Ampersand 11 (2023) 100125

Contents lists available at ScienceDirect

Ampersand

journal homepage: www.elsevier.com/locate/amper

Assessing MT with measures of PE effort

Sergi Alvarez-Vidal *, Antoni Oliver

Universitat Oberta de Catalunya, Spain

A B S T R A C T

Recent improvements in quality obtained by neural machine translation (NMT) have boosted its presence in the translation industry. In many domains and language combinations, translators post-edit raw MT output: they edit and correct the pre-translated text to produce the final translation. However, this process can only produce the expected results if the quality of the raw MT can be assured. MT is usually assessed with automatic metrics, as they are faster and cheaper. However, these metrics are not always good quality indicators and do not correlate to the post-editing effort.

We suggest a two-step evaluation process for MT intended for post-editing. The automatic evaluations are followed by the assessment of the three dimensions of PE effort. This targeted evaluation can ensure a quality of the raw MT which does not jeopardise the final product or compromise the task of post-editors. We include a detailed description of PosEdiOn, an easy-to-use standalone tool which records PE effort, and a use case of its implementation. 18 translators post-edit texts from English into Spanish from the news domain translated with DeepL and an NMT system trained by the authors to gather PE effort metrics. We compare automatic and PE effort metrics to assess which MT system would be more suitable for post-editing.

1. Introduction

With the improvements in machine translation (MT) quality, especially since the widespread use of neural MT (NMT), this technology has exponentially increased its presence in the translation industry. Results of a recent language survey identify post-editing as the second most demanded task among language providers and the activity with the highest growth potential, 64% (European Language Industry Survey, 2022). For many language combinations and domains, translators post-edit the raw MT output, that is, they "edit, modify and/or correct pre-translated text that has been processed by an MT system from a source language into (a) target language(s)" (Allen and Somers, 2003). Yet, translators tend to translate from scratch if the raw MT quality provided is not good enough (Sanchez-Torron and Koehn, 2016), (Parra Escartín and Arcedillo, 2015).

Therefore, assessing the quality of the MT output is an essential step in the post-editing process. However, MT is usually evaluated with automatic metrics to overcome the necessary time and high costs derived from manual evaluation. Moreover, the same metrics are used to evaluate all MT outputs, without taking into account the final use of the translated text. We should devise evaluation methods which can assess the quality of an MT output depending on the task or the function for which the output is intended (Hovy et al., 2002). In the case of post-editing, we need to assess if the quality of the MT output is good enough to post-edit. A logical way to conduct this assessment is taking into account the impact the MT raw output has on post-editing effort.

We suggest a two-step evaluation process, which includes both automatic metrics and measures of PE effort, and we present a use case for its implementation. We want to post-edit English to Spanish texts from the news domain. To do so, we have to choose between two different NMT engines: a commercial solution (DeepL[1]) and a system trained by the authors. In the first step of the evaluation, we use some of the most usual automatic metrics to evaluate the quality of the two systems. We include some traditional scores like BLEU and a more recent metric (COMET) which has yielded very good results in recent MT evaluation campaigns.

In the second step, we calculate PE effort indicators. According to Krings (2001), PE effort consists of three dimensions: technical, temporal and cognitive effort. Temporal effort is related to the time spent post-editing. Technical effort reflects the keys pressed while doing the task, that is, the number of insertions, deletions, replacements and shifts. And cognitive effort refers to the mental processes that take place while post-editing. Even though these three dimensions are closely related (Moorkens et al., 2015), research shows a weak correlation among them (Cumbreño et al., 2021). We selected an indicator of each of the three dimensions for our comparison. For our use case, 18 translators post-edit a short news article translated with each of the systems using PosEdiOn (Oliver et al., 2020), an easy-to-use tool with keylogging functionalities. PosEdiOn records different measures of PE effort and automatically calculates the results. This way, we can compare the automatic metrics

* Corresponding author.
E-mail addresses: salvarezvid@uoc.edu (S. Alvarez-Vidal), aoliverg@uoc.edu (A. Oliver).

[1] https://www.deepl.com/translator.

https://doi.org/10.1016/j.amper.2023.100125
Received 1 February 2023; Received in revised form 31 March 2023; Accepted 25 May 2023
Available online 15 June 2023

with the PE effort metrics produced by PosEdiOn to assess the two MT systems.

2. Background and related work

2.1. Machine translation evaluation

Translations are intended for human users and, as such, human judgements seem clearly to be the right way to assess the quality and the problems presented by translations. Furthermore, the perception and knowledge we have as humans of the world surrounding us allows in many cases to evaluate MT errors and relate them to the severity they have for a specific translation according to the context and situation (Sanders et al., 2011).

However, manual evaluations are costly, both in time and effort, and too often the people who conduct these evaluations have limited knowledge or experience. As a consequence, evaluations can suffer from low inter- and intra-annotator agreements (Snover et al., 2006). Evaluating MT output is a challenging and complex task, which can lead to tiredness among evaluators. Many elements need to be taken into consideration when conducting the evaluation and too often the guidelines delivered to evaluators are not well defined or are prone to different interpretations (Callison-Burch et al., 2007).

