Living on the edge: productivity gain thresholds in machine translation evaluation metrics

Carla Parra Escartín carla.parra@hermestrans.com
Manuel Arcedillo manuel.arcedillo@hermestrans.com
Hermes Traducciones
C/ Cólquide, 6, portal 2, 3.º I
28230 Las Rozas, Madrid, Spain

Abstract

This paper studies the minimum score at which machine translation (MT) evaluation metrics report productivity gains in a machine translation post-editing (MTPE) task. We ran an experiment involving 10 professional in-house translators from our company in which they were asked to carry out a real translation task involving MTPE, translation from scratch and fuzzy-match editing. We then analyzed the results and evaluated the MT output with traditional MT evaluation metrics such as BLEU and TER, as well as the standard used in the translation industry to analyze text similarity in translation memory (TM) matches: the fuzzy score. We report where the threshold between productivity gain and productivity loss lies and contrast it with past experiences in our company. We also compare the productivity of similar segments from MTPE and TM match editing samples in order to gain further insights on their cognitive effort and pricing schemes.

1 Introduction

Over the past few years, translators are experiencing the introduction of MT in their translation workflow. However, it is often difficult for the parties involved to assess if the MT output quality allows any productivity gain and, if applicable, justifies any rate discount. A popular method for evaluating MT output involves using automatic metrics such as BLEU (Papineni et al., 2001) and TER (Snover et al., 2006). However, their estimation may prove technically difficult for general users and their results may be obscure to interpret in terms of productivity due to lack of conclusive research or lack of familiarity with these metrics.

To overcome these challenges, alternative metrics have been proposed which aim at applying the familiarity of translation memory (TM) fuzzy match scores (FMS) to MTPE evaluation as a target-side FMS (Parra Escartín and Arcedillo, 2015b). In this paper, we analyze BLEU, TER and FMS values obtained in an experiment involving ten professional translators with the aim of finding out if a threshold for productivy gain can be found in these metrics. We also discuss how cognitive effort and confidence in MT or TM technologies may impact post-editors' productivity by comparing MTPE and TM match editing throughputs.

The remainder of this paper is structured as follows: Section 2 describes the experiment we carried out. Section 3 registers the productivity (cf. Subsection 3.1) and MT automatic evaluation metrics (cf. Subsection 3.2) obtained. In Section 4 we look at the threshold between productivity gain and productivity loss by correlating the results from Section 3. We then discuss the relation between productivity gains and cognitive effort in Section 5 before summarizing our research and discussing future work in Section 6.

2 Experiment settings

Ten in-house translators were asked to translate the same file from English into Spanish using memoQ,[1] one of their most common computer-assisted translation (CAT) tools. This tool was chosen because it allows to keep track of the time spent in each segment. Translators were only allowed to use the TM, terminology database and MT output included in the hand-off package. We disabled all other memoQ's productivity enhancing features, such as predictive text, sub-segment leverage and automatic fixing of fuzzy matches.[2]

2.1 Test set

The text to be translated had over 7,000 words and was part of a software user guide. It belongs to a real translation request, except we filtered out repetitions and internal leverage segments to avoid skewing due to the inferior effort required to translate the second of two similar segments. All traditional TM fuzzy match bands, exact TM matches and no-match segments were represented in the text. Table 1 shows the word count distribution reported by memoQ.

TM match	Words	Segments	Words/segment
100%	1226	94	13.04
95-99%	231	21	11.00
85-94%	1062	48	22.12
75-84%	696	42	16.57
No Match	3804	263	14.46
Total	**7019**	**468**	**14.99**

Table 1: Word count as computed by memoQ.

We randomly divided the no-match band in two halves using the test set generator included in the *m4loc* package.[3] One half was translated from scratch, without TM or MT suggestion, while the second half was sent to one of our custom MT engines.

