Post-Editing Vs Neural Machine Translation: A Comparative Study of English ↔ Mandarin Translations in Daily Conversations

Yanfu Liu[1(✉)], Cheng Guo[2], and Sourojit Ghosh[3]

[1] Purdue University, West Lafayette, IN 47906, USA
liu4122@purdue.edu
[2] University of California, San Diego, CA 92093, USA
c5guo@ucsd.edu
[3] University of Washington, Seattle, WA 98195, USA
ghosh100@uw.edu

Abstract. Language translation continues to be one of the most common AI-moderated tasks performed in this multicultural world, especially with the rising popularity of large language models such as GPT-4. In this paper, we revisit a human-centered approach to language translation: post-editing. We compare language translations between English and Mandarin – two of the most-spoken languages globally – as performed by the Neural Machine Translator (NMT) employed by Google Translate to one of the state-of-the-art post-editing services in Translate.com. Through a mixed-methods approach combining qualitative analysis of results by 13 bilingual interviewees and quantitative analysis performed by 5 large-language models, we make a case for post-editing and the human element in language translation continuing to stay relevant in this ML-driven age. Though we recognize the cost differential is a significant pain point for users, we demonstrate post-editing producing translations of higher quality over NMT results in almost all cases, especially in conversations containing colloquialisms.

Keywords: Post-editing · NMT · Translation

1 Introduction

In this increasingly multicultural age, language translation has become one of the most-performed tasks each day, with Google Translate recording over a billion users translating 500 million words each day as of April 2021 [25]. Though Google Translate remains the most popular online translation service, the recent uprising of large language models (LLMs) such as GPT-4 through infectiously-popular tools such as ChatGPT, its position as the state-of-the-art machine translation (MT) service is under threat. As Google battles for supremacy over

H. Degen and S. Ntoa (Eds.): HCII 2025, LNAI 15820, pp. 373–387, 2025.
https://doi.org/10.1007/978-3-031-93415-5_22

the AI-moderated language translation market with Bing Translate, OpenAI (the parent company behind GPT-4) and others, there seems to be a steady and concerning reduction of the human element in language translation, which we believe is crucial to the success of translations. This concern is also important for the Human-Computer Interaction (HCI) community, as it highlights the need to balance AI and human involvement in the human-centered AI field. We suggest AI should enhance efficiency in human work rather than replace human's judgment.

In this paper, we contribute towards the growing movement towards human-centeredness in language translation by revisiting a long-existing but under the radar translation procedure: post-editing (PE). PE straddles the space between purely manual and purely machine-driven translation, which operates through human translators conducting machine translations and manually correcting any observed errors [1]. We compare results from English ↔ Mandarin translations between Google Translate's NMT and PE performed by one of the leaders in the field, Translate.com[1], across 4 hand-crafted scenarios that we believe are common across cultures through a mixed-methods study combining 13 qualitative interviews with fluent bilinguals and computational evaluation with 5 LLMs. We make the following novel contributions:

(1) We adopt as the text of our translation tasks a set of four conversations, which are scenarios we believe are experienced across different cultures and contexts without much change in overall tones. This approach is different than other studies, who use for their task excerpts from newspaper articles [15], Wikipedia posts [11], or online text corpora [20], which might be good for research purposes but do not reflect the daily needs of users whose conversations are far more informal than heavily-edited newspaper articles or informative Wikipedia pages. Our approach focuses on conversational sentences and colloquialisms resembling everyday speech which users of translation services are likely to query, and is vindicated through the success of our scenarios towards our goals of comparing translation qualities. Our findings have implications on the future of research into language translation, as we demonstrate the efficacy of choosing source texts that closely correspond to regular use cases.

(2) We do not simply conduct an evaluation into the quality of translations across NMT and PE services, but rather adopt a human-centered approach to make a holistic consideration over which service is a user more likely to use. Through our investigation, we find that while PE might produce superior-quality translations to the ones produced by NMTs, the cost differentials between the two (especially since Google Translate is free of charge) form a significant pain point for users and drives them away from PE services. Rather, we find that users obtain translations from Google Translate, interpret them and manually correct any inaccuracies, thus performing the task of PE themselves. This finding has significant implications in shaping the future of PE research, which hitherto considers it as a translation service to be provided to users [17], rather than a daily, common, active undertaking by end-users themselves. We contribute towards

[1] https://www.translate.com/.

