Raising the TM Threshold in Neural MT Post-Editing: a Case-Study on Two Datasets

Anna Zaretskaya
TransPerfect
Passeig de Gràcia, 11
08007 Barcelona, Spain
azaretskaya@translations.com

Abstract

This study intends to determine whether replacing fuzzy TM matches by suggestions from neural machine translation (NMT) can decrease the post-editing effort. We compare the post-editing distance of TM fuzzy matches and of NMT suggestions based on two datasets. We found that in one of the datasets MT was consistently more useful than TM matches, but in the other dataset it was not. We argue that it is necessary to collect extensive data on PED in TM matches in order to be able to easily optimize the TM threshold for any given project.

1 Introduction

TransPerfect is a large language service provider (LSP) translating about two billion words each year with a strong focus on technology, including machine translation (MT). We provide a variety of different MT services, most of which involve MT post-editing (MTPE). In the past few years, we experienced a steady growth of the share of translations produced with MTPE workflows. This growth can be attributed to the implementation of proprietary neural MT technology (NMT), which has improved the average quality of MT, and consequently increased its benefits and acceptance among our linguist experts community. On average, switching from our previous statistical MT framework to the current neural one decreased the post-editing distance by 9.2%, which means an improvement in quality of approximately 29%.

Our MTPE workflow, similarly to the majority of LSPs, combines translation memory (TM) leverage and MT suggestions. We use the 75% TM threshold, which means that only TM matches of 75% and above are shown to the linguists as draft translations during post-editing, and the rest of the segments are pre-translated by an MT system. This study intends to investigate if the threshold has to be raised considering the increase in MT quality, and if so where the new threshold should lie. In other words, we want to know if the linguists' performance will increase if we use MT suggestions instead of the so-called high fuzzy matches (75-99%), and what it depends on.

We approached this task by measuring the post-editing distance (PED) between the TM matches and the final translation and comparing it to the PED between NMT suggestions and the same final translations. This will show whether the amount of editing that has to be applied to the TM fuzzy matches is greater or smaller than that of NMT output.

For this study we selected two different datasets, which are very similar in regards to their content but differ by language pair: English-Chinese and English-Spanish. This study is intended as an initial stage of a large-scale study that will allow us to draw broader conclusions and create best practices on establishing TM thresholds in NMT post-editing projects.

2 Background

There have been previous studies that compared MT and TM matches from the point of view of post-editing effort as well as linguists' perception. In one of them it was demonstrated that translators were more productive when editing MT suggestions (from a statistical MT system) than editing fuzzy TM matches from the range of 80-90% (Guerberof, 2009). In this experiment translators even produced better quality when editing MT suggestions compared to the quality of edited TM matches. One potential explanation for that was the fact that TM matches are valid sentences in the target language and they read naturally (therefore the errors are easier to miss) while MT errors are more obvious, because they often render sentences ungrammatical.

Two related studies (Moorkens and Way, 2016; Rico et al., 2018) also investigated the potential usefulness of MT suggestions compared to TM matches, concluding that having a reliable MT system and a way to predict its performance in many cases is more beneficial than TM leverage. O'Brien (2006) used eye-tracking techniques to study the cognitive load of post-editors and found that the cognitive activity when editing MT suggestions is similar to the activity observed when editing 80-90% fuzzy matches.

This has been confirmed by other studies on the topic, with evidence showing that, while there are still certain prejudices against MT, using MT suggestions instead of TMs increases translators' performance in certain scenarios. For example, it seems that translators are likely to choose MT suggestions over TM matches during post-editing more often if the origin of the suggestion is unknown (i.e. translators do not know whether it comes from MT or TM) (Sánchez-Gijón et al., 2018). Along the same lines, translators prefer to know whether translation suggestions comes from MT or TM, but they are actually more productive when they are not provided this information (Teixeira, 2014).

