PE effort and neural-based automatic MT metrics: do they correlate?

Sergi Alvarez-Vidal
Universitat Oberta de Catalunya
salvarezvid@uoc.edu

Antoni Oliver
Universitat Oberta de Catalunya
aoliverg@uoc.edu

Abstract

Neural machine translation (NMT) has shown overwhelmingly good results in recent times. This improvement in quality has boosted the presence of NMT in nearly all fields of translation. Most current translation industry workflows include post-editing (PE) of MT as part of their process. For many domains and language combinations, translators post-edit raw machine translation (MT) to produce the final document.

However, this process can only work properly if the quality of the raw MT output can be assured. MT is usually evaluated using automatic scores, as they are much faster and cheaper. However, traditional automatic scores have not been good quality indicators and do not correlate with PE effort. We analyze the correlation of each of the three dimensions of PE effort (temporal, technical and cognitive) with COMET, a neural framework which has obtained outstanding results in recent MT evaluation campaigns.

1 Introduction

In the last decade, MT has steadily increased its presence in all fields of translation. This is mainly due to the improvements in quality following the advances in NMT. 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% (ELIS, 2022). For many language combinations, translators edit, modify and correct the raw MT output to produce a final version. However, this process can only work properly if the quality of the raw MT output can be assured.

To assess the quality of the MT output both manual and automatic metrics are currently used. On the one hand, manual evaluations include sentence ranking, fluency and adequacy, direct assessment (DA) (Graham et al., 2016), and explicit error analysis, such as the ones based on the Multidimensional Quality Metrics (MQM) framework (Freitag et al., 2021a). Even though most of these evaluations produce quite reliable metrics, they have a high cost in time and resources (Papineni et al., 2002), which makes it complicated to use in a daily basis to assess the quality of MT systems. They also suffer from low inter- and intra-annotator agreements (Snover et al., 2006).

On the other hand, automatic evaluations produce quick results. Even though these metrics were originally conceived as a way to compare two systems, in most scenarios they are used as the only means to assess the quality of an MT engine. Automatic scores usually show correlation with human judgments of translation (Coughlin, 2003), even though they have been frequently questioned as a way to assess MT output (Mathur et al., 2020a), especially when they are used to compare high-quality systems (Ma et al., 2019).

The most usual automatic metrics currently used, such as BLEU (Papineni et al., 2002), or TER (Snover et al., 2006) are useful but present clear limitations and do not correlate with PE effort (Shterionov et al., 2019). Since the seminal work by Krings (2001), PE effort includes three dimensions: temporal effort (time spent translating), technical effort (keystrokes and all editing actions) and cognitive effort (mental processes taking place while translating). Even though all three are related, there is not a single measure which includes them all (Moorkens et al., 2015).

In recent times, new automatic metrics based on neural networks, such as BLEURT (Sellam et al., 2020), BERTSCORE (Zhang et al., 2020) and COMET (Rei et al., 2020) have shown outstanding results in recent evaluation campaigns (Mathur et al., 2020b; Freitag et al., 2021b; Freitag et al., 2022) based on MQM evaluations. We analyse if COMET, one of the best-performing metrics in recent campaigns, correlates better with the three dimensions of PE effort and, thus, could be used as a way to predict PE effort.

To do so, we collect PE information from ten translators who post-edited a news article from English into Spanish translated with two different MT engines. Then we study the correlation of each of the PE effort dimensions with COMET using Pearson product-moment correlation.

2 Related Work

2.1 Automatic Metrics

Automatic evaluations were developed as a solution to the slowness and high cost of manual evaluations. The most usual methods compare the MT output (also called hypothesis) with one or more human translations of the same source text (called references). The closer the MT output is to the reference, the better the MT output is considered. However, the main divergence is how they measure the difference between the two.

Some of these measures calculate the edit distance. 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 shift of phrases. All edits have equal cost. WER (Word Error Rate) (Nießen et al., 2000) calculates the Levenshtein distance, which is the minimum number of substitutions, deletions and insertions necessary to convert to hypothesis into the reference translation.

Other measures are precision-oriented. They measure the distance between the hypothesis and the references applying n-gram metrics, which are based on the lexical similarity between an MT output and one or more human references. For example, BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) is currently used as a standard for MT evaluation. 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 in many cases (Coughlin, 2003), some studies have questioned the role of BLEU in MT assessment (Wieting et al., 2019; Mathur et al., 2020a), especially when comparing high-quality systems (Ma et al., 2019).

