Machine Translation Quality and Post-Editor Productivity

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA) Vol. 1: MT Researchers’ Track

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Machine Translation Quality and Post-Editor Productivity

Marina Sanchez-Torron  
msnc017@aucklanduni.ac.nz  
School of Cultures, Languages and Linguistics  
University of Auckland  
Auckland 1010, New Zealand  
Philiip Koehn  
phi@jhu.edu  
Department of Computer Science  
Johns Hopkins University  
Baltimore, MD 21218-2608, USA

Abstract

We assessed how different machine translation (MT) systems affect the post-editing (PE) process and product of professional English–Spanish translators. Our model found that for each 1-point increase in BLEU, there is a PE time decrease of 0.16 seconds per word, about 3-4%. The MT system with the lowest BLEU score produced the output that was post-edited to the lowest quality and with the highest PE effort, measured both in HTER and actual PE operations.

1 Introduction and Related Work

There is a relatively fair amount of empirical research on post-editing machine translation output (PE) focusing on assessing potential time and quality improvements over human translation with or without translation memories (TM). A common finding is that, despite differences among participants, PE is on average faster than unassisted or TM-assisted translation. Some examples of studies finding such speed benefits are those by Guerberof (2009), Flournoy and Duran (2009), Groves and Schmidtke (2009), Plitt and Masselot (2010) and Skadiņš et al. (2011).

In terms of PE quality, studies have shown, through the use of human judgments, that PE leads to quality comparable to (García, 2010) or better than other types of translation (Guerberof, 2009; Fiederer and O’Brien, 2009; Carl et al., 2011; Green et al., 2013; Lääbli et al., 2013). There is therefore strong empirical evidence pointing at the speed and quality benefits of PE.

While many factors may affect productivity in a PE workflow, MT quality is one that is relatively easy to measure. Previous studies to have investigated how MT quality affects PE speed are those by Tatsumi (2009) and O’Brien (2011). They found different levels of correlations between MT quality, measured on a sentence level with automatic metrics, and PE speed. Krings (2001) and De Sutter (2012) too found that human judgments of MT quality correlated to PE speed. Koehn and Germann (2014) found their worst MT system entailed 20% more editing activity than their best one.

To the best of our knowledge, this is the first study to attempt to assess how different MT systems, of the same type but with different quality levels, affect the PE productivity of professional translators. By investigating how MT quality affects both PE time and PE quality, we aim at providing MT users and researchers with another approach to examining the usefulness of MT for PE purposes.
2 Study Setup

Nine translators were hired through ProZ\textsuperscript{1}, self-described as the biggest online translation workplace. Selected participants were offered a fixed compensation based on the standard, general, English–Spanish translation rates displayed on the website. Each translator post-edited four news texts of about 650 words each. Texts had similar complexity levels and were presented to translators in randomized order to dilute possible familiarization or fatigue effects. All four texts were translated into Spanish with nine MT systems (cf. Section 4.1). The output of all nine systems was assigned randomly to participants. As in Cettolo et al. (2013) and Koehn and Germann (2014), to deal with between-participant variability, the following restrictions were implemented:

- All translators post-edited all source sentences.
- No translator post-edited the same source sentence twice.
- All translators were exposed to roughly the same amount of output of all MT systems.

Translators were asked to post-edit to full, human-like quality. They worked remotely on the open-source, web-based, computer aided translation (CAT) tool CASMACAT (Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation)\textsuperscript{2} (Alabau et al., 2013).

3 Translator Profile

All participants were native Spanish, professional translators, with at least 2 years’ experience using CAT tools. All were educated to college level. Except for TR4 and TR5, all participants had degrees in Translation. Table 1 summarizes their main characteristics.

