Quantifying the Influence of MT Output in the Translators' Performance: A Case Study in Technical Translation

Marcos Zampieri
Saarland University
Saarbrücken, Germany
mzampier@uni-koeln.de

Mihaela Vela
Saarland University
Saarbrücken, Germany
m.vela@mx.uni-saarland.de

Abstract

This paper presents experiments on the use of machine translation output for technical translation. MT output was used to produced translation memories that were used with a commercial CAT tool. Our experiments investigate the impact of the use of different translation memories containing MT output in translations' quality and speed compared to the same task without the use of translation memory. We evaluated the performance of 15 novice translators translating technical English texts into German. Results suggest that translators are on average over 28% faster when using TM.

1 Introduction

Professional translators use a number of tools to increase the consistency, quality and speed of their work. Some of these tools include spell checkers, text processing software, terminological databases and others. Among all tools used by professional translators the most important of them nowadays are translation memory (TM) software. TM software use parallel corpora of previously translated examples to serve as models for new translations. Translators then validate or correct previously translated segments and translate new ones increasing the size of the memory after each new translated segment.

One of the great issues in working with TMs is to produce the TM itself. This can be time consuming and the memory should ideally contain a good amount of translated segments to be considered useful and accurate. For this reason, many novice translators do not see the benefits of the use of TM right at the beginning, although it is consensual that on the long run the use of TMs increase the quality and speed of their work. To cope with this limitation, more TM software have provided interface to machine translation (MT) software. MT output can be used to suggest new segments that were not previously translated by a human translator but generated automatically from an MT software. But how helpful are these translations?

To answer this question, the experiments proposed in this paper focus on the translator's performance when using TMs produced by MT output within a commercial CAT tool interface. We evaluate the quality of the translation output as well as the time and effort taken to accomplish each task. The impact of MT and TM in translators' performance has been explored and quantified in different settings (Bowker, 2005; Guerberof, 2009; Guerberof, 2012; Morado Vazquez et al., 2013). We believe this paper constitutes another interesting contribution to the interface between the study of the performance of human translators, CAT tools and machine translation.

2 Related Work

CAT tools have become very popular in the last 20 years. They are used by freelance translators as well as by companies and language service providers to increase translation's quality and speed (Somers and Diaz, 2004; Lagoudaki, 2008). The use of CAT tools is part of the core curriculum of most translation studies degrees and a reasonable level of proficiency in the use of these tools is expected from all graduates. With the improvement of state-of-the-art MT software, a recent trend in CAT research is its integration with machine translation tools as for example the MateCat[1] project (Cettolo et al., 2013).

There is considerable amount of studies on MT post-editing published in the last years (Specia, 2011; Green et al., 2013). Due to the scope of our

[1] www.matecat.com

paper (and space limitation) we will deliberately not discuss the findings of these experiments and instead focus on those that involve the use of translation memories. Post-editing tools are substantially different than commercial CAT tools (such as the one used here) and even though the TMs used in our experiments were produced using MT output, we believe that our experiment setting has more in common with similar studies that investigate TMs than MT post-editing.

The study by Bowker (2005) was one of the first to quantify the influence of TM in translators work. The experiment divided translators in three groups: A, B and C. Translators in Group A did not use a TM, translators in Group B used an unmodified TM and finally translators in group C used a TM that had been deliberately modified with a number of translation errors. The study concluded that when faced with time pressure, translators using TMs tend not to be critical enough about the suggestions presented by the software.

Another similar experiment (Guerberof, 2009) compared productivity and quality of human translations using MT and TM output. The experiment was conducted starting with the hypothesis that the time invested in post-editing one string of machine translated text will correspond to the same time invested in editing a fuzzy matched string located in the 80-90 percent range. This study quantified the performance of 8 translators using a post-editing tool. According to the author, the results indicate that using a TM with 80 to 90 fuzzy matches produces more errors than using MT segments or human translation.

The aforementioned recent work by Morado Vazquez et al. (2013) investigates the performance of twelve human translators (students) using the ACCEPT post-editing tool. Researchers provided MT and TM output and compared time, quality and keystroke effort. Findings of this study indicate that the use of a specific MT has a great impact in the translation activity in all three aspects. In the context of software localization, productivity was also tested by Plitt and Masselot (2010) combining MT output and a post-editing tool. Another study compared the performance of human translators in a scenario using TMs and a commercial CAT tool (Across) with a second scenario using post-editing (Läubli et al., 2013).