2.2 Machine Translation system used

We used Systran[4] to generate the raw MT output. This is the customized rule-based MT engine we normally use with this client. It can be considered a mature engine, since at the time of the experiment it had been subject to ongoing in-house customization for over three years and boasted a consistent record for productivity enhancement. Its customization includes dictionary entries, software settings, and pre- and post-editing scripts. Although Systran includes a statistical component for automatic post-editing, this was not used in our experiment.

2.3 Translators participating in the experiment

Ten professional translators were engaged in the experiment. Two carried out the task as part of a pilot experiment (Parra Escartín and Arcedillo, 2015a) which studied the feasibility of using a target-side FMS as an alternative MT evaluation metric. Eight additional translators were subsequently engaged to perform the same task under the same conditions with the aim of verifying the initial findings and provide deeper insights (Parra Escartín and Arcedillo, 2015b).

The translators were asked to provide their years of experience in translation and MTPE, their experience working in texts from this client, their opinion on MT (positive or negative),

[1] The version used was memoQ 2015 build 3.

[2] For further details on the experiment settings described in this section, see Parra Escartín and Arcedillo (2015b).

[3] `https://code.google.com/p/m4loc`

[4] Systran 7 Premium Translator was used.

and their opinion on CAT tools (positive or negative). Table 2 summarizes the results of our survey. Translators are sorted in descending order according to their combined experience in translation and MTPE. Two translators expressed that they did not like working with MT, despite acknowledging that high quality MT output generally enhances their productivity. This negative bias, however, did not seem to affect the results.

	Trans. exp.	**MTPE exp.**	**Client exp.**	**MT opinion**	**CAT opinion**
TR-1	5	3	Yes	Positive	Positive
TR-2	5	3	Yes	Positive	Positive
TR-3	5.5	2	Yes	Positive	Positive
TR-4	5	1	Yes	**Negative**	Positive
TR-5	5	0	No	Positive	Positive
TR-6	5	0	No	Positive	Positive
TR-7	2	1	Yes	Positive	Positive
TR-8	1.5	1.5	Yes	Positive	Positive
TR-9	2	0	No	**Negative**	Positive
TR-10	1	0	No	Positive	Positive

Table 2: Overview of translator's experience (measured in years) and opinion on MT and CAT.

The translators were provided with a translation package to be opened in memoQ, where all TM and MT suggestions appeared as pre-translated text. They did not have to choose the type of output to post-edit in each segment: they were provided with either a TM or MT suggestion (or no suggestion at all in the case of the translation-from-scratch subsample). These segments were marked according to their standard treatment in memoQ and similar CAT tools: blue and their TM match score in the case of TM suggestions, and a different color and no TM match score in the case of MT suggestions. Whenever editing a TM match, the application would show the difference between the source text stored in the TM and the new source text.

3 Data analysis

The ten packages delivered by the translators were analyzed individually in order to extract the time spent in each segment and calculate automatic evaluation metrics. Even though translators were instructed to perform all necessary edits to achieve the usual quality required by our client, Translator 5's sample showed clear signs of under-editing which would not have reached that goal. Meanwhile, it turned out that Translator 8 enabled memoQ's predictive text feature, thus avoiding an adequate comparison with the rest of samples.[5] This led us to discard the whole sets of Translators 5 and 8.

Of the remaining 3744 segments, we also discarded 6 due to too low or too high productivity (over 40 seconds per word and under 0.2 seconds per word, respectively). These outlier limits were found by computing interquartile ranges.[6] The segments with too long editing times may be due to the translator not closing the editor during pauses or long breaks (in fact, the duration and time of at least two of those segments clearly match translator's lunch breaks). As to the segments left out due to unnaturally high productivity, most of them (28 out of 29) correspond to 100% matches. An explanation for them might be that the translator actually revised the TM suggestion while having the cursor placed in a contiguous segment, and only entered the segment to confirm it. It may also be that translators did not even read those 100% matches

[5] For further details on these issues, see Parra Escartín and Arcedillo (2015b).

