The Impact of Machine Translation Quality on Human Post-editing

Philipp Koehn[◇⋆]
pkoehn@inf.ed.ac.uk
[◇]Center for Speech and Language Processing
The Johns Hopkins University

Ulrich Germann[⋆]
ugermann@inf.ed.ac.uk
[⋆]School of Informatics
University of Edinburgh

Abstract

We investigate the effect of four different competitive machine translation systems on post-editor productivity and behaviour. The study involves four volunteers post-editing automatic translations of news stories from English to German. We see significant difference in productivity due to the systems (about 20%), and even bigger variance between post-editors.

1 Introduction

Statistical machine translation (SMT) has made considerable progress over the past two decades. Numerous recent studies have shown productivity increases with post-editing of MT output over traditional work practices in human translation (e.g., Guerberof, 2009; Plitt and Masselot, 2010; Garcia, 2011; Pouliquen et al., 2011; Skadiņš et al., 2011; den Bogaert and Sutter, 2013; Vazquez et al., 2013; Green et al., 2013; Läubli et al., 2013).

The advances in statistical machine translation over the past years have been driven to a large extent by frequent (friendly) competitive MT evaluation campaigns, such as the shared tasks at the ACL WMT workshop series (Bojar et al., 2013) and IWSLT (Cettolo et al., 2013), and the NIST Open MT Evaluation.[1] These evaluations usually apply a mix of automatic evaluation metrics, most prominently the BLEU score (Papineni et al., 2001), and more subjective human evaluation criteria such as correctness, accuracy, and fluency.

How the quality increases measured by automatic metrics and subjective evaluation criteria relate to actual increases in the productivity of post-editors is still an open research question. It is also not clear yet if some machine translation approaches — say, syntax-based models — are better suited for post-editing than others. These relationships may very well also depend on the language pair in question and the coarse level of MT quality, from barely good enough for post-editing to almost perfect.

The pilot study presented in this paper investigates the influence of the underlying SMT system on post-editing effort and efficiency. The study focuses on translation of general news text from English into German, with translations created by non-professional post-editors working on output from four different translation systems. The data generated by this study is available for download.[2]

We find that the better systems lead to a productivity gain of roughly 20% and carry out in-depth analysis of editing behavior. A significant finding is the high variance in work styles between the different post-editors, compared to the impact of machine translation systems.

2 Related Work

Koponen (2012) examined the relationship between human assessment of post-editing efforts and objective measures such as post-editing time and number of edit operations. She found that segments that require a lot of reordering are perceived as being more difficult, and that long sentences are considered harder, even if only few words changed. She also reports larger variance between translators in post-editing *time* than in post-editing *operations* — a finding that we confirm here as well.

From a detailed analysis of the types of edits performed in sentences with long versus short post-edit times, Koponen et al. (2012) conclude that the observed differences in edit times can be explained at least in part also by the types of necessary edits and the associated cognitive effort. Deleting superfluous function words, for example, appears to be cognitively simple and takes little time, whereas inserting translations for untranslated words requires more cognitive effort

[1] http://www.nist.gov/itl/iad/mig/openmt.cfm

[2] http://www.casmacat.eu/index.php?n=Main.Downloads

Workshop on Humans and Computer-assisted Translation, pages 38–46,
Gothenburg, Sweden, 26 April 2014. ©2014 Association for Computational Linguistics

Table 1: News stories used in the study (size is given in number of sentences)

Source	Size	Title
BBC	49	Norway's rakfisk: Is this the world's smelliest fish?
BBC	47	Mexico's Enrique Pena Nieto faces tough start
CNN	45	Bradley Manning didn't complain about mistreatment, prosecutors contend
CNN	63	My Mexican-American identity crisis
Economist	55	Old battles, new Middle East
Guardian	38	Cigarette plain packaging laws come into force in Australia
NY Times	61	In a Constantly Plugged-In World, It's Not All Bad to Be Bored
NY Times	47	In Colorado, No Playbook for New Marijuana Law
Telegraph	95	Petronella Wyatt: I was bullied out of Oxford for being a Tory

and takes longer. They also compare post-editing styles of different post-editors working on identical post-editing tasks.

Another study by Koponen (2013) showed that inter-translator variance is lower in a controlled language setting when translators are given the choice of output from three different machine translation systems.