Robertson's [27] design of human-centered machine translation by exploring how to help people make use of imperfect translation, by highlighting the amount of work users are currently putting in to deal with such imperfections.

(3) We conduct this study in translations across English and Mandarin, two of the highest-resource languages in the world. It's significant to invest these two languages to repair the communication gaps in business and cultural exchanges. Though we expected translation qualities in both directions across both NMT and PE to be high, this was not the case. Translation qualities were average to low across both languages in both NMTs and PE, which is concerning given how widely-spoken these languages are. This finding also bodes ill for lower-resource languages, and in light of demonstrably poor performance of NMTs and Large Language Models (LLMs) alike in translations into and from low-resource languages (e.g., Fitria, 2021 [9]; Ghosh and Caliskan, 2023 [10]; Prates et al., 2020 [26]), has alarming implications for the field of machine translation (MT).

2 Background

The history of MT has seen the adaptation of three broad approaches: Rule-based, Statistical machine translation (SMT), and NMT. Rule-based MT is an approach based on hard-coded linguistic rules [14]. Originating prior to the advent of machine learning techniques, rule-based translation operated through a collection of rules, a lexicon, and software programs to process the rules [14]. However, such approaches were largely naive and needed to be improved upon by SMTs, which treated language translation as a mathematical problem. SMTs adopt a data-driven approach that uses parallel aligned text corpora, assigning every word/sentence in the source language to probabilities of being correctly translated in the target language where translation accuracy is a function of the probabilities of source-target pairs matching [14]. SMTs demonstrated a marked improvement to rule-based approaches and were the state-of-the-art in the field for a long time.

Over the past two decades, the success and popularity of SMT systems were matched and eventually surpassed by NMTs due to distributed representation and end-to-end learning [18,36]. Pioneered by Sutskever [29] and forming the architecture within Google Translate, NMTs employ an encoder network to map source sentences into a real-valued vector, from which a decoder network produces the translation. Its strength comes from the neural network architecture, which learns from large amounts of data and effectively adapts to new contexts.

In recent months (2022–23), the rise in popularity of LLMs such as GPT-3 and GPT-4 within the massively popular ChatGPT has caused massive stirs in all AI-mediated markets, and MT was no exception. With ChatGPT growing as a language translation tool [10,16], it remains to be seen whether NMTs retain market supremacy.

However, despite their current standing as the state-of-the-art in language translation, NMTs are by no means perfect in their operation. NMTs can and

do still produce a variety of errors during translation, from grammatical errors to syntactic errors, to unnecessary additions or omissions, to errors in lexical or terminological choice, and errors in collocation or style [24]. For example, a sentence for which the correct English translation would be "The house had no running water" could instead be erroneously translated to "The house had no flowing water", due to a collocation error [24]. Such errors might not also be consistent across large sections of text, as NMT-mediated translations for one sentence it might produce a contextually accurate translation, but the next sentence might contain errors in meaning, omissions, additions, or a stylistic problem.

The task of identifying and fixing these errors is called *post-editing* [24]. In the PE workflow, the source document is first translated with MT, and a human translator edits the MT to produce the final translation [30]. PE serves as a best-of-both-worlds solution between fully manual and machine-produced translation, and can be more efficient and accurate than either [11,28].

As can be imagined, the primary hurdle to cross in the context of PE is the human post-editor, since the quality of translation is heavily dependent on them. It is worth noting that simply being a skilled or professional translator does not make someone a good post-editor, since two significant components of PE work are to produce translations at a fast pace and not produce large changes within MT results [24]. Researchers have also tried to 'automate' the PE process, or the automatic incorporation of PE into training data of MT algorithms to generate higher-quality translations [6,19].

Therefore, research into PE has generally been into improving the quality of MT, rather than providing a viable alternative to users seeking translations. In this paper, we make a case for PE being a human-centered alternative to MT.

3 Methods

3.1 Scenario Formation and Translations

To study our stated goals of comparing English ↔ Mandarin translations between NMTs and PE services, we decided to adopt the services of Google Translate and Translate.com as the respective representatives of the two. Google Translate is built upon the Google Neural Machine Translation system (GNMT), which utilizes state-of-the-art training techniques to achieve the largest improvements to date for MT quality [22]. Translate.com is one of the highest rated[2] publicly-available PE services, offering PE and other professional translation services since 2011.

We then designed four scenarios which we believe transcend different countries and cultures, 2 each in English and Mandarin.

