7th International Conference on Higher Education Advances (HEAd'21)
Universitat Politècnica de València, València, 2021
DOI: http://dx.doi.org/10.4995/HEAd21.2021.12789

Post-editing Machine Translation in MateCat: a classroom experiment

Katrin Herget
Department of Languages and Cultures / Centre for Languages, Literatures and Cultures, University of Aveiro, Portugal**.**

Abstract

Advances in machine translation resulted in an increase of both volume and quality of machine-translated texts. However, machine translation still requires humans to post-edit the translation. This paper proposes a product-based approach of a post-editing (PE) experiment that was carried out with a total of 10 MA translation students. The goal of this study comprised both the analysis of the post-editing results performed by student translators involving a machine-translated text in MateCat and the subsequent error markup. By comparing the quality reports obtained at the end of the post-editing process, we analysed the linguistic quality results and observed a heterogeneous error distribution, considerable divergence in severity level ratings and a huge span of TTE (time to edit). This study aims at making a contribution to the integration of post-editing activities into the translation technology classroom for students without prior experience in PE.

Keywords: *Machine translation; post-editing; MateCat; translation technology classes.*

1. Introduction

Advances in Machine Translation (MT) are profoundly changing the translation profession, turning post-editing of machine output into a dynamic and widely-known practice. During the last five years, there has been an enormous upswing through the use of machine learning. Google and DeepL came up with Neural Machine Translation systems that by simulating neurons of the human brain, have established new standards in MT. Nowadays, computer-aided translation tools and translation management systems allow the integration of MT into their translation workflows to enhance productivity. The fast-growing volume of machine-translated documents increasingly makes post-editing (PE) services indispensable, by which a human translator checks a machine-translated outcome for its lingusitic and textual quality.

With regard to an ever more demanding market that values translation quality and speed, Translation Quality Assessment (TQA) has therefore gained a new relevance. There has been a lot of research in the field of TQA linked to MT, trying to establish criteria for measuring quality and error analysis. Due to its huge impact on the translation activity itself, translation technology classes should integrate PEMT activities into their standard curriculum, in order to meet the requirements of today's job market. "In an increasingly competitive market where quality-focused translators come under intense pressure from clients to sustain quality standards while offering more attractive rates and faster turn-around times, models and tools to support translation quality assessment (TQA) are a necessity" (Doherty et al., 2018, p. 95).

2. Human judgement of machine translation

Both industry and research have come up with a variety of quality evaluation metrics that aim at finding a standard for assessing translation quality produced by machines through objective and measurable indicators (cf. Vilar et al., 2006; Lommel et al. 2014). "Automated metrics emerged to address the need for objective, consistent, quick, and affordable assessment of MT output, as opposed to a human evaluation where translators or linguists are asked to evaluate segments manually" (TAUS, s.d.). Due to the fact that human language is a complex, subtle and ambiguous system, it is a rather challenging undertaking to evaluate machine translation. Both the manual and automatic approach aim at establishing valuable parameters for quality evaluation.

In manual evaluation, a machine-translated text is analysed in terms of its grammatical correctness, its semantic adequacy and its suitability for text convention patterns. These parameters, known as *fluency* and *accuracy*, are usually given special attention. Whereas

fluency evaluation is achieved on the basis of the target text, *accuracy* is *assessed* by comparing the source and the machine-translated target text. In automatic evaluation, the MT output is compared to a number of reference translations, assigning them quality scores. In this approach, automatic metric scores correlate with human translations.

3. MateCat

MateCat is a web-based CAT tool which can be used for free by translators, project managers, LSPs and enterprises. MateCat provides matches extracted from a public Translation Memory (MyMemory) that performs machine translation through a combination of Google Translate and Microsoft Translator. Users may also create personal TMs for confidential use. MateCat allows for a quantitative-based method of TQA according to an industry standard that calculates the quality level of a translation with regard to the number of words reviewed. Consequently, the project manager is provided with a quality score on the project's outcome. Based on this score, the project manager may carry out changes to the project to increase productivity (cf. MateCat, s.d.). Figure 1 gives an example showing the post-editing effort for a segment by a participant involved in the experiment.

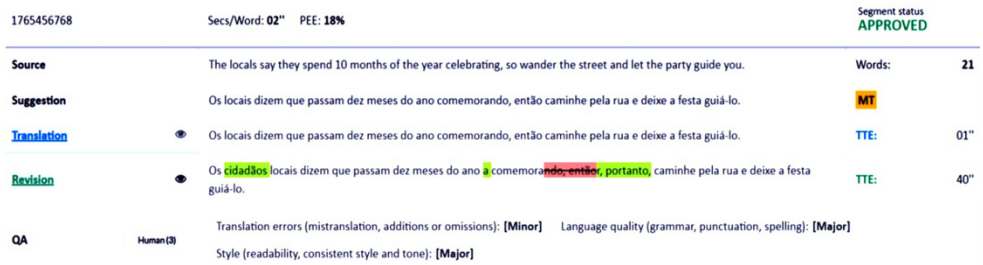

Figure 1. Example of post-editing a segment in MateCat.