In the realm of machine translation research, there has been an increasing interest in the use of MT technology by post-editors. A major push are the two EU-funded research projects MATECAT[3] and CASMACAT[4], which are developing an open source translation and post-editing workbench (Federico et al., 2012; Alabau et al., 2013).

At this point, we are not aware of any study that compares directly the impact of different machine translation systems on post-editor productivity and behaviour.

3 Experimental Design

We thus carried out an experiment on an English–German news translation task, using output from four different SMT systems, post-edited by fluent bilingual native speakers of German with no prior experience in professional translation.

3.1 The Translation Task

The Workshop on Statistical Machine Translation (Bojar et al., 2013) organises an annual evaluation campaign for machine translation systems. The subject matter is translation of news stories from sources such as the New York Times or the BBC. We decided to use output from systems submitted to this evaluation campaign, not only because their output is freely available,[5] but also because it comes with automatic metric scores and human judgements of the translation quality.

The translation direction we chose was English–German, partly due to convenience (the authors of this study are fluent in both languages), but also because this language pair poses special challenges to current machine translation technology, due to the syntactic divergence of the two languages.

We selected data from the most recent evaluation campaign. The subset chosen for our post-editing task comprises 9 different news stories, originally written in English, with a total of 500 sentences. Details are shown in Table 1.

3.2 Machine Translation Systems

A total of 15 different machine translation systems participated in the evaluation campaign. We selected four different systems that differ in their architecture and use of training data:

- an anonymized popular online translation system built by a large Internet company (ONLINE-B)
- the syntax-based translation system of the University of Edinburgh (UEDIN-SYNTAX; Nadejde et al., 2013)
- the phrase-based translation system of the University of Edinburgh (UEDIN-PHRASE; Durrani et al., 2013)
- the machine translation system of the University of Uppsala (UU; Stymne et al., 2013)

In the 2013 WMT evaluation campaign, the systems translated a total of 3000 sentences, and their

[3] http://www.matecat.com/
[4] http://www.casmacat.eu/
[5] http://www.statmt.org/wmt13/results.html

Table 2: Machine translation systems used in the study, with quality scores in the WMT 2013 evaluation campaign.

System	BLEU	SUBJECTIVE
ONLINE-B	20.7	0.637
UEDIN-SYNTAX	19.4	0.614
UEDIN-PHRASE	20.1	0.571
UU	16.1	0.361

output was judged with the BLEU score against a professional reference translation and by subjective ranking. The scores obtained for the different systems on the full test set are shown in Table 2. The first three systems are fairly close in quality (although the differences in subjective human judgement scores are statistically significant), whereas the fourth system (UU) clearly lags behind. The best system ONLINE-B was ranked first according to human judgement and thus can be considered state of the art.

From casual observation, the syntax-based system UEDIN-SYNTAX succeeds more frequently in producing grammatically correct translations. The phrase-based system UEDIN-PHRASE, even though trained on the same parallel data, has higher coverage since it does not have the requirement that translation rules have to match syntactic constituents in the target language, which we presume is the main cause behind the lower BLEU score. The two systems use the same language model.

System UU is also a phrase based system, with a decoder that is able to consider the document level context. It was trained on smaller corpora for both the translation model and the language model.

We do not have any insight into the system ONLINE-B, but we conjecture that it is a phrase-based system with syntactic pre-reordering trained on much larger data sets, but not optimised towards the news domain.

Notice the inconsistency between BLEU score and subjective score for the two systems from the University of Edinburgh. Results from other evaluations have also shown (Callison-Burch et al., 2012) that current automatic evaluation metrics do not as much as human judges appreciate the strengths of the syntax-based system, which builds syntactic structures in the target language during translation. Hence, we were particularly interested how the syntax-based system fares with post-editors.

As mentioned above, the nine documents chosen for the post-editing task analysed in this paper (cf. Table 1) were part of the WMT 2013 evaluation data set. All nine documents had English as the original source language.

3.3 Post-Editors

We recruited four English-German bilingual, native German post-editors. Three were students, staff, or faculty at the University of Edinburgh; the fourth had been previously employed on a contractual basis for linguistic annotation work.[6] The post-editors had no professional experience with translation, and differed in language skills.

3.4 Assignment of MT Output

The goal of this study was to investigate how post-editors' behaviour and productivity are influenced by the quality of the underlying machine translation system. Ideally, we would want to present output from different systems to the same post-editor and see how their observable behaviour changes.