Automatic scores produce quick results and are a necessary tool when developing an MT system, as it is possible to check if the modifications you introduce have had any effect on the product of the translation. One of the main problems is that they usually compare MT outputs (also called hypotheses) with one or more human translations of the same source text (also known as the gold-standard human translation). The closer the MT output is to the reference, the better the MT output is considered. However, there are many possible translations for one single source document. Even though more than one golden reference can be computed in automatic scores, this type of assessment cannot account for the variability in possible correct solutions for one source segment.

Most metrics claim their effectiveness by comparing their performance with other competitive metrics and correlate it with human judgements. However, too often the translation quality of a pair of MT systems relies exclusively on the differences between automatic scores such as BLEU to draw conclusions without performing any further assessment, not even a human evaluation (Marie et al., 2021). Thus, automatic metrics have been increasingly called into question, especially when comparing high-quality systems (Mathur et al., 2020), (Ma et al., 2019), not just for the metric itself but also for the quality and origin of the references used for the assessment (Freitag et al., 2020).

Many different metrics have been suggested and in fact it is still an active research subject. However, the most usual measure is BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002), which is de facto the standard for all MT evaluations. It compares 1 to 4 words from the MT output with multiple references and n-gram precision is modified to eliminate repetitions that occur across sentences. It also includes a brevity penalty that down-scales the score for the MT outputs that are shorter in length than the reference. Even though it has shown correlation with human judgments of translation quality (Coughlin, 2003), it has been questioned many times (Mathur et al., 2020), (Wieting et al., 2019). Furthermore, BLEU is not always reported consistently in the different reports, which produces BLEU scores which are not really comparable due to divergences in the tokenization and normalisation schemes used (Post, 2018).

There are other usual metrics used for automatic evaluation. WER (Word Error Rate) (Nießen et al., 2000) calculates the minimum number of substitutions, deletions and insertions which are necessary to convert the hypothesis into the reference translation. TER (Translation Edit Rate) (Snover et al., 2006) calculates the amount of post-editing necessary to match the reference translation, including insertions, deletions, substitutions and shifts with an equal cost for all edits. NIST (Doddington, 2002) is a variation of BLEU which performs an arithmetic mean instead of a geometric one, takes into account n-grams of length 5 and weighs more heavily n-grams which occur less frequently.

Other measures focus on the lexical recall, which calculates the proportion of lexical units in the reference covered by the MT output. chrF (character n-gram F-score) (Popović, 2015) calculates the n-gram precision and recall arithmetically which is averaged over all n-grams. METEOR (Banerjee and Lavie, 2005) aligns the MT output with the reference translation using synonyms, stems and paraphrases. Then it calculates the candidate-reference similarity taking into account the proportion of aligned words both in the candidate and the reference. Taking into account the type of similarity, it also includes different weights to the word matches. BERTscore (Zhang et al., 2020) is a language generation evaluation metric which is based on pretrained BERT contextual embeddings (Devlin et al., 2019). It computes the similarity of two sentences as a sum of cosine similarities between their tokens' embeddings. COMET (Rei et al., 2020) is an evaluation score which has obtained very good results in recent evaluation campaigns. It is a PyTorch-based framework for training highly multilingual and adaptable MT evaluation models that can function as metrics. Given a sentence embedding for the source, the hypothesis, and the reference, certain combined features are extracted. These combined features are then concatenated into a single vector that serves as input to a feed-forward regressor.

There are also other metrics which evaluate the quality of an MT system without having access to a reference translation. These quality estimation measures assess multiple features of the source and target language to estimate the performance of a system on individual data points (Specia et al., 2018). However, these metrics are out of the scope of our paper and are not as widely used as other traditional metrics.

All these measures provide specific information about some aspect of the MT output and can be useful for certain purposes. However, they are applied regardless of the purpose of the MT output as if there were a universal score of translation quality. In the case of post-editing, we need an evaluation method which takes into account the PE effort necessary to edit the MT output to assess the quality an MT engine produces.

2.2. Post-editing effort

Research has extensively studied how translators work and what elements of the source text generate a challenge when translating. With the widespread use of post-editing, many researchers have focused their attention on how translators post-edit and what is the post-editing effort implied in this task. Since the publication of the seminal work by Krings (2001), it is widely accepted that PE effort includes three dimensions: temporal, technical and cognitive. The main problem is that, even though all three dimensions generate valuable information which needs to be taken into account to calculate the PE effort, research has shown there is not much correlation among the three (Cumbreño et al., 2021).

The first of the three dimensions, temporal effort, is the most used in the translation industry. It is fairly easy to record and it can be directly linked to productivity, and thus used to calculate and establish post-editing rates for translators (Guerberof, 2009). It is also useful to calculate delivery time for the translations in an industry which is always trying to reduce time cycles (Rogers and Sosoni, 2013). Research has consistently shown that PE reduces temporal effort compared to translation from scratch (Jia et al., 2019), (Läubli et al., 2019), even though in some cases it has shown no improvement for general language texts (Screen, 2017). Currently, most commercial CAT tools used to post-edit include time recording mechanisms. However, the temporal dimension alone cannot account for the whole PE effort.

The second dimension is technical effort. It is related to all the editing actions taking place while post-editing the raw MT output. Actions conducted while editing are usually classified into four groups: insertions, deletions, replacements and movements or shifts. The first

two are primary actions which cannot be decomposed. The last two are complex actions which in fact are a combination of deletions and insertions (Do Carmo, 2020). We could also count the individual keys pressed and even include the mouse movements.

