What Can We Really Learn from Post-editing?

Marcis Pinnis marcis.pinnis@tilde.com
Tilde, Riga, Latvia
Rihards Kalnins rihards.kalnins@tilde.com
Tilde, Riga, Latvia
Raivis Skadins raivis.skadins@tilde.com
Tilde, Riga, Latvia
Inguna Skadina inguna.skadina@tilde.com
Tilde, Riga, Latvia

Abstract

The paper describes findings of a large post-editing project in the medical domain carried out by Tilde. It analyzes the efficacy of post-editing of highly technical texts in a specialized domain and provides answers to questions important to localization service providers that consider the introduction of post-editing in their translation workflows. The results show that by carefully analyzing post-editing projects, machine translation providers and language service providers can learn how to boost productivity in localization, save time and optimize resources in the language editing process, as well as to leverage quality post-edits to improve machine translation engines through dynamic learning.

1. Introduction

In order to analyze the efficacy of post-editing of highly technical texts in a specialized domain, Tilde embarked on a project to analyze a large post-editing effort in the medical domain. During the project, Tilde had the unique opportunity to take detailed logs of each activity performed by post-editors. Tilde then analyzed the post-editing results, allowing us to answer important questions like: 1) How effectively do post-editors really work with machine translation (MT)? 2) Do post-editors expend their efforts usefully on editing MT results? 3) How can MT be improved to meet the needs of localization companies that utilize post-editing to boost translation productivity? 4) How does the MT quality affect post-editing performance?

2. MT System

During the course of the project, post-editors used a statistical MT (SMT) system that was based on the phrase-based Moses SMT system (Koehn et al., 2007). The system was trained on the European Medicines Agency (EMEA) parallel corpus from OPUS corpus (Tiedemann, 2009) and latest documents from EMEA website (years 2009-2014) collected by Tilde on the Tilde MT platform (Vasiļjevs et al., 2012). The statistics of the training corpus before and after filtering are given in Table 1. The system's automatic evaluation results are given in Table 2.

Corpus	Sentences before filtering	Sentences after filtering
Parallel	378,869	325,332
Monolingual	378,869	332,652

Table 1: Statistics of the training corpora used to train the SMT system

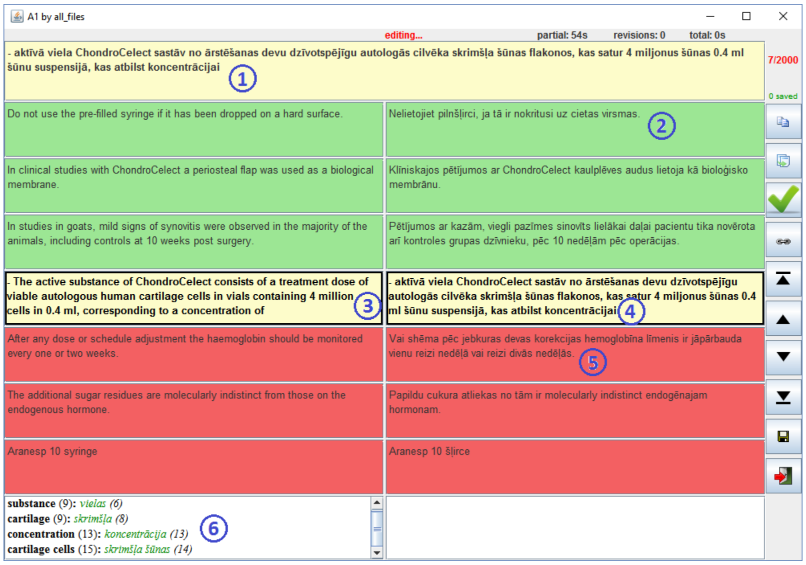
Evaluation scenario	BLEU	NIST	METEOR	TER
Case sensitive	47.42 (45.82-48.88)	9.5300 (9.3469-9.7027)	0.3637	0.3952
Case insensitive	45.79 (44.23-47.26)	9.2735 (9.1036-9.4539)	0.2575	0.4105

Table 2: Automatic evaluation results of the SMT system

3. Post-editing Task

The post-editing task was performed using the tool PET (Aziz et al., 2012), which is able to precisely track the time spent on each segment and all keystrokes that a post-editor performs while post-editing each segment. An example of the graphical user interface of PET as used in the post-editing task is given in Figure 1. The whole post-editing task, which contained 22,500 (360,000 words) sentences, was split into jobs that consisted of 100 sentences. All jobs contained consecutive sentences from the latest documents. All jobs were pre-translated with the SMT system prior to giving the jobs to post-editors, so that translators would not have to wait for the SMT suggestions to appear.

