Measuring User Productivity in Machine Translation Enhanced Computer Assisted Translation

Marcello Federico
Fondazione Bruno Kessler
Via Sommarive, 18
38123 Povo - Trento, Italy
<surname>@fbk.eu

Alessandro Cattelan **Marco Trombetti**
Translated srl
Via Nepal, 29
00144 Rome, Italy
<name>@translated.net

Abstract

This paper addresses the problem of reliably measuring productivity gains by professional translators working with a machine translation enhanced computer assisted translation tool. In particular, we report on a field test we carried out with a commercial CAT tool in which translation memory matches were supplemented with suggestions from a commercial machine translation engine. The field test was conducted with 12 professional translators working on real translation projects. Productivity of translators were measured with two indicators, post-editing speed and post-editing effort, on two translation directions, English–Italian and English–German, and two linguistic domains, legal and information technology. Besides a detailed statistical analysis of the experimental results, we also discuss issues encountered in running the test.

1 Introduction

Worldwide demand of translation services has steadily increased in the last decade, as an effect of market globalization and growth of the information society. Computer assisted translation (CAT) tools are currently the dominant technology in the translation and localization market. These include spell checkers, terminology managers, electronic dictionaries, full-text search tools, concordancers, bitexts, translation memory (TM) managers, and, more recently, machine translation (MT) engines. Recent achievements by the statistical MT approach (Koehn, 2010) have raised new expectations in the translation industry. Research efforts addressing the deployment of statistical MT to improve human translation (HT) productivity generally fall under one of three use cases: (i) *post-editing* of MT outputs (Allen, 2003), where objectives are to predict whether MT output is worth being post-edited or not (Blatz et al., 2004; Quirk, 2004; Specia and Farzindar, 2010), as well as supplying post-editors with efficient editing options (Koehn, 2010); (ii) *interactive MT*, where the goal is developing MT systems that assist HT by predicting words before they are typed (Langlais et al., 2000; Och et al., 2003; Civera et al., 2004; Koehn and Haddow, 2009); (iii) *TM-MT integration*, where the objective is integrating TM matches with MT suggestions (Biçici and Dymetman, 2008; Simard and Isabelle, 2009; He et al., 2010; Koehn and Senellart, 2010).

The MateCat project, which falls under the last case, aims to increase productivity of professional translators by investigating new research issues related to the integration of MT into CAT, namely: (i) self-tuning MT, that performs domain adaptation as soon as translated documents become available, (ii) user-adaptive MT, that performs on-line learning by exploiting user feedback at the segment level, and (iii) informative MT, that supplies the user with confidence scores and alternative translations. The project is also developing an open source Web-based CAT tool integrating new MT functionalities built on top of state-of-the-art MT and CAT technologies, such as Moses(Koehn et al., 2007), IRSTLM (Federico et al., 2008), and MyMemory.[1]

One crucial aspect related to the integration of MT into the HT workflow is how to reliably measure

[1] http://mymemory.translated.net

productivity gains. In this paper we report on a field test in which we tried to measure productivity gains of professional translators working with a commercial CAT tool after its TM was augmented with suggestions coming from a commercial MT engine. In the following sections, we overview previous work on translation productivity evaluation, we describe the scope and structure of the field test, introduce the two productivity indicators we adopted, report the results we found, and finally discuss outcomes and issues of the test.

2 Previous Work

The literature in MT reports several investigations in which HT productivity was measured, either to set a reference for its automatic prediction or to compare HT under different working conditions.