As to our study, we used instead of a post-editing tool, a commercial CAT tool, the SDL Trados Studio 2014 version. A similar setting to ours was explored by Federico et al. (2012) using SDL Trados Studio integrating a commercial MT software. We took the decision of working a commercial CAT tool for two reasons: first, because this is the real-world scenario faced by translators in most companies and language service providers[2] and second, because it allows us to explore a different variable that the aforementioned studies did not substantially explore, namely: MT output as TM segments.

3 Setting the Experiment

In our experiments we provided short texts from the domain of software development containing up to 343 tokens each to 15 beginner translators. The average length of these texts ranges between 210 tokens in experiment 1 to 264 tokens in experiment 3 divided in 15 to 17 segments (average) (see table 2). Translators were given English texts and were asked to translate them into German, their mother tongue. One important remark is that all 15 participants were not aware that the TMs we made available were produced using MT output.

The 15 translators who participated in these experiments are all 3rd semester master degree students who have completed a bachelors degree in translation studies and are familiar with CAT tools. All of them attended at least 20 class hours about TM software and related technologies. Translators who participated in this study were all proficient in English and they have studied it as a foreign language at bachelor level.

As previously mentioned, the CAT tool used in these experiments is the most recent version of SDL Trados, the Studio 2014[3] version. Translators were given three different short texts to be translated in three different scenarios:

1. Using no translation memory.
2. Using a translation memory collected with modified MT examples.
3. Using translation memory collected with unmodified MT examples.

In experiment number two we performed a number of modifications in the TM segments. As

[2] Although the use of MT and post-editing software has been growing, commercial TM software is still the most popular alternative.

[3] http://www.sdl.com/campaign/lt/sdl-trados-studio-2014/

can be seen in table 1, these modifications were sufficient to alter the coverage of the TM, but did not introduce translation errors to the memory.[4] The alterations we performed along with an example of each of them can be summarized as follows:

- Deletion: *'To paste the text currently in the clipboard, use the Edit Paste menu item.' - 'To paste the text, use the Edit Paste menu item.'*
- Modification: *'Persistent Selection is disabled by default.' - 'Persistent Selection is enabled by default.'*
- Substitution: *'The editor is composed of the following components:' - 'The editor is composed of the following elements:'*

Three texts were available per scenario, each of them with different TM coverage scores (see table 1). Students were asked to translate the texts at their own pace without time limitation and were allowed to use external linguistic resources such as dictionaries, lexica, parallel concordancers, etc.

3.1 Corpus and TM

The corpus used for these experiments is the KDE corpus obtained from the Opus[5] repository (Tiedemann, 2012). The corpus contains texts from the domain of software engineering, hence the title: 'a case study in technical translation'. We are convinced that technical translation contains a substantial amount of fixed expressions and technical terms different from, for example, news texts. This makes technical translation, to our understanding, an interesting domain for the use of TM by professional translators and for experiments of this kind.

In scenarios 1, 2 and 3 we measured different aspects of translation such as time and edited segments. One known shortcoming of our experiment design is that unlike most post-editing software the reports available in CAT tools are quite poor (e.g. no information about keystrokes is provided). Even so, we stick to our decision of using a TM software and tried to compensate this shortcoming by a careful qualitative and quantitative data analysis after the experiments.

[4]Modifications were carried out in the source and target languages

[5]http://opus.lingfil.uu.se/

Table number 1 presents the coverage scores for the different TMs and texts used in the experiments. Coverage scores were calculated based on the information provided by SDL Trados Studio. We provided 9 different texts to be translated to German (3 for each scenario), the 6 texts provided for experiments 2 and 3 are presented next.

Text	Experiment	TM Coverage
Text D	2	61.23%
Text E	2	78.16%
Text F	2	59.15%
Average	2	66,18%
Text G	3	88.27%
Text H	3	59.92%
Text I	3	65.16%
Average	3	71,12%

Table 1: TM Coverage

We provided different texts and levels of coverage to investigate the impact of this variable. We assured an equal distribution of texts among translators: each text was translated by 5 translators. This allowed us to calculate average results and to consider the average TM coverage difference of 4,93% between experiment 2 and 3.

4 Results

We observed performance gain when using any of the two TMs, which was expectable. The results varied according to the coverage of the TM. In experiment number 3, texts contained on average over 7 segments with 100% matches[6] and experiment number 2 only 2.68. This allowed translators to finish the task faster in experiment number 3. The average results obtained in the different experiments are presented in table number 2.[7]

Criteria	Exp. 1	Exp. 2	Exp. 3
Number of Segments	15.85	15.47	17.29
Number of Tokens	209.86	202.89	264.53
Context Matches		6.58	6.06
Repetitions			0.18
100%		2.68	7.18
95% to 99%		0.42	0.12
85% to 94%		0.21	
75% to 84%		2.11	0.18
50% to 75%			0.19
New Segments	15.86	5.89	3.24
Time Elapsed (mins.)	37m45s	26m3s	19m21s

Table 2: Average Scores

[6]Translators were allowed to modify 100% and context matches.