It is especially important to ask now more than ever, as we have observed a leap in MT quality in general with the spread of neural MT systems. While the abovementioned studies used statistical machine translation for the experiments, our prediction is that the advantage of MT suggestions over TM matches will be even stronger when neural MT is used. The most recently published study on the topic (Sanchez-Gijón et al., 2019) does use neural MT for the comparison. This experiment carried out on a small dataset follows the authors' previous studies that used SMT: apart from the edit distance, it considers the editing time and the subjective quality perception of the post-editors. The authors come to the conclusion that using NMT reduces the amount of editing, but does not improve the translators' productivity.

In general terms, the results of these and other related studies (He et al, 2010; Yamada, 2011) point to the fact that in many cases MT suggestions are more useful than TM matches, and therefore it is clear that we should ask ourselves whether the widely used TM threshold of 75% still holds. Nevertheless, the specific practical recommendations resulting from these studies are not defined, as they seem to depend on the specific scenario: the way MT quality is measured, how MT suggestions are presented to the user of the translation environment, and of course the specific characteristic of the MT engine. That is why in the long term, our goal is to establish a new universal TM threshold that would suit TransPerfect specific post-editing setup or, if this threshold varies depending on some conditions, identify these conditions and create a simple guideline for establishing a TM threshold on a project basis.

3 Experiment Data and Setup

The datasets used for this study included only translation units (TUs) that, at the moment of their translation, matched with the existing segments in the TMs. We retrieved the source segment, the target segment suggestions from the TM, and the final translation of the same segment. In addition, we produced an NMT suggestion for each of the source segments.

For each TU, we compared the target segment from the TM with the final translation and calculated the PED between them. We will refer to these values as *PED-TM*. We also compared the target produced by the NMT systems with the final translated segment to obtain the values of *PED-MT.*[1] PED is a standard MT quality metric used at TransPerfect and is very common in the translation industry in general. It evaluates the quality of MT from the point of view of the *post-editing effort*, in other words it shows how many changes the linguists make in the initial draft translation in order to produce final translation. It is based on the Levenshtein edit distance, is character-based, and is presented as a percentage of edited characters over all the characters in the sentence. Lower PED means that less post-editing effort required and thus better MT quality.

When talking about the amount of work involved in post-editing, it is common to distinguish technical, temporal and cognitive post-editing effort (Krings, 2001). Even though PED as a method of evaluating post-editing effort is limited only to the technical post-editing effort (i.e. it does not account for the cognitive load of the post-editors, or for the time needed to per-

[1] Even though we call it post-editing distance, in case of segments produced by MT there was no post-editing performed. The final translation used as a reference was not creating by post-editing the corresponding MT output. However, we make this assumption for simplicity of the calculation. We also acknowledge the fact that this way, the PED-MT values might be slightly higher than if the translation were produced by means of actual post-editing.

form post-editing), it allows to obtain objective data on the actual amount of editing needed to reach the final translation, and this is a critical factor in improving translators' performance (Plitt and Masselot, 2010; Federico et al., 2012).

3.1 Datasets

For this experiment we selected post-edited texts from past post-editing projects, two different accounts. Dataset *ENES* contained post-edited files from English into Spanish from the online fashion retail domain. The projects included in the study dated from the time period of January 2018 to March 2019 and were post-edited by 6 different linguists.

Dataset *ENZH* contained data from a different online fashion retail account, post-edited from English into Chinese by 21 different linguists in the time period of February 2018 to March 2019. The data in the two datasets comes from two different accounts, however, the content is very similar (short fashion product descriptions). We deliberately chose the same content type in order to minimize the impact of different content types on the results, but at the same time we were able to compare the results for two different language pairs.

From both datasets we gathered only the translation units (TUs) that are considered high fuzzy matches, i.e. at the time the file was analyzed against a TM, the leverage score of the segments was from 75% to 99% (both included). The number of TUs in the dataset ENES was 8183, with an average source segment length of 5.6 words. The number of TUs in the dataset ENZH was 7521 with an equal average of 5.6 words per source segment. We distributed the TUs in five groups by ranges of TM match scores, a break-down of all the TUs is shown in Table 1.