However, a post-editor who has seen the output from one MT system for a sentence will be at an advantage when post-editing the output from a second system, by having already spent significant time understanding the source sentence and considering the best translation choices.

Hence we used 4 different post-editors, each to post-edit the output in equal amounts from each of the 4 machine translation systems under investigation, so that each post-editor worked on each sentence once and the entire output from all systems was post-edited once by one of the 4 post-editors.

A concern in this setup is that we never know if we measure differences in post-editors or differences in machine translations systems when comparing the behaviour for any given sentence.

Therefore, each post-editor was assigned a translation for each sentence randomly from any of the machine translation systems. This random assignment allows us to marginalise out the dependence on the post-editor when assessing statistics for the different systems.

[6] The ordering here does not reflect the order of post-editors in the discussion later in this paper.

Table 3: Post-editing speed by editor and system.

System	seconds / word					words / hour				
	1	2	3	4	**mean**	1	2	3	4	**mean**
ONLINE-B	2.95	4.69	9.16	4.98	**5.46**	1,220	768	393	723	**659**
UEDIN-PHRASE	3.04	5.01	9.22	4.70	**5.45**	1,184	719	390	766	**661**
UEDIN-SYNTAX	3.03	4.41	9.20	4.97	**5.38**	1,188	816	391	724	**669**
UU	3.11	5.01	11.59	5.58	**6.35**	1,158	719	311	645	**567**
mean per editor	**3.03**	**4.78**	**9.79**	**5.05**		**1,188**	**753**	**368**	**713**	

4 Productivity

The primary argument for post-editing machine translation output as opposed to more traditional approaches is the potential gain in productivity. If translation professionals can work faster with machine translation, then this has real economic benefits. There are also other considerations, for example that post-editing might be done by professionals that are less skilled in the source language (Koehn, 2010).

We measure productivity by time spent on each sentence. This is not a perfect measure. When working on a news story, post-editors tend to speed up when moving down the story since they have already solved some reoccurring translation problems and get more familiar with the context.

4.1 Productivity by MT System

Our main interests is the average translation speed, broken down by machine translation system. The columns labelled "mean" in Table 3 show the results. While the differences are not big for the top three systems, the syntax-based system comes out on top.

We used bootstrap resampling to test the speed differences for statistical significance. Only system UU is significantly worse than the others (at p-level < 0.01), with about 20% lower productivity.

4.2 Productivity by Post-Editor

Post-editing speed is very strongly influenced by the post-editor's skill and effort. Our post-editors were very diverse, showing large differences in translation speed. See the columns labelled 1 to 4 in Table 3 for details.

In particular, post-editor 3 took more than three times as much time as the fastest (PE 1). According to a post-study interview with Post-Editor 3, there were two reasons for this. First, the post-editor was feeling a bit "under the weather" during the study and found it hard to focus. Second, (s)he found the texts very difficult to translate and struggled with idiomatic expressions and cultural references that (s)he did not understand immediately.

4.3 Productivity by System and Post-Editor

While the large differences between the post-editors are unfortunate when the goal is consistency in results, they provide some data on how post-editors of different skill levels are influenced by the quality of the machine translation systems.

Table 3 breaks down translation speed by machine translation system and post-editor. Interestingly, machine translation quality has hardly any effect on the fast Post-Editor 1, and the lower MT performance of system UU affects only Post-Editors 3 and 4. Post-Editor 2 is noticeably faster with UEDIN-SYNTAX — an effect that cannot be observed for the other post-editors. The differences between the other systems are not large for any of the post-editors.

Statistically significant — as determined by bootstrap resampling — are only the differences in post-editing speed for Post-Editor 3 with system UU versus ONLINE-B and UEDIN-PHRASE at p-level < 0.01, and against UEDIN-SYNTAX at p-level <0.02, and for Post-Editor 4 for UU versus UEDIN-PHRASE at p-level < 0.05. Note that the absence of statistical significance in our data has much to do with the small sample size; more extensive experiments may be necessary to ensure more solid findings.

5 Translation Edit Rate

Given the inherent difficulties in obtaining timing information, we can also measure the impact of machine translation system quality on post-editing effort in terms of how much the post-editors change the machine translation output, as done, for example in Cettolo et al. (2013).