4. Method and Study Design

The goal of this study is to analyse the usability of MatCat for introducing students to PEMT in the translation technology classroom. We want to find out to what extent student translators are able to define and categorise MT errors and assess them on the basis of a severity scale. In this experiment, students had to identify translation errors on their own and to decide which category the error could be allocated to. The error was then rated according to its impact factor as *neutral*, *minor* or *major.* The idea behind the experiment was to carry out an authentic PEMT activity, where the participants were asked to make all the necessary decisions on their own. The study was conducted with a group of 10 MA students of Specialised Translation who are familiar with MT and CAT tools. The

participants all being native speakers of Portuguese, were given an English source text (ST) of 447 words from the domain of tourism, which was machine-translated through MateCat. The ST did not contain any specific terminology, to make sure that the students were able to master this activity without using additional research tools. Students were then asked to evaluate the translation quality on the basis of the integrated issue grid that MateCat provides:

a) Style (readability, consistent style and tone);
b) Tag issues (mismatches, whitespaces);
c) Translation errors (mistranslation, additions or omissions);
d) Terminology and translation consistency;
e) Language quality (grammar, punctuation, spelling).

According to the severity levels of *neutral*, *minor* and *major*, participants had to rate each segment of the machine-translated text.
Figure 2 represents the example of a quantitative total error analysis performed by one of the participants.

QR Job summary

ID: 3602282	English US > Portuguese	100%	Reviewed Words 447	Translator Not assigned	Time to edit 00:15:19	PEE 14%	55.93 Quality score Fail Threshold 20

Issues	Neutral Weight: 0	Major Weight: 2	Minor Weight: 0.5	Total Error Points 25
Style (readability, consistent style and tone)	1	2	4	6
Tag issues (mismatches, whitespaces)				0
Translation errors (mistranslation, additions or omissions)	3	7	2	15
Terminology and translation consistency		1		2
Language quality (grammar, punctuation, spelling)		1		2

Figure 2: Total Error Analysis in MateCat

5. Results

The comparison of PE results obtained through a classroom experiment involving 10 MA student translators revealed the following data in terms of time to edit (TTE), post-editing effort (PEE) and error evaluation. The calculation of TTE, based on the time a translator spends on each segment, amounted in an average of 9min2s, showing, however, a huge variance among the students, ranging from 3min23s to 29min48s. PEE refers to the effort in

updating machine translation suggestions, and is automatically calculated based on the formula "total error points * 1000) / reviewed words".

The challenge of this PE experiment consisted not only in post-editing a machine-translated text, but also in defining and categorising errors. The final quality report revealed a huge span among students' error evaluation and the respective ratings. Since some segments contained several errors that overlapped, students had difficulties in assigning such overlappings to different categories, resulting in a huge variance of the total error grid. Another difficulty encountered was the delimitation of the severity-levels: *neutral*, *minor* and *major*. Given the generic character of these categories, students had more problems in classifying the errors that are less obvious, leading to a number of divergent answers. Table 1 gives a detailed overview of students' error analysis, revealing a rather heterogeneous classification result.

Table 1. Overview of Students' Error Analysis.

Participant	Style			Tag issues			Translation errors			Terminology and translation			Language quality			Total error points
	NE	M	MI	NE	MA	MI	NE	MA	MI	NE	MA	MI	NE	MA	MI	
1	2	5			1			3		1	6	1		3		36.5
2	1	2	4				3	7	2		1			1		25
3								8	1		5	2	1	2		31.5
4							1	2		1	4			1	1	14.5
5		1		1			1	4	1		4					18.5
6							1				11	3		3	7	33
7			4				1	5	7	2		2			2	17.5
8	2		1		1		3	1	2		1	8				25.73
9		1						2	2		1				3	10.5
10					1	1		12			5	3		2	2	43

Participants were faced with several challenges when carrying out the PE activity. Although the total error distribution was fragmented, we observed a rather homogeneous error distribution in a few segments. The following machine-translated segment was identified as a major translation problem by 80% of the participants involved in this experiment. The MT output was considered inappropriate and categorized as major translation error. 20% classified the translation problem as *Terminology and Translation consistency error* with minor and major severity-level ratings, respectively. The PE results obtained show that students had no difficulty in detecting a semantic translation error, as shown in the following example. However, they revealed difficulties in trying to correct the MT

suggestion. Examples P7 and P8 show that the participants did not succeed in conveying the original meaning, coming up with wrong translations.