Furthermore, there is a lack of consistency in the reporting of BLEU scores. That is, the parameters introduced in this metrics can have many variations and the resulting scores are not really comparable, due basically to the different tokenization and normalization applied to the reference (Post, 2018). Besides, it can also be affected by the outliers and sample size (Mathur et al., 2020a).

NIST (Doddington, 2002) is another precision-oriented measure. The main difference with BLEU is that NIST performs an arithmetic mean instead of a geometric one. It also takes into account n-grams of length 5 and weights more heavily n-grams which occur less frequently.

Some other measures combine lexical precision and recall. For example, chrF (character n-gram F-score) (Popović, 2015) calculates n-gram precision and recall arithmetically averaged over all n-grams. METEOR (Banerjee and Lavie, 2005) aligns the MT output to the reference translation using stems, synonyms, and paraphrases, besides exact word matching, and then computes candidate-reference similarity based on the proportion of aligned words in the candidate and in the reference.

Another possible approach is to use the post-edited version as the hypothesis. It is a quick way to obtain a proxy measure for technical effort, as it measures the modifications introduced into the final post-edited version, although it does not take into account the real post-editing process. HTER (Snover et al., 2006) is the most used human-targeted metric in machine translation and it is commonly employed as a gold standard in assessment of quality estimation (Graham et al., 2016), but we could also use other human-targeted metrics such as HBLEU.

To solve the problems many traditional automatic metrics have to assess the quality of current NMT models (Shterionov et al., 2018), neural models have been suggested. They are based on Quality Estimation (Specia et al., 2018) and include certain key features to produce an estimating model. For example, 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.

BERTSCORE (Zhang et al., 2020) leverages contextual embeddings from pre-trained transformers in order to create soft-alignments between words in candidate and reference sentences using cosine similarity. Based on the alignment matrix, it returns a precision, recall and F1 score. YISI-1 (Lo, 2019) measures the semantic similarity between a machine translation and human references. It aggregates the IDF-weighted lexical semantic similarities based on the contextual embeddings extracted from pre-trained language models. BLEURT (Sellam et al., 2020) is a learned metric that is fine-tuned to produce a DA for a given translation by encoding it jointly with its reference.

2.2 Post-editing effort

All research on PE effort has been based on the seminal work by Krings (Krings, 2001), which includes three dimensions of effort: temporal, technical and cognitive effort. Even though these three dimensions are related, there is not a single measure which includes them all (Moorkens et al., 2015; Aranberri and Gibert, 2019).

Temporal effort, which is the time spent post-editing the translation, is the most used dimension when analyzing PE effort in the translation industry, as it has a direct correlation to productivity. Research has consistently showed it improves when compared with translation from scratch (Läubli et al., 2019; Jia et al., 2019).

Technical effort is related to the editing process conducted by the translator while post-editing. It refers to all the keys and mouse movements a translator uses to modify the raw MT output to produce the final version. It is usually measured with keystroke analysis or key-logging data. It is often measured using indirect metrics such as HTER (Snover et al., 2006),

Cognitive effort is directly linked to cognitive demand and has been used as part of the cognitive load theory mainly in educational psychology (Paas et al., 2003). This dimensions of effort cannot be measured directly and different indirect proxy measures are used, such as think aloud protocol (TAP) (Vieira, 2016), eye-tracking (Carl et al., 2011; Doherty, 2013), choice network analysis (Campbell, 1999) and pause analysis (Lacruz et al., 2012).

Pauses have also shown to be good indicators of cognitive effort in post-editing. Lacruz et al. (2012; 2014) suggested a measure of pauses that counted clusters of short pauses while post-editing. Results showed a very good correlation with PE effort and established the pause threshold at 300 ms.

Translation industry has often used time as a measure of PE effort (Guerberof, 2009; Parra Escartín and Arcedillo, 2015), as it focuses in productivity. Post-editing is usually compared to translation from scratch, but PE between different MT models for different domains and language combinations do not always produce a clear improvement (Castilho et al., 2017; Screen, 2017; Bentivogli et al., 2018) and show a lack of correlation between post-editing productivity gains and MT quality metrics collected for the same NMT systems (Sarti et al., 2022). HTER is currently used as the main indirect automatic measure to study PE effort. However, correlation between general automatic scores and PE effort indicators do not shed light to its possible correlation (Shterionov et al., 2018; Alvarez et al., 2019).

