Ghosts in the machine: Can adaptive MT help reclaim a place for the human in the loop?

Hanna Martikainen (hanna-julia.martikainen@sorbonne-nouvelle.fr)

École Supérieure d'Interprètes et de Traducteurs, Université Sorbonne Nouvelle – Paris 3, France

CLESTHIA - Langage, systèmes, discours - EA 7345

Abstract

While the productivity gains brought about by machine translation (MT) can help translators meet ever-tighter deadlines and respond to pressing demands for publishing content simultaneously in different languages, these tools also impose a workflow that tends to reduce the human translator's role to simply correcting mistakes made by the machine in a one-way process with no real interaction. Thus, although more cost-effective, post-editing of MT output also appears a less creative and enjoyable task than translation. Adaptive MT, on the other hand, has been advertised as a way to recenter the translation process on the human and foster more genuine interaction with the machine. Said to have been developed for professional translation workflows, the technology enables a dynamic work process that is supposedly very different from the repetitive task that post-editing static MT output can be. This paper presents an experiment with adaptive MT conducted during the 2020-2021 academic year. As part of a course on MT and post-editing, second-year master's students carried out group projects on the Lilt platform. In this paper, students' views on the MT engine are analyzed, with a focus on their interaction with the technology. While students recognize the potential of adaptive MT for empowering the human in the loop, MT quality and CAT (computer-assisted translation) ergonomics in general appear to have a greater influence on usability than interaction with the machine.

Keywords

adaptive machine translation; human-machine interaction; ergonomics of post-editing; translator training

INTRODUCTION

The anguish over machines replacing humans is nothing new, and the fear has certainly not died down with the advent of neural machine translation (NMT). As early as 2016, Google's NMT engine was supposedly "bridging the gap between human and machine translation" [Wu et al., 2016], and just a couple of years later, Microsoft AI was claimed to have indeed achieved "human parity" [Hassan et al., 2018]. Today, NMT has supposedly surpassed the human translator altogether, with Facebook AI declared "super-human" at the 2019 WMT evaluation campaign [Barrault et al., 2019: 23]. Although such claims are largely explained by WMT evaluation conditions [Läubli et al., 2018: 4795; Toral et al., 2018: 121], the hype surrounding NMT is not waning. It is therefore easy to forget that machine translation (MT) technologies, although increasingly integrated into the professional translator's toolkit today[1], were not initially developed with professional translators in mind. Government-funded and designed in the context of the cold war, MT technologies are still in high demand in the military and defense sector, alongside the automotive and healthcare industries[2]. Research on

[1] See, for instance, the Elis Survey 2021 (https://euatc.org/elis2021/) or the 2021 Nimdzi 100 (https://www.nimdzi.com/nimdzi-100-top-lsp/).

[2] See for instance https://www.psmarketresearch.com/market-analysis/machine-translation-market.

MT is dominated by corporate players, i.e. online tech giants developing search engine, e-commerce or social media solutions, and authored by computer science scholars instead of linguistics or translation studies scholars, which entails a shift of perspective on translation as a whole [Larsonneur, 2019: 4-5]. Lines are getting increasingly blurred between human and machine translation, a confusion voluntarily fed by MT engines being dubbed 'translators' [Rossi, 2019].

While the productivity gains brought about by MT [Screen, 2019: 135-136] can help translators meet ever-tighter deadlines and respond to pressing demands for publishing content simultaneously in different languages, these tools also impose a workflow that tends to reduce the human translator's role to simply correcting mistakes made by the machine, in a one-way process with no real interaction. Thus, although more cost-effective, post-editing of MT output also appears a less creative and enjoyable task than translation [Sakamoto, 2019: 206]. Adaptive MT, on the other hand, has been advertised as "Machine Translation for Human Translators" [Denkowski et al., 2015], a way to recenter the translation process on the human and foster more genuine interaction with the machine, combining "the best of both worlds"[3].

This paper presents first observations from an experiment with adaptive MT conducted during the 2020-2021 academic year. As part of a course on MT and post-editing, second-year master's students carried out group projects on the Lilt platform to explore the potential of adaptive MT in the translation workflow. The following research questions were of interest in this qualitative exploratory study: Do students feel more empowered, or more in control of the translation process when working with adaptive MT instead of intervening at the very end of the production chain through postediting of static MT output? Which parameters influence usability and enjoyability of adaptive MT?