To compute all these calculations, keylogging data is usually used, which requires the use of a specific software. An indirect way to calculate the edits introduced into the translated text is to compare the raw MT output and the final translated text. Even though it does not account for all the modifications introduced in the different stages of translation and revision of the text, it can give information about the changes present in the final version.

The most common of these indirect measures is TER (Snover et al., 2006). It allows block movement of words, called shifts. These movements have the same cost as insertions, deletions or substitutions. It uses "a greedy search to select the words to be shifted, as well as further constraints on the words to be shifted. These constraints are intended to simulate the way in which a human editor might choose the words to shift" (Snover et al., 2006).

The third dimension is cognitive effort. It has its origin in cognitive psychology and relates to all the mental processes that take place while translators post-edit. It includes reading the texts, thinking about the translation from the source and studying the suggested MT solutions, correcting all the errors or mistranslations detected in the raw MT output, and revising the final version of the translation produced. Cognitive processes are always present while post-editing even if no keys are pressed and no corrections are introduced into the MT output.

Due to its nature, it is the most complex dimension to measure. In fact, only indirect measures can be used to trace this effort dimension. Krings (2001) suggested Think-Aloud Protocol (TAP) to study the different cognitive processes involved while post-editing. This technique is useful to understand the mental processes which take place while translating but it has received some criticism because it is difficult for translators to explain in words all their thoughts (House et al., 2019) and because the narrating process interferes with the translation process (Toury, 2012). However, Vieira (2016) found there was a strong correlation between TAP ratings and other measures of effort.

Eye-tracking has also been used to measure cognitive effort (Carl et al., 2011), (Doherty, 2013). It counts the number and duration of fixations, when the eyes are relatively still (Moorkens et al., 2018). To obtain more reliable results, it has been used in combination with pause analysis (O'Brien, 2009), and retrospective think-aloud protocols (Alves, 2003).

Pauses have been used to study speech and writing production (Schilperoord, 1996), (Alamargot et al., 2007) as well as second-language learning (Zulkifli, 2013) as they have been considered evidence of the cognitive processes that take place in the brain. They have also shown to be good indicators of cognitive effort in post-editing (Lacruz et al., 2012). O'Brien (O'Brien, 2006) suggested pause ratio as a way to measure cognitive effort in post-editing. It divided the total pause time for a specific segment by the total PE time, and was used to study negative translatability indicators. However, her research did not show any significant correlation. Lacruz et al. (2012), (Lacruz et al., 2014) suggested another measure of pauses that counted clusters of short pauses instead of the whole pause time. Results showed a clear correlation with PE effort and established the pause threshold at 300 ms.

3. First step: automatic evaluation

To assess the importance of using a two-step evaluation method which takes into account the PE effort necessary, we devised a use case which included the comparison of two NMT systems. We compared a widely known commercial MT engine (DeepL) with an MT engine trained by the authors in the news domain for the English-Spanish language combination. Both DeepL and our tailored MT engine were trained for a generalistic domain. The idea was to reproduce a real PE scenario. We wanted to post-edit news from English into Spanish, and we needed to select the MT system which produced the best quality to reduce the PE effort of the translators who would post-edit the texts.

For our NMT engine, we compiled a parallel corpus from Global Voices.[2] To do so, we downloaded all the pieces of news in English which had a translated version into Spanish, from the year 2004–2022. For the alignment of the texts we used MTUOC-aligner,[3] following the SBERT strategy. In this strategy, all the texts in English and Spanish for a given year are segmented and aligned, regardless of the piece of news in which they appear. The task is in fact a search of translated segments in comparable corpora. Afterwards, a cleaning process was performed and a parallel corpus of 791,959 unique parallel segments was obtained.[4] Since this number of segments is not sufficient to train a neural MT system, we used MTUOC-corpus-combination[5] to select 20,000,000 segments from the Paracrawl v9 English-Spanish corpus. The selection is based on a language model computed from the source segments of the compiled Global Voices corpus, so the selected segments are expected to be similar segments to those found in the news domain. This combination resulted in a training corpus of 20,781,959 segments. From the compiled Global Voices corpus, we reserved 5000 segments for validation and 5000 segments for evaluation. In this way, the training was performed using a combination of the Global Voices corpus and selected segments from Paracrawl, but the validation and the evaluation was carried out using segments from the Global Voices corpus.

The corpus was processed using SentencePiece (Kudo and Richardson, 2018) with the following parameters: joining languages: True; model type: bpe; vocabulary size 64,000; vocabulary threshold: 50. The (sub)word alignments of the training corpus have been calculated using eflomal (Östling and Tiedemann, 2016) in order to use guided-alignment in the training.

The NMT system was trained using the Marian-nmt toolkit[6] (Junczys-Dowmunt et al., 2018) with a transformer configuration. Two validation metrics were used: bleu-detok and cross-entropy. The early-stopping criterion was set to 5 on any of the metrics, and the validation frequency was set to 5000.

In the first step of the MT evaluation, we used the most usual automatic metrics, and also COMET, which has yielded good results in recent evaluation campaigns. We only included these measures as an example of possible automatic scores and in no case did we intend to include all possible automatic measures. Even though this is the most frequent way to evaluate MT quality in industrial scenarios, we have already seen it has been repeatedly questioned, especially when comparing high-quality systems. However, these metrics can give interesting information and can be used as an approximate evaluation of the MT results that we will later assess using PE effort. In Table 1, we can see that all metrics used yield better results for the NMT system trained by the authors except for COMET. For BLEU and NIST, the higher the value for the

Table 1
Automatic metrics for the MT engines used.