In (O'Brien, 2011), post-editing productivity is measured along two dimensions: processing speed and cognitive effort. The two aspects are quantified by measuring, respectively, time intervals taken to post-edit single segments, and fixation intervals, in which the eyes and attention of the user is directed to some part of the segment. Fixations intervals, which can be detected with an eye tracking system, have shown in previous translation studies to correlate well with the degree of difficulty experienced by the user. Processing speed is finally expressed in number of words post-edited per second. According to (O'Brien, 2011) improving processing speed is indeed the primary interest of translation industry as this figure can be directly related to the cost of the translation. In (Specia and Farzindar, 2010), post-editing productivity is instead measured in terms of human-targeted translation edit rate (HTER), a metric introduced by (Snover et al., 2006) that measures the edit distance between an MT output and its minimally post-edited version produced by a human translator. Hence, the smaller the HTER is for a segment, the less post-editing human labour is assumed, and the higher is the productivity. Notice that although the edit distance used in HTER accounts for insertions, deletions, substitutions, and shifts of words, as well as substitutions of synonims, HTER is clearly opaque to the effort and time taken to translate difficult and easy words. This approach indeed recalls productivity measures originally used in interactive MT (Langlais et al., 2000), which compared keystroke counts needed to produce the translation from scratch and with the interactive system. Also worth noticing from the MT interactive scenario are the measurement of activity intervals and the analysis of pauses (Koehn and Haddow, 2009) in order to infer different types of cognitive efforts by the translators.

In the following, we survey two works from the recent literature that have tried to measure productivity gains by professional translators when TM matches were integrated with MT suggestions.

In (Guerberof, 2009), eight professional translators were asked to translate a fixed number of segments (791 source words) from English into Spanish, one third of which from scratch, one third from TM matches and one third from MT suggestions. TM matches were selected to be in the 80–90 percent fuzzy match range. A commercial statistical MT engine was trained on the content of the TM plus a core glossary. The translators used a Web-based post-editing tool, supplied with a core glossary, to translate/post-edit all segments but without knowing their origin. The tool measured the time taken for each segment. Besides measuring and comparing productivity in terms of speed, (Guerberof, 2009) carried out a detailed analysis of the quality of the produced translations.

In (Plitt and Masselot, 2010), twelve professional translators were involved in an experiment comparing human translation versus post-editing productivity when MT outputs are provided. The test was performed on information technology documentation, on four translation directions and by employing three translators per direction. Under all conditions a total of 144,648 source words were processed. The MT engine was a specifically trained Moses engine, while the post-editing tool was inspired by the CAITRA tool (Koehn and Haddow, 2009). Post-editing productivity was measured in terms of processing speed (words per second) and edit distance. A pause analysis was carried out to compare keyboard and pause times of translation versus post-editing. Finally, a a blind test was conducted to compare the quality of the segments produced with the two modalities.

Our contribution is very related to the last two works although it departs from them in one fundamental aspects. The impact of MT on productivity is evaluated with a popular commercial CAT tool which seamlessly integrates MT suggestions within TM matches and all the other standard features. Translators were asked to translate full documents, rather than isolated segments, without changing their working routine [2]. Indeed, MT suggestions were provided just in addition to TM suggestions and translators were left free to decide whether to translate segments from scratch or to post-edit the provided matches. Finally, the origin of each suggestion, TM or MT, was shown to the user. [3]

Overall we believe that our experimental setting, though less controlled than the previous ones, can provide more realistic figures about the potential benefits of enhancing CAT with MT.

3 Objectives and Methods

The aim of the field test was to establish a reference baseline for the web-based CAT tool that will be developed in the MateCat project. The considered reference is a commercial CAT tool (SDL Trados Studio) integrating a commercial MT engine (Google Translate) and the same TM technology (MyMemory) that will be employed in the MateCat tool. In particular, we tried to automatically measure productivity of human translators to estimate the utility of suggestions coming from the MT engine. Moreover, the test also served the purpose to check the overall evaluation procedure and to spot potential technical issues.

We extended a standard version of SDL Trados Studio with a publicly available MyMemory plug-in, designed to provide the translator with matches from the TM server (MyMemory) and eventually MT suggestions from Google Translate whenever a TM match is not present. The plug-in also allows collecting information from the user, such as the time spent editing a segment and the match similarity of any supplied hint to the translated segment. The plug-in records actions such as the opening and saving of a segment, the content of each source segment, the best ranking suggestion provided and the target segment saved by the translator. In fact, our collected data do only record information about the drafting phase of the translation. Moreover with our setting it is not possible to detect whether a translator was effectively working on a segment or had stopped editing it, nor to detect if the translator used any information from other external sources.