Scenario 1 (English): Two friends run into each other on the street, after having not interacted much over the past month. They greet each other, share a joke, and discuss how they have both been busy with school and internships. The conversation is brief since they are going their separate ways.

[2] https://www.translate.com/reviews.

Scenario 2 (English): Three friends are walking down the street as they try to decide upon a restaurant for lunch. They consider an option that they spot on their way, and though they agree upon it, they balk at the sight of a long queue and wait time. One of them suggests another option across the street and vouches for its quality, but the group instead agrees on a restaurant right next door because they are tired of walking.
Scenario 3 (Mandarin): A son answers his phone, and his parents are on the other side. His parents ask why he hadn't called in a while and upon hearing how busy he is with work, further inquire if he has been eating and resting well. Their son assures them he is fine and taking care of himself, but needed to hang up because he had to go.
Scenario 4 (Mandarin): Two friends meet at one of their apartments, and set about playing video games. The host picks a game to play that the visitor is new to, they are excited to play together as they are both good at video games. The host starts the game and shows their friend some of the features, missions list, and in-game inventory.

The text of the above scenarios is constructed to mirror as closely as possible to real conversations that all the authors have personally experienced, with casual tone and containing colloquialisms. The detailed text of the scenarios, both in English and Mandarin.

We then prepared the NMT and PE translations of the four scenarios. For NMT translations, we fed the scenarios into Google Translate, supplying the translator with the entire scenario so as to provide the system with as much context as possible. To mirror the translations that potential users would receive if they performed the same task, we do not edit the NMT versions of the translations at all.

For the PE translations, we contacted Translate.com and set up an account, following which we supplied it with the scenarios. Translate.com priced the initial translations at $0.07 per word, though any edits were free of charge. Though we did not intend to comment on the quality of the PE translations, we were forced to do so on account of the first round of PE results being exactly the same as MT. We requested a second round of translations, and found those free of typos, though we do not otherwise comment on the quality. The PE process through Translate.com cost $67.84.

3.2 Recruitment, Participants and Interviews

Having designed the scenarios, we set about recruiting users of translation services to conduct our study. We recruited users through our social and personal networks, screening in individuals fluent in both English and Mandarin who self-identified as regularly conducting translations, either manually or through online services such as Google Translate, Bing Translate, ChatGPT or others. Through our efforts, we recruited 13 interviewees, hereafter referred to as P1-13. Interviews were conducted either in person or via Zoom, based on interviewee preferences.

In our interviews, we asked participants about what languages they spoke and how fluent they were in such languages, with a specific focus on their fluency in Mandarin and English. We also asked how they performed translations (manually or through services) and how often they did so. We specifically asked them for instances and situations where they needed to perform translations and which services, if any, they did so. We asked about their overall satisfaction with translation qualities, if and how often they encountered errors in produced translations, and how they dealt with such errors. Finally, we specifically asked them about their usage and experience with the two services we use here: Google Translate and Translate.com.

The second stage proceeded with us posing them the four scenarios (in random order) and associated translations, and asking for them to comment on translation quality and comparisons across the two. Participants were not informed which of the provided translations were from which service. Participants were also asked to identify what, if any, they perceived to be the deterministic difference.

Finally, the interview concluded with the information to participants about the cost of PE on Translate.com, something which was new information to all our participants. We asked them a question about their decision on which service, out of Translate.com, Google Translate or others, they would choose in the future. The study was approved by the authors' home institution's Institutional Review Board.

3.3 Analysis

We began our analysis of interview results with a grounded theory approach [5]. Researchers individually coded interviews on the online coding tool Taguette[3], and clustered observed codes into themes. We also tabulated interviewee preferences (PE vs. NMT) for each scenario.

We supplement our qualitative analysis with evaluation of translations through five LLMs. These LLMs were chosen on the criteria of being open source, trained on large corpora of multilingual data, and effective at predicting and determining the quality of segments of text in conversations. The chosen models are: BERTScore [35], BERT [8], GPT-Neo [4], XLNet [33], and Yoso [34]. The quality of translated content was evaluated by determining how 'natural' the output sounded i.e. how likely is it that the translated output/conversation resembled human language. Since the source text was entirely natural language, the ideal translations should also correspond closely to natural language.