[6] Federico et al. (2012) also establish outlier thresholds in their experiment. In their case, the upper threshold was 30 seconds per word, while the lower was 0,5 seconds per word.

and instead directly confirmed them, either because they trusted the TM suggestion or because they intended to revise them at a later stage during self-revision. However, we lack a way to verify these hypotheses.

3.1 Productivity report

Table 3 reports the results obtained for each translator in each band. The average words per hour across translators is provided in the last column, while the last row shows the productivity gain achieved by the MTPE band over translation from scratch. This MTPE gain is calculated according to Equation 1, where PE_Throughput is the productivity achieved by one translator in words per hour when post-editing MT output, and TRA_Throughput is the productivity achieved by that same translator when translating from scratch.

$$Gain = (\frac{PE\_Throughput}{TRA\_Throughput} - 1) * 100 \quad (1)$$

	Seg.	Words	TR-1	TR-2	TR-3	TR-4	TR-6	TR-7	TR-9	TR-10	Avg.
100%	94	1226	3277	2942	1894	1767	1579	**4039**	2798	1395	**2461**
95-99%	21	231	**2642**	2625	1476	1299	963	2011	1133	1543	**1477**
85-94%	48	1062	**2960**	2248	1660	1678	1232	2164	2429	1012	**1923**
75-84%	42	696	**1592**	**1592**	1372	1140	1019	1342	1576	741	**1297**
MTPE	131	1890	**1804**	1743	1369	1141	922	1481	1433	739	**1329**
Trans.	132	1914	**1701**	1319	993	916	933	1236	1223	473	**1099**
MTPE gain %	-	-	6.06	32.15	37.78	24.62	-1.23	19.80	17.20	56.34	**24.09**

Table 3: Productivity achieved per translator and match band in words per hour.

The throughputs obtained seem a bit high when compared to the usual reference values of 313–375 words per hour (2500–3000 words per day) for translation tasks. While it is true that the content we usually receive from this client often allows our translators to translate faster than their regular throughput, the values reported here are still exceptionally high. To explain them, it must be noted that the productivity in Table 3 only reflects the time spent by translators with the cursor placed inside a segment in memoQ's editor. Other tasks such as reading of instructions, file management, escalation of queries, self-revision, terminological investigation, quality assurance and communication with project managers and/or other parties involved in the project are not reflected here, but are all part of a regular project. Care should also be taken when transferring productivity gains into rate discounts, as translation rates often include tasks which are not affected by the introduction of MTPE (such as project management, file engineering or revision by a second linguist).

Examples 2–4 are samples of text to be translated in the project. As may be observed, the text is written in plain English and the sentences are not too complex.

(2) {1}Activate trial version of the application{2}.

(3) When the trial license expires, the trial version of the application cannot be activated for a second time.

(4) This check box enables / disables the option that prevents Firewall from stopping until the operating system shuts down completely.

As shown in the last row of Table 3, all translators except one (Translator 6, who experienced 1.2% productivity loss when facing MTPE) were faster in MTPE than translating from scratch. However, the productivity gains have a great variability ranging from 6.06% to 56.34%. Other researchers studying MTPE have also reported this great variation in their experiments

(Guerberof Arenas, 2009; Plitt and Masselot, 2010; Koehn and Germann, 2014). For instance, productivity gains between 20% and 131% were reported by Plitt and Masselot (2010).

A possible explanation for the slight productivity loss experienced by Translator 6 might be that this translator had experienced an extensive period of inactivity and had barely used memoQ. Also, the biggest productivity gain was achieved by the least experienced translator (Translator 10), while the smallest productivity gain corresponds to the most experienced one (Translator 1). However, this trend cannot be confirmed by the rest of results.[7]

3.2 MT evaluation metrics

Since collecting and comparing translation-from-scratch and MTPE throughputs in each project is time-consuming and may not always be feasible, it is common practice to use automated metrics as an indicator of productivity gains. We calculated document-level and segment-level values for BLEU,[8] TER and target-side FMS taking the segments belonging to each band as separate documents. BLEU and TER were obtained using Asiya (Giménez and Màrquez, 2010), whereas the FMS was computed using Okapi framework's Rainbow application via its Translation Comparison feature.[9] This target-side FMS is based on the Sørensen-Dice coefficient (Sørensen, 1948; Dice, 1945) using character 3-grams. One difference with both BLEU and TER is that this FMS is applied to character rather than word sequences.

