Estimating post-editing effort: a study on human judgements, task-based and reference-based metrics of MT quality

Carolina Scarton[1], Mikel L. Forcada[2], Miquel Esplà-Gomis[2], Lucia Specia[1,3]

[1] Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
[2] Dept. Llenguatges i Sist. Inform., Universitat d'Alacant, 03690 St. Vicent del Raspeig, Spain
[3] Department of Computing, Imperial College London, London SW7 2AZ, UK.

c.scarton@sheffield.ac.uk, mlf@ua.es, mespla@dlsi.ua.es, l.specia@imperial.ac.uk

Abstract

Devising metrics to assess translation quality has always been at the core of machine translation (MT) research. Traditional automatic reference-based metrics, such as BLEU, have shown correlations with human judgements of adequacy and fluency and have been paramount for the advancement of MT system development. Crowd-sourcing has popularised and enabled the scalability of metrics based on human judgments, such as subjective *direct assessments* (DA) of adequacy, that are believed to be more reliable than reference-based automatic metrics. Finally, task-based measurements, such as post-editing time, are expected to provide a more detailed evaluation of the usefulness of translations for a specific task. Therefore, while DA averages adequacy *judgements* to obtain an appraisal of (perceived) quality independently of the task, and reference-based automatic metrics try to objectively estimate quality also in a task-independent way, task-based metrics are *measurements* obtained either during or after performing a specific task. In this paper we argue that, although expensive, task-based measurements are the most reliable when estimating MT quality in a specific task; in our case, this task is post-editing. To that end, we report experiments on a dataset with newly-collected post-editing indicators and show their usefulness when estimating post-editing effort. Our results show that task-based metrics comparing machine-translated and post-edited versions are the best at tracking post-editing effort, as expected. These metrics are followed by DA, and then by metrics comparing the machine-translated version and independent references. We suggest that MT practitioners should be aware of these differences and acknowledge their implications when deciding how to evaluate MT for post-editing purposes.

1. Introduction

Assessing the quality of the output of machine translation (MT) systems has been a widely explored topic in the last two decades. As with other applications outputting language (e.g. text summarisation), quality assessment of MT is challenging and highly dependent on the purpose of the translation. Therefore, the quality of machine translated (MT'ed) texts may depend on their usage. Table 1 shows two machine translations into English, and their respective post-edited versions. In Example 1, the MT'ed version has a different meaning from that of the original sentence: readers may be led to believe that *the product* is good and they should buy it, whilst the correct recommendation is against buying the product. Although this sentence would be problematic for an end user, it is rather simple to correct by a post-editor (only one word needs to be added). Example 2, on the other hand, shows a sentence where an end user can understand the MT'ed version with little effort, even though it contains multiple errors. A post-editor, however, would need to perform at least five word-level edit operations in order to transform the machine translation into the post-edited version.

These examples illustrate how sensitive MT evaluation is to purpose. Nirenburg [1] argues that MT can be classified into two groups according to its purpose: dissemination or assimilation. MT for dissemination is expected to be either ready as is or adequate for post-editing, since the purpose is publication. In contrast, MT for assimilation has the purpose of communication: the MT'ed text does not need to be grammatically correct as long as the reader can understand its message. In this paper we will focus on MT for dissemination; more specifically, when MT is used for post-editing (PE). PE is the task of editing MT'ed texts, a common practice among translation providers, where the aim is to improve productivity and, consequently, reduce translation costs. However, when MT'ed sentences contain too many problems, it may be easier to translate from scratch than to post-edit MT (this is indeed often reported by translators [2]).

Some recent work has claimed that state-of-the-art MT can be used for dissemination without supervision, with a well-known example claiming to have achieved *human parity* for Chinese-to-English translation [3]. However, in-depth revisions of this work [4, 5] suggest that we are still far from achieving human performance and, therefore, PE will still remain a key task in the translation industry. Therefore, finding ways of estimating the quality of MT'ed texts in terms of PE effort is a highly desirable feature (it would, for example, al-

	Example 1	Example 2
Reference	Do **not** buy this product, it's their craziest invention!	The **battery** **lasts** **6** **hours** and it can be **fully** **recharged** in **30** **minutes**.
MT	Do buy this product, it's their craziest invention!	**Six-hours** **battery**, **30** **minutes** to **full** **charge** **last**.

Table 1: Examples of MT'ed sentences and their PE'ed versions

low accurate budgeting of a translation job) and it is also a relevant topic for research in MT.

According to Krings [6], PE effort has three dimensions: temporal, cognitive and technical. The temporal dimension is the one most easily related to professional productivity or throughput: one just has to directly measure the time spent by the post-editor in transforming the MT output into an adequate PE'ed version. PE time for a sentence may be expected to increase roughly linearly with the total number of words in a sentence; therefore, PE time is normalised by dividing it by the number of words in the MT'ed sentence. The measured ratio between PE time and the number of words in the segment (PETpW) can be directly used to assess the effort of post-editing a segment. The main drawback of extracting PE time is that it is relatively more expensive and requires post-editors to avoid breaks during the editing of a given sentence.