<table>
<thead>
<tr>
<th>Translator</th>
<th>Translation experience (years)</th>
<th>PE certification</th>
<th>PE experience (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR1</td>
<td>2 to 5</td>
<td>No</td>
<td>&lt;2</td>
</tr>
<tr>
<td>TR2</td>
<td>2 to 5</td>
<td>No</td>
<td>2 to 5</td>
</tr>
<tr>
<td>TR3</td>
<td>5 to 10</td>
<td>Yes</td>
<td>5 to 10</td>
</tr>
<tr>
<td>TR4</td>
<td>&gt;10</td>
<td>No</td>
<td>2 to 5</td>
</tr>
<tr>
<td>TR5</td>
<td>5 to 10</td>
<td>No</td>
<td>None</td>
</tr>
<tr>
<td>TR6</td>
<td>&gt;10</td>
<td>Yes</td>
<td>&lt;2</td>
</tr>
<tr>
<td>TR7</td>
<td>5 to 10</td>
<td>No</td>
<td>5 to 10</td>
</tr>
<tr>
<td>TR8</td>
<td>2 to 5</td>
<td>No</td>
<td>2 to 5</td>
</tr>
<tr>
<td>TR9</td>
<td>&gt;10</td>
<td>No</td>
<td>2 to 5</td>
</tr>
</tbody>
</table>

Table 1: Translators’ background

Translators’ perceptions of PE and MT were elicited through eight questions answered with a Likert scale. Because of the small sample size, we grouped the answers from the five initial levels into three: Strongly disagree and Disagree were grouped into a new level No; and Agree and Strongly agree into a new level Yes. Neutral was left as such. Table 2 displays a summary of participants’ PE and MT perceptions.

\textsuperscript{1}\url{http://www.proz.com}
\textsuperscript{2}\url{http://www.casmacat.eu/}
Table 2: Translators’ perceptions of PE and MT

Answers show therefore a mix of opinions towards PE. Overall, translators are comfortable post-editing but do not prefer it to editing a human translation.

4 Variables of interest

4.1 Machine Translation Quality

Initiatives providing free resources for MT development and evaluation help boost MT research collaboration and efforts. One of these efforts is the WMT evaluation campaign of the Association for Computational Linguistics (ACL). Among the data they freely provide are training and test data sets. As part of their 8th Workshop on Statistical Machine Translation (WMT 2013)\(^3\), the organizers released a test set comprised of 3,000 source English sentences and their corresponding Spanish human reference translations.

We trained nine MT systems with training data from the European Parliament proceedings, News Commentary, Common Crawl, and United Nations. The systems are phrase-based Moses systems (Koehn et al., 2007) with hierarchical lexicalized reordering (Galley and Manning, 2008), operation sequence model (Durrani et al., 2013), and sparse lexical features, using all available language model data (target side of the full parallel corpus, plus the provided monolingual news corpus and LDC Gigaword). The best system reaches comparable quality to the best system participating in the WMT 2013 evaluation campaign.

We iteratively halved the parallel training corpus to obtain systems of inferior quality. The quality of the MT systems was measured with case-sensitive BLEU (Papineni et al., 2002) on the official WMT 2013 test set. Table 3 summarizes MT systems’ quality and training corpus size.

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>Training sentences</th>
<th>Training words (English)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT1</td>
<td>30.37</td>
<td>14,700k</td>
<td>385M</td>
</tr>
<tr>
<td>MT2</td>
<td>30.08</td>
<td>7,350k</td>
<td>192M</td>
</tr>
<tr>
<td>MT3</td>
<td>29.60</td>
<td>3,675k</td>
<td>96M</td>
</tr>
<tr>
<td>MT4</td>
<td>29.16</td>
<td>1,837k</td>
<td>48M</td>
</tr>
<tr>
<td>MT5</td>
<td>28.61</td>
<td>918k</td>
<td>24M</td>
</tr>
<tr>
<td>MT6</td>
<td>27.89</td>
<td>459k</td>
<td>12M</td>
</tr>
<tr>
<td>MT7</td>
<td>26.93</td>
<td>230k</td>
<td>6.0M</td>
</tr>
<tr>
<td>MT8</td>
<td>26.14</td>
<td>115k</td>
<td>3.0M</td>
</tr>
<tr>
<td>MT9</td>
<td>24.85</td>
<td>57k</td>
<td>1.5M</td>
</tr>
</tbody>
</table>