I MT technologies in professional translation workflows

1.1 Pros and cons of post-editing for professional translators

Static MT use in professional translation workflows has been investigated in different settings. [Cadwell et al., 2018] compare MT use and translators' attitudes towards the technologies in an institutional and a commercial setting. While acknowledging the potential benefits MT can bring about, chief among which are productivity gains obtained by increased speed and reduced typing effort, and providing inspiration, professional users also fear its potential negative impact on the translator's abilities, and recognize that MT can reduce the creativity and enjoyability of the translation task, devalueing the translator's work [Cadwell et al., 2018: 310-311].

The ideal scenario for translator well-being would seem to be integration of MT in the usual translation workflow, wherein MT is just another potentially useful tool available for the professional translator through the CAT interface. This best practice workflow, often dubbed 'augmented translation'[4], is notably adopted in the institutional setting, for instance at the European Commission Directorate-General for Translation. In this configuration, an "intricate interweaving of technology and human intervention", MT suggestions are available to the user in the usual workflow as "easily dismissed segments", which limits the impact of MT in the process [Rossi and Chevrot, 2019: 180]. This freedom to choose is a key factor in the high rate of acceptance of MT technologies at the DGT [Rossi and Chevrot, 2019: 184]. In the same spirit, the future profile of UN translators as defined during the 7th United Nations MoU Universities Conference consists in "digitally-fluent" subject-matter specialists, working as expert revisers in an "augmented translation environment supported by integrated systems and

[3] https://martech.zone/lilt-neural-machine-translation/

[4] https://insights.csa-research.com/reportaction/305013226/Marketing

artifical intelligence" [Elizalde and Bondonno, 2021]. This ideal configuration is rarely encountered, however, in the commercial sector, where translators working for LSPs more often perceive MT as being imposed on them [Cadwell et al., 2018: 310].

1.2 Adaptive MT

Said to have been developed for professional translation workflows, adaptive MT technologies allow for a dynamic and interactive work process that is supposedly very different from the repetitive task that post-editing static MT output can be. Initially seen as the "latest big leap forward"[5] in MT, the technology does not seem to have gained quite the predicted momentum, and has only attracted limited attention from researchers [i.e., Denkowski et al., 2014; Bentivogli et al., 2015; Daems and Macken, 2019]. Interestingly, a recent CSA survey showed that, although professional translators mostly do not appreciate working with MT, those who do use these technologies will rather work with adaptive systems such as Lilt, rather than post-editing raw MT output [Pielmeier, H. and O'Mara, P., 2020: 42-43]. This would suggest that adaptive or interactive MT might indeed hold promise as a tool designed for and adapted to professional translation workflows.

II CONTEXT OF THE EXPERIMENT

2.1 MT & post-editing course

The experiment was conducted as part of a course in MT and post-editing at the École Supérieure d'Interprètes et de Traducteurs (ESIT) that took place during the second semester of the academic year 2020-2021 in a remote setting. A total of 70 second year master's students were enrolled in the course. The course was divided into four modules, for a total of 19,5 hours over 13 sessions. The first module consisted in an introduction to machine translation post-editing or MTPE and covered topics such as MT history, industry standards for post-editing, productivity and quality in MTPE. In the second module, students experimented with MT quality assessment using metrics such as fluency and accuracy or edit distance, and through error assessment grids. Module 3 dealt with best practices in post-editing on the LSP market, i.e. MT integration into CAT in an augmented translation environment. Finally, in the last module, students acquainted themselves with adaptive MT over three sessions through group projects using Lilt.

2.2 Lilt[6]

With Lilt, the MT engine adapts at two different levels. Like most MT engines today, Lilt learns from the human user and produces better solutions over time. Unlike static MT engines, however, interactive (or 'real-time adaptive') MT systems also adapt during the translation process. Every word that gets translated is immediately taken into account in an effort to adapt the engine's output and try to predict the phrase the human user is going for. Thus, if the initial MT suggestion is no good, a new one will be available to the user after typing just a couple of words from the intended translation. Contrarily to most existing systems, MT output is not presented to the user in the space devoted to the target text. Instead, for each segment, the system shows the source text, followed by a blank space where the user inserts or types the translation, and only under this blank space, the suggested MT output that adapts to the target text being produced. The user interface is simple and intuitive, and allows for easy insertion of translation memory matches, MT suggestions, and glossary terms.