TM range	# of TUs in ENES	# of TUs in ENZH
75-79%	3243	2801
80-84%	1956	1811
85-89%	1401	1446
90-94%	420	361
95-99%	1163	1102
Total	8183	7521

Table 1. Breakdown of TM match scores in the experiment data.

The biggest range in terms of segment count is the lowest range of 75-79%. In the ENES dataset, it constitutes 40% of all segments, and in the ENZH it constitutes 37% of all segments.

3.2 Neural Machine Translation

The MT systems used for the experiment were proprietary neural MT systems. Both systems are the ones that are currently used in the post-editing projects in the two accounts. The ENES system was a generic one, i.e. it was created using a generic training corpus and did not undergo any kind of customization using client data. The ENZH system had been customized using the client TM.

The average post-editing distance of the ENES system on the account content in general (on all segments that were actually post-edited in real projects) was 25.86%, and the average PE distance of this system measured on the dataset selected for this experiment was 30.30%. The average PE distance of the ENZH system on all the post-edited in the account was 23.09%, while the PE distance measured on our dataset was 15.17%.

4 Results

The results of the comparison of the PED-MT and PED-TM values for the two datasets were strikingly different. In dataset ENES, PED-MT was consistently higher than PED-TM (Figure 1).

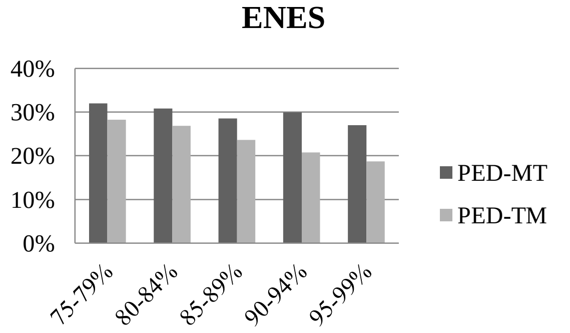

Figure 1. Comparison of PED-MT and PED-TM in different ranges of fuzzy matches in the ENES dataset.

The picture in the ENZH dataset was almost exactly the opposite: in all the TM ranges except for one we observe lower PED-MT and higher PED-TM (Figure 2).

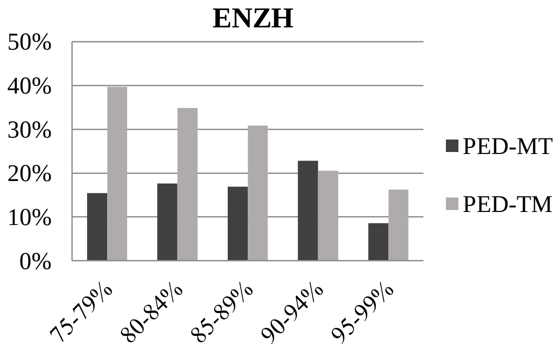

Figure 2. Post-editing distance in different ranges of fuzzy matches compared to the post-editing distance of MT segments.

This result was not unexpected, considering the difference in the performance of the two MT systems: the average PED-MT in the ENES dataset was significantly higher than PED-MT in the ENZH dataset. This is of course due to the system customization. We have seen that a customized system can improve the PED by up to 20% compared to a baseline generic system. In fact, we have confirmed this by calculating the PED-MT value on the same ENZH dataset, but using a generic NMT system, and the result was 31.31%, which is significantly higher that the PED-MT value of the customized system (16.61%).

In addition, almost a half of the MT segments in the ENZH dataset (42.5%) were exactly the same as the final translation, i.e. PED-MT was equal to 0% and these segments did not need any editing. (Table 2).

	PED-MT = 0%
ENES	1200 (8.3%)
ENZH	3378 (42.5%)

Table 2. Number of segments with PED-MT equal to 0 in both datasets.

Based on these results, the ENZH account is, without a doubt, a good candidate for replacing fuzzy matches by NMT suggestions. In fact, we have received feedback from one of the post-editors working on the account, who confirmed our observations and pointed out the following:

"Funny thing is for these files, fuzzy matches take much more time than MT, because the changes in high fuzzy matches need to be carefully identified, but some of the MT is perfect."