Previous work has proposed several ways to address this issue. For example, the shared task on quality estimation (QE) of MT organised yearly as part of WMT conferences started with the purpose of training models to predict perceived PE effort, moving later to predicting more accurate measurements such as actual PE time and the *translation edit rate* (TER) [7] observed when comparing a MT'ed sentence and its post-edited (PE'ed) counterparts, called *human-targeted TER* or HTER [8]. HTER gives an indirect indication of the effort needed to transform a MT'ed sentence into its PE'ed version.

Despite its popularity, HTER has been subject to criticism: Graham et al. [9] criticise this metric and contend that subjective direct assessments (DA) of adequacy are more reliable than HTER measurements [10, 11]. They define adequacy as the degree to which the MT'ed segment expresses the meaning of the reference segment in the target language. Adequacy is therefore assessed in the target language, monolingually. Their DA is a combination of many independent human judgements of adequacy for a given sentence (in a 0%–100% scale) into a single score —standardised to zero mean and unit standard deviation after low-quality assessments are filtered out.

As discussed above, MT should be evaluated according to its purpose. Nevertheless, previous work has disregarded this assumption by using DA as gold standard for tasks where PE effort is the aspect of quality to be assessed [9, 12, 13]. Although we agree that DA may be a useful and reasonably cheap way of assessing subjectively the adequacy of MT output, in this paper we provide an in-depth analysis of ways to assess PE effort, with a focus on reducing PE time, a highly desirable feature by the translation industry.

We propose to assess the usefulness of metrics according to their ability to *rank* translations based on the time that would be required to post-edit them. This has a very practical application in the translation industry, where knowing which segments are easier to post-edit and which are the most difficult would allow a project manager to select post-editors accordingly, perhaps sending segments estimated as "easier" to less experienced (or cheaper) translators and/or sending the "most difficult" segments to experienced translators or to be translated from scratch.

Our main contributions are:

- a comprehensive review of task-specific (PE-based) metrics (e.g. HTER), reference-based metrics (e.g. TER) and DA, where the goal is to rank MT'ed segments according to PE time;
- the release of a dataset with source, MT'ed, reference, and PE'ed texts; detailed information about five independent post-editing jobs for each MT'ed text, and DA annotations;[1]
- a new ranking score for MT'ed segments called *split-averaged, time-ratio assessment* (SATRA).

In Section 2 we present the dataset created and used for this paper. Section 3 presents our ranking analysis using all evaluation metrics available in our dataset. In section 4 we discuss related work. The paper ends with concluding remarks (Section 5).

2. Dataset and annotators

We extend the dataset made available by the WMT 2016 shared task on document-level quality estimation [14]. This dataset contains 1,047 segments totalling 26,875 words, MT'ed by 41 different systems —with an average of 26 segments per system— extracted from the test sets of English–Spanish WMT translation shared tasks between 2008 and 2012. Existing MT'ed segments were crowd-annotated (via Amazon Mechanical Turk) using DA scores made available by Graham et al. [12].[2]

Although the aim of the work presented in [12] was to generate DA scores at the document level, they first assessed each segment independently. Each segment has then a DA score and these are the values used in our experiments. The DA value of each segment is obtained by averaging the assessment of various annotators (previous work recommend at

[1] https://github.com/carolscarton/iwslt2019
[2] https://github.com/ygraham/eacl2017

	ANN0		ANN1		ANN2		ANN3		ANN4		ALL	
	Mean	st. dev.	Mean	st. dev.	Mean	st. dev.	Mean	st. dev	Mean	st. dev	Mean	st. dev
HTER	0.32	0.17	0.27	0.25	0.25	0.16	0.30	0.18	0.30	0.18	0.29	0.19
HBLEU	0.49	0.21	0.60	0.26	0.57	0.21	0.52	0.22	0.53	0.21	0.54	0.23
HMETEOR	0.65	0.16	0.72	0.25	0.72	0.16	0.67	0.17	0.68	0.16	0.69	0.19
Keys/char	0.43	0.33	0.44	0.40	0.46	0.41	0.55	0.42	0.42	0.35	0.46	0.39
PETpW (sec/word)	3.88	2.91	2.42	2.78	3.66	2.71	3.58	2.31	4.23	4.22	4.23	4.22

Table 2: Statistics (mean and standard deviation) of task-specific (PE-based) metrics in our dataset

least 15 annotations per segment [15], and the study in [12] follows the same protocol).