Automatic evaluation	DeepL	Tailored NMT
BLEU	0.382	**0.409**
NIST	7.981	**8.147**
WER	0.495	**0.47**
%EdDist	36.088	**34.689**
TER	0.459	**0.442**
COMET	**0.7475**	0.654

[2] https://globalvoices.org/.
[3] https://github.com/aoliverg/MTUOC-aligner.
[4] The existing Global Voices corpus published in Opus Corpus has a total of 355,143 segments for the English-Spanish language pair.
[5] https://github.com/aoliverg/MTUOC-corpus-combination.
[6] https://marian-nmt.github.io/.

automatic metric, the better is the MT quality considered. In the case of WER, edit distance and TER, a lower value states a higher MT quality. For COMET, we used the model wmt-20-comet-da and the higher the value, the better the quality of the MT engine.

4. Second step: evaluation of post-editing effort

Once the training of the MT engines was completed and once the evaluation with automatic metrics had been conducted, we moved to the second step of the evaluation process. For this second step, we evaluated the PE effort. As mentioned above, PE effort consists of three separate but interrelated dimensions. However, usual CAT tools used by professional translators are not able to record all three dimensions. Moreover, our goal was to use a stand-alone tool which did not depend on any proprietary program currently used for post-editing, as different translators and translation companies use some of the multiple programmes available on the market.

For this reason, we used PosEdiOn v2 (Oliver et al., 2020), a simple stand-alone tool that allows post-editing of MT output and records information of the post-editing effort (time, keystrokes and mouse actions) at sentence-level. It does not require any installation at all. The translator receives a package containing the editor program, a configuration file where certain parameters of the user interface such as the font size can be customised, and the text that has to be post-edited. The PosEdiOn editor program is distributed as a Python v3 code, and as executable files for Windows, Mac and Linux. The executable version does not require any additional installation or configuration. Once the program is running, a simple user interface displays the source and target segments that have to be post-edited. The interface displays a chronometer,[7] and the current and total number of segments (see Fig. 1).

The program stores in a database all the actions performed by the user (pressed keys, mouse movements) along with its timestamp. It also detects and stores when the editor loses focus, that is, when the user is performing a task in another application. Users can click on the PAUSE button to pause the task and the chronometer is stopped. When a segment is validated using Enter, its background turns green. There are further colour indicators for different stages of the translation process: orange (revision needed) or red (problem detected), that can be activated with several keyboard shortcuts. Users also have different options to move between segments. In summary, PosEdiOn has a very intuitive and easy-to-use interface, and the results it yields do not depend on the user's knowledge of a commercial CAT tool.

Once the translation is finished, the post-editor returns the task so it can be evaluated. The user can send the folder containing the project once compressed again, or just the SQLite database generated by PosEdiOn. Once the file is received, it can be analysed using the companion program PosEdiOn-analyzer. This tool offers a wide range of scores to evaluate the post-editing process: number of insertions, deletions, reordering operations, long pauses (pauses longer than a given threshold, 300 ms. by default), HBLEU, HNIST, HTER (Snover et al., 2006), HWER and HEditDistance. It also implements some of the scores proposed by Barrachina et al. (2009): KSR (keystroke ratio), MAR (mouse-action ratio) and KSRM (keystroke and mouse action ratio).

For our use case, the PE data was collected from a total of 18 translation students. They were all enrolled in the Degree of Translation and Interpreting Studies at the Universitat Oberta de Catalunya (UOC). As part of the curriculum, they translate, post-edit and correct for the Virtual Translation Agency, a simulated translation bureau (Buysschaert et al., 2018), with tasks that resemble professional jobs. For all the tasks they do for the Virtual Translation Agency, they all translate using a commercial online CAT tool called Phrase.[8] For most of them, this is the only tool to which they have access. This term, they were all working on different articles from Wikipedia from English into Spanish. For the post-editing task using PosEdiOn, they were given detailed instructions regarding the main characteristics of the tool they would use to post-edit, and the final publishable quality they were expected to deliver. They also had four days to familiarise themselves with the software used for the task and practise their post-editing skills with a test task.

For the evaluation task in PosEdiOn, we divided the 18 participants into two groups of nine people. Each group post-edited the same two texts. We selected a news article of 878 words from *The Guardian* published on 8th January 2023 that explained new procedures in foetal surgery for babies with spina bifida conducted in the United Kingdom, and we divided the text in two halves. For the first group, we translated the first text with DeepL and the second text with our NMT system. For the second group, we translated the first text with our NMT system and the second text with DeepL. That way, both texts were post-edited by the same number of participants after having been machine translated by both engines. Thus, each translator received two different compressed folders without any reference to the MT engine containing each one of the texts ready to post-edit in PosEdiOn.

