Quantitative Analysis of Post-Editing Effort Indicators for NMT

Sergi Alvarez
Universitat Pompeu Fabra
salvarezvid@uoc.edu

Antoni Oliver
Universitat Oberta de Catalunya
aoliverg@uoc.edu

Toni Badia
Universitat Pompeu Fabra
toni.badia@upf.edu

Abstract

The recent improvements in machine translation (MT) have boosted the use of post-editing (PE) in the translation industry. A new MT paradigm, neural MT (NMT), is displacing its corpus-based predecessor, statistical machine translation (SMT), in the translation workflows currently implemented because it usually increases the fluency and accuracy of the MT output. However, usual automatic measurements do not always indicate the quality of the MT output and there is still no clear correlation between PE effort and productivity. We present a quantitative analysis of different PE effort indicators for two NMT systems (transformer and seq2seq) for English-Spanish in-domain medical documents. We compare both systems and study the correlation between PE time and other scores. Results show less PE effort for the transformer NMT model and a high correlation between PE time and keystrokes.

1 Introduction

The use of machine translation (MT) systems for the production of drafts that are later post-edited has become a widespread practice in the translation industry. Research has concluded that post-editing of machine translation (PEMT) is usually more efficient than translating from scratch (Plitt and Masselot, 2010; Federico et al., 2012; Green et al., 2013). Thus, it has been included in the translation workflow because it increases productivity when compared with human translation (Aranberri et al., 2014) and reduces costs (Guerberof, 2009) without having a negative impact on quality (Plitt and Masselot, 2010). Post-editors "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, 2003, p. 296).

In recent years, neural machine translation (NMT) has produced promising results in terms of quality, for example in WMT 2019 (Barrault et al., 2019). This has increased the interest in this new paradigm for the translation industry, which has begun to substitute its corpus-based predecessor, statistical machine translation (SMT), with new NMT models. It has also boosted the incorporation of PEMT in many translation workflows. In the 2018 Language Industry Survey,[1] 37% of the respondents reported an increase of MT post-editing and an additional 17% indicated that they had started implementing this practice.

Given the improved-quality performance of NMT and its widespread use in industrial scenarios, it is necessary to study the potential this approach can offer to post-editing. One of the main problems is that automatic scores give a general idea of the MT output quality but do not always correlate to post-editing effort (Koponen, 2016; Shterionov et al., 2018). However, many professional translators state that if the quality of the MT output is not good enough, they delete the remaining segments and translate everything from scratch (Parra Escartín and Arcedillo, 2015).

One of the main goals both of industry and research is to establish a correlation between the quality measurements of the MT output and translators' performance. Regarding post-editing ef-

[1] http://fit-europe-rc.org/wp-content/uploads/2019/05/2018-Language-Industry-Survey-Report.pdf?x77803

fort, all research uses the three separate but interrelated, dimensions established by Krings (2001): temporal, technical and cognitive. Temporal effort measures the time spent post-editing the MT output. Technical effort makes reference to the insertions and deletions applied by the translator and is usually measured with keystroke analysis, HTER (Snover et al., 2006) or Levenshtein distance (edit distance). Cognitive effort relates to the cognitive processes taking place during post-editing and has been measured by eye-tracking or think-aloud protocols. Krings (2001) claimed that post-editing effort could be determined as a combination of all three dimensions. Even though no current measure includes them all, cognitive effort was found to correlate with technical and temporal PE effort in a study by Moorkens et al. (2015).

In this paper we present a preliminary comparative quantitative analysis of different post-editing effort indicators (technical and temporal) for two NMT systems for English-Spanish in-domain medical documents. First of all, we trained a transformer and seq2seq model and compared them with Google Translate and an SMT engine (check section 4.1 for further detail on the results). As the NMT systems produced better quality results, we used them to translate three English-to-Spanish medical texts. Then, two different translators post-edited each version with PosEdiOn,[2] a post-editing tool developed mainly to collect information on different direct and indirect effort indicators (technical and temporal effort).

In Section 2 we analyse some of the previous work on post-editing effort. We explain the different NMT architectures in Section 3. In Section 4 we detail the MT systems and corpora used. We explain the experimental settings in Section 5 and we present the results in Section 6.

