standard" of the filed patent. Each translator is given two seperate tasks: (1) He/she makes a choice between the two alternative translations in the context of the example sentence given (2 Alt.), and (2) He/she is given additional information on the usage frequency distribution of the two alternate renditions from the massive PatentLex database (2 Alt.+PAT.). The translator's choice is assessed to the "Gold standard". The accumulative results of the two sets are compared to see if the additional PatentLex information provided has helped the translator's final choice in line with the "Gold standard".

The study was conducted among year 3 university students interested in translation from a university among the top 200 out of over 3,000 tertiary institutions of education in Mainland China. They also provided information on their English score and time spent on the exercise.

Table 3. Comparison of Terminological Enhancement in Translation[10]

	2 Alt. (%)	Time (min)	2 Alt. + PAT. (%)	Time (min)	CET4	CET6
C to E	52	17.05	70	17.03	498	430
	65.1	30.17	48.6	26	480	497
	2 Alt. (%)	Time (min)	2 Alt. + PAT. (%)	Time (min)	CET4	CET6
E to C	42.1	18.65	62.9	19.17	496	439
	64.3	17.53	30	18.16	485	497

The student responses show two kinds of opposing tendencies: In C to E tasks, about half of the students, when given the two alternate renditions and additional PatentLex information such as distributional frequencies improve their performance from 52% to 70% with reference to the "Gold standard". For the remaining ones their performance dropped from 65% to 48.6% under similar conditions. Such a contrasting trend is unusual and invites attention and explanation. One could be from the putative observation on the possible correlations between the students' survey performance and their mandatory test score on English: CET4 and CET6. CET (College English Test) is required of university students in China for graduation. It is taken upon entry at university and CET6 is taken subsequently before graduation. CET6 has higher requirements than CET4, and the threshold is usually 425 for most institutions.

Some generizations may be made from Table 3. The set of technical terms used in the exercise contain relatively uncommon words for the students and the results show that those

[10] Alt: alternatives; +: additional usage frequency information among the altenatives in PatentLex; CET: China's College English Test.

students relatively weaker in English readily relied on the additional PatentLex information provided when doing the C to E exercise. Their improvement is from 50% to 70%, and it is northworthy that the time they spent on both tasks is about the same and equal to 17 minutes.

On the other hand, the students relatively stronger in English seemed more engaged with the task and materials given. They spent considerably more time, averaging 26 to 30 minutes (compared to only 17 minutes for the same task by the other group). Despite their strenuous efforts, these year 3 students' limited repertoire in English did not allow them to benefit substantially within the time allotted.

The corresponding exercise on the translation of English terms to Chinese took place after the exercise on Chinese to English term translation. A similar trend may be observed for students relatively weaker in English (as indicated by CET scores) i.e., the extra PatentLex information benefited only one type of students: those relatively weak in English.

In general Students relatively stronger in English seem unable to benefit from the extra information provided. In both cases there could be additional contributing factors which could account for the variations. The CET scores are only for English and not for Chinese, and it shows only one aspect of a student's bilingual ability which is more than simply the combined knowledge of each language. The biggest drop of 34% follows from the introduction of additional PatentLex information in the last part of the exercise. This could be due to the fact that the students whose native language was Chinese were insufficiently stimulated by PatentLex input to help them resolve monolingual lexical issues even in Chinese because they were beyond the students' usual repertoire. This was not dissimilar to the case of Chinese translation to English terms discussed earlier. However, because of the added pressure of time, the stimulation effect of the extra information brought about greater uncertainty and the biggest drop from 64.3% to 30%.

This study has been useful in several ways. It draws attention to the positive impact which the additional PatentLex information could make when directed at appropriately motivated post-editors and that its impact on truly experienced translators as well as the need for the broader process of post-editing to be investigated.

6. Conclusion

Despite the increasingly impressive MT performance of ChatGPT over MT providers, its superior performance is neither exclusive nor consistent. This will likely give rise to the imminent deployment of a new generation of MT post-editors able to make judicious comparisons and revisions, and to replace front-line human translators. The development of AI mediated MTPE systems which provide both important terminological alternatives and the relevant contexts is an crucial direction of development. Our preliminary study shows that well-motivated measures could demonstratably enhance the productivity of suitable MTPE practitioners, and that MTPE can improve speed and quality, as well as versatility (i.e., range of unfamiliar subjects for those already familiar with the grammatical structure of the target language). MTPE will help translators identify sentences that require additional editorial work and will facilitate their subsequent efforts. At the same time, the process of post-editing among different practitioners in terms of experience, linguistic knowledge and attitude deserve further attention just as the capabiities of LLMs and other technologies deserve to be further explored.

We also note the limitations of using BLEU scores to assess translations. BLEU scores are based on matching a translated sentence with reference translations, which are assumed to be the best possible. In reality, there often may be better alternates, especially over time. This means the training databases should be updated regularly and special measures should be made to sensitize the MTPE practioners proactively to the developments, which no robotic system is capable of doing in the foreseeable future.

Acknowledgments

We gratefully acknowledge support from: (1) Hong Kong's Innovation and Technology Commission (ITC) grant to Chilin (HK) Ltd for the PaTTA project (ITC/ESS Project: B/E019/20) and (2) CityU Strategic Research Grant (projects #7005709 and #7005803).