Once they had finished and had sent back all the post-editing tasks done with PosEdiOn, we analysed the results and selected the PE effort metrics we wanted to compare. There are multiple scores which can be taken into account as indicators of PE effort. However, in order to simplify the interpretation of the results, we only selected one of the most usual metrics for every dimension of PE effort (Cumbreño et al., 2021). For the temporal effort, we counted the total time spent post-editing and normalised the value by the total number of tokens. The technical effort was calculated using the total number of keystrokes normalised by the number of tokens. For the cognitive effort, we used the number of pauses as a proxy metric. We also normalised the value by the total number of tokens. Based on the results of previous research (Lacruz et al., 2014), we counted the number of pauses longer than 300 ms and normalised it by the total number of words. We calculated the arithmetic mean for all the data of all nine translators for each of the values we wanted to study.

As it can be seen in Table 2, all evaluated metrics were better for DeepL for all the different dimensions except for the time spent post-editing Text B. As it has been pointed out before, the three dimensions of effort do not always correlate. In this case, while time is higher, the other two indicators of PE effort are slightly lower for the DeepL engine, which would suggest that in fact PE effort is lower. As a whole, results show that post-editors need a lower effort to post-edit de MT output translated with DeepL. However, while keystrokes and pauses show a variation of approximately 50% between the two engines, time only varies in 13% for Text A and 21% for Text B.

One of the main problems when collecting data from multiple participants is the great variation among them. In order to compensate for great divergences among participants, we also chose to prune the results. The pruning is based on a maximum value of normalised time, keystrokes and pauses. These maximum values are calculated with the mean value and two times the standard deviation. All segments with a normalised time, keystrokes and pauses greater than the maximum are not taken into account to calculate the pruned values of all scores. In these calculations, pauses were normalised by segment.

As it can be seen in Table 3, results confirm DeepL yields better results, even for the normalised time. This could be due to errors in the registration process of time (long pauses of participants that were unaccounted for). Keystrokes and pauses are doubled for our custom NMT engine, while the time difference is much lower.

As a further step, we could study the different MT errors produced by the different MT engines. However, error analysis is out of the scope of this paper. Furthermore, our goal is to focus on the effort translators produce for the different MT outputs without accounting for the number of errors each machine-translated version presents.

Even though in general effort indicators are lower for DeepL, we can

[7] The chronometer can be disabled and hidden using the configuration file.
[8] https://phrase.com/.

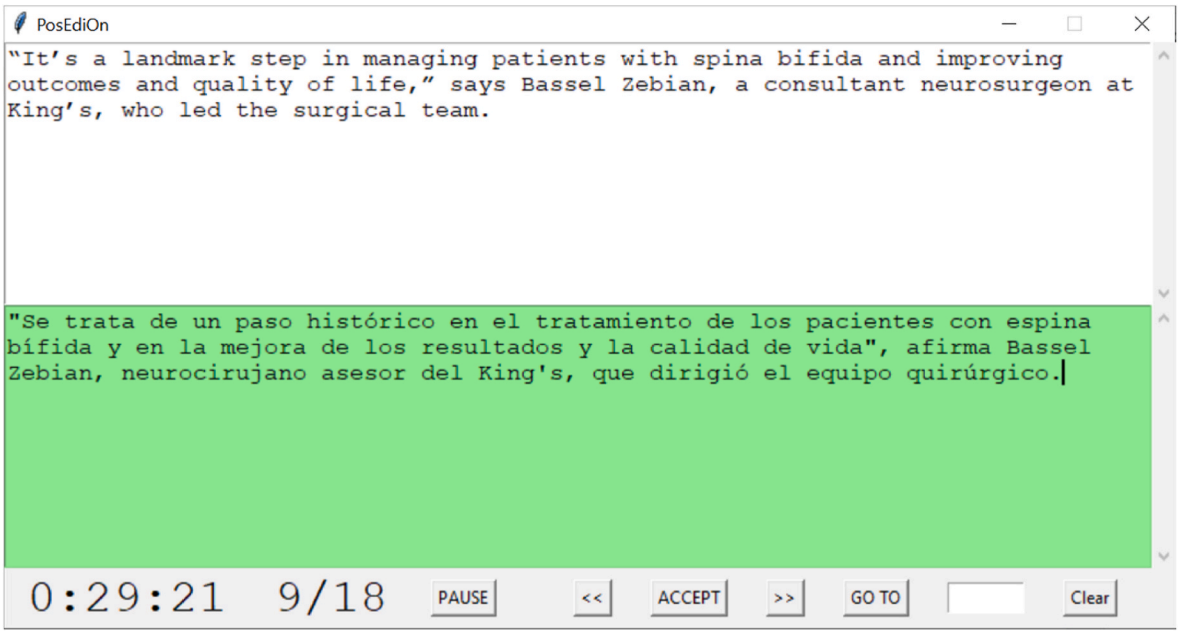


Fig. 1. Interface of PosEdiOn.

Table 2
Metrics of PE effort for each text and MT engine.

PE effort metrics	Text A		Text B	
	Custom NMT	DeepL	Custom NMT	DeepL
Total time	42.92	**36.20**	**44.95**	53.38
Normalised time	5.17	**4.47**	**4.35**	5.27
Total keystrokes	1458	**641.67**	1132.44	**540.78**
Normalised keystrokes	2.91	**1.37**	1.82	**0.89**
Total pauses	499.11	**260.78**	472.44	**246.44**
Normalised pauses	1	**0.53**	0.76	**0.40**

Table 3
Total pruned values for PE effort dimensions.