2 Previous Work

NMT is not a new architecture, but it can only be applied once the computational limitations have been solved (Cho et al., 2014; Bahdanau et al., 2015). The promising results obtained in automatic metrics such as BLEU (Papineni et al., 2002) have been paired with excellent scores in human evaluation of NMT (Wu et al., 2016; Junczys-Dowmunt et al., 2016; Isabelle et al., 2017) when compared to SMT, which has been the predominant MT architecture so far.

Once the improvement in quality has been determined, it was necessary to analyse its benefits for post-editing. One of the first complete papers studying the impact of SMT and NMT in post-editing was (Bentivogli et al., 2016). They carried out a small scale study on post-editing NMT and SMT outputs of English to German translated TED talks. They conclude that NMT in general terms decreases the post-editing effort, but degrades faster than SMT with sentence length. One of the main strengths of NMT is reordering of the target sentence.

Toral and Sánchez-Cartagena (2017) increase the initial scope of the study by Bentivogli et al. (2016) by increasing the language combinations and the metrics. One of the main conclusions is an improvement in quality when using NMT, although it is not the same for all the language combinations.

Castilho et al. (2017) report on a comparative analysis of phrase-based SMT (PBSMT) and NMT. They compare four language pairs and different automatic metrics and human evaluation methods. General results show a quality increase for NMT, although it also highlights some of the weaknesses of this new system. It focuses on post-editing and uses the PET interface (Aziz et al., 2012) to compare educational domain outputs from both systems using different metrics. NMT is shown to reduce word order errors and improve fluency. However, even if keystrokes are reduced, temporal PE effort exhibits no significant reduction.

Koponen et al. (2019) present a comparison of PE changes performed on NMT, rule-based MT (RBMT) and SMT output for the English-Finnish language combination. A total of 33 translation students participate in this English-to-Finnish PE experiment. It outlines the strategies participants adopt to post-edit the different outputs, which contributes to the understanding of NMT, RBMT and SMT approaches. It also concludes that PE effort is lower for NMT than for SMT.

In industrial scenarios, Shterionov et al. (2018) show that NMT systems obtain higher rankings by human reviewers than phrased-based SMT in all cases. They highlight that automatic measures such as BLEU, F-measure (Chinchor, 1992) and TER scores do not always correlate with NMT quality. Rather, they usually tend to underestimate it. Even in closely-related languages, which

[2]https://sourceforge.net/projects/posedion/

System	BLEU	NIST	WER	DA
Marian S2S	0.3601	7.6142	0.6893	64
Marian Transformer	0.3616	7.3863	0.6334	68
Moses	0.3942	7.8146	0.7386	46
Google Translate	0.3304	7.1197	0.7788	56

Table 1: Automatic and DA evaluation figures

are traditionally post-edited with RBMT systems, NMT systems with worse automatic metrics show better results in human evaluation (Costa-Jussà, 2017; Alvarez et al., 2019).

Regarding PE effort indicators, PE time is one of the most commonly-used elements to study MT quality, although research shows considerable variation among translators (Koponen et al., 2019). HTER is another measure frequently used in the industry due to its theoretical correlation to PE effort (Specia and Farzindar, 2010). However, research has shown it does not always correspond to translators' perception of quality (Koponen, 2012; Graham et al., 2016). In fact, some authors suggest new ways of measuring PE effort taking into account different scores (Scarton et al., 2019) or a multidimensional approach that combines some of the currently existing measures (Aranberri et al., 2014).

Given the undeniable improvements in quality NMT offers for post-editing, we study two different NMT systems and how they affect different indicators of post-editing effort. We also analyse the correlation of PE time with different direct and indirect measures of technical effort (keystrokes, HBLEU, HTER and edit distance). As far as we are aware, there are no studies comparing how two different NMT outputs affect post-editing for English to Spanish in-domain texts.

3 NMT architectures

The basic architecture of NMT models (Cho et al., 2014; Sutskever et al., 2014) consists of an encoder and a decoder. First of all, each word included in the input sentence is introduced as a separate element into the encoder so that it can encode it into an internal fixed-length representation called the context vector. It contains the meaning of the whole sentence. Then, the decoder decodes the context vector and predicts the output sequence.