3 Experimental Set-up

3.1 MT engines

To compare the PE effort measures and automatic scores, we decided to collect information from two different MT engines to avoid any bias produced by the MT model. We use a known commercial MT engine (DeepL) [1] and an MT engine trained by the authors to translate from English into Spanish two different fragments from a news article.

For the NMT engine trained by the authors, we first compiled a parallel corpus originated from Global Voices. In order to do so, we downloaded

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

all the news articles written in English which had a known translated version into Spanish from 2004 until 2022. In order to align all the texts, we used MTUOC-aligner[2], which is based on the SBERT strategy. That is, we segment and align all the texts written in English and Spanish for a specific year without taking into account the news article in which they appear. Thus, the task is a search of translated segments in comparable corpora. The next step includes a cleaning process to produce a parallel corpus of 791,959 unique parallel segments.

Since this number of segments is not enough to train a neural MT system, we selected 20,000,000 million segments from the Paracrawl v9 English-Spanish corpus using MTUOC-corpus-combination[3]. This 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. Using this combination, we produced a final training corpus of a total of 20,781,959 segments. From the corpus, we reserved 5,000 segments for validation and 5,000 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.

We used SentencePiece (Kudo and Richardson, 2018) to process the corpus using 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 (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 5,000.

We assessed the quality of the two NMT systems using some of the most frequently-used automatic metrics. For the evaluation, we used MTUOC-eval[4], a tool offering a wide range of automatic evaluation metrics. In Table 1, we can see the results of the evaluation. For COMET using references, we used the model wmt-20-comet-da and for COMET with no references we used the model wmt21-comet-qe-mqm.

	Marian	DeepL
BLEU	**0.401**	0.382
NIST	**8.056**	7.981
WER	**0.478**	0.495
%EdDist	**35.189**	36.088
TER	**0.448**	0.459
COMET (ref.)	0.654	**0.7475**
COMET (no ref.)	0.115	**0.1211**

Table 1: Evaluation of the MT systems with automatic metrics.

As we can observe in table 1, all the *classical* automatic metrics (BLEU, NIST, WER, %EdDist and TER), obtain better results for the Marian system trained for the experiments. However, both versions of COMET assign a better quality to DeepL. Even though the assessment of the raw MT quality is out of the scope of this paper and we are only focusing on metrics of PE effort, we can see that different automatic metrics do not coincide on the quality evaluations when comparing two different systems.

3.2 Methodology

To collect information on PE effort so that we could later compare the different PE effort indicators with results of automatic scores, we had the help of ten student translators. They were all enrolled in the Degree of Translation and Interpreting Studies at the Universitat Oberta de Catalunya (UOC). All of them were at the last year of their university studies and had previous experience translating from English into Spanish for the news domain.

They all conducted the post-editing task using PosEdiOn[5] v2 (Oliver et al., 2020), a simple standalone tool that allows post-editing of MT output and records information of the post-editing effort (time, keystrokes and mouse actions) at sentence-level. The PosEdiOn editor program is distributed as a Python v3 code, and as executable files for

[2] https://github.com/aoliverg/MTUOC-aligner

[3] https://github.com/aoliverg/MTUOC-corpus-combination

[4] https://github.com/aoliverg/MTUOC-eval

[5] https://github.com/aoliverg/PosEdiOn

Windows, Mac and Linux, and does not require any type of installation.

When working with PosEdiOn, translators receive a package with the program and the text which needs to be post-edited. Once the program is executed, they access a simple interface which can be partially customized. The interface displays a chronometer, and the current and total number of segments. 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.

There are certain shortcuts translators can use while post-editing. Users can also click on the PAUSE button to pause the task and stop the chronometer. When a segment is validated, its background turns green. There are also additional colors that can be used to indicate the different steps of the translation process for the current segment: orange (revision needed) or red (problem detected). Translators can access this and other options with shortcuts explained in the documentation[6].

For the post-editing task using PosEdiOn, all translators were given detailed instructions about the tool. They had a one-week period to test the tool, practise its use with a test text, read the documentation and ask all necessary questions. After the trial period, they were sent the files to post-edit.

The ten translators post-edited two different machine translated texts. Each of texts was about 400 words and was a fragment extracted from the same news article, published on The Guardian on 8th January 2023. Both fragments had an equivalent lexical variety, measured with type-token ratio. The text explained new procedures in foetal surgery for babies with spina bifida conducted in the United Kingdom. It included some medical terminology which could generate difficulties for the MT engines. The first text was translated with DeepL and the second one with our NMT system. Once translated with the different NMT engines, we prepared a compressed file ready to post-edit in PosEdiOn. We sent each translator both compressed files without stating any further information about the MT engines used.