PE effort metrics	Custom NMT	DeepL
Normalised time	4.209	**3.579**
Normalised keystrokes	2.225	**0.918**
Normalised pauses	23.922	**11.162**

calculate the statistical significance for each indicator. To do so, we have randomly resampled the total 340 segments into samples of 50 segments. For these calculations, we have worked with the pruned segments. A total number of 10,000 resamplings have been performed. For each new sample, we have calculated the means of each indicator for each MT system, and counted how many times DeepL is performing better than Marian. Then the percentages have been calculated and can be observed in Table 4. Statistical significance for normalised keystrokes and total of long pauses are very high, so we can be sure that DeepL is performing much better for this indicator. The statistical significance for normalised time is moderate.

Further interesting data we studied was the number of unmodified segments for each of the translations, which is also a good indication of the quality of the raw MT output. As can be seen in Table 5, the number of unmodified segments is much lower for the custom NMT system, while for DeepL 29.01% and 34.34% of the segments were left without modification. That is, for the raw MT output produced using DeepL, approximately one third of the segments were already of publishable quality and were left as they were, without introducing any modifications.

All in all, the evaluation of PE effort showed PE effort was clearly lower when post-editing the raw MT output produced with DeepL.

Table 4
Statistical significance for each PE effort indicator.

PE effort indicators	Statistical relevance
Normalised time	82.75%
Normalised keystrokes	99.97%
Long pauses	100%

5. Conclusions and recommendations

In an industrial scenario, there is a need for a quick turnaround, which also includes quick evaluation methods. Automatic metrics can provide an easy way to assess the quality of MT output. However, automatic metrics such as BLEU were intended to be used as a development tool and we cannot blindly use them to assess MT systems without taking into account the final use of the translated text. In the case of post-editing, we suggest PE effort should be taken into account to assess the difficulty or complexity of the raw MT output once a translator needs to modify it to produce a publishable quality final text.

In the use case we conducted, automatic metrics showed a slightly better performance of our customised NMT engine. However, all metrics of PE effort showed much better results for the raw MT output translated with DeepL. As such, in a post-editing scenario, we should choose DeepL to translate texts from the news domain from English into Spanish.

Our analysis shows that automatic metrics are an insufficient indicator of the quality of raw MT output for post-editing and should be complemented with other evaluation metrics, preferably ones which take into account the three dimensions of PE effort.

Table 5
Unmodified segments for each NMT system.

	Text A		Text B	
	Custom NMT	DeepL	Custom NMT	DeepL
Total number of segments	18	18	22	22
Unmodified segments	0.33	5.22	1.33	7.56
% of unmodified segments	1.85	**29.01**	6.06	**34.34**

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Alamargot, D., Dansac, C., Chesnet, D., Michel, F., 2007. Parallel processing before and after pauses: a combined analysis of graphomotor and eye movements during procedural text production. In: Studies in Writing. https://doi.org/10.1163/9781849508223_003.

Allen, J.H., 2003. In: Somers, H. (Ed.), 'Post-editing', in Computers and Translation: A Translator's Guide. John Benjamins Publishing Company, pp. 297–317. https://doi.org/10.1075/btl.35.19all.

Alves, F., 2003. Tradução, cognição e contextualização: triangulando a interface processo-produto no desempenho de tradutores novatos. DELTA Doc. E Estud. Em Linguística Teórica E Apl. 19 (3). https://revistas.pucsp.br/index.php/delta/article/view/38328. (Accessed 10 December 2022).

Banerjee, S., Lavie, A., 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. Ann Arbor, Michigan. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation And/or Summarization, pp. 65–72. https://aclanthology.org/W05-0909. (Accessed 8 January 2023).

Barrachina, S., et al., 2009. Statistical approaches to computer-assisted translation. Comput. Ling. 35 (1), 3–28. https://doi.org/10.1162/coli.2008.07-055-R2-06-29.

Buysschaert, J., Fernandez-Parra, M., Kerremans, K., Koponen, M., Egdom, G.-W., 2018. Embracing digital disruption in translator training: technology immersion in simulated translation bureaus. Tradumàtica Tecnol. Trad. 125. https://doi.org/10.5565/rev/tradumatica.209.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., Schroeder, J., 2007. (Meta-) evaluation of machine translation. Prague, Czech Republic. In: Proceedings of the Second Workshop on Statistical Machine Translation, pp. 136–158. https://aclanthology.org/W07-0718. (Accessed 22 December 2022).

Carl, M., Dragsted, B., Elming, J., Hardt, D., Jakobsen, A.L., 2011. The process of post-editing: a pilot study. Cph. Stud. Lang. 131–142.

Coughlin, D., 2003. Correlating automated and human assessments of machine translation quality. In: *Proceedings Of Machine Translation Summit IX: Papers*, New Orleans, USA. https://aclanthology.org/2003.mtsummit-papers.9. (Accessed 9 December 2022).

Cumbreño, C., Aranberri, N., 2021. What do you say? Comparison of metrics for post-editing effort. In: Carl, M. (Ed.), Explorations in Empirical Translation Process Research. Springer International Publishing, Cham, pp. 57–79. https://doi.org/10.1007/978-3-030-69777-8_3.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: pre-training of deep bidirectional transformers for language understanding. NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171–4186. https://doi.org/10.18653/v1/n19-1423 *(Long and Short Papers)*.

Do Carmo, F., 2020. Editing actions: a missing link between translation process research and machine translation research. In: Explorations in Empirical Translation Process Research. Springer.