Table 4 reports the average results for each metric and the average gain obtained. In each segment, we computed the automatic metrics and productivity gain for each translator individually using their specific throughputs and text output. This means that in no case we calculated the metrics using multiple references. The gain is calculated according to Equation 1 above.

	BLEU	**TER**	**FMS**	**Gain**
100%	92.68	4.10	97.48	127.38%
95-99%	85.35	9.19	92.12	66.20%
85-94%	82.31	11.99	91.38	76.82%
75-84%	70.70	20.98	85.60	22.52%
MTPE	66.07	20.90	87.91	24.09%

Table 4: Automatic scores and productivity gain for each band.

A surprising finding is the average gain reported for the 95–99% band. It would be logical to expect the gain of this band being somewhere in between the 100% and the 85–94% band values, as the three automatic metrics actually indicate, but it turned out to be inferior to the 85–94% band. This can be explained by the fact that the vast majority of edits required in the 95–99% band involved dealing exclusively with inline tags. Although the impact of these operations has not been researched enough, these results show that they can have a big impact in terms of productivity, slowing down the translator more than it would be expected.

Moreover, when calculating automated metrics, inline tags are generally deleted to avoid their division in different tokens. Instead of deleting them, when calculating the automated metrics reported in this paper we converted each tag into a unique token. Although this operation brought the 95–99% values closer to the 85–94% band, it was not enough to compensate all the effort put into tag handling, as hinted by the productivity gain values. More research is needed

[7] For further discussion on the impact of experience in the throughputs reported in the experiment, see Parra Escartín and Arcedillo (2015b).

[8] As stated in Asiya's documentation, BLEU scores are estimated using the NIST's script used for evaluation campaigns and available at: `http://www.itl.nist.gov/iad/mig//tools/`.

[9] We used Rainbow (`http://okapi.opentag.com`) because it natively supports the most common bilingual formats in the industry, already computes a target-side FMS and is freely available, thus potentially improving transparency and usability of MTPE evaluation.

in this area to find out the appropriate weight or penalty that automated metrics should assign to inline tags. For this reason, we opted to treat the 95–99% band as an outlier and ignore it from further analysis.

4 Productivity gain thresholds

With the aim of establishing where the productivity gain threshold lies, we crossed productivity values with automatic MT evaluation metrics. For each translator in the experiment, we estimated the sentence-level BLEU, TER and FMS values in the MTPE and TM match samples.

Figures 1, 3 and 5 show the overall results for each evaluation metric, while Figures 2, 4 and 6 show the number of segments for each band and metric. The gain values can be expected to be more reliable the higher the number of segments which fell in that band.

As can be observed in Figure 1, the last BLEU band which reported productivity gains is the 45–50 band. Between 35 and 45, the productiviy is slightly inferior to the translation-from-scratch average. Below 35, the productivity decreases more drastically. This finding is interesting, as it is normally said that a BLEU score of 30 reflects understandable translations, while scores over 50 can be considered good and fluent translations (Lavie, 2010). Even though a BLEU score of 30–45 may be useful for other MT applications (such as gisting), in the case of MTPE it does not seem to yield any productivity increase.

In our experience in past projects, BLEU values above 50 are a clear indicator that productiviy gains can be achieved, while gains in the 45–50 range are common, but cannot be taken for granted. Below 45, we have never experienced productivity gains, so the findings reported here seem to support our past experiences. These experiences cover a broad range of domains, all types of MT systems (generic, customized, rule-based, statistical, hybrid, etc.) and diverse quality of MT output. However, they only encompass English into Spanish tasks and different results are likely to be experienced by other language pairs.