For the post-editing task, we hired five professional translators with experience in PE (hereafter referred to as ANN0, ANN1, ANN2, ANN3 and ANN4), who generated PE'ed versions for each segment of this dataset. The annotators used the PET tool [16], which records the edit operations performed during the PE task, including the time elapsed to post-edit a segment. We use PE time normalised by the length of the target segment in words, PETpW).

We then calculated the following task-specific (PE-based) metrics:

- HTER, HBLEU and HMETEOR, respectively the TER, BLEU [17], and METEOR [18] scores of the MT'ed segment using the PE'ed version as reference,[3] and
- Keys/char: ratio between the number of keys pressed by an annotator and the number of characters in the MT'ed segment.

Since we have access to the references from the WMT datasets, we also calculated standard reference-based BLEU, METEOR and TER scores.

Table 2 shows some statistics of the task-specific (PE-based) metrics extracted for this dataset. We show statistics per annotator and also the averaged values for all annotators (ALL). Statistics for DA, and reference-based metrics are shown in Table 3. All averages in both tables are weighted by the number of MT'ed words, as post-editing time —the measurement we want to track— is expected to grow linearly with sentence length. As may be seen, the values of quality indicators show a rather wide range.

	Mean	st. dev.
DA	-0.02	0.61
TER	0.57	0.21
BLEU	0.24	0.16
METEOR	0.42	0.16

Table 3: Statistics (mean and standard deviation) of DA and reference-based metrics in our dataset

3. Comparing the ranking ability of metrics

In order to analyse the performance of evaluation metrics as a proxy for PETpW, we propose experiments that look at how these metrics rank MT'ed segments. We argue that looking at the rankings gives a reliable perspective of the usefulness of the metrics for the PE task, since it gives us the relative differences, in terms of effort, among the segments to be PE'ed.

3.1. Ranking correlation

In this experiment, we try to identify which metric produces rankings that are closest to PETpW rankings. Firstly, we calculate Spearman's ρ rank correlation coefficient between PETpW and all metrics. In addition to ρ, we also compute a new ranking score called *split-averaged, time-ratio assessment* (SATRA) for a ranking R as follows:

$$\mathrm{SATRA}(R) = \frac{1}{N-1}\sum_{j=1}^{N-1}\frac{\tau_1^j(R)}{\tau_{j+1}^N(R)}$$

with

$$\tau_m^n(R) = \sum_{j=m}^{n} T(R_j)\left(\sum_{i=m}^{n} L(R_j)\right)^{-1},$$

the average measured PETpW for segments R_m to R_n (those ranked m-th to n-th), where $T(R_j)$ is the total PE time and $L(R_j)$ the total length in (MT'ed) words for segment R_j (ranked j-th). The value of $\mathrm{SATRA}(R)$ should be close to 1 for a random ranking (the average PETpW above any split of the ranking and that below the split should roughly be the same), smaller than 1 for a good ranking (one that would rank easier-to-post-edit segments better than hard-to-post-edit ones), and the minimum possible for a ranking based on the measured PETpW. These two scores are used to measure how close two ranked distributions are.[4]

Table 4 shows Spearman's ρ correlation coefficients and SATRA scores between all metrics and the individual PETpW of all annotators and the averaged values of all annotators (ALL). The last line of the table provides the scores obtained by an oracle using the actual PETpW as the ranking metric; this helps to interpret SATRA scores as, unlike Spearman's ρ, they do not have a fixed lower bound. DA shows moderate Spearman's ρ values across all annotators and for ALL, which are considerably smaller than those achieved by HTER, HBLEU, HMETEOR and Keys/char. SATRA shows similar results: DA presents larger (worse) values than the task-specific PE-based metrics. Following previous work [21], we calculate the statistical significance difference

[3] These metrics were calculated using the Asiya toolkit [19].

[4] SATRA is similar to DeltaAVG [20] but has a simpler interpretation in terms of the average PE time per word above and below any split of the rank.

	ANN0		ANN1		ANN2		ANN3		ANN4		ALL	
	ρ	S	ρ	S	ρ	S	ρ	S	ρ	S	ρ	S
TER	.24*	.78	.32*	.67	.26	.73	.23	.81	.20	.83	.30	.77
BLEU	.25*	.74	.33*	.64	.29	.70	.30*	.75	.23*	.77	.33	.72
METEOR	.25*	.74	.34	.63	.31	.67	.30*	.76	.23*	75	.35	.71
DA	.38	.68	.48	.59	.44	.66	.45	.70	.43	.62	.52	.64
HTER	.58	.53	.62	.47	.71	.47	.67	.54	.61	.49	.69	.53
HBLEU	.54*	.54	.60*	.49	.67	.48	**.68**	.54	.58*	.50	.68	.53
HMETEOR	.53*	.55	.61*	.48	.69	.47	.65	.54	.59*	.50	.68	.54
Keys/char	**.63**	**.48**	**.75**	**.37**	**.74**	**.45**	**.68**	**.52**	**.63**	**.43**	**.76**	**.49**
PETpW	1.0	.31	1.0	.25	1.0	.32	1.0	.38	1.0	.26	1.0	.39

Table 4: Spearman's ρ (↑) and SATRA (S, ↓) scores for all metrics using PETpW as gold standard. The best results are shown in bold and * means no statistically significant difference between the metrics according to Williams test with $p < 0.01$.