Doddington, G., 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the Second International Conference on Human Language Technology Research. Mar., San Francisco, CA, USA, pp. 138–145.

Doherty, S., 2013. Investigating the Effects of Controlled Language on the Reading and Language on the Reading and Comprehension of Machine Translated Texts: A Mixed-Methods Approach Using Eye Tracking. Dublin Ciry University.

ELIS, 'European Language Industry Survey, 2022. Trends, Expectations and Concerns of the European Language Industry'. ELIS Research, 2022. https://elis-survey.org/.

Freitag, M., Grangier, D., Caswell, I., 2020. BLEU might be Guilty but References are not Innocent. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 61–71. https://doi.org/10.18653/v1/2020.emnlp-main.5. Online, Nov.

Guerberof, A., 2009. Productivity and Quality in MT Post-editing. Proc. MT Summit XII, pp. 8–13.

House, J., 2019. Suggestions for a new interdisciplinary linguo-cognitive theory in translation Studies. In: Li, D., Lei, V.L.C., He, Y. (Eds.), Researching Cognitive Processes of Translation. Springer, Singapore, pp. 3–14. https://doi.org/10.1007/978-981-13-1984-6_1.

Hovy, E., King, M., Popescu-Belis, A., 2002. Principles of context-based machine translation evaluation. Mach. Translat. 17 (1), 43–75. https://doi.org/10.1023/A:1025510524115. Mar.

Jia, Y., Carl, M., Wang, X., 2019. How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study. J. Spec. Transl. 60–86.

Junczys-Dowmunt, M., et al., 2018. Marian: fast neural machine translation in C++. In: Proceedings of ACL 2018. System Demonstrations, Melbourne, Australia, pp. 116–121. https://doi.org/10.18653/v1/P18-4020.

Krings, H.P., 2001. Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes. The Kent State University Press, Kent, Ohio & London. https://benjamins.com/catalog/target.15.2.15jak. (Accessed 2 November 2022).

Kudo, T., Richardson, J., 2018. SentencePiece: a simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Nov, Brussels, Belgium, pp. 66–71. https://doi.org/10.18653/v1/D18-2012.

Lacruz, I., Shreve, G.M., Angelone, E., 2012. Average pause ratio as an indicator of cognitive effort in post-editing: a case study. San Diego, California, USA. In: Workshop on Post-Editing Technology and Practice. https://aclanthology.org/2012.amta-wptp.3. (Accessed 28 July 2022).

Lacruz, I., Denkowski, M., Lavie, A., 2014. Cognitive demand and cognitive effort in post-editing. In: Proceedings of the 11th Conference of the Association for Machine Translation in the Americas. Canada, Vancouver, pp. 73–84. https://aclanthology.org/2014.amta-wptp.6. (Accessed 1 November 2022).

Läubli, S., Amrhein, C., Düggelin, P., Gonzalez, B., Zwahlen, A., Volk, M., 2019. Post-editing productivity with neural machine translation: an empirical assessment of speed and quality in the banking and finance domain. Dublin, Ireland Proceedings of Machine Translation Summit XVII: Research Track 267–272. https://aclanthology.org/W19-6626. (Accessed 28 July 2022).

Ma, Q., Wei, J., Bojar, O., Graham, Y., 2019. Results of the WMT19 metrics shared task: segment-level and strong MT systems pose big challenges. Shared Task Papers, Day 1 Proceedings of the Fourth Conference on Machine Translation 2, 62–90. https://doi.org/10.18653/v1/W19-5302. Florence, Italy.

Marie, B., Fujita, A., Rubino, R., 2021. Scientific credibility of machine translation research: a meta-evaluation of 769 papers. Long Papers. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, vol. 1, pp. 7297–7306. https://doi.org/10.18653/v1/2021.acl-long.566. Online.

Mathur, N., Baldwin, T., Cohn, T., 2020. Tangled up in BLEU: reevaluating the evaluation of automatic machine translation evaluation metrics. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4984–4997. https://doi.org/10.18653/v1/2020.acl-main.448. Online.

Moorkens, J., O'Brien, S., da Silva, I.A.L., de Lima Fonseca, N.B., Alves, F., 2015. Correlations of perceived post-editing effort with measurements of actual effort. Mach. Translat. 29 (3–4), 267–284. https://doi.org/10.1007/s10590-015-9175-2.

Moorkens, J., 2018. Eye tracking as a measure of cognitive effort for post-editing of machine translation. In: Walker, C., Federici, F.M. (Eds.), Eye Tracking and Multidisciplinary Studies on Translation. John Benjamins Publishing Company, pp. 55–70. https://doi.org/10.1075/btl.143.04moo.

Nießen, S., Och, F.J., Leusch, G., Ney, H., 2000. An evaluation tool for machine translation: fast evaluation for MT research. Athens, Greece. In: Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00). http://www.lrec-conf.org/proceedings/lrec2000/pdf/278.pdf. (Accessed 8 January 2023).

Oliver, A., Alvarez, S., Badia, T., 2020. PosEdiOn: post-editing assessment in PythOn. Lisboa, Portugal. In: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, pp. 403–410. https://aclanthology.org/2020.eamt-1.43. (Accessed 1 November 2022).