Instead of encoding the input sequence into a single fixed context vector, attention (Bahdanau et al., 2015) is proposed as a solution to the limitation of the encoder-decoder model encoding the input sequence to one fixed length vector. It develops a context vector that is filtered specifically for each output time step.

Transformer (Vaswani et al., 2017) follows mainly the encoder-decoder model with attention passed from encoder to decoder. It employs a self-attention mechanism that allows the encoder and decoder to account for every word included in the entire input sequence. Transformer proposes to encode each position, apply self-attention in both decoder and encoder, and enhance the idea of self-attention by calculating multi-head attention. This improves performance expanding the model's ability to focus on different positions and gives the attention layer multiple sets of weight matrices. There are no recurrent networks, only a fully connected feed-forward network.

4 MT systems and training corpora

4.1 MT systems

For the experiments, we used Marian[3] (Junczys-Dowmunt et al., 2018) to train two NMT systems. For the first one (1) we used an RNN-based encoder-decoder model with attention mechanism (s2s), layer normalization, tied embeddings, deep encoders of depth 4, residual connectors and LSTM cells. For the second one (2), the transformer, we used the configuration in the example of the Marian documentation,[4] that is, 6 layer encoder and 6 layer decoder, tied embeddings for source, target and output layer, label smoothing, learn rate warm-up and cool down.

To establish a comparison baseline, we trained a Moses model with the same corpus, and also used Google translate. We assessed the resulting engines with standard automatic metrics (see Table 1). The best scores for BLEU were obtained by the Moses engine, even though WER was better for the two NMT systems. This is in line with the

[3]https://marian-nmt.github.io
[4]https://github.com/marian-nmt/marian-examples/tree/master/transformer

Corpus	Segments/Entries	Tokens eng	Tokens spa
BMTR	816,544	14,726,693	16,836,428
Medline Abstracts	100,797	1,772,461	1,964,860
UFAL	258,701	3,202,162	3,437,936
Kreshmoi	1,500	28,454	32,158
IBECS	72,168	13,575,418	15,014,299
SciELO	741,407	17,464,256	19,305,165
MedLine	140,479	1,649,869	1,846,374
MSD Manuals	241,336	3,719,933	4,467,906
EMEA	366,769	5,327,963	6,008,543
Portal Clinic	8,797	159,717	169,294
Glossary MeSpEn	125,645	-	-
ICD10-en-es	5,202	-	-
SnowMedCT Denom.	887,492	-	-1
SnowMedCT Def.	4,268	177,861	184,574
Total	**4,430,765**	**66,147,518**	**74,663,550**

Table 2: Size of the corpora and glossaries used to create the corpus to train the MT systems

results of recent research, which has shown certain automatic metrics tend to underestimate NMT systems (Shterionov et al., 2018; Alvarez et al., 2019).

Additionally, we conducted a manual evaluation of a 30-segment sample for the three MT outputs employing monolingual direct assessment (DA) of translation adequacy (Graham and Baldwin, 2014; Graham and Liu, 2016). We used this DA setup because it simplifies the task of translation assessment (usually done as a bilingual task) into a simpler monolingual assessment task. We obtained the results averaging the assessment of two annotators and the NMT systems received higher marks.

As it can be seen in Table 1, DA classified Moses as the worst rated. Therefore, we decided to include only the two NMT systems for the post-editing tasks.

4.2 Corpora

To train the system we have used several publicly available corpora in the English-Spanish pair:

- Biomedical translation repository (BMTR)[5]
- Medline abstracts training data provided by Biomedical Translation Task 2019[6]
- The UFAL Medical Corpus[7] v1.0.
- The Khresmoi development data[8]
- The IBECS[9] (*Spanish Bibliographical Index in Health Sciences*) corpus.
- The SciELO corpus[10]
- The EMEA[11] (*European Medicines Agency*) corpus.

We have also created several corpora from websites with medical content:

- Medline Plus[12]: we have compiled our own corpus from the web and we have combined this with the corpus compiled in MeSpEn.
- MSD Manuals[13] English-Spanish corpus, compiled for this project under permission of the copyright holders.
- Portal Clínic[14] English-Spanish corpus, compiled by us for this project.