Table 4: Edit rate and types of edits per system

System	HTER	ins	del	sub	shift	wide shift
ONLINE-B	35.7	4.8	7.4	18.9	4.6	5.8
UEDIN-PHRASE	37.9	5.5	7.4	20.0	5.0	6.6
UEDIN-SYNTAX	36.7	4.7	7.6	19.8	4.6	5.7
UU	43.7	4.6	11.4	21.9	5.8	7.2

Table 5: Edit rate and types of edits per post-editor

P-E	HTER	ins	del	sub	shift	wide shift
1	35.2	5.4	6.7	18.7	4.4	5.3
2	43.1	4.1	10.4	23.1	5.4	6.9
3	37.7	5.9	7.9	18.8	5.0	6.6
4	37.5	4.3	8.5	19.6	5.1	6.4

There are two ways to measure how much the machine translation output was edited by the post-editor. One way is to compare the final translation with the original machine translation output. This is what we will do in this section. In Section 6, we will consider which parts of the final translation were actually changed by the post-editor and discuss the difference.

5.1 HTER as Quality Measure

The edit distance between machine translation output and human reference translation can be measured in the number of insertions, deletions, substitutions and (phrasal) moves. A metric that simply counts the minimal number of such edit operations and divides it by the length of the human reference translation is the *translation edit rate*, short TER (Snover et al., 2006).

If the human reference translation is created from the machine translation output to minimise the number of edit operations needed for an acceptable translation, this variant is called *human-mediated* TER, or HTER. Note that in our experiment the post-editors are not strictly trying to minimise the number of edit operations — they may be inclined to make additional changes due to arbitrary considerations of style or perform edits that are faster rather than minimise the number of operations (e.g., deleting whole passages and rewriting them).

5.2 Edits by MT System

Table 4 shows the HTER scores — keep in mind our desiderata above — for the four systems. The scores are similar to the productivity number, with the three leading systems close together and the trailing system UU well behind.

Notably, we draw more statistically significant distinctions here. While as above, UU is significantly worse than all other systems (p-level $<$ 0.01), we also find that ONLINE-B is better than UEDIN-PHRASE (p-level $<$ 0.01).

Hence, HTER is a more sensitive metric than translation speed. This may be due to the fact that the time measurements are noisier than the count of edit operations. But it may also because HTER and productivity (i.e., time) do not measure the exactly the same thing. For instance, edits that require only a few keystrokes may be cognitively demanding (e.g., terminological choices), and thus take more time.

We cannot make any strong claim based on our numbers, but it is worth pointing out that post-editing UEDIN-SYNTAX was slightly faster than ONLINE-B (by 0.08 seconds/word), while the HTER score is lower (by 1 point). A closer look at the edit operations reveals that the post-edit of UEDIN-SYNTAX output required slightly fewer short and long shifts (movements of phrases), but more substitutions. Intuitively, moving a phrase around is a more time-consuming task than replacing a word. The benefit of a syntax-based system that aims to produce correct syntactic structure (including word order), may have real benefits in terms of post-editing time.

5.3 Edits by Post-Editor

Table 5 displays the edit rate broken down by post-editor. There is little correlation between edit rate and post-editor speed. While the fastest Post-Editor 1 produces translations with the smallest edit rate, the difference to two of the others (included the slowest Post-Editor 3) is not large. The