	ANN0		ANN1		ANN2		ANN3		ANN4	
	ρ	S	ρ	S	ρ	S	ρ	S	ρ	S
DA	.52	.63	.51	.65	.51	.64	**.61**	.64	.52	.65
HTER	**.59**	.59	.45	.71	**.60**	**.58**	.57	.59	**.62**	**.57**
HBLEU	.57	.59	.45	.73	.57	.59	.56	.60	.60	.58
HMETEOR	.57	.59	.42	.72	.58	.59	.55	.60	.60	.58
Keys/char	**.59**	**.58**	**.54**	**.62**	.57	.60	.59	**.58**	.60	.58
PETpW	.58	.53	.62	.57	.61	.57	.62	.55	.63	.55

Table 5: Spearman's ρ (↑) and SATRA (↓) scores for all metrics using PETpW as gold standard for the leave-one-out experiment

between all metrics using Williams' test over the Spearman's ρ scores ($p < 0.01$).[5] The large majority of the results are statistically different.

The best overall metric is Keys/char, which achieves the highest Spearman's ρ scores and the lowest SATRA scores for all annotators individually and for ALL. The only annotator where the Keys/char metric is not so salient is ANN3. Our hypothesis is that this annotator may have interacted more with the mouse,[6] instead of with the keyboard. HTER, HBLEU and HMETEOR do not show significant differences among them. This is in line with the results reported by previous work [9] that found no difference between these metrics when correlating them to DA. Finally, independent-reference-based metrics show the worst ranking scores with respect to PETpW.

In a real-world scenario, the PETpW of one annotator could be estimated based on the PETpW of other annotator(s). In order to simulate this case and evaluate whether the results from Table 4 would still stand, we performed leave-one-out experiments. In this case, SATRA and Spearman's ρ scores are calculated between each one of the studied metrics for one annotator and the averaged PETpW of all other annotators. For example, for ANN0, the PETpW is the average PETpW of ANN1 to ANN4, and its correlation with DA and the HTER, HBLEU, HMETEOR, Keys/char and PETpW for ANN0 post-edits. Table 5 shows the results of this experiment. As expected, the difference between Spearman's ρ and SATRA scores for HTER, HBLEU and HMETEOR and for Keys/char is lower than in Table 4, since we are not dealing with the individual PETpW of each annotator. SATRA scores for PE-based metrics are better (lower) than DA (except for ANN1), and similarly for Keys/char (except for ANN1). In addition, Keys/char is not the best metric overall anymore, although it still shows the best SATRA in three out of five cases. In general, PE-based approaches still outperform DA in most cases. For reference, the last row of Table 5 shows Spearman's ρ and SATRA for the PETpW of each annotator versus the leave-one-out PETpW.

It is worth mentioning that, with only five annotators, it is difficult to devise a model that would be a good estimator of quality for new annotators. In fact, after doing an analysis using the distribution-agnostic Kolmogorov–Smirnov test[7] over the PETpW distributions (considering $p < 0.05$), we identified three clusters of annotators where their PETpW measurements come from the same distribution. Basically, ANN0, ANN2 and ANN4 could be clustered together, whilst ANN1 and ANN3 would have their own clusters. This may be impacting our results, but a deeper analysis of the effect of such clusters is left for future work.

3.2. Analysis of tails

The experiments in this section aim to obtain a closer view of how the metrics studied perform for the best and the worst segments, by performing an analysis of the tails of the PETpW distribution. In other words, we want to analyse how the task-specific PE-based metrics, reference-based metrics,

[5] Williams test is calculated using `mt-qe-eval`: `https://github.com/ygraham/mt-qe-eval`.

[6] PET records keyboard actions, but not mouse actions.