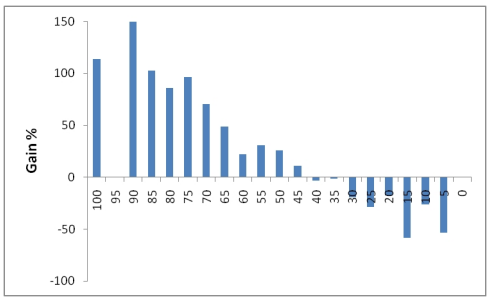

Figure 1: BLEU scores vs. productivity gain.

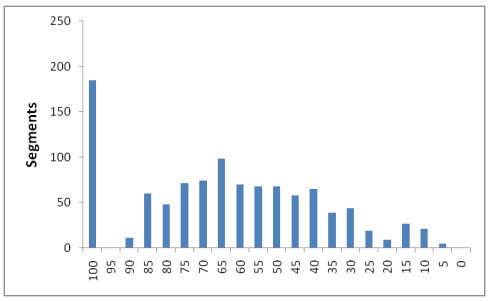

Figure 2: Segments per BLEU band.

As far as the productivity gains for each band are concerned, it is also noticeable that the BLEU 90–95 band obtains a greater productivity gain than the 100 mark. The few number of segments within this band (11), significantly lower than segments in the 100 and the 85–90 groups, may have skewed the results for this particular band. Even though there are other examples of inconsistencies with contiguous bands (such as the productivity of the 60–65 band being a bit lower than expected and the one for 75–80 being a bit higher), the trend seems clear and our results point to a productivity gain threshold around 45 BLEU score. Apart of the few number of segments in the 90–95 range, it is also noticeable that no segment fell into the 95–99 BLEU band, despite the relatively large number of segments in the 100 and 85–90 ranges.

The TER results are more difficult to interpret (Figure 3). The productivity gain drops slightly below translation-from-scratch average at the 30–35 band, rises above this average at 35–40 and drops again for good at the 40–45 mark. According to this data, the last TER range with clear productivity increase would be the 25–30 band, with the tipping point somewhere between 30 and 40. In our past experiences, the threshold for productivity increase was also situated around 35, although its variability proved to be higher than BLEU's.

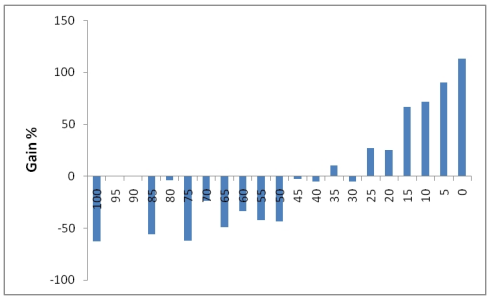

Figure 3: TER scores vs. productivity gain.

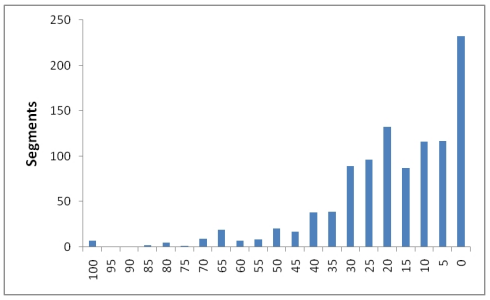

Figure 4: Segments per TER band.

The FMS, on the other hand, shows a more consistent monotonic trend (Figure 5). Also, the distribution of the number of segments within the different bands is closer to a standard normal distribution than the other metrics evaluated (Figure 6). The last band where productivity gain is reported is the 75–80 range. The productivity for 70–75 falls slightly below translation-from-scratch average, while below 70 the productivity loss is more dramatic. Therefore, the productivity gain threshold for FMS seems to lie at the 75% mark. It is interesting that these target-side FMS results match the traditional industry practice of not offering discounts for TM matches below a source-side FMS of 75%, following the general assumption that they do not yield any productivity increase.