2.3 Student projects

[5] https://termcoord.eu/2017/06/the-latest-big-leap-forward-in-machine-translation-adaptive-mt/
[6] https://lilt.com/

Each student group translated a text in the Lilt interface and analysed the experience in a written report. The group projects had a set of fixed parameters. The texts chosen for translation had to have a certain length according to the number of translators in the group (i.e., 500-1,000 words per translator), and a minimum of three translators had to work on the text. One member in each group needed to perform project management tasks: project setup, resource creation including importing a glossary of a dozen terms; task assignment. The project manager was also asked to pretranslate the text in the Lilt interface for later comparison with the text resulting from interactive translation. Other than this comparison between the pretranslation and the final translation, the final report had to include group productivity statistics. Students were also invited to reflect in their analysis on the task of working with interactive MT. Other project parameters were open: students were free to choose their groups and the tasks to be performed by each member in the group, as well as the texts to be translated (domain, degree of specialization). Group projects could be conducted in any of the working languages (A, B, C) of the students, within the limits imposed by Lilt (i.e., English required as source or target language for several language pairs).

2.4 Study limitations

The biggest limitation to the relevance of this study was the short amount of time the groups had to work in the Lilt interface and exploit the possibilities of adaptive MT. Although the projects allowed for exploration of the interactive aspect, i.e. MT output for each segment adapting in real time during translation, most groups did not work on long enough extracts to get to experiment with longer-term adaptation of MT output and propagation of corrections to subsequent segments in order to learn more deeply about the interactions with the system. Time constraints were also an issue in terms of the learning curve that can be quite steep when working with interactive MT before getting to the point of reaping benefits from the process.

The scope of the experiment was also limited by the choice of the MT engine. Although open-source interactive and adaptive MT solutions have been developed in academia (e.g. Denkowski et al., 2014), their use requires technical comptences not available in the translation classroom context. Lilt is, to the author's knowledge, the only commercially-available and ready-to-use tool that is not only adaptive but also interactive (i.e., real-time adaptive).

III RESULTS AND DISCUSSION

3.1 Project languages & groups

A total of 14 groups were formed, ranging from 3 to 7 members per group. Most groups (8/14) translated from English to French, while one group translated in each of the following language combinations: English-Arabic; English-Italian; English-Portuguese; English-Spanish; English-Chinese; Russian-English. The texts chosen for translation represented various genres (e.g., general press, web pages, UN report, data sheet, EU brochure…) and dealt with different topics (e.g., human rights, youth job program, environment and economy, meteorology, cooking, cancel culture…). Length of translated extracts varied from about 1,000 to 5,000 words.

Main observations from the student reports are presented in this section. All translations are by the author and the original comments in French are given in footnotes.

3.2 Main observations: positive aspects

Most student groups participating in the experiment found adaptive/interactive MT technology promising indeed, with fast adaptation and adequate adjustment: *“Lilt adapts very*

fast. The suggested corrections are adequate, and real-time adaptation of the output is certainly an advantage."[7]
Students appreciated reduced typing effort with functional real-time adaptation, when only minor changes were required for the MT output to adapt adequately to obtain the desired output and translation choices were automatically propagated: *"Having the choice to enter the translation yourself or to use the MT output (or even doing both at once) is an interesting pratical option"*[8].
Most groups were appreciative of the simplicity of project management in the Lilt interface compared with other tools they were familiar with. Lilt's intuitive functioning and ease of use, for instance the click-and-add term insertion function, was often mentioned in the reports.
Finally, some students also felt less constrained by MT suggestions presented in separate space from blank target segment: *"Lilt output is not automatically inserted in the target zone but in a separate space under it, and the translator is therefore less influenced by the suggestion."*[9]