[7] `https://en.wikipedia.org/wiki/Kolmogorov-Smirnov_test`

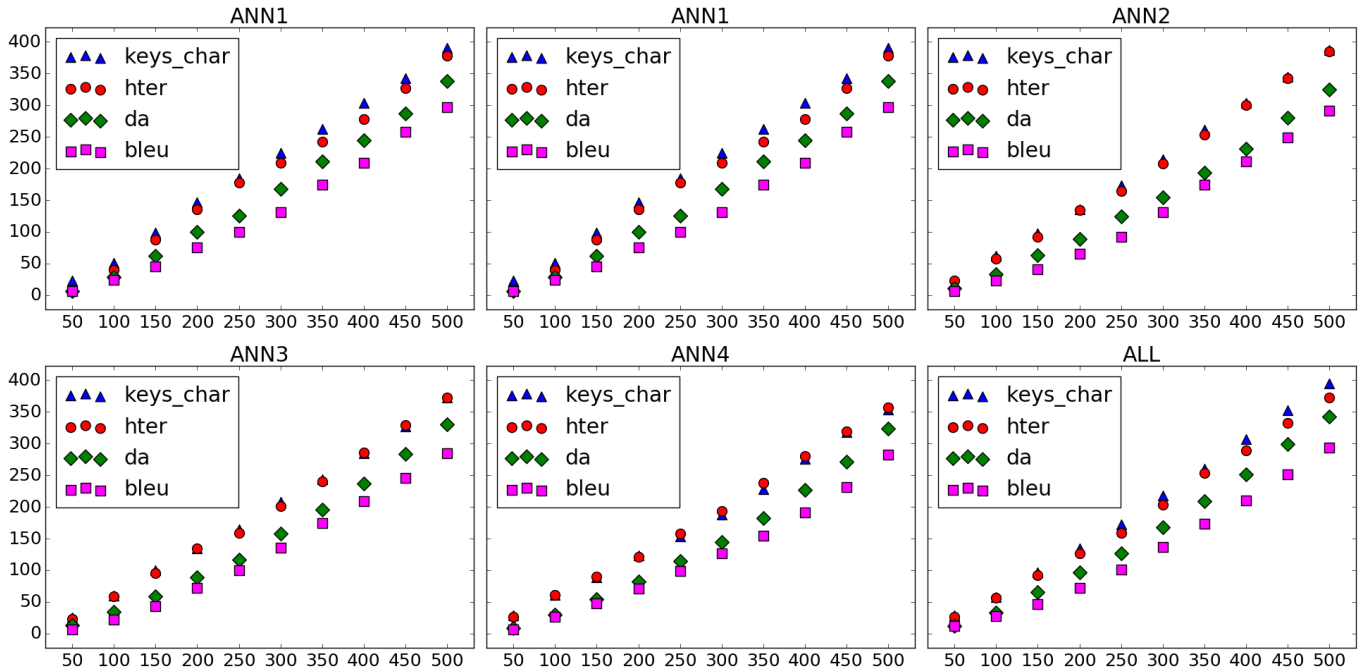

Figure 1: Number of segments shared between the 500 best sentences acording to PETpW and the other metrics

and DA perform on the best and worst segments according to PETpW. Our experiment consists in counting the number of common segments between different cuts of the PETpW ranking and the rankings obtained with each different metric.

Best segments: firstly, we look at the first 500 sentences in the PETpW ranking, that is, the 500 easiest-to-post-edit sentences, and compare to the first 500 sentences in the rankings according to other metrics. We split the rankings in sets of 50 to show the performance of the metrics and the differences among them. Figure 1 shows the results of this experiment for all annotators individually and for the ALL case. For clarity we only show four metrics: Key/char, HTER, DA and BLEU. One can clearly identify three groups of metrics:

- BLEU, TER and METEOR rankings behave similarly and show the lowest number of segments in common to the PETpW ranking;
- HTER, HBLEU, HMETEOR, and Keys/char rankings are the best, sharing the largest number of segments with the PETpW ranking;
- DA ranking is better than the reference-based metrics, but worse than the task-specific PE-based metrics.

These findings are in agreement with those obtained when ranking all segments.

Figure 2 shows scatter plots for DA vs. PETpW and HTER vs. PETpW averaged over all five annotators. The top two graphs show the scatter plots for the entire dataset. In this case, both metrics look similar in comparison to PETpW, although DA seems to show more outliers. The bottom two graphs show the scatter plots for the best 500 segments according to PETpW. In this case, HTER shows a clear tendency, where the majority of the values have a low HTER

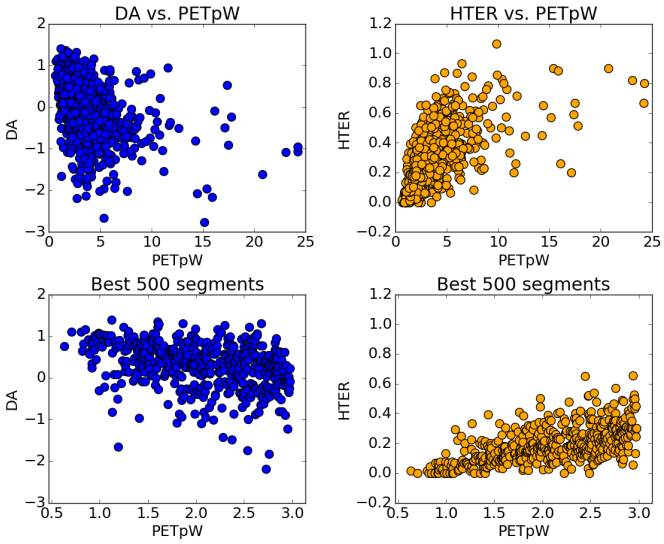

Figure 2: Scatter plots for DA vs. PETpW and HTER vs. PETpW

score and a low PETpW. DA, on the other hand, shows a much sparser graph.