The results from each model were interpreted based on metrics unique to them. For BERTScore, we observe the F1 score, where a higher score implies that the text bears high resemblance to natural language [35]. For the other four models, we measure Perplexity (PPL), one of the most common metrics for evaluating the quality of texts. PPL is a well-established information-theoretic measure [21], which is adapted to estimate the average uncertainty of predicting

[3] https://www.taguette.org/.

the next word in a sequence given the previous words where lower perplexity implies that the text is closer to natural (human) language [31]. Notably, we do not use BiLingual Evaluation Understudy (BLEU), the commonly preferred metric for evaluating MT, because BLEU fails to account for meaning-preserving lexical and compositional diversity [35].

4 Findings

4.1 Comparing NMT and PE Results

Our analysis compares translation results across NMT and PE for each scenario, through a combination of computational and qualitative evaluation.

Scenario 1. Scores from BERT ($PPL_{NMT} = 1.89 > 1.83 = PPL_{PE}$), GPT-Neo ($PPL_{NMT} = 3.48 > 3.32 = PPL_{PE}$), and XLNet ($PPL_{NMT} = 1.64 > 1.56 = PPL_{PE}$) demonstrated the superior quality of the PE translation, while BERTScore ($F1_{NMT} = 0.71 = 0.71 = F1_{PE}$) and Yoso ($PPL_{NMT} = 1.12 = 1.12 = PPL_{PE}$) had no preference. These results indicate that the PE translation in Scenario 1 is superior to the NMT version, which our 13 participants unanimously agreed upon.

> *The NMT version seems to be going very word-by-word. It translates 'not much, bro' into '兄弟不多', which actually translates to 'there's a few brothers' in Mandarin, and it translates 'see you, bro' into '兄弟见', which are all weird sentences.* - P2
>
> *Just take the second sentence as an example. In the original one, it is 'not much bro. How's it going?' And in the [NMT] translation, it is translating directly to '兄弟不多'. But in [PE] translation, it translates pretty well to '没什么特别的', which is the very precise translation of "not much bro". So, in this context, I think it [PE] is very good.* - P3

We thus observe that our participants indicate a strong preference towards the PE translation over the NMT one, especially given the former being able to infer and accurately translate colloquialisms present within the conversation.

Scenario 2. Evaluation from the 5 LLMs indicate that BERTScore ($F1_{NMT} = 0.74 > 0.73 = F1_{PE}$), BERT ($PPL_{NMT} = 1.99 < 2.01 = PPL_{PE}$), and XLNet ($PPL_{NMT} = 1.63 < 1.73 = PPL_{PE}$) prefer the NMT translation over the PE, whereas GPT-Neo ($PPL_{NMT} = 2.95 > 2.77 = PPL_{PE}$) and Yoso ($PPL_{NMT} = 1.14 > 1.13 = PPL_{PE}$) disagree. These results do not indicate a clear winner and neither do interviews, because 5 each of our participants prefer one over the other and 3 indicated no preference.

> *In the [PE] one, 'there's too many people waiting ahead of us' is translated to 'if there are too many people waiting for us in the front', which is kind of weird. But also, 'we can try another one' should be translated to '我们可以试试另一家', where '另一家' is a very native-speaker way to say 'another restaurant'. For the [NMT] version, '另一个' is not that native.* - P4

Neither translation is quite in context. In the [NMT] one, '我想是这样的' doesn't make sense, because it does not understand that 'this way' means where to go. For the [PE] one, "如果前面有太多的人在等我们' doesn't also make sense since they're not waiting for B. - P12

Our participants found errors in both NMT and PE translations for Scenario 2, and were generally dissatisfied across the board.

Scenario 3. In this case, the models BERTScore ($F1_{NMT} = 0.69 < 0.70 = F1_{PE}$), BERT ($PPL_{NMT} = 1.53 > 1.14 = PPL_{PE}$), and GPT-Neo ($PPL_{NMT} = 7.24 > 6.34 = PPL_{PE}$) all indicate a preference for PE, while XLNet ($PPL_{NMT} = 1.48 = 1.48 = PPL_{PE}$) and Yoso ($PPL_{NMT} = 1.14 = 1.14 = PPL_{PE}$) have no preference. While this might indicate a strong preference for PE here, results from qualitative interviews are far less conclusive, with 7 participants preferring the PE version while 6 appreciated the NMT version more.