Nevertheless, there was one TM range (90-94%) where TM matches had lower PED than MT suggestions. The analysis of the segments revealed one possible reason for this, which is the segment length. The average number of words in the segments of this TM match range was 10.54, which is significantly higher than the average for the dataset (5.6). Our assumption is that this MT system performs worse on longer segments.

We further investigated this assumption on the ENZH dataset. Table 3 shows the average PED-MT and the average segment length in each of the TM match ranges.

TM range	Avg. PED-MT	Avg. Length
75-79%	15.43%	4.64
80-84%	17.63%	5.62
85-89%	16.86%	6.02
90-94%	22.79%	10.54
95-99%	8.60%	6.65

Table 3. Average PED-MT and average source segment length in different TM ranges in the ENZH dataset.

Even thought we observed only weak correlation between the segment length and PED-MT (*r*=0.32), there is a clear association as the range 90-94% seems to be an exception both in terms of segment length and PED-MT. The reason for that might be that the longer sentences are more challenging for MT to handle. In the retail product descriptions, longer sentences usually constitute a more creative part of a description, which requires substantial modifications in the target language in order for it to sound natural. Shorter sentences, on the other hand, are very straight forward, not creative, and only list the characteristics of the product that normally come from a limited set.

As expected, there was observed an association between the fuzzy match score and the PED-TM value: the correlation was stronger in the ENZH dataset (*r*=-0.40) and weaker in the ENES dataset (*r*=-0.20). This means that the higher the fuzzy match the less it needs to be edited. However, MT performs relatively similar in all fuzzy match ranges. This has an implication when choosing a new TM threshold: while some fuzzy matches require more editing than others, MT suggestions require the same amount of editing on average.

Another interesting observation was the difference in the average PED-TM in the two datasets: 25.37% in ENES and 32.24% in ENZH. This is due to the difference in writing systems and the way PED is calculated. The average number of characters in Chinese sentences is lower, and since the PED is calculated as a percentage on the total number of characters, the PED will always be higher. For this reason, if we base our TM threshold strategy uniquely on the PED we should treat the languages with character-based writing systems like Chinese and Japanese differently than European languages. This issue will be discussed in more detail in the following section.

5 Discussion

The difference in the results obtained for the two datasets demonstrate the importance of the initial high performance of the MT system that is needed in order to provide high-quality segments that will potentially replace fuzzy matches. The main difference between the two MT systems was the fact that one of them was generic and the other one was customized for the client content. NMT system customization with a large amount of high-quality data can significantly improve the system performance. An experiment that had been conducted at TransPerfect showed that a customization with additional 100 000 new translation units yields about 4% increase of the PE distance over the baseline system, and the quality grows exponentially when adding more data. Depending on the initial performance and the quality of the data, customization can boost the performance by up to 20% of PED.

This study has shown that, when the performance of the MT system is sufficiently good, replacing fuzzy matches (or at least some of them) reduces the overall post-editing distance, or in other words, the post-editing effort. The challenge lies in establishing the definition of the sufficiently good performance for this specific purpose.

We suggest that one simple approach is comparing the average PED of the MT system on the content type to the post-editing distance required to edit the TM matches, similar to what was done in this study. If we know the average PED-TM for each TM range, we will be able to determine if the MT output requires less or more editing than fuzzy matches, and if so we can raise the TM threshold to the corresponding TM range. For this we need, however, to determine if the average PED-TM values are consistent across all languages and content types. Thus, we have already mentioned that these values can depend on the writing system of the target language: in the TM match range of 75-80%, the average PED-TM in the ENES dataset was around 28% while in the ENZH it was approximately 40%. We need to carry out a large-scale comparison that would include other languages and content types in order to have a full picture of PED-TM.

Then, we will be able to compare it to the PED-MT in each specific case. For example, if we have an account where NMT is used for post-editing from English into Spanish, and we know that the average PED is 18%, we must be able to say with a high degree of certainty that this value is lower than the average PED-TM of TM matches between 75% and 79%, and only then we can raise the TM threshold to 80%. As mentioned in Section 3.1, in both our datasets, 75-79% fuzzy matches constituted about one third of all the TUs, so replacing them by NMT suggestions means improving the quality of approximately one third of all fuzzy matches in post-editing projects.