Worst segments: a similar trend is shown when we analyse the 500 worst segments (due to space constraints, Figure 3 only shows results for ALL), although the gap between DA and task-specific PE-based metrics is smaller. One hypothesis is that, for the worst segments, where the quality is very low, differences in adequacy track differences in PE time better.

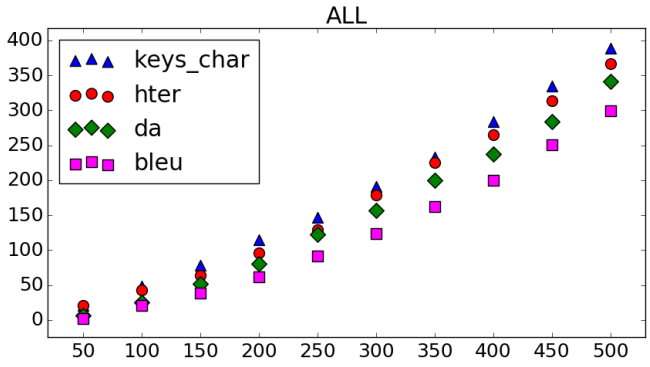

Figure 3: Number of segments shared between the 500 worst sentences according to PETpW and the other metrics for all translators.

4. Related work

In what follows we present previous work on human task-based evaluation that targeted PE effort and on the use of DA for the same purpose.

PE time is a straightforward indicator of MT quality: segments that take longer to be PE'ed are considered worse than segments that can be quickly corrected. Koponen et al. [22] argue that PE time is the most effective way of measuring cognitive aspects of the PE task and relate them to the quality of the translations. Plitt and Masselot [23] use PETpW (actually, its converse: words per hour) to measure the gain in productivity when post-editing MT'ed text —in a real translation workflow— over the productivity when performing translation from scratch.

Perceived PE effort: humans are asked to give a score for the MT'ed sentences according to a Likert [24] scale representing perceived PE effort [25]. This type of score can be given with or without actual post-editing and it represents a judgement on how difficult it would be (or it was) to fix the given MT'ed sentence. Perceived PE effort scores were used in the WMT 2012 [20] and WMT 2014 [26] QE shared task editions.

Eye-tracking: previous work have also relied on eye-tracking to evaluate PE effort. O'Brien [27] measures fixation time and correlates it with GTM (a similarity metric between the machine translation and the reference sentence based on precision, recall and F-measure [28]). Low GTM scores show correlation with high fixation time. PE pauses (extracted from keystroke logs) can also be viewed as an indirect measure of cognitive effort [29]. Long pauses are associated with segments that demand more cognitive PE effort.

Edit distance and n-gram-based scores: PE effort can also be evaluated indirectly, by using a metric that takes into account edit operations. HTER [8] is an example of such a metric, which computes the minimum number of edits to transform the machine translation into the PE'ed version. Task-specific, PE-based (*human* or *H-*) variants of commonly-used reference-based similarity measures have also been studied, such as HBLEU and HMETEOR. However, HTER is the most widely used as an indirect measurement of PE effort [14, 26, 30, 31, 32, 33]

DA: Graham et al. [10, 11] propose the use of DA for MT evaluation. According to the authors, the biggest advantage of their approach in comparison to early practices of adequacy judgements is that they can reliably crowd-source the annotations. Graham et al. [9] also express a strong criticism of HTER on the grounds that it does not show high Pearson r correlation scores with DA. In another work [12], the same authors also criticise a variant of HTER for document-level QE, suggesting that DA is a more adequate metric to compare different QE systems. Recently, Bentivogli et al. [13] evaluate HTER and mTER (multi-reference TER) against DA scores and conclude that mTER is a better proxy for PE effort because it shows higher correlation scores with DA than HTER. However, our analysis on a real-world measurement of productivity (PETpW) show that PE-based metrics (including HTER) are the most adequate metrics to approximate PETpW, outperforming DA. Therefore, we argue that if mTER and HTER were compared using their correlations to PETpW, the results could be different (this analysis is left for future work).

5. Concluding remarks

The advancement and adoption of MT depends more than ever on the availability of reliable metrics to evaluate its quality. Averaged subjective *direct assessment* (DA) of MT quality, which is performed independently of purpose and may easily be crowd-sourced, has become very popular. However, in an important application of MT, namely dissemination via post-editing, it is only natural to use actual *measurements* that are obtained after performing post-editing. It is also natural for quality estimation models to target such metrics.