Table 3: MT Systems’ quality and training corpus size

\(^3\)http://www.statmt.org/wmt13/
4.2 Human-mediated Translation Edit Rate

Human-mediated Translation Edit Rate (HTER; Snover et al., 2006) measures the minimum number of operations (insertions, deletions, substitutions and shifts) needed to convert a machine translation into its post-edited version. It is computed by dividing the number of operations by the number of words in the post-edited version. HTER can be used as a measure of MT quality: the fewer the changes that need to be applied to the machine translation, the more similar it is to the post-edited, reference translation and therefore the higher the MT quality. Likewise, HTER can also be used as a measure of technical PE effort: the fewer changes necessary to convert the machine translation into its post-edited version, the less the effort exerted by the translator.

In this study, HTER scores were computed for all nine systems, based on the machine translated texts and their non-minimally post-edited versions, with the freely available tercom software.

4.3 Actual Edit Rate

HTER is concerned about the PE product, not the process. It therefore does not measure translators’ actual edit operations, which may involve going back and applying corrections to previously post-edited parts of the text. We are interested in examining how actual edit operations vary across systems. Edit operations (i.e., insertions and deletions) are measured as the keystroke and/or mouse combinations leading to the insertion or deletion event, which do not necessarily correspond to the number of characters inserted or deleted. For instance, if a translator deletes a word by selecting it with the mouse and pressing backspace, it counts as one deletion event. If they move a word by cutting it and pasting it somewhere else, it constitutes one deletion and one insertion event. Insertion and deletion events were normalized by the number of words in the machine translated text to obtain a what we call here the Actual Edit Rate, henceforth AER.

4.4 Post-editing Time

Mean PE time per word per system was calculated by dividing the time spent post-editing each MT system’s output by the number of source words translated by each system. Table 4 shows time measurements as PE time in seconds per word (spw) and as PE speed in words per hour (wph).

<table>
<thead>
<tr>
<th>System</th>
<th>Mean PE time (spw)</th>
<th>Mean PE speed (wph)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT1</td>
<td>4.06</td>
<td>887</td>
</tr>
<tr>
<td>MT2</td>
<td>4.38</td>
<td>822</td>
</tr>
<tr>
<td>MT3</td>
<td>4.23</td>
<td>851</td>
</tr>
<tr>
<td>MT4</td>
<td>4.54</td>
<td>793</td>
</tr>
<tr>
<td>MT5</td>
<td>4.35</td>
<td>828</td>
</tr>
<tr>
<td>MT6</td>
<td>4.36</td>
<td>826</td>
</tr>
<tr>
<td>MT7</td>
<td>4.66</td>
<td>773</td>
</tr>
<tr>
<td>MT8</td>
<td>4.94</td>
<td>729</td>
</tr>
<tr>
<td>MT9</td>
<td>5.03</td>
<td>716</td>
</tr>
</tbody>
</table>

Table 4: Systems’ mean PE time (seconds per word) and PE speed (words per hour)

4https://github.com/jhclark/tercom
4.5 Post-editing Quality

Error-based quality assessment frameworks allow for the quantification of translation quality according to the type and severity of errors present in the translation. Previous PE studies (Guerberof, 2009; Temizöz, 2013) have used LISA’s Quality Model 3.1, one such framework traditionally used in localization settings, to evaluate the quality of the post-edited texts. This model, however, is not officially available anymore. We therefore assessed PE quality with another error-based model, a quality metric compliant with QTLaunchPad’s Multidimensional Quality Metric (MQM) framework.

We used the decision trees and guidelines provided in Burchardt and Lommel (2014) as a reference during the issue annotation process. Quality was assessed along the dimensions of Accuracy (Mistranslations, Omissions, Untranslated and Additions) and Fluency (Grammar, Style and Typography). We excluded Spelling from our quality metric as some translators pointed out that the spell checker did not work. Following LISA’s standard weight scale, minor errors were given a weight of 1 and major errors were given a weight of 5.