References

Chan, Elsie K. Y., Lee, John S. Y., Cheng, C., and Tsou, Benjamin K. (2023). Post-editing of Technical Terms based on Bilingual Example Sentences. In *Proc. 19th Machine Translation Summit*.

Green, S., Heer, J., and Manning, C.D. (2013). The Efficacy of Human Post-Editing for Language Translation. In *Proc. CHI*.

Goto Isao, Chow, K.P., Lu Bin, Sumita Eiichiro, Tsou Benjamin K. (2013). Overview of the Patent Machine Translation Task at the NRCIR-10 Workshop. *Proceedings of the 10th NTCIR Conference* 18-21 Jun 2013 Tokyo, Japan. pp. 260 -286.

Lee, John S. Y., Tsou, Benjamin K., and Cai, Tianyuan. (2020). Using Bilingual Patents for Translation Training. In *Proc. 28th International Conference on Computational Linguistics* (COLING).

Lu, Bin, Benjamin K. Tsou, Jingbo Zhu, Tao Jiang, and Oi Yee Kwong. (2009). The Construction of an English-Chinese patent parallel corpus. *MT Summit XII: Third Workshop on Patent Translation*, Ottawa (Canada), pp. 17-24.

Tsou Benjamin K., Chow, K.P., Lee, John, Yip, K.F. Ji, Y.X. and Wu, Kevin. (2020). "Bilingual Multiword Expressions, Multiple-correspondence, and their Cultivation from Parallel Patents: The Chinese-English Case". In Minh Le Nguyen, Mai Chi Luong, and Sanghoun Song (Eds.). In *Proceeding of 34th PACLIC Workshop on MWEA*, Hanoi, Vietnam. pp. 589-602.

Tsou, Benjamin. (2022). "Translation 4.5: The Age of Post-Editing Technology" Keynote Speech, 50th Anniversary Translation Lecture Series 2021-22 (5), 9 Apr, HK Translation Society: Hong Kong.

Tsou Benjamin K., Yiu, Elvis, and Mak, Kelly. (2022). "Machine Translation Post-editing and the Role of Big Data in Technical Translation". International Congress on English Language Education and Applied Linguistics (ICELEAL 2022), Dec 6-9, The Education University of Hong Kong.

Chilin PatentLex Bilingual sentence data is available at the University of Pennsylvania Linguistic Data Consortium (LDC) site. https://www.ldc.upenn.edu/language-resources/data, and on the TAUS Data Marketplace https://datamarketplace.taus.net/

"The State of Machine Translation 2023 (inten.to)" https://inten.to/machine-translation-report-2023/ (free with registration) 20 June 2023.

Appendix A. Comparison of the technical translation by GPT-4 and GPT 3.5.

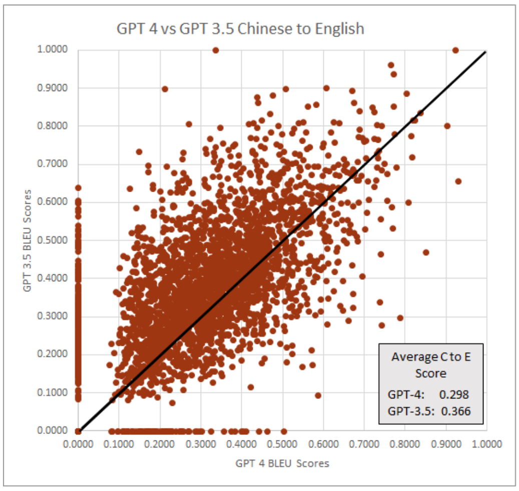


Figure 7A. GPT-4 vs GPT 3.5: Chinese to English

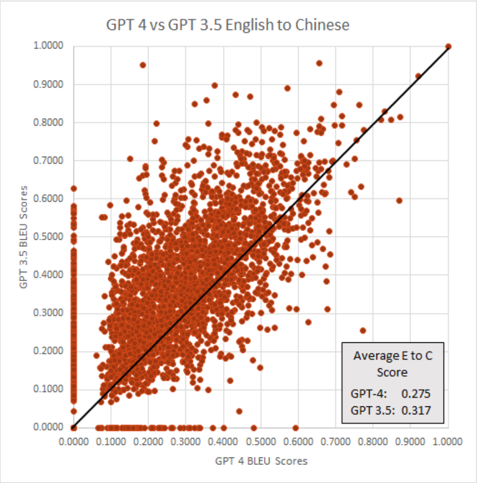


Figure 7B. GPT-4 vs GPT 3.5: English to Chinese

It is notable that GPT-3.5 outperforms GPT-4. On the Chinese to English translation of the source of 3,000 technical sentences between the two are significant. These interesting findings are only tentative but draw attention to the need to better understand the linguistic and related background of those doing translation and post-editing under time constraints**.** ChatGPT is focused on improvement in the generation of human-like text based on prompts and not strictly on translation. The lower BLEU scores of GPT-4 shows that there is likely a gap between the improved "human-like" output and the practised wisdom embedded in the accumulative database of Patents because of the much larger language models used in the training of the former. This is not surprising because the PatentLex database is retrospective while any successive attempt at human-like output would include innovations and suitable stylistic perferences.