They had a week to post-edit both texts. They received detailed instructions of the publishable-quality expected. Once they had finished, they returned the compressed files. PosEdiOn includes a small additional program which enables a quick analysis and produces a wide range of automatic scores to assess 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), It also includes COMET (Rei et al., 2020) and HCOMET. The former measure include the pretrained features and the latter uses the post-edited text as the reference.

4 Results

We used all the data collected while each of the translators post-edited using PosEdiOn to calculate the PE effort indicators. For each segment of the post-edited texts, we calculated the three dimensions of PE effort.

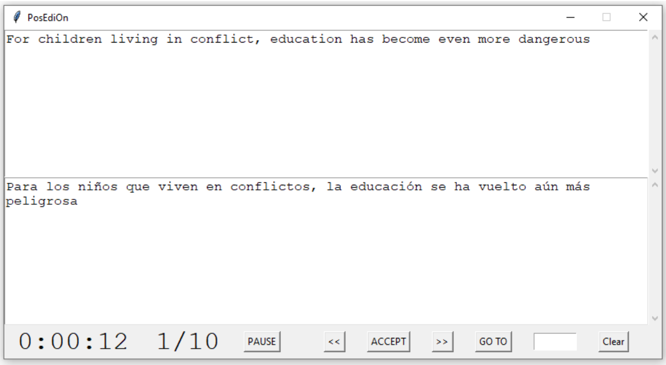
For the temporal effort, we calculated the total time per segment normalised by the number of tokens. For the technical effort, we calculated the number of keystrokes normalized by the number of tokens. For the cognitive effort, we calculated the number of pauses longer than 300 ms plus one (the initial pause for each segment) following the research results suggested by Lacruz et al. (2014). In table 2 we can observe the average values for each MT engine.

	Marian	DeepL
long pauses	22.07	**12.57**
norm. time	4.71	**3.05**
norm. keystrokes	1.74	**1.36**

Table 2: Average values for the different PE effort indicators.

According to these indicators, all three dimensions of effort were reduced when using DeepL, which would seem to show a correlation with the results of the automatic evaluation metrics for COMET (see table 1). However, we wanted to study the correlation of the automatic metrics at a segment level. To do so, we used the same three measures of each of the PE effort indicators and correlate them with four automatic metrics (HBLEU, HTER, HCOMET and COMET)

[6] https://github.com/aoliverg/PosEdiOn/wiki



For children living in conflict, education has become even more dangerous

Para los niños que viven en conflictos, la educación se ha vuelto aún más peligrosa

Figure 1: GUI interface of PosEdiOn

		Marian		DeepL	
		CORREL	STEYX	CORREL	STEYX
Long pauses	HBLEU	**-0.663**	**13.444**	-0.496	17.755
Long pauses	HTER	0.637	13.841	**0.635**	**15.792**
Long pauses	HCOMET	-0.358	16.767	-0.497	17.743
Long pauses	COMET	-0.552	14.975	-0.275	19.660
Norm. time	HBLEU	**-0.336**	**3.839**	-0.524	3.487
Norm. time	HTER	0.324	3.857	**0.572**	**3.358**
Norm. time	HCOMET	-0.257	3.939	-0.399	3.755
Norm. time	COMET	-0.303	3.884	-0.120	4.065
Norm. keystrq.	HBLEU	-0.753	0.953	-0.640	1.711
Norm. keystrq.	HTER	**0.769**	**0.925**	**0.655**	**1.682**
Norm. keystrq.	HCOMET	-0.468	1.280	-0.344	2.090
Norm. keystrq.	COMET	-0.419	1.346	-0.021	2.225

Table 3: Correlation between effort indicator and automatic measures

calculated segment by segment with PosEdiOn-analyzer. HBLEU, HTER and HCOMET are calculated using the machine translated segment as the hypothesis and the post-edited segment as the reference. Even though they do not account for the translation process, they compare the final PE resulting final with the raw MT output. COMET does not need a reference as it uses a pre-trained model. For HCOMET we have used the model wmt20-comet-da, and for COMET without references we have used the model wmt21-comet-qe-mqm.