sec	current_translation
0.000	Norwegen ist es nicht oft auf die globale Agenda - und die meisten Nachrichten, wie es scheint.
0.508	Norwegen ist esnicht oft auf die globale Agenda - und die meisten Nachrichten, wie es scheint.
0.733	Norwegen ist enicht oft auf die globale Agenda - und die meisten Nachrichten, wie es scheint.
0.970	Norwegen ist enicht oft auf die globale Agenda - und die meisten Nachrichten, wie es scheint.
0.975	Norwegen ist nicht oft auf die globale Agenda - und die meisten Nachrichten, wie es scheint.
3.317	Norwegen ist nicht oft auf di globale Agenda - und die meisten Nachrichten, wie es scheint.
3.413	Norwegen ist nicht oft auf d globale Agenda - und die meisten Nachrichten, wie es scheint.
3.524	Norwegen ist nicht oft auf de globale Agenda - und die meisten Nachrichten, wie es scheint.
3.588	Norwegen ist nicht oft auf der globale Agenda - und die meisten Nachrichten, wie es scheint.
5.116	Norwegen ist nicht oft auf der globalen Agenda - und die meisten Nachrichten, wie es scheint.
17.986	Norwegen ist nicht oft auf der globalen Agenda - und die meisten , wie es scheint.
19.386	Norwegen ist nicht oft auf der globalen NachrichtenAgenda - und die meisten , wie es scheint.
20.116	Norwegen ist nicht oft auf der globalen Nachrichtengenda - und die meisten , wie es scheint.
20.196	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten , wie es scheint.
20.298	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten , wie es scheint.
29.596	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten scheint.
31.459	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten schein.
31.659	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten scheine.
31.796	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten scheinen.
32.060	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten scheinen .
34.283	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten scheinen d.
34.380	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten scheinen da.
34.443	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten scheinen das.
34.636	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten scheinen das .
35.507	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten scheinen das s.
35.637	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten scheinen das so.
35.744	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten scheinen das so .
35.949	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten scheinen das so z.
36.053	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten scheinen das so zu.
36.166	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten scheinen das so zu .
36.805	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten scheinen das so zu w.
38.011	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten scheinen das so zu wo.
38.394	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten scheinen das so zu wol.
38.699	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten scheinen das so zu woll.
38.795	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten scheinen das so zu wolle.
38.947	Norwegen ist nicht oft auf der globalen Nachrichtenagenda - und die meisten scheinen das so zu wollen.
char	mmmmmmmmmmmmmmmmmmmmmmmmmmmmmmttmmmmmmmmtmpppppppppppptmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmttttttttttttttttttttttm
word	m m m m m x x x m m m m x t t t t m

Figure 1: Analysis of the post-editing process: Most editing actions that result in changes in the translation are adding or deleting of a character by a keystroke. However, notice the cut (17.986 sec) and paste (19.386 sec) action. Each character can be traced either to the original machine translation output (m), a typing action of the post-editor (t), or a pasting action of the post-editor (p). This then allows tokens in the output classified as either original MT (m), typed (t), pasted (not in figure) or partially edited (x).

outlier here is Post-Editor 2, whose output has a much larger edit rate.

6 Editing Actions

The HTER is an analysis of the *product* of post-editing. The final translation is compared to the original machine translation output. In this section, we examine how the *process* of post-editing is influenced by the machine translation system.

Our post-editing workbench provides detailed logging of each HCI interaction (key strokes, mouse clicks, etc.). This allows us to reconstruct how a translation was generated. See Figure 1 for an example how a translated was edited.

Table 6: Character provenance by system

System	MT	typed	pasted
ONLINE-B	68.3	28.0	3.3
UEDIN-PHRASE	62.9	31.3	5.2
UEDIN-SYNTAX	65.9	29.1	4.5
UU	56.1	37.9	5.6

6.1 Character Provenance by System

If we follow the editing actions, we can trace the origin of each letter in the final output: was it part of the original MT output, was it typed in by the user, or moved in a cut and paste action? Table 6 breaks down the characters in the final translations

Table 7: Token provenance by system

System	MT	typed	pasted	edited
ONLINE-B	65.2	21.4	2.3	10.8
UEDIN-PHRASE	60.5	24.7	3.9	10.6
UEDIN-SYNTAX	62.6	22.4	3.4	11.3
UU	53.2	31.0	4.0	11.7

by origin for each system. The numbers correspond to the HTER scores, with a remarkable consistency ranking for typed and pasted characters.

6.2 Token Provenance by System

We perform a similar analysis on the word level, introducing a fourth type of provenance: words whose characters are of mixed origin, i.e., words that were partially edited. Table 7 shows the numbers for each machine translation system. The suspicion from the HTER score that the syntax-based system UEDIN-SYNTAX requires less movement is not confirmed by these numbers. There are significantly more words moved by pasting (3.4%) than for ONLINE-B (2.3%). In general, cutting and pasting is not as common as the HTER score would suggest: the two types of shifts moved 10.3% and 10.2% of phrases, respectively. It seems that most words that could be moved are rather deleted and typed again.

6.3 Behaviour By Post-Editor

The post-editors differ significantly in their behaviour, as the numbers in Table 8 illustrate. Post-Editor 1, who is the fastest, leaves the most characters unchanged (72.9% vs. 57.7–64.4% for the others). Remarkably, this did not result in a dramatically lower HTER score (recall: 35.2 vs. 37.5–43.1 for the others).