ST: Terceira is Portuguese for "third", and fittingly this is the Azores' third largest island […]
MT: Terceira é a palavra portuguesa para "terceira", e apropriadamente esta é a terceira maior ilha dos Açores [...]
PE P1: É a terceira maior ilha dos Açores [...]
PE P2: A Terceira é, tal como o nome indica, a terceira maior ilha dos Açores [...]
PE P3: Terceira, tal como nome indica, é a terceira maior ilha dos Açores […]
PE P4: Terceira, apropriadamente é a terceira maior ilha dos Açores [...]
PE P5: Terceira é, tal como o nome indica, a terceira maior ilha dos Açores […]
PE P6: Terceira é apropriadamente a terceira ilha dos Açores [...]
PE P7: Terceira tem um sentido especial no seu nome: é a terceira maior ilha dos Açores [...]
PE P8: "Third" é a palavra portuguesa para "terceira", e apropriadamente esta é a terceira maior ilha dos Açores [...]"
PE P9: Terceira, tal como nome indica, é a terceira maior ilha dos Açores […]
PE P10: Terceira é apropriadamente a terceira maior ilha dos Açores [...]

The analysis of the results also showed that 30% of the participants let themselves influence and intimidate by the machine results, accepting translations based on the Brazilian variant of Portuguese. As can be observed from the following examples, the machine translation suggested the use of the pronoun você that in European Portuguese is usually avoided in a written text, whereas it is widely used in Brazil. Although familiar with these sociolinguistic differences, some students did not consider post-editing the segment, violating text convention patterns. We assume that this would not have occurred when translating the text without machine support.

ST: If you want to slip into a gentle pace of life, this is the island for you.
MT: Se você quer entrar em um ritmo de vida tranquilo, esta ilha é para você.
PE P1: Se quiser entrar num ritmo de vida tranquilo, esta ilha é para si.
PE P5: Se quer entrar em um ritmo de vida tranquilo, esta ilha é para você.
PE P8: Se você quer experienciar um ritmo de vida tranquilo, esta ilha é para você.
PE P9: Se quer entrar em um ritmo de vida tranquilo, esta ilha é para você.

It could also be observed that in segments with many errors, students were more attentive towards the correction of lexical errors, often neglecting grammatical aspects, such as the use of articles, or keeping accented characters from the Brazilian Portuguese.

6. Discussion

Due to the impact of MT on today's translation practice, it is of utmost importance to train future translators on PE. Over the last years, more and more Higher Education institutions have implemented PE modules/courses into their standard curricula, in an attempt to keep

track with industry innovations and changes. By preparing future translators for their professional life through a set of project-based activities involving translation technology tools, they learn "to make their own informed decisions about the type of work they can and want to do in the future, and to negotiate rates or deadlines with possible clients or employers" (Guerberof Arenas & Moorkens, 2019, p. 232). The present study collected PE data for the language pair English - European Portuguese and aimed at making a contribution to the research on PE activities for a subsequent integration into the translation technology classroom.

Acknowledgments

The author would like to thank the MA students who participated in this experiment for their time and feedback.

References

Doherty, S., Moorkens, J., Gaspari, F., Castilo, S. (2018). On Education and Training in Translation Quality Assessment. In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation quality assessment: From principles to practice*, (pp. 95-106). Springer: Heidelberg & Berlin.

Guerberof Arenas, A., & Moorkens, J. (2019). Machine Translation and Post-editing Training as part of a Master's Programme. *Journal of Specialised Translation*, 31, 217-238.

Lommel, A. R., Burchardt, A., & Uszkoreit, H. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. Tradumàtica: tecnologies de la traducció, 0 (12). 455–463. Retrieved from: https://pdfs.semanticscholar.org/774e/33248ef2bc5971e17702a5ad6308a16de551.pdf?_ga=2.159057203.806587727.1612786425-425445379.1612786425

MateCat (s.d.). Retrieved from: https://site.matecat.com/.

TAUS (s.d.) Translation Automation User Society. Retrieved from: https://blog.taus.net/knowledgehub/automated-mt-evaluation-metrics

Vilar, D., Xu, J., D'Haro, L., & Ney, H. (2006). Error analysis of statistical machine translation output. In Proceedings of the *Fifth International Conference on Language Resources and Evaluation (LREC)*, pp. 697–702, Genoa, Italy.