In table 3, we can observe the correlation (CORREL) calculated with Pearson product-moment correlation and the standard error of the lineal regression (STEYX) for each PE effort metric and all four automatic scores. The higher the value of CORREL, the better the correlation, with a maximum value of 1. Values from 0.7 to 0.1 show a high correlation; 0.5 to 0.7 point to a moderate correlation; 0.3 to 0.5 are a sign of a low correlation, and 0 to 0.3 show no correlation. A correlation of 1 indicates a perfect positive correlation, and a value of -1 indicates a perfect negative correlation. At the same time, the lower the STEYX value, the better the correlation.

For cognitive effort calculated with long pauses, the best values are obtained by HBLEU for the Marian set, and HTER for the DeepL set. Both measures show a moderate correlation with a high standard error. For the temporal effort calculated with the normalized time, the same two metrics yield again the best results, even though they show a low and moderate correlation with a much lower STEYX value. For the technical effort calculated with normalized keystrokes, the best values are obtained by HTER, which show a high and moderate correlation with a very low standard error.

It is important to note that neither HCOMET nor COMET perform well in terms of correlation with effort indicators when calculated segment by segment. We must keep in mind, however, that the values of COMET related measures are dependent on the models used, and different models can score differently on the same data. Furthermore, results differ for each of the PE effort indicators but also for the two MT engines used. Even so, measures which take into account the PE version as hypothesis seem to show a moderate correlation, which could suggest they can give an approximate indication of the PE effort necessary.

5 Conclusions and future work

In this paper we have presented the results of an experiment aiming to assess the correlation of several automatic metrics with the three dimensions of post-editing effort: temporal effort, technical effort and cognitive effort. The main goal was to check whether a relatively new neural-based metric, COMET, correlates better than other widely used metrics, such as HBLEU and HTER, and could be used as a predictor for PE effort.

The limitations of this paper include the length of the text post-edited and the total number of translators who have participated in the PE task. However, the results obtained from this small sample show that COMET does not correlate for any of the PE effort indicators. HBLEU and HTER show a moderate to strong correlation for some of the indicators, but low for others. This would confirm the results of previous research stating the lack of correlation between all three dimensions of effort. The variability depending on the MT model could point to the types of errors produced by the MT engines and the different PE effort implied in correcting them.

For future experiments, we will collect data from a larger number of translators and larger texts, and we will train COMET models which can correlate better with some or all the PE effort indicators. The final goal would be to obtain a measure which could predict better PE effort than the current automatic measures used in the translation industry.

Acknowledgments

This work is partially supported by the project *TAN-IBE: Neural Machine Translation for the romance languages of the Iberian Peninsula*, founded by the Spanish Ministry of Science and Innovation Proyectos de generación de conocimiento 2021. Reference: PID2021-124663OB-I00.

References

Alvarez, Sergi, Antoni Oliver, and Toni Badia. 2019. Does NMT make a difference when post-editing closely related languages? the case of spanish-catalan. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 49–56. European Association for Machine Translation.

Aranberri, Nora and Ona de Gibert. 2019. Estrategia multidimensional para la selección de candidatos de traducción automática para posedición. *Linguamática*, 11(2):3–16. Number: 2.

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72. Association for Computational Linguistics.

Barrachina, Sergio, Oliver Bender, Francisco Casacuberta, Jorge Civera, Elsa Cubel, Shahram Khadivi, Antonio Lagarda, Hermann Ney, Jesús Tomás, Enrique Vidal, and Juan-Miguel Vilar. 2009. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28. Place: Cambridge, MA Publisher: MIT Press.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2018. Neural versus phrase-based MT quality: An in-depth analysis on english–german and english–french. *Computer Speech & Language*, 49:52–70.

Campbell, Stuart. 1999. A cognitive approach to source text difficulty in translation1. *Target. International Journal of Translation Studies*, 11(1):33–63.

Carl, Michael, Barbara Dragsted, Jakob Elming, Daniel Hardt, and Arnt Lykke Jakobsen. 2011. The process of post-editing: A pilot study. *Copenhagen Studies in Language*, (41):131–142. Publisher: Samfundslitteratur.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Yota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio, Antonio Valerio Miceli Barone, and Maria Gialama. 2017. A comparative quality evaluation of PBSMT and NMT using professional translators. In *Proceedings of Machine Translation Summit XVI: Research Track*.

Coughlin, Deborah. 2003. Correlating automated and human assessments of machine translation quality. In *Proceedings of Machine Translation Summit IX: Papers*.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 138–145. Morgan Kaufmann Publishers Inc.

Doherty, Stephen. 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. *PhD Thesis*.