While post-editing texts, translators were asked to evaluate the quality of each MT suggestion, marking it as one of the following: “near perfect,” “very good,” “poor,” and “very poor.” If the translator did not apply changes to the MT suggestion, the post-editing tool automatically rated it as “unchanged”, which means that the MT suggestion was perfect and did not require any post-editing.



A1 by all_files

editing... partial: 54s revisions: 0 total: 0s

- aktīvā viela ChondroCelect sastāv no ārstēšanas devu dzīvotspējīgu autologās cilvēka skrimšļa šūnas flakonos, kas satur 4 miljonus šūnas 0.4 ml šūnu suspensijā, kas atbilst koncentrācijai

7/2000

0 saved

Do not use the pre-filled syringe if it has been dropped on a hard surface.	Nelietojiet pilnšļirci, ja tā ir nokritusi uz cietas virsmas.
In clinical studies with ChondroCelect a periosteal flap was used as a biological membrane.	Klīniskajos pētījumos ar ChondroCelect kaulplēves audus lietoja kā bioloģisko membrānu.
In studies in goats, mild signs of synovitis were observed in the majority of the animals, including controls at 10 weeks post surgery.	Pētījumos ar kazām, viegli pazīmes sinovīts lielākai daļai pacientu tika novērota arī kontroles grupas dzīvnieku, pēc 10 nedēļām pēc operācijas.
- The active substance of ChondroCelect consists of a treatment dose of viable autologous human cartilage cells in vials containing 4 million cells in 0.4 ml, corresponding to a concentration of	- aktīvā viela ChondroCelect sastāv no ārstēšanas devu dzīvotspējīgu autologās cilvēka skrimšļa šūnas flakonos, kas satur 4 miljonus šūnas 0.4 ml šūnu suspensijā, kas atbilst koncentrācijai
After any dose or schedule adjustment the haemoglobin should be monitored every one or two weeks.	Vai shēma pēc jebkuras devas korekcijas hemoglobīna līmenis ir jāpārbauda vienu reizi nedēļā vai reizi divās nedēļās.
The additional sugar residues are molecularly indistinct from those on the endogenous hormone.	Papildu cukura atliekas no tām ir molecularly indistinct endogēnajam hormonam.
Aranesp 10 syringe	Aranesp 10 šļirce

substance (9): *vielas (6)*
cartilage (9): *skrimšļa (8)*
concentration (13): *koncentrācija (13)*
cartilage cells (15): *skrimšļa šūnas (14)*

Figure 1: Example of the user interface of the PET post-editing tool showing: 1) the MT suggestion; 2) previous context; 3) the source text of the segment that is being post-edited; 4) the target editing field (showing the MT suggestion); 5) the further context; and 6) the entries of the term collection that are found in the source text of the current segment

A total of five professional translators worked on the post-editing project full-time for approximately five weeks in total. The post-editors were also asked not to spend excessive amounts of time on each segment, as the quality expectations were not “human translation quality” but rather “post-editing quality.” To assist post-editing, post-editors were provided with an in-domain term collection that was integrated in the post-editing tool and automatically showed translation suggestions for known terms.

The detailed logs of each translator's work recorded the timing of each keystroke, measuring the time spent on post-editing in three distinct intervals: the amount of time that elapsed between the appearance of a MT segment and the first click, or "reading time"; the amount of time between the first edit and approving the segment, or "editing time"; and the amount of time spent between approving the segment and completing the assessment of the quality by clicking the "Finish" button, referred to as "assessment time."

4. Preliminary Results

The results showed that the use of custom MT resulted in a considerable boost in overall translation productivity (see Table 3). The translators' average translation speed for human translation of medical domain texts is approximately 800-900 tokens per hour (pure translation time, not counting pauses between sentences). But MT succeeded in boosting the average translation productivity to 2,694 tokens per hour – an approximately 200% increase.

This strong boost in productivity came about thanks to the high translation quality of the MT system used by post-editors (BLEU score of 47.42). The analysis showed that the MT system produced a majority of MT segments – over 37% – that were marked "unchanged," demanding no editing time at all from the post-editors.