Post-Editor 3, while taking the longest time, does not change the most number of characters. However, (s)he uses dramatically more cutting and pasting. Is this activity particularly slow? One way to check is to examine more closely how the post-editors spread out their actions over time.

Table 8: Character provenance by post-editor

Post-Editor	MT	typed	pasted
1	72.9	22.9	3.5
2	57.7	39.4	2.7
3	58.9	29.5	10.7
4	64.4	33.5	1.9

7 Editing Activities

Koehn (2009) suggests to divide up the time spent by translators and post-editors into intervals of the following types:

- initial pauses: the pause at the beginning of the translation, if it exists
- end pause: the pause at the end of the translation, if it exists
- short pause of length 2–6 seconds
- medium pauses of length 6–60 seconds
- big pauses longer than 60 seconds
- various working activities (in our case just typing and mouse actions)

When we break up the time spent on each activity and normalise it by the number of words in the original machine translation output, we get the numbers in Table 9, per machine translation system and post-editor.

The worse quality of the UU system causes mainly more work activity, big medium pauses. Each contributes roughly 0.3 seconds per word. The syntax-based system UEDIN-SYNTAX may pose fewer hard translation problems (showing up in initial and big pauses) than the HTER-preferred ONLINE-B system, but the effect is not strong.

We noted that ONLINE-B has a statistically significant better HTER score than UEDIN-PHRASE. While this is reflected in the additional working activity for the latter (2.41 sec./word vs. 2.26 sec./word), time is made up in the pauses. Our data is not sufficiently conclusive to gain any deeper insight here — it is certainly a question that we want to explore in the future.

The difference in post-editors mirrors some of the earlier findings: The number of characters and words changed leads to longer working activity, but the slow Post-Editor 3 is mainly slowed down by initial, big and medium pauses, indicating difficulties with solving translation problems, and not slow cutting and pasting actions. The faster Post-Editor 1 rarely pauses long and is quick with typing and mouse movements.

8 Conclusion

We compared how four different machine translation systems affect post-editing productivity and behaviour by analysing final translations and user

Table 9: Time spent on different activities, by machine translation system (top) and post-editor (bottom).

System	initial pause	big pause	med. pause	short pause	end pause	working
ONLINE-B	0.37 s/w	0.61 s/w	1.88 s/w	0.30 s/w	0.00 s/w	2.26 s/w
UEDIN-PHRASE	0.32 s/w	0.55 s/w	1.74 s/w	0.32 s/w	0.00 s/w	2.41 s/w
UEDIN-SYNTAX	0.32 s/w	0.50 s/w	1.90 s/w	0.31 s/w	0.00 s/w	2.30 s/w
UU	0.28 s/w	0.74 s/w	2.14 s/w	0.34 s/w	0.00 s/w	2.75 s/w

Post-Editor	initial pause	big pause	med. pause	short pause	end pause	working
1	0.35 s/w	0.01 s/w	0.63 s/w	0.27 s/w	0.00 s/w	1.76 s/w
2	0.04 s/w	0.19 s/w	1.13 s/w	0.35 s/w	0.00 s/w	3.06 s/w
3	0.91 s/w	1.85 s/w	3.99 s/w	0.29 s/w	0.00 s/w	2.53 s/w
4	0.02 s/w	0.36 s/w	1.94 s/w	0.35 s/w	0.00 s/w	2.33 s/w

activity data. The best system under consideration yielded abut 20% better productivity than the worst, although the three systems on top are not statistically significantly different in terms of productivity.

We noted differences in metrics that measure productivity and edit distance metrics. The latter allowed us to draw more statistically significant conclusions, but may measure something distinct. Productivity is the main concern of commercial use of post-editing machine translation, and we find that better machine translation leads to less time spent on editing, but more importantly, less time spent of figuring out harder translation problems (indicated by pauses of more than six seconds).

Finally, an important finding is that the differences between post-editors is much larger than the difference between machine translation systems. This points towards the importance of skilled post-editors, but this finding should be validated with professional post-editors, and not the volunteers used in this study.

Acknowledgements

This work was supported under the CASMACAT project (grant agreement N^{o} 287576) by the European Union 7th Framework Programme (FP7/2007-2013).