The [PE] sentences don't sound polite, when it says 'you guys take care of yourselves', which is not how I speak to my parents. The [NMT] version said that 'you also take care of your health', which sounds more fair. - P1
The [PE] translation is not like a real conversation between a mother and son, because of 'it's okay, just so-so'. - P5
Both translations are really good and I can understand every sentence, but the [PE] one is more like real human conversation. - P7
The [PE] translation says 'stressed about your studies', which is a better fit for "紧张' than 'nervous' as the [NMT] one says. - P11

Thus, we do not observe a clear pattern of preferences for NMT or PE translations in this scenario.

Scenario 4. We observe that 4 LLMs (BERTScore: $F1_{NMT} = 0.72 < 0.73 = F1_{PE}$, BERT: $PPL_{NMT} = 1.22 > 1.19 = PPL_{PE}$, GPT-Neo: $PPL_{NMT} = 7.03 > 5.45 = PPL_{PE}$, XLNet: $PPL_{NMT} = 1.59 > 1.50 = PPL_{PE}$, and 13 participants unanimously prefer the PE version of the translation, with Yoso ($PPL_{NMT} = 1.14 = 1.14 = PPL_{PE}$) having no preference.

The [PE] one says 'let's see how skilled you are', whereas [NMT] says 'look at your show operation'. There is a clear winner. - P8
[PE] uses 'pro' which is better because 'pro' has a specific meaning in gaming, and is more like the usual language people would use and is less formal than [NMT]. - P10

Our participants have a strong preference for the PE version of the translation because of the preservation in the translations of the colloquialisms and informal tone present in the original conversation.

Thus, we demonstrate how translations produced by PE outperform their NMT counterparts across 4 scenarios. Full results are summarized in Table 1.

Table 1. Summary of findings comparing translation qualities across NMT and PE translations, through LLMs and qualitative interviews

Scenario	LLM Finding	Interview Finding
1	PE	PE
2	No Preference	No Preference
3	PE	No Preference
4	PE	PE

4.2 Price of PE as a Significant Pain Point

Beyond comparisons of translations, a significant pain point that emerged through our interviews was the price of PE, once we revealed that it cost us almost $70 to generate the PE texts for this study.

> *$70 for translating a few paragraphs is just too much. That's just a waste of money* - P1.
> *$66? That's too much, the price is super high. Sure, there's labor so it will be more expensive, but it is too high.* - P4.
> *$60 for the translations I just saw? I think this is too expensive.* - P6.

Participants were genuinely shocked to find out how much it cost to generate the PE translations. They unanimously felt that it was unreasonably expensive for day-to-day usage and, barring scenarios of high importance, they would be loathe to pay such prices, as perhaps best summarized by P7.

> *I think if I'm just using translations in my daily life, like any of these scenarios, I don't think you need to pay so much to translate them. Now, if I'm writing something scientific, like a grant or proposal, I might consider using post-editing* - P7.

In particular, participants (such as P3, P4, P6, and P11) took issue with the fact that despite the high price point of PE translations, there were still errors within generated results. The consensus was that since it represented a pricier alternative to free services such as Google Translate, PE results should be completely error-free. As P10 put it best,

> *You have those silly mistakes, which should not happen with a human translator. The machine translation can have those, and the human should fix them but if they are not, it shouldn't cost that much.* - P10

Participants thus expressed that while their experiences with NMT or other MTs (through services such as DeepL Translator, Baidu, Youdao, or ChatGPT) have been poor-to-average, they found it acceptable because such services were free of cost.

I mean that's a waste of money. I just think that the translation for the human is just worth like $1. If we consider the cost, of course Google is much better. - P8
That's kind of expensive and I feel like the translation provided by Google presents ideas pretty well. Yeah, there are some wordings that are kind of iffy, but you can get the idea, you're just missing some small parts. - P9
If I have the choice of paying $70.00 to get the human one or the free machine translation, I wouldn't pay $70.00. Since I can understand [MT] already, there's no justification. - P12

4.3 Participants Doing Their Own Post-Editing

In exploring our participants' decision to accept sub-par results from MT and NMTs due to it being free, we also uncovered the pattern that they regularly found themselves using generated MTs to gain a baseline understanding and editing provided results as they deemed fit.

Considering the quality, I would always choose PE, but the cost is too much. At that cost, I can be a post editor. - P7
Using Google Translate, even though there are some inaccuracies, it does not affect my understanding of the whole meaning, and I can correct those myself. - P10
When I'm studying, I will always choose machine translation because I can get it quickly and correct any mistakes I see myself. - P13

Our participants thus acted as their own post editors, as they performed a round of translation on services such as Google Translate and then applied their own understanding of the source and target languages to identify and resolve errors. This was yet another reason why participants were loathe to pay high prices for PE services.