We have also used several glossaries and glossary-like databases treating them as corpora. These resources contain a lot of useful terms and expressions in the medical domain. Namely, we have used the English-Spanish glossary from MeSpEn, the 10th revision of the International Statistical Classification of ICD and SnowMedCT. With all the corpora and glossaries we have created an in-domain training corpus of 4,430,765 segments and entries.

[5] https://github.com/biomedical-translation-corpora/corpora
[6] http://www.statmt.org/wmt19/biomedical-translation-task.html
[7] https://ufal.mff.cuni.cz/ufal_medical_corpus
[8] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2122
[9] http://ibecs.isciii.es
[10] https://sites.google.com/view/felipe-soares/datasets
[11] http://opus.nlpl.eu/EMEA.php
[12] https://medlineplus.gov/
[13] https://www.msdmanuals.com/
[14] https://portal.hospitalclinic.org

	T1 (S2S)		T2 (S2S)		T3 (T)		T4 (T)	
	mean	st. dev.	mean	st. dev.	mean	st. dev.	mean	st. dev.
HTER	0.16	0.12	0.11	0.09	0.17	0.17	0.12	0.17
HBLEU	0.53	0.27	0.65	0.27	0.56	0.29	0.67	0.33
HEd	1.28	1.19	0.84	0.94	1.56	2.04	1.09	2.07
Keys/tok	6.36	28.25	3.38	5.25	7.53	27.62	5.91	25.59
PETpT	9.19	33.97	4.61	8.56	4.57	12.22	3.03	8.69

Table 3: PE-based metrics (mean and standard deviation) for the task

	S2S NMT		Transf. NMT	
	mean	st. dev.	mean	st. dev.
HTER	0.13	0.10	**0.11**	0.09
HBLEU	0.59	0.27	**0.65**	0.27
HEd	1.06	1.06	**0.84**	0.94
Keys/tok	4.87	16.75	**3.38**	5.25
PETpT	6.90	21.26	**4.61**	8.56

Table 4: Total PE-based metrics for each NMT model

In Table 2 the size of all corpora and glossaries used for training the MT systems is shown. Figures are calculated eliminating all the repeated source segment-target segment pairs in the corpora.

5 Experiment

We used the two NMT systems (transformer and s2s) trained with the corpora described above to translate from English into Spanish three texts (1468, 631 and 2247 words respectively) from the medical domain.

Four professional translators with at least one year of post-editing experience carried out the task: two of them post-edited the s2s output (T1 and T2) and the other two, the transformer output (T3 and T4). They were asked to produce publishable quality translations. As we wanted to reduce the external variables as much as possible, they all used PosEdiOn[15], a computer-assisted translation tool specifically designed for assessing post-editing effort, which logs both post-editing time and edits (keystrokes, insertions and deletions, that is, technical effort). The main characteristics of the post-editing tool were also explained to them before starting the task.

In order to avoid any bias, translators never post-edited the same text twice. However, they were told that an NMT system was used to produce the output. They received previous information on the tool and a three day period to test it before doing the task. They were paid their usual rate and had a two-week deadline. Two of them expressed concerns about the tool, as they preferred to work with their usual tools. However, they did not think it would affect the final quality of their job or their usual working speed. While post-editing, they could search for all the required information in order to produce the final translation. They could also pause the post-editing task whenever they wanted.

[15]https://sourceforge.net/projects/posedion/

6 Results

6.1 PE effort indicators

Once translators finished post-editing, we calculated the following task-specific (PE based) metrics (showed in Table 3):

- **PETpT**, PE time in seconds normalised by the length of the target segment in tokens.
- **HTER**, the TER value comparing the raw MT output with the post-edited segment.
- **HBLEU**, the BLEU score obtained by comparing the raw MT output with the post-edited segment.
- **HEd**, an edit distance value (Levenshtein distance) calculated comparing the raw MT output with the post-edited segment.
- **Keystrokes** normalized by the number of tokens.