MT suggestion assessment	**Total editing time**	**Total source length in tokens**	**Segment count**	**Productivity (tokens post-edited in one hour)**
0. Unchanged	14:15:43	83,661	5,488	5,865
1. Near perfect	12:12:01	46,108	2,458	3,779
2. Very good	44:11:31	102,309	4,962	2,315
3. Poor	26:40:50	37,956	1,717	1,422
4. Very poor	04:13:39	3,582	175	847
Grand Total	101:33:46	273,616	14,800	2,694

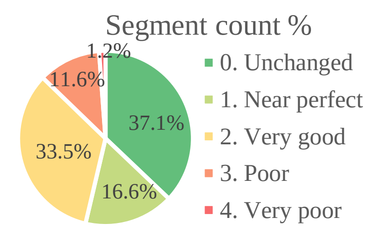


Table 3: Sum of editing time and productivity gains

Not surprisingly, the average amount of post-editing time for each segment rose incrementally as the quality of the MT result (as marked by the post-editor) declined (see Figure 2). A "near perfect" segment had an average of 19.23 seconds of post-editing; a "very good" segment had an average of 34.80 seconds; a "poor" segment demanded 59.55 seconds of the post-editor's time; and a "very poor" segment clocked in at 90.98 seconds of work.

Though the editing time grew incrementally, the total "reading time" on the part of post-editors grew more gradually. Even an "unchanged" segment demanded an average of 10 seconds of reading time from post-editors. This figure should be kept in mind by localization companies that want to increase their MT use: even an unchanged segment, with perfect MT quality, demands 10 seconds of a post-editor's time for review.

The most surprising results of the study came when graphing the relationship between the quality of an MT suggestion and the PE quality of the segment *vi-à-vis* a reference human translation (see Figure 3). We found that segments with "near perfect" or "very good" MT quality led to the creation of post-edited texts with Translation Edit Rate (TER) scores that ranked fairly consistently in relation to the reference translation. For instance, an MT segment with "near perfect" quality produced, on average, a post-edit that had TER scores of 0.22-0.30 in relation to the reference translation.

However, when MT produced “poor” and “very poor” segments, the post-editing quality was furthest from the reference translation. Post-edits of “poor” and “very poor” segments could have TER scores that ranged from 0.50 to 1.00. One possible reason for the wide range in scoring was that the target language in the project was a morphologically rich, highly inflected language with relatively free word order. Therefore, thanks to the liberal syntax, post-editors had a wider range of options for constructing post-edited sentences. This led to grave inconsistency in results for post-edits that demanded the most efforts, namely, edits of “poor” and “very poor” segments.

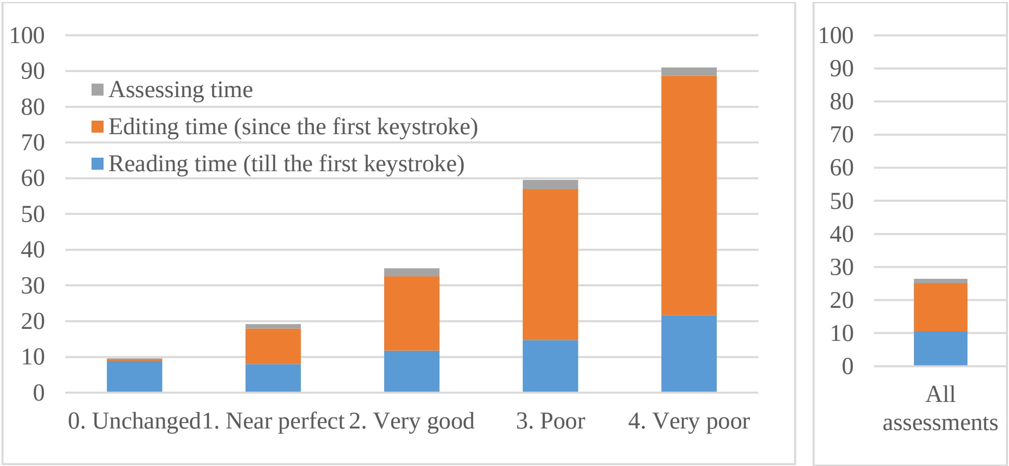


Figure 2: Average of reading, editing, and assessment times for segments with different MT suggestion quality assessments (left) and all segments (right)