It is tempting to take this analogy further and try to apply the TM match pricing schemes to MTPE tasks (i. e., apply to unedited MT segments the same discount as 100% TM matches, etc.). In the next section we compare throughputs of TM matches and MTPE segments in order to evaluate the appropriateness of this approach.

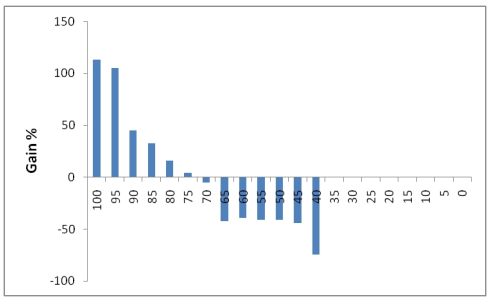

Figure 5: FMS vs. productivity gain.

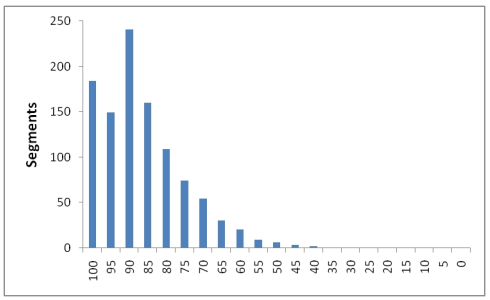

Figure 6: Segments per FMS band.

5 Productivity and cognitive effort in TM matches and MTPE segments

Since in our experiment a significant amount of TM fuzzy matches were post-edited alongside MTPE segments, exact matches and no-match segments, we can compare the productivity of segments which required equal amount of text editing but belong to different categories. For example, we can compare the productivity of MTPE segments which required no editing with unmodified 100% TM matches, or MTPE segments with 75–84% target-side FMS with its equivalent segments from the TM sample.[10] Our hypothesis is that MTPE segments are slower to post-edit because they require a higher cognitive effort[11] to identify the parts of the sentence that need (or do not need) editing, while when post-editing TM matches those parts are already highlighted for the user by the CAT tool.

Another factor is that translators may consider TM suggestions more reliable than MT output. In this experiment, the TMs provided to the translators were the actual ones used in production with this particular client. All segments contained in these TMs have been self-revised by the translator, revised by a second linguist and validated by the client at least once over the past few years. Therefore, the matches retrieved can be considered reliable and representative of the quality expected by the client. If this same experiment is re-run using less reliable TMs (for example, ones with no terminological, syntactical or orthotipographic consistency), it is possible that the effort put into achieving structural and terminological homogeneity, or even correcting translation errors, would negate part or all the productivity gains reported below of TM matches over high-quality MT output.

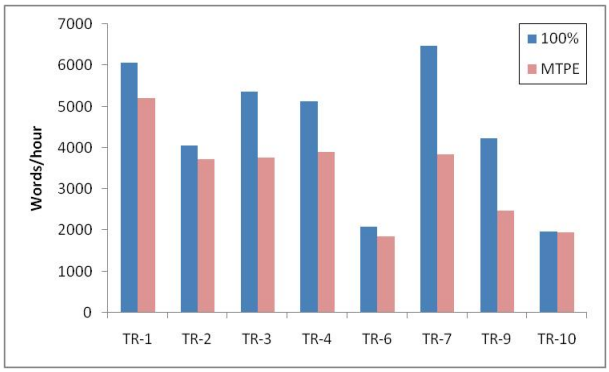

Figure 7: Productivity of unedited 100% matches and unedited MTPE segments.