Before starting, translators were provided with simple instructions providing recommendations about how to organize their work flow and after the experiments the collected data were post-processed to remove irrelevant or inconsistent data. The following sections provide further information on these aspects.

3.1 Translators, Tasks, and Languages

The experiment involved 12 professional translators all of which with a strong professional record and very familiar with the employed CAT tool. Translators were split over two translation directions, English to German (EN>DE) and English to Italian (EN>IT), and two domains, information technology (IT) and legal. In other words, each translator processed documents of only one domain and one target language. All translators working on the same domain were assigned the same set of documents. First, half of the documents was translated by only relying on TM matches, while for the second half translators were provided with suggestions from TM and MT ranked as follows. MyMemory assigns a fixed 15% match penalty to MT suggestions. By consequence MT suggestion have a 85% match score while an exact TM match has 100%. This implies that any TM suggestion with a score higher than 85% will win over MT. This methodology and the chosen penalty value (15%) are default parameters shared among most professional CAT tools.

For the Legal domain, two different documents were used, that contained English text extracted from a call for tender by European institutions that describes the contract binding the tenderer (requirements, selection and exclusion requirements, payments, etc.). As these were standard documents from European institutions, portions of the source text (standard wording) were already available on line.

For the information technology domain, several files from a software user manual in English were

[2] In fact, we only asked translators to process segments as sequentially as possible.

[3] Although this information might bias the user's behaviour, it can be helpful in the presence of terminology, whose translation must generally comply with the TM content.

used. The manual was not publicly available on line in English nor in any other language.

Notice that for both domains, TM matches accounted only for a small fraction of words. In particular, 75–90% fuzzy matches only accounted for about 10% of the total number of words to be translated. Notice also that for reasons specified later, 100% TM matches were excluded from our measurements.

3.2 Productivity Indicators

Two key performance indicators were considered in the experiment:

1. *Post-editing speed*, which is the average number of words processed by the translator in one hour.

2. *Post-editing effort*, which is the average percentage of word changes applied by the translator on the suggestions provided by the CAT tool.

The first indicator directly expressed the time labour required by the translators, hence improvements on this figure are directly related to cost savings.

The second indicator measures the quality of the matches provided by the TM and MT. This corresponds to computing a distance score between matches provided by the system and the post-edited version submitted by the user. The indicator is indeed an estimate from below of the percentage of edit operations performed in the whole set of translated segments.

4 Data Collection and Issues

Translators were instructed to follow some rules in order to reduce measurement errors. They were asked to create a distinct project package in SDL Trados Studio for each test condition: TM and TM+MT. The project package contained the file(s) to translate and a single TM or MT provider, namely our plug-in. Translators were required to not add any additional TM or MT provider but the supplied plug-in, which gets both TM and MT matches from the MyMemory server.

Translators were provided with only TM matches for the first part of the test, while for the second part they received both TM and MT matches. In fact, TM or TM+MT matches were generated from the MyMemory server, based on the type of test and on the translators user-name and IP.

	EN>DE	EN>IT	Total
Legal (TM)	7,221	7,041	14,262
Legal (TM+MT)	8,568	13,087	21,655
IT (TM)	18,425	8,553	26,978
IT (TM+MT)	19,972	9,791	29,763
Total	54,186	28,472	92,658

Table 1: Number of words available for the statistical analysis, after removing segments due to protocol violations and applying the time-threshold filtering.

Concerning their work modality, translators were asked to translate segments as sequentially as possible, that is to not move to a new segment without having completed and saved the current one. This requirement was meant to avoid issues such as measuring editing time of overlapping segments (i.e. segments enclosing entirely or partially other segments).

While translators were interacting with the CAT tool, the following data and statistics were automatically recorded for each processed segment:

- Matches provided by the TM server (if any),
- Matches provided by the MT engine (if any),
- Matches used by the translator as a basis for their translation (if any),
- Target segments edited by the translator,
- Time intervals needed to edit each segment.