6 Conclusion and Future Work

This study was the first step in defining optimal TM threshold for MTPE projects with neural MT. Our hypothesis was that using NMT suggestions instead of TM fuzzy matches can reduce, at least in some cases, the post-editing effort. In order to confirm it, we have compared the PED of NMT suggestions with PED of TM fuzzy matches of different ranges. The results obtained in the two datasets were very different in two aspects. First, the general quality of the NMT systems used varied significantly. When the PED-MT values were low (meaning good MT performance), the MT suggestions required less editing than the TM matches, and so in this case we could see the benefits of replacing them by MT. However, when the general MT quality is lower (for example, when the MT system is generic and not customized for the content type), the TM matches continue to me the best source of draft translation.

Given these results, the next step in optimizing the MTPE workflow will consist in gathering data on the average PED of different ranges of TM matches in a wide variety languages and content types. This will allow us to compare the PED-MT with PED-TM for any given project.

Along with the post-editing distance, there are other metrics we use to measure post-editing effort, the most common being post-editing time. PE time and distance do not always correlate. Post-editing activity involves time spent on understanding the source segment and the MT/TM suggestion and assessing if the latter is usable. In fact, studies of linguists' behavior during post-editing have shown that they mostly spend time on contemplating the changes than executing them (Koehn, 2009). As part of future work, we are planning to compare the time required to edit NMT suggestions with the time it takes to edit TM fuzzy matches.

References

Federico, Marcello, Cattelan, Alessandro and Trombetti, Marco, 2012. Measuring User Productivity in Machine Translation Enhanced Computer Assisted Translation. *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA).*

Guerberof, Ana, 2009. Productivity and quality in MT post-editing. *MT Summit XII. The twelfth Machine Translation Summit*. Ottawa, Canada.

He, Yifan, Ma, Yanjun, Roturier, Johann, Way, Andy, and van Genabith, Josef, 2010. Improving the Post-Editing Experience Using Translation Recommendation: A User Study. *AMTA 2010 - 9th Conference of the Association for Machine Translation in the Americas.*

Koehn, Philipp, 2009. A process study of computer-aided translation. *Machine Translation* 23(4): 241-263.

Krings, Hans P., 2001. *Repairing Texts*. Kent State University Press, Ohio, USA.

Moorkens, Joss, and Way, Andy 2016. Comparing Translators Acceptability of TM and SMT Outputs. *Baltic J. Modern Computing*, 4(2):141-151.

O'Brien, Sharon, 2006. Eye-tracking and translation memory matches. *Perspectives: Studies in Translatology*, 14(3), 185-205.

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

Rico, Celia, Sánchez-Gijón, Pilar, and Torres-Hostench, Olga, 2018. The Challenge of Machine Translation Post-editing: An Academic Perspective. *Trends in E-Tools and Resources for Translators and Interpreters*, Brill: 203-218.

Sánchez-Gijón, Pilar, Moorkens, Joss and Way, Andy. 2018. Perception vs. Acceptability of TM and SMT Output: What do translators prefer? *EAMT 2018, 21st Annual Conference of the European Association for Machine Translation*, Alicante, Spain, 331.

Sánchez-Gijón, Pilar, Moorkens, Joss and Way, Andy. 2019. Post-Editing Neural Machine Translation versus Translation Memory Segments. Machine Translation, 33(1-2), 31-59.

Teixeira, Carlos S.C., 2014. Perceived vs. measured performance in the post-editing of suggestions from machine translation and translation memories. *WPTP-3, Third Workshop on Post-Editing Technology and Practice*. Vancouver, Canada, 45-60.

Yamada, Masaru, 2011. *Revising text: An empirical investigation of revision and the effects of integrating a TM and MT system into the translation process* (Doctoral thesis). Rikkyo University.