Post-editor	Unmodified seg.
T1 (S2S)	22
T2 (S2S)	31
T3 (T)	19
T4 (T)	58

Table 5: Unmodified segments after post-editing

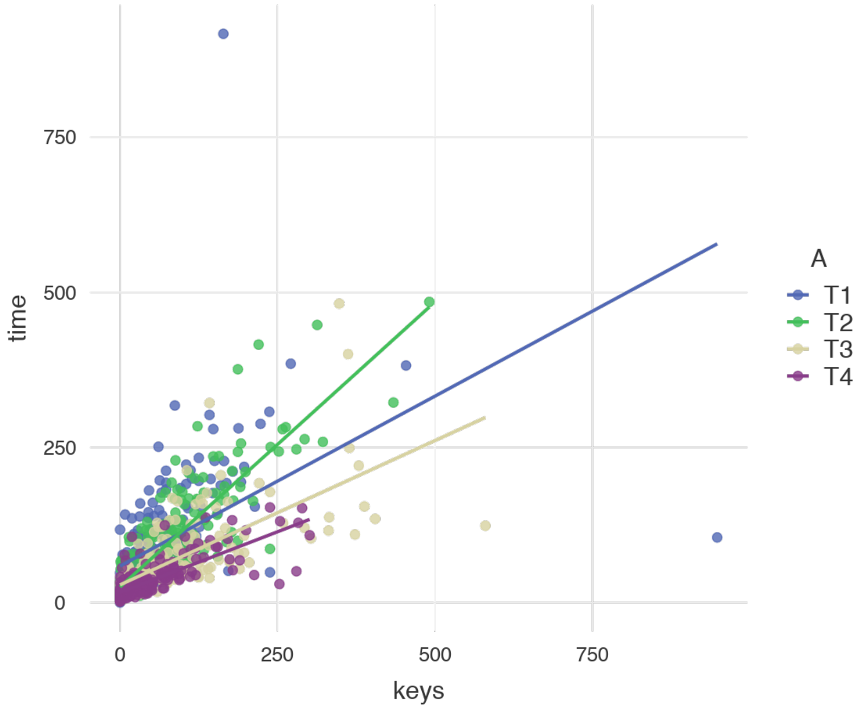

Figure 1: Scatter plot of keystrokes and time for all of the translators

In order to avoid the maximum number of outliers, we did not include those segments in which (normalized) time or (normalized) keystrokes doubled the mean plus the standard deviation of the total time or number of keystrokes. As usually happens in these types of tasks, post-editing effort indicators show a considerable variation among different translators. For the seg2seg model, translators showed a difference of 4.58 PETpT between them. This difference was reduced to 1.54 in the case of the transformer model. However, if we check the total figures for each of the systems (see Table 4), post-editing time is clearly reduced for the transformer model, as well as all the other scores.

We also used the distribution-agnostic Kolmogorov–Smirnov test to compare the distribution of PETpT for the two translators of each NMT model. We found there was no clear distribution (considering $p < 0.05$). This would seem to indicate the need to increase the number of translators for any given post-editing test to obtain a more representative mean.

Another interesting figure to understand PE effort is the number of unmodified segments. Even though that does not mean those segments imply no PE effort, it could give an indication of MT output. Table 5 shows the number of unmodified segments per translators from a total of 224 segments. There is not a clear tendency for any MT system, but rather a preference corresponding to the individual translator, especially T4, who didn't modify a high number of segments, which correlates to the low PE time recorded.

We also checked PETpT related to segment length, as research has shown longer segments tend to imply higher PE effort (Bentivogli et al., 2016). We studied segments with more than 35 tokens to see if PETpT or any other PE effort indicator increased. We could find no statistically sig-

	T1 (S2S)	T2 (S2S)	T3 (T)	T4 (T)	ALL
HTER	0.309*	0.545*	0.418*	0.00705*	0.49*
HBLEU	-0.072	-0.209	-0.148	-0.370*	-0.21*
HEd	0.043*	0.706	0.0770*	0.809*	0.66
Keys	**0.823***	**0.868***	**0.824***	**0.822***	**0.82***

Table 6: Spearman's correlation with time as a gold standard for different effort indicators (*p<0.001)