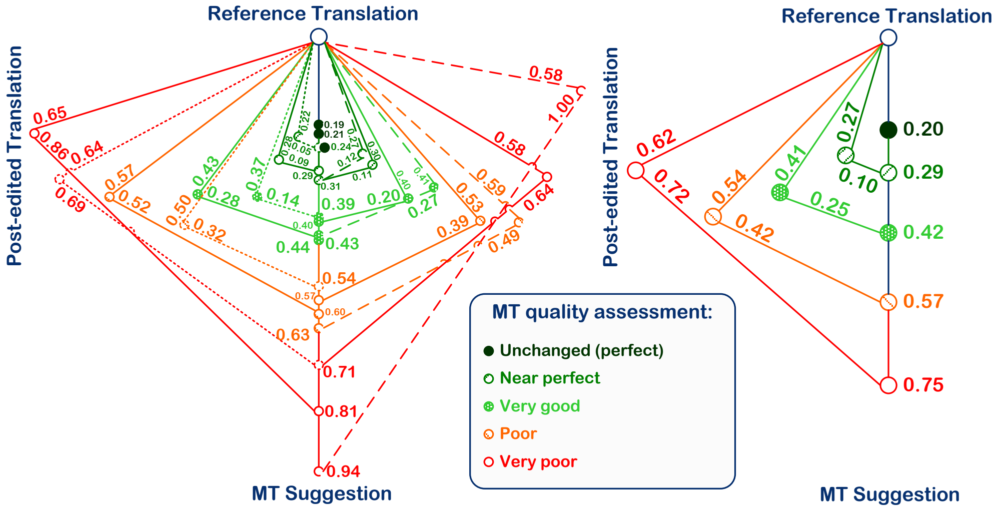


Figure 3: Translation edit rates between reference translations, MT suggestions, and post-edited translations (left – for 4 individual translators, right – for all translators) depending on different MT suggestion quality assessments

Inconsistent post-editing quality poses a serious problem for language editors and linguists at localization companies, who must expend extra effort in establishing linguistic uniformity in a text. This problem is further compounded for post-editing projects that utilize multiple post-editors.

5. Key Findings

By carefully tracking the work of post-editors during this project, we were able to make four concrete findings:

- Post-editing resulted in a huge increase in translation productivity (200%)
- The amount of time spent on post-editing a segment is, on average, directly proportional to the quality of the MT
- High quality MT results lead to relatively consistent post-editing quality
- Poor quality MT leads to a high degree of inconsistency between post-editing quality and a perfect human translation

6. Conclusion: What Does This Mean for Language Service Providers?

Most importantly, however, these results have allowed us to offer several recommendations for localization service providers utilizing MT in the post-editing process.

First, post-editing projects must be carefully tracked in the CAT tool environment. By asking post-editors to rank the quality of a segment – an activity that takes up only a fractional interval of editing time, approximately 3% – much insight can be gained and then applied to the final language editing process.

As many segments in post-editing will remain "unchanged" or just slightly changed – that is, if the quality of the MT system is high – language editors who are alerted to the quality of segments can safely accept these post-edits without expending any additional language editing efforts. However, taking into account the great inconsistency in post-editing of "poor" and "very poor" segments, language editors should expend extra effort on editing these segments in order to ensure that they conform to the overall stylistic quality of the text.

Second, the results of the finding also illuminate the ways in which Dynamic Learning can improve MT quality. MT results that are marked "near perfect" and "very good" produce relatively high quality post-edits, therefore these post-edits can safely be used to dynamically improve MT engines through the Dynamic Learning function. However, inconsistent post-editing quality, as produced from "poor" quality MT can severely pollute the quality of a MT system and should be removed from dynamic improvement to the MT engine.

Therefore, logging a post-editor's quality assessment or editing time of MT suggestions can also help improve the quality of engines with Dynamic Learning. LSPs and their MT vendors should only allow post-edited segments of "near perfect" and "very good" quality to be used to dynamically improve the underlying MT engine.

By carefully analyzing post-editing projects, MT providers and LSPs can learn how to boost productivity in localization, save time and optimize resources in the language editing process, as well as to leverage quality post-edits to improve MT engines through Dynamic Learning. Only in this way will LSPs be enabled to meet the booming volumes of translation – up to a 67% increase (Lommel, 2016) – that they can expect from enterprises in the next few years.

Acknowledgements

The work within the QT21 project has received funding from the European Union under grant agreement n° 645452.

References

Aziz, W.; Sousa, S. C. M.; Specia, L. (2012). PET: A Tool for Post-editing and Assessing Machine Translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC 2012), Istanbul, Turkey. May 2012.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., … Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (pp. 177--180), Stroudsburg, PA, USA: Association for Computational Linguistics.

Tiedemann, J. (2009). News from OPUS-A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (Vol. 5, pp. 237-248).

Vasiļjevs, A., Skadiņš, R., & Tiedemann, J. (2012). LetsMT!: A Cloud-Based Platform for Do-It-Yourself Machine Translation. In *Proceedings of the ACL 2012 System Demonstrations* (pp. 43–48), Jeju Island, Korea: Association for Computational Linguistics.

Lommel, A., DePalma, D. (2016). "Europe's Leading Role in Machine Translation: How Europe Is Driving the Shift to MT." Common Sense Advisory Report. http://cracker-project.eu/csa-mt-report/