5 Data Filtering

As said before, our plug-in is not able to detect all the user operations while she is processing a segment, nor the change of focus from one segment to another. The plug-in does only record opening and saving of segments. Every time a segment is opened in SDL Trados Studio, a GET request is sent to the server in order to retrieve matches from MyMemory. Once the segment is saved, a SET request is sent to the server with the translated segment to be stored. Segment overlaps occurs when a translator opens a

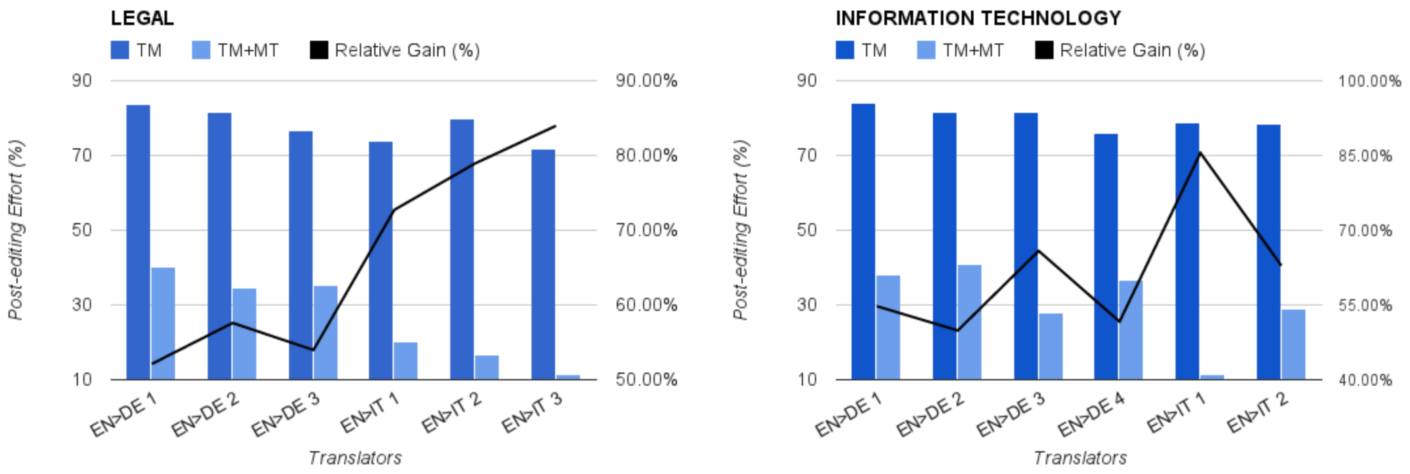

Figure 1: Post-editing effort by each translator, for each suggestion mode, and corresponding relative gains.

segment (GET) and then she moves to another segment without saving the first opened segment (no SET is issued). As segment overlaps do not permit to correctly measure the processing time of the respective segments, two specific types of overlapping conditions were identified and removed from the recorded data:

- Enclosures: e.g. GET A – GET B – SET B - - SET A
- Pipelines: e.g. GET A – GET B – SET A – SET B

Moreover, in order to remove unreliable measurements, we assumed that time intervals shorter or longer than two given thresholds were probably not related to the translation work flow, but were more likely caused by interaction errors – e.g. the translator stopped working without saving the segment she was editing.

In particular, our performance analysis is limited to segments whose processing time per word passes the following two (empirically set) thresholds:

- $\leq$ 30s per word: translation times over 30 seconds/word for a drafting of the translation are assumed to be dependent on factors unrelated to the complexity of the source text and more likely dependent on software errors or the translator's behaviour (pauses, distractions, etc.).
- $\geq$ 0.5s per word: translation times below 0.5 seconds/word are assumed to be unrealistic for most segments and are probably due to accidental interactions with the software (e.g. saving a segment without reading or editing it).

Collected data was also filtered to remove all 100% TM matches and repetitions, given that the time to edit for those segments is irrelevant and that SDL Trados Studio automatically translates perfect matches provided by the TM.