[7]According to the Trados Studio documentation, a *repetition* occurs every time the tool finds the exact same segment in another (or the same) file the user is translating

As to the time spent per segment, experiments indicate a performance gain of over 52% in experiment number 3 and over 28% in experiment number 2.

Criteria	Exp.1	Exp. 2	Exp. 3
Time Segment (mins.)	2m22s	1m41s	1m07s
Average gain to 1		+28.87%	+52.82%
Average gain to 2			+33.77%

Table 3: Time per Segment

Apart from the expectable performance gain when using TM, we also found a considerable difference between the use of the modified and unmodified TM. Translators completed segments in experiment number 3, on average, 33.77% faster than experiment two. The difference of coverage between the two TMs was 4,93%, which suggests that a few percentage points of TM coverage results on a greater performance boost.

We also have to acknowledge that the experiments were carried out by translators in the same order in which they are presented in this paper. This may, of course, influence performance in all three experiments as translators were more used to the task towards the end of the experiment. One hypothesis is that the poor performance in experiment 1, could be improved if this task was done for last and conversely, the performance boost observed in experiment 3, could be a bit lower if this experiment was done first. This variable was not explored in similar productivity studies such as those presented in section two and, to our understanding, inverting the order of tasks could be an interesting variable to be tested in future experiments.

As a general remark, although all translators had experience with the 2014 version of Trados Studio, we observed a great difficulty in performing simple tasks with Windows for at least half of the group. Simple operations such as copying, renaming and moving files or creating folders in the file system were very time consuming. Trados interface also posed difficulties to translators. For example, the generation of reports through batch tasks in a different window was for most translators confusing. These operations could be simplified as it is in other CAT tools such as memoQ.[8]

[8]http://kilgray.com/products/memoq

4.1 A Glance at Quality Estimation

One of the future directions that this work will take is to investigate the quality of human translations. Our initial hypothesis is that it is possible to apply state-of-the-art metrics such as BLEU (Papineni et al., 2002) or METEOR (Denkowski and Lavie, 2011) to estimate the quality of these translations regardless of how they are produced.

For machine translation output, quality nowadays is measured by automatic evaluation metrics such as the aforementioned IBM BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Denkowski and Lavie, 2011), the Levensthein (1966) distance based WER (word error-rate) metric, the position-independent error rate metric PER (Tillmann et al., 1997) and the translation error rate metric TER (Snover et al., 2006) with its newer version TERp (Snover et al., 2009).

The most frequently used one is IBM BLEU (Papineni et al., 2002). It is easy to use, language-independent, fast and requires only the candidate and reference translation. IBM BLEU is based on the n-gram precision by matching the machine translation output against one or more reference translations. It accounts for adequacy and fluency through word precision, respectively the n-gram precision, by calculating the geometric mean. Instead of recall, in IBM BLEU the brevity penalty (BP) was introduced.

Different from IBM BLEU, METEOR evaluates a candidate translation by calculating the precision and recall on unigram level and combining them in a parametrized harmonic mean. The result from the harmonic mean is than scaled by a fragmentation penalty which penalizes gaps and differences in word order.

For our investigation we applied METEOR on the human translated text. Our intention is to test whether we can reproduce the observations from the experiments: is the experiment setting 3 better than the setting of experiment 2? Therefore, METEOR is used here to investigate whether we can correlate it with our experiments and not to evaluate the produced translations. Table number 4 presents the scores obtained with METEOR.

	Exp. 2	Exp. 3
Average Score (mean)	0.14	0.41
Best Result	0.35	0.58
Worst Result	0.11	0.25

Table 4: METEOR Scores

In experiment number 3 we have previously observed that the translators' performance was significantly better and that translators could translate each segment on average 33.77% faster than experiment 2 and 52.82% faster than experiment 1. By applying METEOR scores we can also observe that experiment 3 achieved higher scores which seems to indicate more suitable translations than experiment number 2. Quality estimation is one of the aspects we would like to explore in future work.

5 Conclusion

This paper is a first step towards the comparison of different TMs produced with MT output and their direct impact in human translation. Our study shows a substantial improvement in performance with the use of translation memories containing MT output used trough commercial CAT software. To our knowledge this experiment setting was not tested in similar studies, which makes our paper a new contribution in the study of translators' performance. Although the performance gain seems intuitive, the quantification of these aspects within a controlled experiment was not substantially explored.