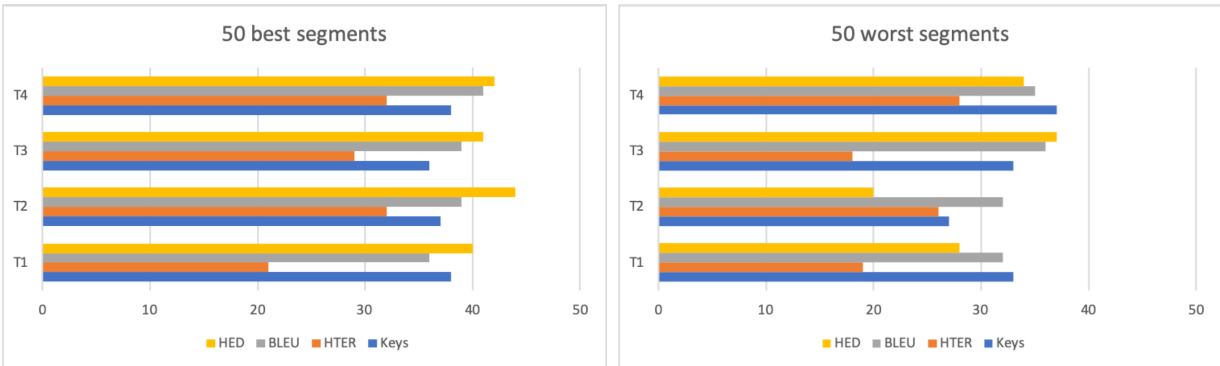

Figure 2: Correlation for best and worst segments

nificant evidence linking segment length to translators' effort in our experiments. This could indicate newer NMT models do not always reduce MT quality in longer segments.

Our results with a limited number of translators confirm previous studies (Castilho et al., 2017; Shterionov et al., 2018; Alvarez et al., 2019) and further, more extensive experimentation is needed in order to obtain meaningful indicators of MT output quality.

6.2 Correlation between scores

Once we established the overall results per each model, we tried to identify which metric produced scores that were closest to the total time spent per segment. We calculated Spearman's correlation coefficient between the total amount of time and all other metrics.

As can be seen in Table 6, the best overall correlation is found with the number of keys (see Figure 1) for all translators as well as for the total, followed by the calculated edit distance. Most of the results obtained show a statistically significant correlation, especially those figures relating to the number of keystrokes (*p<0.001).

These results are in line with the conclusions reported by previous work (Graham et al., 2016; Scarton et al., 2019) that found no clear correlation between temporal effort and the most frequent metrics, even though the number of keystrokes was the metric more closely related.

6.3 Tails distribution

There was a lack of correlation between the distribution of PE time among translators, and between this indicator and the others. We wanted to take a closer look at the best and worse segments to analyse if the correlation improved. We counted the number of common segments between the 50 best and worst time segments and all other metrics calculated.

As can be seen in Figure 2, there is a better correlation for the segments in which less time was spent. Furthermore, the edit distance shows the best correlation in these cases. For the segments with the higher time recorded, correlation is notably reduced in all cases and the edit distance and the number of keystrokes show a higher correlation.

7 Concluding remarks

There is a need for reliable metrics to evaluate MT quality in order to produce outputs which translators can post-edit without too much effort. Our experiments have shown that no single PE indicator can provide all the information necessary to assess the quality of the MT output. PE time provides a useful measure, even though it does not always correspond with other PE metrics and includes a great variation among translators. The only score that seems directly related to temporal effort are keystrokes (technical effort), but not

HTER or HBLEU.

In industrial scenarios, the quality of a certain MT output is usually linked to PE time. The results of our experiments suggest that the analysis of temporal effort can indicate the quality of the MT output, but we believe a multidimensional approach that includes different effort indicators would be a safer path to assess to convenience of post-editing a certain MT output.

Our future work will study further indicators of MT quality for post-editing in depth, mainly the characterization of source text to assess PE effort.

Acknowledgements: This work was supported by Universitat Pompeu Fabra (grant COMPLEMENTA 2019).

The training of the neural machine translation systems has been possible thanks to the NVIDIA GPU grant programme.

References

Allen, Jeffrey H. 2003. Post-editing. In Sommer, Harold, editor, *Computers and Translation: A translator's guide*, pages 297–317. John Benjamin, Amsterdam.

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 Volume 2: Translator, Project and User Tracks*, pages 49–56, Dublin, Ireland, August. European Association for Machine Translation.