3.3 Main observations: negative aspects

In many instances, however, the real-time adaptation function was found not to be up to the task, resulting in degraded quality and introduction of errors not seen at the pre-translation stage: *"Generally speaking, each change to the suggested MT output resulted in either a grammatically incorrect suggestion or a change in meaning."*[10]
Also, Lilt MT engine was observed to suffer from the usual NMT issues of calque, omission of unrecognized words, terminological variation and ungrammatical suggestion (e.g. problems with articles and word gender). Hallucinations or non-existing words were frequently produced by the engine, e.g. "dried rose petals: *Pétathé*". Syntax was also frequently an issue for the Lilt engine, on sentence level ("women burn survivors: *les femmes brûlent les survivants*") as well as within complex noun phrases ("small women's organizations: *organisations de petites femmes*").

3.3.1 MT quality as key parameter in user satisfaction

Importantly, the final reports mostly suggest that, in the context of this experiment, interaction with the MT engine appeared less important in terms of usability than plain MT quality. Quality of the initial output was often lacking dramatically for the English-French language pair and mostly compared negatively with competition, i.e. freely available static NMT engines such as DeepL or Google Translate. In some instances, the initial MT output was the kind of 'word salad' encountered with the first statistical MT engines some decades ago:

Noix de musinfuser aux les États-Unis est une épice classique à la cuisson mais elle peut être fraîchement râpée en saucescrème, crème crème, crème fouetcrème chantilly et plats rôti de

[7] « Lilt s'adapte très vite. Les corrections qu'il apporte à mesure de la traduction sont pertinentes, et le fait que la phrase proposée se modifie à mesure de la traduction est un vrai plus. »
[8] « Avoir le choix entre taper toute la traduction soi-même ou se servir des suggestions (ou même travailler des deux façons à la fois) est une option intéressante et pratique »
[9] « Les propositions de Lilt ne s'ajoutent pas directement dans la zone de traduction, mais dans une zone en dessous de celle-ci. Cela permet d'être moins influencés par le logiciel. »
[10] « De manière générale, chaque modification conduisait soit à une suggestion grammaticalement incorrecte, soit à une suggestion qui consistait en un déplacement de sens. »

légumes, mijoté légumes verts ou infusée en noix de muscade mouluépices de Noël noix de muscade en mule, thé ou café.[11]

The main issue impacting MT quality was integration of glossary terms in the output. Each group started with a glossary of at least a dozen terms, and some ended up with glossaries of well over a hundred terms. Glossary terms were frequently either not used in the MT output or inserted randomly into MT output. As illustrated in the example above, insertion of glossary terms often resulted in degraded MT quality[12]. Integration of fixed terminology tends to be challenging in NMT because of the lack of grammatical rules, which makes terminology insertion difficult in many languages such as French.

The groups' working language pair was found to influence students' assessment of MT output. Most notably, for the combinations involving Brazilian portuguese and Arabic, user satisfaction was markedly higher, as Lilt MT quality was found superior to competition: *"All the members in our group agreed that Lilt MT quality was better than DeepL or Google for example* [for the English to Portuguese language pair]. *"*[13]

3.3.2 Impact of CAT ergonomics on user satisfaction

Moreover, usability of the technology depends not only on MT quality but also – and perhaps even more so – on ergonomics of the CAT tool, which often compared negatively with competition. Among the various CAT ergonomics issues mentioned in the reports were: unusable hyperlinks, undetected errors during spell check, non-modifiable revised translation, improperly handled tags, etc. Also, keyboard shortcuts for different operations were not always working properly. It should be noted that students also had quite high expectations for the CAT interface, and were sometimes expecting functionalities that could only be possible thanks to a robust AI, such as automatic declination of lemmatized glossary terms when they are inserted in the translation or automatic currency conversion.

Conclusion

The experiment conducted with master's students suggests that interactive/adaptive MT indeed does have the potential to offer a better user experience than post-editing static MT. Although a rather steep learning curve is to be expected before being at ease with interactive translation, usability of the technology actually appears to have more to do with CAT and MT quality. Students in training programs are usually familiar with industry standards for translation environment ergonomics, and tend to expect the same high level of performance from all the tools they will integrate into their future professional practice. Students are also trained to put in practice a workflow inspired by best practice, i.e. integrating different MT engines into the CAT environment of their choice so as to maintain the benefits offered by CAT tools, particularly for ensuring coherence, and limit MT interference. Therefore, while interested in the interactive translation process, many of the students considered that, for the technology to be useful in professional practice, it would need to be implemented in the CAT workflow and offer better quality MT from the go.