5 Analysis and Results

Sentences with a logged period of inactivity of 2 minutes or more were excluded from analysis as such long pauses are likely not indicative of difficulties posed by the underlying MT system. Also excluded were sentences inadvertently skipped by translators and sentences with unreliable measurements\(^5\). In total, out of 1233 observations, 1202 were considered valid and submitted for analysis.

5.1 Post-Editing Time by Machine Translation System

For each MT system, we plotted mean PE time against BLEU score:

\(^5\)CASMACAT logs editing time on a segment basis as the interval between the opening and closing of each segment. When translators access the PE interface, the first segment in the document is opened automatically. Translators do not usually start post-editing right away, instead they scroll through the document. Once acquainted with its contents, they go up the first segment, post-edit it and, when finished, close it. Most editing times logged for first segments in our study are in fact spent browsing through the whole document and are therefore unreliable.
A linear regression was applied to predict PE time based on MT quality. Our model was significant ($F(1, 7) = 33.62$), with an $R^2$ of .828. It describes the effect of system’s BLEU on mean PE time as following a decreasing linear relationship. Specifically, for every 1-point increase in BLEU, there is a decrease in PE time of approximately 0.16 seconds per word.

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
\text{Estimate} & \text{Std. error} & \text{t value} & \text{p value} & \text{95\% confidence interval} \\
\hline
\text{intercept} & 8.88 & 0.76 & 11.74 & <.001 & [7.37, 10.39] \\
\text{slope} & -0.16 & 0.03 & -5.80 & <.001 & [-0.21, -0.10] \\
\hline
\end{array}
\]

Table 5: Linear regression results

Assumption checks confirm the validity of the results: the plot of residual versus fitted values shows some noise but no distinctive pattern, and although residuals show a slight departure from normality, this is expected in small samples.

5.2 HTER and AER by Machine Translation System

We investigate how both HTER and AER vary between MT systems. Table 6 displays HTER scores and mean AER by MT system:

\[
\begin{array}{|c|c|c|}
\hline
\text{System} & \text{HTER} & \text{AER} \\
\hline
\text{MT1} & 40.75 & 3.36 \\
\text{MT2} & 40.85 & 3.05 \\
\text{MT3} & 42.41 & 3.03 \\
\text{MT4} & 41.57 & 3.58 \\
\text{MT5} & 42.29 & 3.61 \\
\text{MT6} & 43.57 & 3.66 \\
\text{MT7} & 44.79 & 3.33 \\
\text{MT8} & 46.15 & 3.57 \\
\text{MT9} & 50.30 & 4.20 \\
\hline
\end{array}
\]

Table 6: Systems’ HTER (%) and AER (events per word)

As expected, we see an almost continuous gradual increase in HTER as the quality of the MT system decreases. In contrast, our data does not allow us to establish any significant association between AER and MT quality. Keyboard activity may just not be as sensitive to MT quality as PE time. Nevertheless, MT9, the system with the lowest BLEU score has both the highest HTER and AER of all systems, representing an increase of 23.43% and 22% respectively over MT1, the best system.

5.3 Post-Editing Quality by Machine Translation System

Using an error-based framework to assess text quality usually involves determining an arbitrary pass/fail threshold for textual units. There is not a theoretical body of literature concerning these frameworks so we set the minimum sentence quality acceptance level at a percentage commonly referred to in industry documents, i.e., 95%\(^6\).