Table 1 reports the number of words available for the statistical analysis after removing segments with overlaps, 100% matches, and after applying the the time-threshold conditions. Overall, this resulted in the removal of roughly 30% of the translated words. In particular, by cascading the filtering conditions, we get the following progressive reductions:

- 18% from overlapping segments
- 9% from 100% TM matches[4]
- 2% from too fast editing ($<$ 0.5s per word)
- 4.5 % from too slow editing ($> 30s$ per word)

The amount of available data to carry out our analysis varies significantly among translators and conditions, from a minimum of 1767 words (110 seg-

[4] This matches result from incremental updates for the TM.

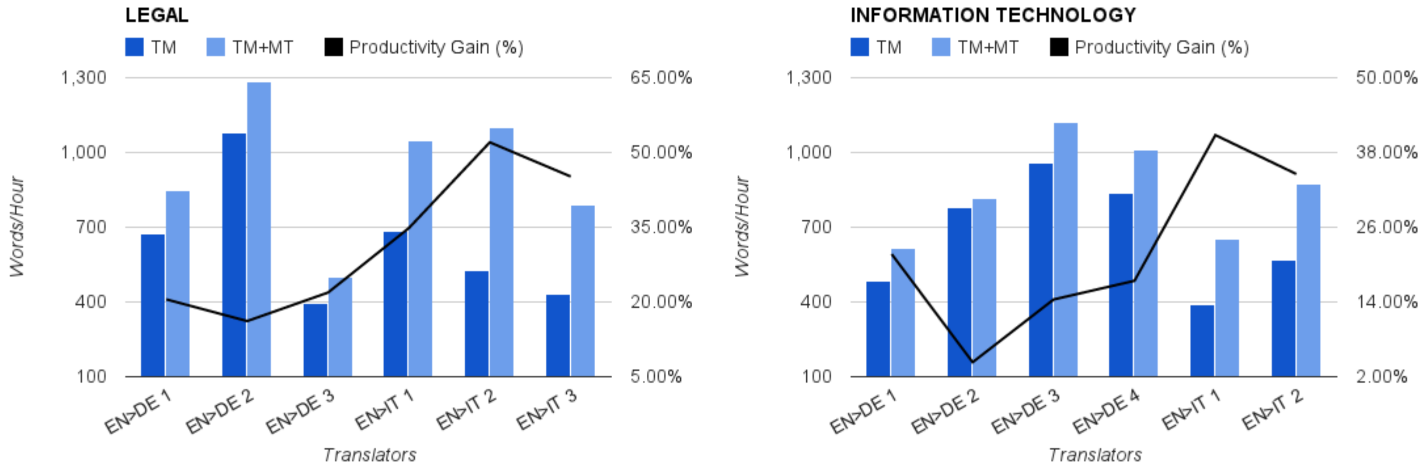

Figure 2: Average post-editing speed of all translators, for each suggestion mode, and relative time gains.

ments) to a maximum of 5244 words (626 segments). Nevertheless, the data resulted sufficient to perform rather accurate comparisons.

6 Results

6.1 Post-Editing Effort

This indicator aims at defining the quality of the matches provided by the TM and MT systems. We measured the percentage of words edited in a segment by comparing the match provided by the system and the final segment submitted by the translator. An enhanced edit-distance function was used to compare segments pairs, which simulates that used by Trados SDL. The function computes a similarity of segments with the algorithm in (Oliver, 1993), by reserving a special treatment to formatting tags, casing and punctuation marks. The similarity match can be interpreted as an indication of the quality of the suggestions provided by the TM and MT systems. Conversely, an estimate of the involved post-editing effort can be simply computed by taking the complement of the similarity match (100% -SimilarityMatch).

As shown in Figure 1, the post-editing effort decreases significantly for all translators when MT matches are supplied in addition the TM matches. Even though this may be considered an obvious consequence of doubling the sources for the matches, the extent to which the post-editing effort drops proves the effectiveness of the MT engine used in the test. Indeed, all individual translators took advantage from suggestions coming from the MT engine. On the legal domain, post-editing effort with only TM was on average 80.7% for EN>DE and 75% for EN>IT. With the availability of MT suggestions, these figures dropped on average, respectively, to 36.7% and 16.15%. This corresponds to a post-editing effort reduction on the legal domain of 54.6% for EN>DE and 78.5% for EN>IT. On the information technology documents, post-editing effort with only TM was on average 80.9% for EN>DE and 78.6% for EN>IT. With the availability of MT suggestions, the corresponding figures dropped to 35.9% and 20.2%, respectively. Hence, the relative gains on the two translation directions were 55.5% and 74.2%.