5 Discussion

5.1 Post-Editing as the Future of Machine Translation

In this study, we observe a marked difference between NMT and PE translations across 4 different scenarios (particularly salient for Scenarios 1 and 4), both from computational and qualitative evidence, as we demonstrate how PE outperforms NMT translations. Despite significant advances in MT research over the past decade, including the development of NMTs into a dominant market force and novel language models such as GPT-3/4 trying to compete but still containing the same errors as Google Translate [10,16], our findings demonstrate that the human element in translation remains irreplaceable.

Having said that, this study also demonstrates that PE is far from perfect, and we call for a higher investment of resources and research into making PE

the best version of itself. Though current research into PE focuses on reducing the amount of time or generally how to support the post-editors themselves [11,12,30], we believe that there are user-side improvements to also consider. For instance, when P1 mentioned how the PE translation of Scenario 3 didn't sound polite enough according to their understanding of parent-son conversations, they would not have had any avenue to explain this reason for being dissatisfied with the translation quality to Translate.com. Translate.com does allow users to comment on the quality of translations, but does not allow for much commentary on why. We believe that the first and most important improvement to PE services should be for a more flexible and dynamic way to provide feedback to post-editors about translation quality.

Furthermore, current PE services such as Translate.com are not live, and have wait-times that can range from anywhere between a few hours to a few days. If PE is to be the future of MT, then there needs to be the development of a live and on-demand PE service. We recognize this would require significant investment towards the infrastructure and hiring competent post-editors on fair wages, and hope that it becomes a reality someday.

5.2 Poor Performance of NMTs

Our finding that NMTs are less preferred over PE and the generally high number of inaccuracies and outright errors in NMT results gives us cause for concern, especially given the studied languages (English and Mandarin) being two of the highest-resource languages i.e. languages for which there exist large amounts of data publicly available for training language models [7]. English ↔ Mandarin is one of the most common use-cases for language translation, especially by researchers of MT (e.g. Beh and Canty, 2015 [3]; Ho et al., 2019 [13]; Mathias and Byrne, 2006 [23]). To see such a poor performance, both by qualitative and quantitative evaluations, represents the need for designers at Google to take a long, hard look at themselves.

If the performance of the state-of-the-art translation technology in two of the highest-resource languages is poor to average across the board, this also bodes ill for lower-resource languages. Especially in the aforementioned context of people being their own post-editors, if translations across two high-resource languages are producing errors and inaccuracies necessitating human/end-user correction, then it might not be too much of a stretch to imagine that the number and degree of such errors might increase as languages in question become lower resource. Evidence in languages such as Bahasa [2], Bengali [10], Malay [26], Tagalog [9] and others reveals a significant gap in the performance of Google Translate and user expectations. In recognition of the labor of PE and the possibility of it increasing with a language's decrease down the resource tree, we call for stronger work in improving the quality of language translation services such as Google Translate.

5.3 Conversational, Commonplace Scenarios As Translation Use Cases

The primary metrics along which our participants differentiated the quality of NMT and PE based translations were the presence of colloquialisms, and the formal/informal tones that were necessary in different contexts, given the conversations. In particular, participants preferred PE translations over NMTs for the specific reasons that PE accurately carried over the colloquialisms and adequately preserved the tone of the text.

Since we specifically designed scenarios to contain such colloquialisms and variations in tone as they are part of daily conversations, participant engagement with them was heartening. For Scenario 3 and the conversation between parents and their son, we presented it as a different context to all of the others, in which the conversations were always between friends. In evaluating the translations, participants pointed out how the PE version of the conversation saw the son being impolite in his responses, with P1 even going so far as to say that they would not talk to their own parents in the way that the PE translation suggested. That participants related so closely to the provided scenarios and placed themselves in the shoes of the speakers in the conversations reflects their strong engagement with the content.

That participants comment on the tone of the translation also raises a point on the subjective nature of translations, where 'correctness' cannot be measured only through computational models and even tallying preferences across participants should be done carefully. In our study, we collected participants who are all born in China and are evaluating the quality of English ↔ Mandarin translations, which establishes a shared baseline across them. These results would be different were we to include participants from other countries or cultures, cultures where referring to one's parents in less formal language might be more acceptable.