Östling, R., Tiedemann, J., 2016. Efficient word alignment with Markov chain Monte Carlo. Prague Bull. Math. Linguist. 106 https://doi.org/10.1515/pralin-2016-0013.

O'Brien, S., 2006. Pauses as Indicators of Cognitive Effort in Post-editing Machine Translation Output. Cult., Lang https://doi.org/10.1556/ACR.7.2006.1.1.

O'Brien, S., 2009. An empirical investigation of temporal and technical post-editing effort. Transl. Interpret. Stud. 2 (1), 83–136. https://doi.org/10.1075/tis.2.1.03ob.

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318. https://doi.org/10.3115/1073083.1073135. Philadelphia, Pennsylvania, USA, Jul.

Parra Escartín, C., Arcedillo, M., 2015. A fuzzier approach to machine translation evaluation: a pilot study on post-editing productivity and automated metrics in commercial settings. In: Proceedings of the Fourth Workshop on Hybrid Approaches to Translation. HyTra), Beijing, Jul, pp. 40–45. https://doi.org/10.18653/v1/W15-4107.

Popović, M., 2015. chrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation. Portugal, Lisbon, pp. 392–395. https://doi.org/10.18653/v1/W15-3049.

Post, M., 2018. A call for clarity in reporting BLEU scores. arXiv, Sep 12. https://doi.org/10.48550/arXiv.1804.08771.

Rei, R., Stewart, C., Farinha, A.C., Lavie, A., 2020. COMET: a neural framework for MT evaluation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2685–2702. https://doi.org/10.18653/v1/2020.emnlp-main.213. Online, Nov.

Rogers, M.A., Sosoni, V., 2013. Translation in an age of austerity: from riches to pauper, or not? MTm J 5. https://openresearch.surrey.ac.uk/esploro/outputs/book/Special-Issue-of-mTm-on-Translation/99514675602346. (Accessed 6 November 2022).

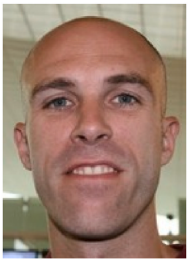
Sanchez-Torron, M., Koehn, P., 2016. Machine translation quality and post-editor productivity. Austin, TX, USA. In: Conferences of the Association for Machine Translation in the Americas: MT Researchers' Track, pp. 16–26. https://aclanthology.org/2016.amta-researchers.2. (Accessed 22 December 2022).

Sanders, Gregory, Przybocki, Mark, Madnani, Nitin, Snover, Matthew, 2011. Human subjective judgments. In: Handbook of Natural Language Processing and Machine Translation. Springer, pp. 750–759.

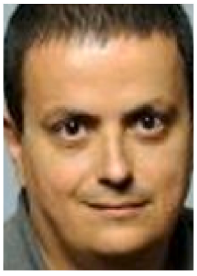
Schilperoord, J., 1996. It's about time: temporal aspects of cognitive processes in text production. Utrecht Stud. Lang. Commun. ume 6. https://brill.com/view/title/31102. (Accessed 30 July 2022).

Screen, B., 2017. Machine translation and Welsh: analysing free statistical machine translation for the professional translation of an under-researched language pair. J. Spec. Transl. 28, 218–244.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J., 2006. A study of translation edit rate with targeted human annotation. Cambridge, Massachusetts, USA. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, pp. 223–231. https://aclanthology.org/2006.amta-papers.25. (Accessed 6 November 2022).
Specia, L., Scarton, C., Paetzold, G.H., 2018. Quality Estimation for Machine Translation. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-02168-8.
Toury, G., 2012. Descriptive Translation Studies – and beyond. John Benjamins Publishing Company. https://benjamins.com/catalog/btl.100. (Accessed 8 December 2022).
Vieira, L.N., 2016. Cognitive Effort in Post-Editing of Machine Translation: Evidence from Eye Movements, Subjective Ratings, and Think-Aloud Protocols. University of Bristol.
Wieting, J., Berg-Kirkpatrick, T., Gimpel, K., Neubig, G., 2019. Beyond BLEU:training neural machine translation with semantic similarity. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4344–4355. https://doi.org/10.18653/v1/P19-1427. Florence, Italy.
Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y., 2020. BERTScore: evaluating text generation with BERT. arXiv. https://doi.org/10.48550/arXiv.1904.09675. Feb. 24.
Zulkifli, P.A.M.B., 2013. Applying Pause Analysis to Explore Cognitive Processes in the Copying of Sentences by Second Language Users. Ph.D., University of Sussex. http://sro.sussex.ac.uk/id/eprint/45933/. (Accessed 30 July 2022).



Sergi Alvarez-Vidal holds a PhD in Translation and Language Sciences from Universitat Pompeu Fabra. He is currently an Adjunct Professor at Universitat Oberta de Catalunya (UOC). He has worked as a freelance translator for more than 15 years, specialized in technical translation and localization. His research focuses on how MT can affect translations and translators, mainly studying post-editing and its effect on the translation process.



Antoni Oliver holds a PhD in Linguistics and a degree in Slavic Language and Literature from the Universitat de Barcelona, as well as a foundation degree in Telecommunications from the Universitat Politècnica de Catalunya.

He is an Associate Professor at the Universitat Oberta de Catalunya (UOC). He directs the Translation and Technologies Master programme and lectures in the Arts and Humanities Department, where he coordinates subjects related to language technologies.