PE quality is measured for the whole post-edited output, via both the MQM score and the count of sentences falling in the Fail and Pass categories, and reported by MT system. Also reported are normalized issue counts classified according to their severity (note that no critical errors were found in any of the texts).

\[\text{MQM Score (%) = 100 - \left( \frac{5 \times \text{Issues}_{\text{Minor}} + 10 \times \text{Issues}_{\text{Major}}}{\text{Sentence length}} \right) \times 100}\]

\[\text{A 28-word sentence with 2 minor issues and 1 major issue would have therefore a score of 100-} (2+5)/28*100 = 75\%\]
<table>
<thead>
<tr>
<th>System</th>
<th>MQM Score</th>
<th>Fail</th>
<th>Pass</th>
<th>Minor issues/k</th>
<th>Major issues/k</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT1</td>
<td>97.86</td>
<td>15</td>
<td>117</td>
<td>12.39</td>
<td>1.72</td>
</tr>
<tr>
<td>MT2</td>
<td>97.19</td>
<td>20</td>
<td>113</td>
<td>12.83</td>
<td>3.12</td>
</tr>
<tr>
<td>MT3</td>
<td>98.01</td>
<td>16</td>
<td>117</td>
<td>12.59</td>
<td>1.75</td>
</tr>
<tr>
<td>MT4</td>
<td>96.76</td>
<td>26</td>
<td>109</td>
<td>15.32</td>
<td>3.40</td>
</tr>
<tr>
<td>MT5</td>
<td>97.84</td>
<td>18</td>
<td>116</td>
<td>11.93</td>
<td>2.04</td>
</tr>
<tr>
<td>MT6</td>
<td>98.63</td>
<td>10</td>
<td>124</td>
<td>9.88</td>
<td>1.02</td>
</tr>
<tr>
<td>MT7</td>
<td>97.31</td>
<td>17</td>
<td>115</td>
<td>11.39</td>
<td>3.45</td>
</tr>
<tr>
<td>MT8</td>
<td>97.47</td>
<td>22</td>
<td>112</td>
<td>10.48</td>
<td>3.38</td>
</tr>
<tr>
<td>MT9</td>
<td>95.81</td>
<td>32</td>
<td>103</td>
<td>14.28</td>
<td>4.76</td>
</tr>
</tbody>
</table>

Table 7: Indicators of translation quality of post-edited translations by MT system

As expected, given that participants are professional translators post-editing to high quality, sentence-level MQM scores follow a left-skewed distribution (867 of the 1202 sentences score 100%). A Pearson’s chi-square test found differences in the Fail/Pass proportions between MT systems ($\chi^2(8)= 19.40, p <.05$), with a post-hoc pairwise comparison with Holm’s adjustment finding significant the differences between the MT6-MT9 pair. While the Fail/Pass ratio or the MQM scores are not significantly different for all the other pairs, MT9, the system with the lowest BLEU score, produced the output that ended up with the lowest MQM score and the highest Fail/Pass ratio: 1 in 4 sentences were post-edited to below the acceptable 95% MQM score.

In terms of error categories, minor issues are more or less equally divided into Fluency and Adequacy issues across systems, while major issues are practically all Adequacy issues, mostly Mistranslations.

5.4 Post-Editing Quality vs. Post-Editing Time by Machine Translation System

We plotted the Fail/Pass ratio of the post-edited output against the mean PE time for each MT system:

![Figure 2: Scatter plot of systems Fail/Pass ratios against mean PE times with regression line](image)

Our model describes a positive linear relationship between PE time and Fail/Pass ratio: the more time is spent post-editing, the higher the Fail/Pass ratio of the post-edited output. The
model is significant \( F(1,7) = 8.06 \), with an \( R^2 \) of .535. Such a low \( R^2 \) points at additional factors affecting the quality of the post-edited texts. This will be investigated in further studies.

5.5 Post-Editing Time and Quality by Translators

Table 8 shows that differences in HTER, AER and PE time between translators are more pronounced than between MT systems.