ELIS. 2022. European language industry survey 2022. trends, expectations and concerns of the european language industry.

Freitag, Markus, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Freitag, Markus, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774. Association for Computational Linguistics.

Freitag, Markus, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68. Association for Computational Linguistics.

Graham, Yvette, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi. 2016. Is all that glitters in machine translation quality estimation really gold? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3124–3134, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Guerberof, Ana. 2009. Productivity and quality in MT post-editing. *Proceedings of MT Summit XII*, pages 8–13.

Jia, Yanfang, Michael Carl, and Xiangling Wang. 2019. How does the post-editing of neural machine translation compare with from-scratch translation? a product and process study. *The Journal of Specialised Translation*, pages 60–86.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in c++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121. Association for Computational Linguistics.

Krings, Hans P. 2001. *Repairing texts: Empirical investigations of machine translation post-editing processes*. The Kent State University Press.

Kudo, Taku and John Richardson. 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*, pages 66–71. Association for Computational Linguistics.

Lacruz, Isabel, Gregory M. Shreve, and Erik Angelone. 2012. Average pause ratio as an indicator of cognitive effort in post-editing: A case study. In *Workshop on Post-Editing Technology and Practice*. Association for Machine Translation in the Americas.

Lacruz, Isabel, Michael Denkowski, and Alon Lavie. 2014. Cognitive demand and cognitive effort in post-editing. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, pages 73–84. Association for Machine Translation in the Americas.

Lo, Chi-kiu. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. arxiv preprint. DOI:10.18653/v1/W19-5358.

Läubli, Samuel, Chantal Amrhein, Patrick Düggelin, Beatriz Gonzalez, Alena Zwahlen, and Martin Volk. 2019. Post-editing productivity with neural machine translation: An empirical assessment of speed and quality in the banking and finance domain. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 267–272. European Association for Machine Translation.

Ma, Qingsong, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90. Association for Computational Linguistics.

Mathur, Nitika, Timothy Baldwin, and Trevor Cohn. 2020a. 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*, pages 4984–4997, Online, July. Association for Computational Linguistics.

Mathur, Nitika, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020b. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725. Association for Computational Linguistics.

Moorkens, Joss, Sharon O'Brien, Igor A. L. da Silva, Norma B. de Lima Fonseca, and Fabio Alves. 2015. Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation*, 29(3):267–284.

Nießen, Sonja, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An evaluation tool for machine translation: Fast evaluation for MT research. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. European Language Resources Association (ELRA).

Oliver, Antoni, Sergi Alvarez, and Toni Badia. 2020. PosEdiOn: Post-editing assessment in PythOn. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 403–410. European Association for Machine Translation.

Paas, Fred, Juhani E. Tuovinen, Huib Tabbers, and Pascal W. M. Van Gerven. 2003. Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, 38(1):63–71.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Parra Escartín, Carla and Manuel Arcedillo. 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)*, pages 40–45. Association for Computational Linguistics.

Popović, Maja. 2015. chrF: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. Association for Computational Linguistics.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702. Association for Computational Linguistics.

Sarti, Gabriele, Arianna Bisazza, Ana Guerberof Arenas, and Antonio Toral. 2022. DivEMT: Neural machine translation post-editing effort across typologically diverse languages.

Screen, Ben. 2017. Machine translation and welsh: Analysing free statistical machine translation for the professional translation of an under-researched language pair. *The Journal of Specialised Translation*, 28:218–244.

Sellam, Thibault, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: Learning robust metrics for text generation. arXiv preprint.

Shterionov, Dimitar, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'Dowd, and Andy Way. 2018. Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation*, 32(3):217–235.

Shterionov, Dimitar, Félix Do Carmo, Joss Moorkens, Eric Paquin, Dag Schmidtke, Declan Groves, and Andy Way. 2019. When less is more in neural quality estimation of machine translation. an industry case study. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 228–235. European Association for Machine Translation.

Snover, Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231. Association for Machine Translation in the Americas.

Specia, Lucia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. *Quality Estimation for Machine Translation*. Synthesis Lectures on Human Language Technologies. Springer International Publishing.

Vieira, Lucas Nunes. 2016. Cognitive effort in post-editing of machine translation: Evidence from eye movements, subjective ratings, and think-aloud protocols. *PhD Thesis*.

Wieting, John, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355. Association for Computational Linguistics.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. arxiv preprint DOI: 10.48550/arXiv.1904.09675.

Östling, Robert and Jörg Tiedemann. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics*, 106.