The results of our experiments on a dataset that includes PE indicators collected for five translators show that DA *judgements* provide a reasonable approximation of relevant, measurable aspects of MT usefulness in a dissemination task, such as PE time; however, as expected, task-specific metrics comparing MT'ed and PE'ed text – such as HTER or the number of keystrokes per raw MT character – are better trackers of PETpW. DA does however perform better than metrics such as BLEU, TER or METEOR with respect to an independent reference translation.

These results lead us to recommend that MT practitioners should use task-specific metrics wherever this is possible, and non-expert subjective judgements such as DA only when specific, measurable metrics are not available or feasible for a task.

Acknowledgements: Work supported by the Spanish government through project EFFORTUNE (TIN2015-69632-R) and grant PRX16/00043 for MLF, and by the European Commission through project GoURMET (No. 825299). CS and MLF are both first authors with equal contribution.

6. References

[1] S. Nirenburg, *Progress in Machine Translation*. Amsterdam, Netherlands: IOS B. V., 1993.

[2] M. Sanchez-Torron and P. Koehn, "Machine translation quality and post-editor productivity," in *Proceedings of the Conference of the Association for Machine Translation in the Americas Vol. 1: MT Researchers' Track*, Austin, TX, 2016, pp. 16–26. [Online]. Available: https://amtaweb.org/wp-content/uploads/2016/10/AMTA2016_Research_Proceedings_v7.pdf

[3] H. Hassan, A. Aue, C. Chen, V. Chowdhary, J. Clark, C. Federmann, X. Huang, M. Junczys-Dowmunt, W. Lewis, M. Li, S. Liu, T. Liu, R. Luo, A. Menezes, T. Qin, F. Seide, X. Tan, F. Tian, L. Wu, S. Wu, Y. Xia, D. Zhang, Z. Zhang, and M. Zhou, "Achieving human parity on automatic chinese to english news translation," *Computing Research Repository*, vol. arXiv:1803.05567, 2018. [Online]. Available: http://arxiv.org/abs/1803.05567

[4] S. Läubli, R. Sennrich, and M. Volk, "Has machine translation achieved human parity? a case for document-level evaluation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018, pp. 4791–4796. [Online]. Available: http://aclweb.org/anthology/D18-1512

[5] A. Toral, S. Castilho, K. Hu, and A. Way, "Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation," in *Proceedings of the Third Conference on Machine Translation (WMT), Volume 1: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 113–123. [Online]. Available: http://www.statmt.org/wmt18/pdf/WMT012.pdf

[6] H. P. Krings, *Repairing texts: Empirical investigations of machine translation post-editing process*. Kent, OH: The Kent State University Press, 2001.

[7] M. Snover, N. Madnani, B. Dorr, and R. Schwartz, "TERp system description," in *Proceedings of the MetricsMATR workshop*, vol. 34, no. 67, 2008, p. 108.

[8] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," in *Proceedings of the Seventh biennial conference of the Association for Machine Translation in the Americas*, Cambridge, MA, 2006, pp. 223–231.

[9] Y. Graham, T. Baldwin, M. Dowling, M. Eskevich, T. Lynn, and L. Tounsi, "Is all that glitters in machine translation quality estimation really gold?" in *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan, 2016, pp. 3124–3134. [Online]. Available: http://aclweb.org/anthology/C16-1294

[10] Y. Graham, T. Baldwin, A. Moffat, and J. Zobel, "Is machine translation getting better over time?" in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, 2014, pp. 443–451. [Online]. Available: http://www.aclweb.org/anthology/E14-1047

[11] Y. Graham, A. Baldwin, Timothy Moffat, and J. Zobel, "Can machine translation systems be evaluated by the crowd alone," *Natural Language Engineering*, vol. 23, no. 1, pp. 3–30, 2016.

[12] Y. Graham, Q. Ma, T. Baldwin, Q. Liu, C. Parra, and C. Scarton, "Improving evaluation of document-level machine translation quality estimation," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, 2017, pp. 356–361. [Online]. Available: http://www.aclweb.org/anthology/E/E17/E17-2057.pdf

[13] L. Bentivogli, M. Cettolo, M. Federico, and C. Federmann, "Machine translation human evaluation: an investigation of evaluation based on post-editing and its relation with direct assessment," in *Proceedings of the 15th International Workshop on Spoken Language Translation*, Bruges, Belgium, 2018, pp. 62–69. [Online]. Available: https://workshop2018.iwslt.org/downloads/Proceedings_IWSLT_2018.pdf

[14] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, V. Logacheva, C. Monz, M. Negri, A. Neveol, M. Neves, M. Popel, M. Post, R. Rubino, C. Scarton, L. Specia, M. Turchi, K. Verspoor, and M. Zampieri, "Findings of the 2016 Conference on Statistical Machine Translation," in *Proceedings of the First Conference on Statistical Machine Translation*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 131–198. [Online]. Available: http://aclweb.org/anthology/W16-2301