We believe that our choice of regular conversations across relatable scenarios as material for conducting and evaluating translation procedures is noteworthy. While source material such as novels [30], Wikipedia articles [11,32], newspaper articles [15], or text corpora available online [20] might be relevant to users as material they might translate, material that more closely reflects daily conversations might resemble stronger use cases and adopt more subjective understandings of translation quality. From personal experience as international students, some of the authors themselves regularly use translation services to navigate conversations with professors and peers alike, conversations that contain turns-of-phrase and subtle colloquialisms that they often miss as they are lost in translation. Were MT services more attuned to infer and convey such colloquialisms, they would better serve their users. We hope that future researchers in the field would consider materials from daily conversations that are relatable across different cultures as texts of study, and strongly believe this would lead to the development of better MT services.

5.4 Implications for Human-Centered Machine Translation

Finally, our finding that end-users of MT services are regularly functioning as their own post-editors, which we imagine will be relatable to our readers who use MT, have implications on future directions of human-centered machine translation. We make progress towards an answer to one of Robertson's [27] questions towards the design of human-centered machine translation, which asks 'how can we design systems that help people make use of imperfect translation?' Our finding shows that users of MT services already have their own approaches to dealing with imperfect translations, and recommend that future designers of MT services consider incorporating stronger feedback mechanisms for participants to describe why provided translations are imperfect and how they can be improved. While services such as Google Translate technically do currently have such feedback options, our personal experience suggests that it is not very findable and is limited in the ways in which it accepts feedback. (As an exercise to our users, can you locate the Feedback button for Google Translate, and how satisfied are you with the feedback it allows you to provide?) Pursuant to Ghosh's [10] suggestions on human-centered (re)design of machine translation, we believe that this would also open up an avenue for increased user participation in the design of MT services such as with model training and interpretation, rather than simply consuming and rating translations.

6 Conclusion

In this paper, we compare the translation quality across the statedly state-of-the-art machine translations provided by Google Translate's Neural Machine Translations and post-editing between English ↔ Mandarin translations of 4 handcrafted conversations from daily-life scenarios that transcend linguistic and cultural boundaries, through a combination of computational evaluation through 5 language models and analysis of qualitative interviews from 13 participants fluent in both languages. We demonstrate the superiority of PE over NMT almost entirely, although the high cost of PE would inhibit our participants from abandoning their currently used free MT services. We provide recommendations on improving PE as the future of MT, and towards a more human-centered approach towards automated language translation.

7 Limitations

Given our goals around comparing the quality of PE and NMT translations, one limitation of our work is that although we did not intend to, we were forced to comment on the low quality of the first round of PE translations through Translate.com on account of there being a lot of typos. Therefore, we forced a level of quality control upon the PE results, which might have influenced our results on quality comparisons. However, we do not believe that this limitation undermines our findings, because were users of PE services to receive the same typo-laden translations that we received, we believe that they too would request a second round of translations.

References

1. Allen, J.: Post editing. In: Computers and Translation, pp. 297–318. John Benjamins BV (2003)
2. Amilia, I.K., Yuwono, D.E.: A study of the translation of google translate. Lingua J. Ilmiah **16**(2), 1–21 (2020)
3. Beh, T., Canty, D.: English and Mandarin translation using google translate software for pre-anaesthetic consultation. Anaesth. Intensive Care **43**(6), 792 (2015)
4. Black, S., Gao, L., Wang, P., Leahy, C., Biderman, S.: GPT-Neo: large scale autoregressive language modeling with Mesh-TensorFlow. **58**(2) (2021)
5. Charmaz, K.: Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis. Sage (2006)
6. Chollampatt, S., Susanto, R.H., Tan, L., Szymanska, E.: Can automatic post-editing improve NMT? In: Findings of the Association for Computational Linguistics: EMNLP 2020 (2020)
7. Cieri, C., Maxwell, M., Strassel, S., Tracey, J.: Selection criteria for low resource language programs. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 4543–4549 (2016)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
9. Fitria, T.N.: Gender bias in translation using google translate: problems and solution. Lang. Circle J. Lang. Lit. **15**(2) (2021)
10. Ghosh, S., Caliskan, A.: ChatGPT perpetuates gender bias in machine translation and ignores non-gendered pronouns: findings across Bengali and five other low-resource languages. In: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (2023). https://doi.org/10.1145/3600211.3604672
11. Green, S., Heer, J., Manning, C.D.: The efficacy of human post-editing for language translation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 439–448 (2013)
12. Herbig, N., Pal, S., van Genabith, J., Krüger, A.: Multi-modal approaches for post-editing machine translation. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–11 (2019)
13. Ho, S.S., Holloway, A., Stenhouse, R.: Analytic methods' considerations for the translation of sensitive qualitative data from mandarin into English. Int. J. Qual. Methods **18**, 1609406919868354 (2019)
14. Hutchins, W.J.: Machine translation: a brief history. In: Concise History of the Language Sciences, pp. 431–445. Elsevier (1995)
15. Jia, Y., Carl, M., Wang, X.: Post-editing neural machine translation versus phrase-based machine translation for English-Chinese. Mach. Transl. **33**(1–2), 9–29 (2019)
16. Jiao, W., Wang, W., Huang, J.T., Wang, X., Tu, Z.: Is ChatGPT a good translator? A preliminary study. arXiv preprint arXiv:2301.08745 (2023)
17. Koponen, M.: Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. J. Spec. Transl. **25**(2) (2016)
18. Koponen, M., Sulubacak, U., Vitikainen, K., Tiedemann, J.: MT for subtitling: user evaluation of post-editing productivity. In: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, pp. 115–124. European Association for Machine Translation, Lisboa, Portugal (2020). https://aclanthology.org/2020.eamt-1.13