We opted for the use of a state-of-the-art commercial CAT tool as this is the real-world scenario that most translators face everyday. In comparison to translating without TM, translators were on average 28.87% faster using a modified TM and 52.82% using an unmodified one. Between the two TMs we observed that translators were on average 33.77% faster when using the unmodified TM. As previously mentioned, the order in which this tasks were carried out should be also taken into account. The performance boost of 33.77% when using a TM that is only 4,93% better is also an interesting outcome of our experiments that should be looked at in more detail.

Finally, in this paper we used METEOR scores to assess whether it is possible to correlate translations' speed, quality and TM coverage. The average score for experiment number 2 was 0.14 and for experiment number 3 was 0.41. Our initial analysis suggests that a relation between the two variables exists for our dataset. Whether this relation can be found in other scenarios is still an open question and we wish to investigate this variable more carefully in future work.

5.1 Future Work

We consider these experiments as a pilot study that was carried out to provide us a set of variables that we wish to investigate further. There are a number of aspects that we wish to look in more detail in future work.

Future experiments include the aforementioned quality estimation analysis by applying state-of-the-art metrics used in machine translation. Using these metrics we would like to explore the extent to which it is possible to use automatic methods to study the interplay between quality and performance in computer assisted translation. Furthermore, we would like to perform a qualitative analysis of the produced translations using human annotators and inter annotator agreement (Carletta, 1996).

The performance boost observed between scenarios 2 and 3 should be looked in more detail in future experiments. We would like to replicate these experiments using other different TMs and explore this variable more carefully. Another aspect that we would like to explore in the future is the direct impact of the use of different CAT tools. Does the same TM combined with different CAT tools produce different results? When conducting these experiments, we observed that a simplified interface may speed up translators' work considerably.

Other directions that our work will take include controlling other variables not taken into account in this pilot study such as: the use of terminological databases, spelling correctors, etc. How and to which extent do they influence performance and quality? Finally, we would also like to use eye-tracking to analyse the focus of attention of translators as it was done in previous experiments (O'brien, 2006).

Acknowledgments

We thank the students who participated in these experiments for their time. We would also like to thank the detailed feedback provided by the anonymous reviewers who helped us to increase the quality of this paper.

References

Lynne Bowker. 2005. Productivity vs quality? a pilot study on the impact of translation memory systems. *Localisation Reader*, pages 133–140.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Mauro Cettolo, Christophe Servan, Nicola Bertoldi, Marcello Federico, Loic Barrault, and Holger Schwenk. 2013. Issues in incremental adaptation of statistical mt from human post-edits. In *Proceedings of the MT Summit XIV Workshop on Post-editing Technology and Practice (WPTP-2)*, Nice, France.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT 2002, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Marcello Federico, Alessandro Cattelan, and Marco Trombetti. 2012. Measuring user productivity in machine translation enhanced computer assisted translation. In *Proceedings of Conference of the Association for Machine Translation in the Americas (AMTA)*.

Spence Green, Jeffrey Heer, and Christopher D Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Ana Guerberof. 2009. Productivity and quality in the post-editing of outputs from translation memories and machine translation. *Localisation Focus*, 7(1):133–140.

Ana Guerberof. 2012. *Productivity and Quality in the Post-Edition of Outputs from Translation Memories and Machine Translation*. Ph.D. thesis, Rovira and Virgili University Tarragona.

Elina Lagoudaki. 2008. The value of machine translation for the professional translator. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas*, pages 262–269, Waikiki, Hawaii.

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

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, (8):707–710, February.

Lucia Morado Vazquez, Silvia Rodriguez Vazquez, and Pierrette Bouillon. 2013. Comparing forum data post-editing performance using translation memory and machine translation output: a pilot study. In *Proceedings of the Machine Translation Summit XIV)*, Nice, France.

Sharon O'brien. 2006. Eye-tracking and translation memory matches. *Perspectives: Studies in Translatology*, 14:185–204.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, AMTA.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or hter? exploring different human judgments with a tunable mt metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation at the 12th Meeting of the European Chapter of the Association for Computational Linguistics*, EACL 2009.

Harold Somers and Gabriela Fernandez Diaz. 2004. Translation memory vs. example-based mt: What is the difference? *International Journal of Translation*, 16(2):5–33.

Lucia Specia. 2011. Exploiting objective annotations for measuring translation post-editing effort. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 73–80, Leuven, Belgium.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Christoph Tillmann, Stephan Vogel, Hermann Ney, Alexander Zubiaga, and Hassan Sawaf. 1997. Accelerated dp based search for statistical translation. In *European Conference on Speech Communication and Technology*, EUROSPEECH 1977, pages 2667–2670.