[15] Y. Graham, T. Baldwin, and N. Mathur, "Accurate Evaluation of Segment-level Machine Translation Metrics," in *Proceedings of the Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*. Dever, CO: Association for Computational Linguistics, 2015, pp. 1183–119. [Online]. Available: https://www.aclweb.org/anthology/N15-1124

[16] W. Aziz, S. C. M. Sousa, and L. Specia, "PET: a tool for post-editing and assessing machine translation," in

Proceedings of the 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey, 2012, pp. 3982–3987. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2012/pdf/985_Paper.pdf

[17] K. Papineni, S. Roukos, T. Ward, and W. jing Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA: Association for Computational Linguistics, 2002, pp. 311–318. [Online]. Available: https://www.aclweb.org/anthology/P02-1040.pdf

[18] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*. Ann Harbor, MI: Association for Computation Linguistics, 2005, pp. 65–72. [Online]. Available: http://aclweb.org/anthology/W05-0909

[19] J. Giménez and L. Màrquez, "Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation," *The Prague Bulletin of Mathematical Linguistics*, no. 94, pp. 77–86, 2010.

[20] C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia, "Findings of the 2012 Workshop on Statistical Machine Translation," in *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Montréal, Canada: Association for Computational Linguistics, June 2012, pp. 10–51. [Online]. Available: http://www.aclweb.org/anthology/W12-3102

[21] Y. Graham, "Improving Evaluation of Machine Translation Quality Estimation," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Beijing, China: Association for Computational Linguistics, 2015, pp. 1804–1813. [Online]. Available: http://www.aclweb.org/anthology/P15-1174

[22] M. Koponen, W. Aziz, L. Ramos, and L. Specia, "Post-editing time as a measure of cognitive effort," in *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice*, San Diego, CA, 2012, pp. 11–20.

[23] M. Plitt and F. Masselot, "A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context," *The Prague Bulletin of Mathematical Linguistics*, vol. 93, pp. 7–16, 2010. [Online]. Available: https://ufal.mff.cuni.cz/pbml/93/art-plitt-masselot.pdf

[24] R. Likert, "A technique for the measurement of attitudes." *Archives of psychology*, vol. 140, pp. 1–55, 1932.

[25] L. Specia, N. Hajlaoui, C. Hallet, and W. Aziz, "Predicting machine translation adequacy," in *Proceedings of the Machine Translation Summit XIII*, Xiamen, China, September 2011, pp. 19–23.

[26] O. Bojar, C. Buck, C. Federman, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut, L. Specia, and A. Tamchyna, "Findings of the 2014 Workshop on Statistical Machine Translation," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, MD: Association for Computational Linguistics, June 2014, pp. 12–58. [Online]. Available: http://www.aclweb.org/anthology/W/W14/W14-3302

[27] S. O'Brien, "Towards predicting post-editing productivity," *Machine Translation*, vol. 25, pp. 197–215, 2011.

[28] J. P. Turian, L. Shen, and I. D. Melamed, "Evaluation of Machine Translation and its Evaluation," in *Proceedings of the Machine Translation Summit IX*, New Orleans, LA, 2003, pp. 386–393.

[29] I. Lacruz, M. Denkowski, and A. Lavie, "Cognitive Demand and Cognitive Effort in Post-Editing," in *Proceedings of the Third Workshop on Post-Editing Technology and Practice*, Vancouver, Canada, 2014, pp. 73–84.

[30] O. Bojar, C. Buck, C. Callison-Burch, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia, "Findings of the 2013 Workshop on Statistical Machine Translation," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 1–44. [Online]. Available: http://www.aclweb.org/anthology/W13-2201

[31] O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, M. Huck, C. Hokamp, P. Koehn, V. Logacheva, C. Monz, M. Negri, M. Post, C. Scarton, L. Specia, and M. Turchi, "Findings of the 2015 Workshop on Statistical Machine Translation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, 2015, pp. 1–46. [Online]. Available: http://aclweb.org/anthology/W15-3001.pdf

[32] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, S. Huang, M. Huck, P. Koehn, Q. Liu, V. Logacheva, C. Monz, M. Negri, M. Post, R. Rubino, L. Specia, and M. Turchi, "Findings of the 2017 Conference on Machine Translation

(WMT17)," in *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Copenhagen, Denmark: Association for Computational Linguistics, September 2017, pp. 169–214. [Online]. Available: http://www.aclweb.org/anthology/W17-4717

[33] L. Specia, F. Blain, V. Logacheva, R. Astudillo, and A. F. T. Martins, "Findings of the WMT 2018 shared task on quality estimation," in *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, October 2018, pp. 702–722. [Online]. Available: http://www.aclweb.org/anthology/W18-6452