[11] English source text: "Nutmeg in the US is a classic baking spice but can be freshly grated into cream sauces, custard, eggnogs, whipped cream, roasted vegetable dishes, stewed greens or infuse ground nutmeg into mulling spices, tea or coffee."

[12] Glossary terms: nutmeg (*noix de muscade*), sauce (*sauce*), custard (*crème pâtissière*), whipped cream (*crème chantilly*), mulling spices (*épices de Noël*).

[13] « D'après l'avis de tous les membres du groupe, la traduction automatique de Lilt s'est avérée meilleure que celles de Google Translate et de DeepL, par exemple. »

The author wishes to thank the organisers and the participants of robotrad2021 for the exchanges that enriched the initial presentation of this work, and 2020-2021 M2 students at ESIT (Sorbonne Nouvelle) for their participation in the experiment. The author would also like to express their gratitude to the reviewers and editors of this volume for their thoughtful suggestions.

References

Barrault, L., Bojar, O., Costa-Jussà, M. R., Federmann, C., Fishel, M., Graham, Y., ... & Zampieri, M. (2019, August). Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)* (pp. 1-61).

Bentivogli, L., Bertoldi, N., Cettolo, M., Federico, M., Negri, M., & Turchi, M. (2015). On the evaluation of adaptive machine translation for human post-editing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(2), 388-399.

Cadwell, P., O'Brien, S., & Teixeira, C. S. (2018). Resistance and accommodation: factors for the (non-) adoption of machine translation among professional translators. *Perspectives*, 26(3), 301-321.

Daems, J., & Macken, L. (2019). Interactive adaptive SMT versus interactive adaptive NMT: a user experience evaluation. *Machine Translation*, 33(1), 117-134.

Denkowski, M., Lavie, A., Lacruz, I., & Dyer, C. (2014, April). Real time adaptive machine translation for post-editing with cdec and TransCenter. In *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation* (pp. 72-77).

Denkowski, M., Lavie, A., Dyer, C., Carbonell, J., Shreve, G., & Lacruz, I. (2015) Adaptive MT Systems for Post-Editing Tasks: Machine Translation for Human Translators. MTMA 2015, May 10th-15th, 2015.

Elizalde, C., & Bondonno, N. (2021). Leveraging language technologies and machine translation: the new profile of professional translators. 7th UN MoU Universities Conference, ESCWA, 17-19 May, 2021.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., ... & Zhou, M. (2018). Achieving human parity on automatic Chinese to English news translation. *arXiv preprint arXiv:1803.05567*.

Larsonneur, C. (2019). The Disruptions of Neutral Machine Translation. *spheres: Journal for Digital Cultures*, (5), 1-10. DOI: https://doi.org/10.25969/mediarep/134

Läubli, S., Sennrich, R., & Volk, M. (2018). Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4791–4796), Brussels, Belgium. Association for Computational Linguistics.

Pielmeier, H. & O'Mara, P. (2020). The state of the linguist supply chain: Translators and interpreters in 2020 (Report). Boston: CSA Research.

Rossi, C. (2019, October). Les usages actuels de la traduction automatique. In *Atelier Digit-Hum 2019:" les humanités numériques en langues"*.

Rossi, C., & Chevrot, J. P. (2019). Uses and perceptions of machine translation at the European Commission. *The Journal of specialised translation (JoSTrans)*.

Sakamoto, A. (2019). Why do many translators resist post-editing? A sociological analysis using Bourdieu's concepts. *The Journal of Specialised Translation*, 31, 201-216.

Screen, B. (2019). What effect does post-editing have on the translation product from an end-user's perspective. *The Journal of Specialised Translation*, 31, 133-157.

Toral, A., Castilho, S., Hu, K., & Way, A. (2018). Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers* (pp. 113-123), Brussels, Belgium. Association for Computational Linguistics.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.