Aranberri, Nora, Gorka Labaka, Arantza Ilarraza, and Kepa Sarasola. 2014. Comparison of Post-Editing Productivity between Professional Translators and Lay Users. In *Proceedings of the Third Workshop on Post-Editing Technology and Practice (WPTP - 3)*, Vancouver, Canada.

Aziz, Wilker, Sheila C. M. De Sousa, and Lucia Specia. 2012. PET: A Tool for Post-editing and Assessing Machine Translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3982–3987.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In Bengio, Yoshua and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Barrault, Loïc, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, August. Association for Computational Linguistics.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267. Association for Computational Linguistics.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Yota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Miceli Barone, and Maria Gialama. 2017. A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators. In *Proceedings of MT Summit XVI, vol.1: Research Track*, pages 116–131, 9.

Chinchor, Nancy. 1992. MUC-4 Evaluation Metrics. In *Proceedings of the Fourth Message Understanding Conference*, pages 22–29.

Cho, Kyunghyun, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, Doha, Qatar. Association for Computational Linguistics.

Costa-Jussà, Marta R. 2017. Why Catalan-Spanish Neural Machine Translation? Analysis, Comparison and Combination with Standard Rule and Phrase-based Technologies. *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 55–62.

Federico, M., A. Catttelan, and M. Trombetti. 2012. Measuring User Productivity in Machine Translation Enhanced Computer Assisted Translation. In *Proceedings of the 10th Conference of the AMTA*, pages 44–56. AMTA.

Graham, Yvette and Timothy Baldwin. 2014. Testing for Significance of Increased Correlation with Human Judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176, Doha, Qatar, October. Association for Computational Linguistics.

Graham, Yvette and Qun Liu. 2016. Achieving Accurate Conclusions in Evaluation of Automatic Machine Translation Metrics. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, CA. 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.

Green, Spence, Jeffrey Heer, and Christopher D. Manning. 2013. The Efficacy of Human Post-editing for Language Translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI 13*. ACM Press.

Guerberof, Ana. 2009. Productivity and Quality in MT Post-editing. In *Proceedings of MT Summit XII*, pages 8–14. Association of Machine Translation.

Isabelle, Pierre, Colin Cherry, and George F. Foster. 2017. A Challenge Set Approach to Evaluating Machine Translation. *CoRR*, abs/1704.07431.

Junczys-Dowmunt, Marcin, Tomasz Dwojak, and Hieu Hoang. 2016. Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. *CoRR*, abs/1610.01108.

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++. *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, July.

Koponen, Maarit, Leena Salmi, and Markku Nikulin. 2019. A Product and Process Analysis of Post-editor Corrections on Neural, Statistical and Rule-based Machine Translation Output. *Machine Translation*, (33, pages 61–90).

Koponen, Maarit. 2012. Comparing Human Perceptions of Post-editing Effort with Post-editing Operations. *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 181–190.

Koponen, Maarit. 2016. Is Machine Translation Post-editing Worth the Effort? A Survey of Research into Post-editing and Effort. *The Journal of Specialised Translation*, pages 131–148.

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

Papineni, Kishore, Salim Roukos, Todd Ward, and Wj Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. Number July, pages 311–318.

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. *Proceedings of the ACL 2015 Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, 1(2010):40–45.

Plitt, Mirko and François Masselot. 2010. A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics NUMBER*, 93:7–16.

Scarton, Carolina, Mikel L. Forcada, Miquel Esplà-Gomis, and Lucia Specia. 2019. Estimating Post-editing Effort: A Study on Human Judgements, Task-based and Reference-based Metrics of MT Quality. In *Proceedings of IWSLT 2019*, volume abs/1910.06204, Hong Kong, China.

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, Septembre.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of Association for Machine Translation in the Americas*, (August):223–231.

Specia, Lucia and Atefeh Farzindar. 2010. Estimating Machine Translation Post-editing Effort with HTER. *AMTA-2010 Workshop Bringing MT to the User: MT Research and the Translation Industry*, page 33–41.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In Ghahramani, Z., M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Toral, Antonio and Víctor M. Sánchez-Cartagena. 2017. A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, Long Papers, pages 1063–1073, East Stroudsburg. Association for Computational Linguistics (ACL).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS)*.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv*, abs/1609.08144.