6.2 Post-editing Speed

The charts in Figure 2 show post-editing speed, expressed in number of words processed per hour, by each translator for each suggestion mode, and relative time gains. On both domains and language pairs most of the translators were able to achieve substantial time savings when passing from the TM to the TM+MT suggestion mode. However, post-editing speed figures vary significantly across translators, languages, and domains. High variance across translators was also reported in (Plitt and Masselot, 2010) and some possible explanations for it are provided in

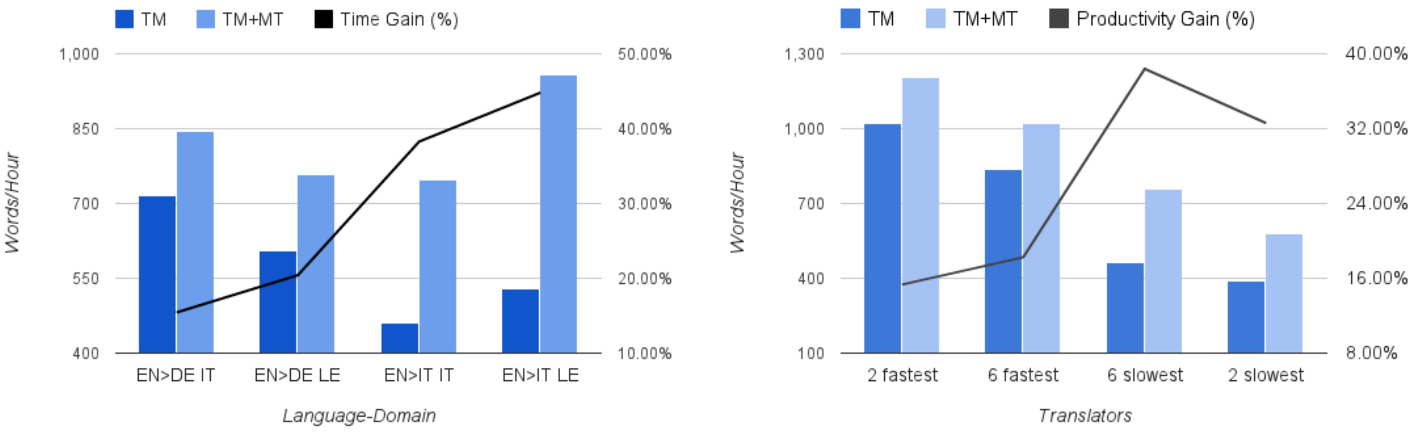

Figure 3: Absolute and relative time-to-edit gains on different languages and domains (left), and of slow and fast translators (right).

the Discussion section. Relative time gains[5] range indeed from 4% to 52% with an average of 27%. By applying a randomized (or permutation) significance test (Riezler and Maxwell, 2005) on each single translator, we found that the average time reductions were all significant at level $p < 0.01$ for all translators,[6] but Legal EN>DE 2 ($p < 0.04$) and Legal EN>DE 3 ($p < 0.17$).

In the following we perform some further analysis of collected data, also for the sake of comparison with the results reported in the similar study conducted by (Plitt and Masselot, 2010).

Languages and Domains

In Figure 3 Left, we show post-editing speed versus translation direction and domain. Major benefits from MT suggestions were achieved on the two "slowest" tasks with TM-only matches, that is EN>IT on legal and IT documents. This observation is confirmed by the corresponding post-edit effort gains shown in Figure 1. In particular, the average relative time savings range from 15.43% for EN>DE IT domain to 44.81% for EN>IT Legal domain.

Fast and Slow Translators

In Figure 3 Right, we compare post-editing speed figures by the slowest and fastest translators of our pool. Largest relative gains in productivity were achieved by the slower translators, while fast translators showed smaller margins for improvement. This results are in line with with the trends reported in (Plitt and Masselot, 2010).