Figure 7 shows the throughput of segments which were left unedited from the 100% TM match and MTPE bands. All translators except one worked faster with these unedited 100% matches than with the MTPE sub-sample. The exception is Translator 10, the least experienced translator, whose productivity in both categories was almost the same. Since in both categories the amount of editing was the same (in this case, no edits were performed), the difference may lie in the higher cognitive effort invested in the MTPE segments, or that the translator had less confidence in the MT suggestion and spent more time checking it.

To study if this behaviour replicates when text editing is involved, we compared segments

[10] It should be noted that different CAT tools use different implementations of the FMS. Thus, the CAT tool used may also have an impact on the results obtained. Replicating the experiment reported here using different CAT tools would be need to shed light on this issue.

[11] Previous studies (Koponen et al. (2012)) have already successfully applied post-editing time as a way to assess the cognitive effort involved in MTPE.

with similar FMS values both from the MTPE and 75–84% TM match sets. We selected that TM match set because it is the one closest to the MTPE sample in terms of number of words per segment, productivity and automatic scores (Tables 1, 3 and 4 respectively). We then selected only the segments from those bands with a target-side FMS of 75–84 to be sure that both sets involved the same amount of target text editing. Figure 8 shows the productivity obtained by each translator in both sets. Again, all translators except the least experienced one (Translator 10) worked slower with the MTPE sub-sample than when post-editing TM matches.

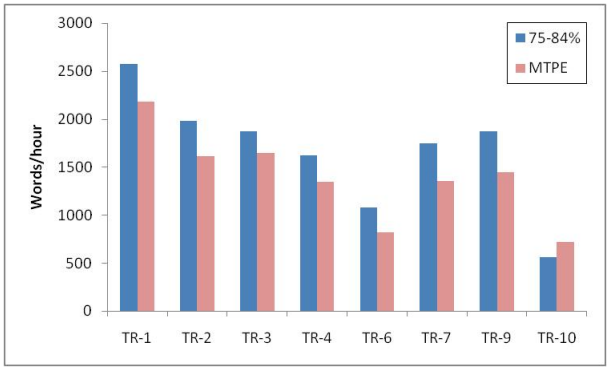

Figure 8: Productivity of 75–84% TM matches and MTPE segments with 75–84 FMS.

These results indicate that, even though the edits performed in both samples were the same, when facing MTPE samples translators need to invest more cognitive effort deciding which parts of the sentence need correction. When leveraging TM segments, however, these parts are already highlighted by the CAT tool, thus making them faster to post-edit. It may be worth researching a similar feature for MTPE segments, where the parts of the sentence with lowest confidence are highlighted for the user.[12]

One of the conclusions that can be drawn from these results is that the pricing system of source-side FMS due to TM leverage cannot be directly applied to MTPE. The different productivity gains reported for these two kinds of translation technologies indicate that MT segments are slower to post-edit than equivalent TM matches, even where no editing is involved. Therefore, it would not be adequate to apply the 100% match rate discount to MTPE segments that require no editing, or the low fuzzy match rate discount to equivalent MTPE segments.

We are aware that other researchers have reported equivalences between MTPE and TM match editing in the past. For example, Guerberof Arenas (2009) concludes that MTPE is equal to 85–94% TM fuzzy editing. However, in these experiments TM matches were not used in their usual settings, as the translator did not know if the suggestion came from MT or TM, and therefore no highlighting of the text differences was offered to the translator. For this same reason, the conclusion that the quality of post-edited output from MT is higher than post-edited TM output cannot be applied to a real commercial environment as the one we tried to replicate here.

6 Conclusion and future work

In this paper, we reported an experiment where ten professional translators with diverse experience in translation and MTPE complete the same translation and post-editing task within their everyday work environment using files from a real translation request. The more than 7,000

[12]Nayek et al. (2015) have already considered this approach in the context of the development of a post-editing tool.

words of the file to be translated included a significant amount of TM fuzzy matches, TM exact matches and no-match segments. Half of the no-match segments was randomly selected for MTPE, while the other half was translated from scratch. The MT output for the MTPE sample was generated using one of our customized RBMT engines.