<table>
<thead>
<tr>
<th>Translators</th>
<th>HTER</th>
<th>AER</th>
<th>Mean PE time (spw)</th>
<th>MQM Score</th>
<th>Fail</th>
<th>Pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR1</td>
<td>44.79</td>
<td>2.29</td>
<td>4.57</td>
<td>98.65</td>
<td>10</td>
<td>124</td>
</tr>
<tr>
<td>TR2</td>
<td>42.76</td>
<td>3.33</td>
<td>4.14</td>
<td>97.13</td>
<td>23</td>
<td>102</td>
</tr>
<tr>
<td>TR3</td>
<td>34.18</td>
<td>2.05</td>
<td>3.25</td>
<td>96.50</td>
<td>26</td>
<td>106</td>
</tr>
<tr>
<td>TR4</td>
<td>49.90</td>
<td>3.52</td>
<td>2.98</td>
<td>98.10</td>
<td>17</td>
<td>120</td>
</tr>
<tr>
<td>TR5</td>
<td>54.28</td>
<td>4.72</td>
<td>4.68</td>
<td>97.45</td>
<td>17</td>
<td>119</td>
</tr>
<tr>
<td>TR6</td>
<td>37.14</td>
<td>2.78</td>
<td>2.86</td>
<td>97.43</td>
<td>24</td>
<td>113</td>
</tr>
<tr>
<td>TR7</td>
<td>39.18</td>
<td>2.23</td>
<td>6.36</td>
<td>97.92</td>
<td>18</td>
<td>112</td>
</tr>
<tr>
<td>TR8</td>
<td>50.77</td>
<td>7.63</td>
<td>6.29</td>
<td>97.20</td>
<td>19</td>
<td>117</td>
</tr>
<tr>
<td>TR9</td>
<td>39.21</td>
<td>2.81</td>
<td>5.45</td>
<td>96.48</td>
<td>22</td>
<td>113</td>
</tr>
</tbody>
</table>

Table 8: Indicators of translation quality of post-edited translations by translator

The inter-subject variability reported in Table 8 mirrors the findings of previous empirical PE studies such as those mentioned in Section 1.

The slowest translator, TR7 has nevertheless both low HTER and AER. TR7’s log shows they left the CASMACAT interface (by accessing another tab in the browser) and re-accessed CASMACAT an average of over 100 times per text. Without a screen recorder, we do not know what the translator was doing outside CASMACAT, but it is likely that they were engaged in translation-related web searches, possibly because the texts posed comparatively more difficulties for them.

TR8, the second slowest post-editor, has the second highest HTER and the highest AER. Comparing both variables among translators, we see that TR8 has a HTER comparable to that of TR4, yet TR8’s AER is more than double that of TR4. This indicates that TR8 did, on average, considerable overwriting before settling on a final post-edited version, likely slowing them down in the process.

In our set, the two fastest translators, TR6 and TR4, left the CASMACAT interface the fewer number of times of all (5 and 3 times per text, respectively, on average), an indication that they did not need to consult many online translation resources. TR6 and TR4 are not only the fastest but also the most experienced translators of all participants, considering experience both in terms of length (>10 years for both) and translation volume in the preceding 12 months (40,000-55,000 and 25,000-39,900 words, respectively).

The two translators with industry PE certifications, TR6 and TR3, were the first and third fastest post-editors. They produced the texts with the two lowest HTER scores. While differences in the quality of the post-edited output are not statistically significant for translators, TR6 and TR3 produced two of the three translations with the highest Fail/Pass ratios. The second highest Fail/Pass ratio was produced by TR2, the less experienced translator, both in length of experience (2 to 5 years) and translation volume in the 12 preceding months (<10,000 words).

Lastly, we investigated the relationship between translators’ PE time and PE quality by plotting Fail/Pass ratio against mean PE times. We did not find any association between translators’ PE time and PE quality, as evidenced by the lack of pattern in the scatter plot in Figure 3:
6 Conclusions

We presented a study that measured the impact of machine translation quality on post-editor speed and final translation quality. We found a linear relationship between machine translation quality, as measured by the BLEU score of the system, and post-editing speed of about 0.16 seconds/word post-editing time decrease per BLEU point increase. This is about a 3–4% speed increase for each BLEU point. We also found that worse machine translation output ultimately led to worse translation quality after post-editing. As future lines of research, we suggest investigating whether these findings extend to other language pairs.

Acknowledgements

This research was supported by a grant from The University of Auckland’s Faculty of Arts Doctoral Research Fund.

References