References

Alabau, Vicent, Ragnar Bonk, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Jesús González, Philipp Koehn, Luis Leiva, Bartolomé Mesa-Lao, Daniel Ortiz, Herve Saint-Amand, Germán Sanchis, and Chara Tsoukala. 2013. "CASMACAT: An open source workbench for advanced computer aided translation." *The Prague Bulletin of Mathematical Linguistics*, 100:101–112.

Bojar, Ondřej, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. "Findings of the 2013 Workshop on Statistical Machine Translation." *Proceedings of the Eighth Workshop on Statistical Machine Translation*, 1–44. Sofia, Bulgaria.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. "Findings of the 2012 workshop on statistical machine translation." *Proceedings of the Seventh Workshop on Statistical Machine Translation*, 10–48. Montreal, Canada.

Cettolo, Mauro, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2013. "Report on the 10th IWSLT evaluation campaign." *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

den Bogaert, Joachim Van and Nathalie De Sutter. 2013. "Productivity or quality? Let's do both." *Machine Translation Summit XIV*, 381–390.

Durrani, Nadir, Barry Haddow, Kenneth Heafield, and Philipp Koehn. 2013. "Edinburgh's machine translation systems for European language pairs." *Proceedings of the Eighth Workshop on Statistical Machine Translation*, 114–121. Sofia, Bulgaria.

Federico, Marcello, Alessandro Cattelan, and Marco Trombetti. 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)*.

Garcia, Ignacio. 2011. "Translating by post-editing: is it the way forward?" *Machine Translation*, 25(3):217–237.

Green, Spence, Jeffrey Heer, and Christopher D. Manning. 2013. "The efficacy of human post-editing for language translation." *ACM Human Factors in Computing Systems (CHI)*.

Guerberof, Ana. 2009. "Productivity and quality in mt post-editing." *MT Summit Workshop on New Tools for Translators*.

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

Koehn, Philipp. 2010. "Enabling monolingual translators: Post-editing vs. options." *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 537–545. Los Angeles, California.

Koponen, Maarit. 2012. "Comparing human perceptions of post-editing effort with post-editing operations." *Pro-*

ceedings of the Seventh Workshop on Statistical Machine Translation, 227–236. Montreal, Canada.

Koponen, Maarit. 2013. "This translation is not too bad: an analysis of post-editor choices in a machine-translation post-editing task." *Proceedings of Workshop on Post-editing Technology and Practice*, 1–9.

Koponen, Maarit, Wilker Aziz, Luciana Ramos, and Lucia Specia. 2012. "Post-editing time as a measure of cognitive effort ." *AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*, 11–20. San Diego, USA.

Läubli, Samuel, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, and Martin Volk. 2013. "Assessing post-editing efficiency in a realistic translation environment." *Proceedings of Workshop on Post-editing Technology and Practice*, 83–91.

Nadejde, Maria, Philip Williams, and Philipp Koehn. 2013. "Edinburgh's syntax-based machine translation systems." *Proceedings of the Eighth Workshop on Statistical Machine Translation*, 170–176. Sofia, Bulgaria.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. *BLEU: a Method for Automatic Evaluation of Machine Translation*. Tech. Rep. RC22176(W0109-022), IBM Research Report.

Plitt, Mirko and Francois Masselot. 2010. "A productivity test of statistical machine translation post-editing in a typical localisation context." *Prague Bulletin of Mathematical Linguistics*, 93:7–16.

Pouliquen, Bruno, Christophe Mazenc, and Aldo Iorio. 2011. "Tapta: A user-driven translation system for patent documents based on domain-aware statistical machine translation." *Proceedings of th 15th International Conference of the European Association for Machine Translation (EAMT)*, 5–12.

Skadiņš, Raivis, Maris Puriņš, Inguna Skadiņa, and Andrejs Vasiļjevs. 2011. "Evaluation of SMT in localization to under-resourced inflected language." *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT)*, 35–40.

Snover, Matthew, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. "A study of translation edit rate with targeted human annotation." *5th Conference of the Association for Machine Translation in the Americas (AMTA)*. Boston, Massachusetts.

Stymne, Sara, Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2013. "Tunable distortion limits and corpus cleaning for SMT." *Proceedings of the Eighth Workshop on Statistical Machine Translation*, 225–231. Sofia, Bulgaria.

Vazquez, Lucia Morado, Silvia Rodriguez Vazquez, and Pierrette Bouillon. 2013. "Comparing forum data post-editing performance using translation memory and machine translation output: A pilot study." *Machine Translation Summit XIV*, 249–256.