We compared the productivity gain achieved due to MTPE with automated metrics such as BLEU, TER and FMS in order to find out the threshold at which each metric starts to consider that productivity gains can be achieved. According to our results, BLEU scores above 45, TER values below 30 and FMS values above 75 mean that productivity increases can be obtained. We have also detected a grey area in TER values between 30 and 40, in which the productivity increase is difficult to interpret. These thresholds agree with past experiences in our company, although only the pair English into Spanish has been considered and different performance by other pairs is to be expected.

A comparison between equivalent segments from the MTPE and TM matching samples has shown that, where equivalent text editing is involved, MTPE segments are slower to post-edit than TM matches, even where the TM/MT output was left unmodified. This can be explained by the higher cognitive effort required in MTPE segments, where the translator needs to identify the parts of the sentence which need editing, as opposed to having the CAT tool highlight these parts automatically in TM leveraging. Translators' confidence in TM suggestion may also play a role here. A side effect of these findings is that MT output cannot be assigned the same pricing system as TM fuzzy matching, although using fuzzy scores for MTPE evaluation may help to map these two systems and create intuitive analogies.

As future work, it may be interesting to find out if highlighting the parts of the sentence which the MT system feels less confident about would reduce the cognitive effort involved in MTPE. We also plan to apply quality estimation techniques to this same data set in order to prevent poor MT output from being post-edited, thus eliminating cases of productivity loss when compared to translation from scratch.

Acknowledgments

The research reported in this paper is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement nº 317471.

We would like to thank our colleagues at Hermes: The project managers, for allocating the time and resources that allowed us to run the experiment, and the translators who participated in it and provided us with the data analysed in this paper. We would also like to thank the anonymous reviewers for their valuable feedback to improve this paper and for the ideas for future work they have provided us with.

References

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecological Society of America*, 26(3):297–302.

Federico, M., Cattelan, A., and Trombetti, M. (2012). Measuring User Productivity in Machine Translation Enhanced Computer Assisted Translation. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*, San Diego, CA. AMTA.

Giménez, J. and Màrquez, L. (2010). Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86.

Guerberof Arenas, A. (2009). Productivity and quality in the post-editing of outputs from translation memories and machine translation. *The International Journal of Localisation*, 7, Issue 1:11–21.

Koehn, P. and Germann, U. (2014). The Impact of Machine Translation Quality on Human Post-editing. In *Proceedings of the Workshop on Humans and Computer-assisted Translation*, pages 38–46, Gothenburg, Sweden. Association of Computational Linguistics.

Koponen, M., Aziz, W., Ramos, L., and Specia, L. (2012). Post-editing time as a measure of cognitive effort. In *Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice (WPTP)*, pages 11–20.

Lavie, A. (2010). *Evaluating the Output of Machine Translation Systems*. AMTA, Denver, Colorado, USA.

Nayek, T., Naskar, S. K., Pal, S., Zampieri, M., Vela, M., and van Genabith, J. (2015). CATaLog: New Approaches to TM and Post Editing Interfaces. In *Proceedings of the RANLP 2015 workshop on Natural Language Processing for Translation Memories*, pages 36–42, Hissar, Bulgaria.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). Bleu: a Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598.

Parra Escartín, C. and Arcedillo, M. (2015a). A fuzzier approach to machine translation evaluation: A pilot study on post-editing productivity and automated metrics in commercial settings. In *Proceedings of the ACL 2015 Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, pages 40–45, Beijing, China. ACL.

Parra Escartín, C. and Arcedillo, M. (2015b). Machine translation evaluation made fuzzier: A study on post-editing productivity and evaluation metrics in commercial settings. In *Proceedings of the MT Summit XV*, Miami (Florida). International Association for Machine Translation (IAMT).

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

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (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*, pages 223–231, Cambridge, Massachusetts, USA.

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab*, 5 (4):1–34.