19. Kranias, L., Samiotou, A.: Automatic translation memory fuzzy match post-editing: a step beyond traditional TM/MT integration. In: LREC (2004)
20. Lagarda, A.L., Alabau, V., Casacuberta, F., Silva, R., Diaz-de Liano, E.: Statistical post-editing of a rule-based machine translation system. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, pp. 217–220 (2009)
21. Lakew, S.M., Karakanta, A., Federico, M., Negri, M., Turchi, M.: Adapting multilingual neural machine translation to unseen languages. arXiv preprint arXiv:1910.13998 (2019)
22. Le, Q.V., Schuster, M.: A neural network for machine translation, at production scale. Google AI Blog **27** (2016)
23. Mathias, L., Byrne, W.: Statistical phrase-based speech translation. In: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, vol. 1. IEEE (2006)
24. O'Brien, S.: How to deal with errors in machine translation: postediting. Mach. Transl. Everyone Empowering Users Age Artif. Intell. **18**, 105 (2022)
25. Pitman, J.: Google translate: one billion installs, one billion stories. In: The Keyword (2021)
26. Prates, M.O., Avelar, P.H., Lamb, L.C.: Assessing gender bias in machine translation: a case study with google translate. Neural Comput. Appl. **32**, 6363–6381 (2020)
27. Robertson, S., et al.: Three directions for the design of human-centered machine translation. Google Research (2021)
28. Shenoy, R., Herbig, N., Krüger, A., van Genabith, J.: Investigating the helpfulness of word-level quality estimation for post-editing machine translation output. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 10173–10185 (2021)
29. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. Adv. Neural Inf. Process. Syst. **27** (2014)
30. Toral, A., Wieling, M., Way, A.: Post-editing effort of a novel with statistical and neural machine translation. Front. Digital Hum. **5**, 9 (2018)
31. Vasilatos, C., Alam, M., Rahwan, T., Zaki, Y., Maniatakos, M.: HowkGPT: investigating the detection of ChatGPT-generated university student homework through context-aware perplexity analysis. arXiv preprint arXiv:2305.18226 (2023)
32. Yamada, M.: The impact of google neural machine translation on post-editing by student translators. J. Spec. Transl. **31**(1), 87–106 (2019)
33. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: XLNet: generalized autoregressive pretraining for language understanding. Adv. Neural Inf. Process. Syst. **32** (2019)
34. Zeng, Z., Xiong, Y., Ravi, S., Acharya, S., Fung, G.M., Singh, V.: You only sample (almost) once: linear cost self-attention via Bernoulli sampling. In: International Conference on Machine Learning, pp. 12321–12332. PMLR (2021)
35. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: evaluating text generation with BERT. In: The International Conference on Learning Representations (2020)
36. Zhao, Y., Zhang, J., He, Z., Zong, C., Wu, H.: Addressing troublesome words in neural machine translation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 391–400. Association for Computational Linguistics, Brussels, Belgium (2018).https://doi.org/10.18653/v1/D18-1036