Length of Segment

Finally, we analysed productivity figures for different lengths of segments. First, we clustered source segments into short (1–10 words), medium (11–20 words), and long (>20 words) segments. In Figure 4 Left we report minimum, average, and maximum time-to-edit (seconds/word) by translators exploiting MT suggestions. As can seen, MT suggestions increase productivity more on medium to long segments rather than on short segments. This results can be explained with the fixed time costs incurred by the translators when post-editing every segment – e.g. time to position the mouse on the word to correct –, whose relative impact is larger on short segments.

To get a more detailed analysis of the productivity gains on long segments, we also compared productivity directly at the segment level, rather than at the word level, for increasing segment lengths. By reasonably assuming that post-editing time in-

[5] Time gain is defined in terms of speed gain by: $\text{TimeGain} = 1 - (1 + \text{SpeedGain})^{-1}$.

[6] i.e. the probability of these two measurements being the same (p) is below 0.01.

creases linearly with the segment length, we estimated the time-to-edit trend for each suggestion modality through simple linear regressions models, which are plotted in Figure 4 and whose parameters are reported in the respective caption. The trends reported in Figure 4 show that providing HT with MT suggestions significantly lowers their time-to-edit rate. The different offsets of the two lines, 3.6 seconds for TM and 4.0 seconds for TM+MT, could be explained by the fact that in the TM mode suggestions are accepted less frequently than in the TM+MT mode. This operation has indeed a fixed cost, which in the TM+MT mode has an higher incidence.

Finally, also these findings are in line with those reported in (Plitt and Masselot, 2010), by duly taking into account that their comparison was in fact between MT post-editing versus translating from scratch.

7 Discussion

Even though all translators translated the same content and were provided with the same instructions and information, the results for the two productivity indicators show a certain degree of variation in terms of post-editing speed and post-editing effort. The variation in general depends on two factors: the quality of the matches provided by the plug-in and the performance of each translator.

The quality of the matches from the MyMemory TM server depends on the amount of translated segments that it contains for each language pair and domain. There are indeed some differences in the number of segments for EN>DE and EN>IT. Also, MT matches tend to be of higher quality for the EN>IT language pair than for EN>DE.

All translators were supposed to deliver a drafting of the target text. However, it is generally difficult to assess objectively the quality of a translation and translators are not capable of determining when their translations are "good enough" for a drafting. Some translators may consider it appropriate to deliver a translation that is semantically correct while poor in style. Some others may put in more effort in order to provide not only a semantically correct translation, but also one they consider more appropriate from a linguistic (i.e. grammar, style) and a terminological point of view.

The different approach by each translator played a role in the variations we can see in terms of post-editing speed and post-editing effort. Some translators accepted MT matches without much editing because they considered such matches to be semantically correct and appropriate for a drafting of the text. Others spent more time on each segment editing more words because they felt they needed to provide a higher quality target text (improving on style and language quality). Although we do not expect that single professional translators are letting their quality standard be influenced by the received suggestions, in future field test we will nevertheless perform some quality checks on the post-edited segments.

Moreover, post-editing speed is also be influenced by the way translators use the software (SDL Trados Studio). While all translators were required to use the same settings for the project package, we could not force them to use a specific setting for the UI[7]. How UI elements are arranged can affect performance: translators may have to perform some extra actions in order to view the matches from the TM or MT (if the translation matches window is too small, translators are required to scroll through the results using a mouse or touchpad), they may have to activate the preview feature to see the text they are translating in context (although this may not be too important when working on a draft). Also, some translators may be used to move from one segment to another using keyboard shortcuts, while others may use the mouse or touchpad. Even though such activities do not account for significant changes in terms of overall productivity on a daily basis, they can certainly affect the time to edit by half seconds per segment.

8 Conclusions and Future Work

We have presented the set-up and outcomes of a field test that measured productivity of professional translators working with a commercial CAT tool embedding state-of-the-art TM technology. We compared productivity of translators before and after supplementing TM matches with suggestions coming from

[7] In SDL Trados Studio the UI elements can be re-arranged to match the translators preferences

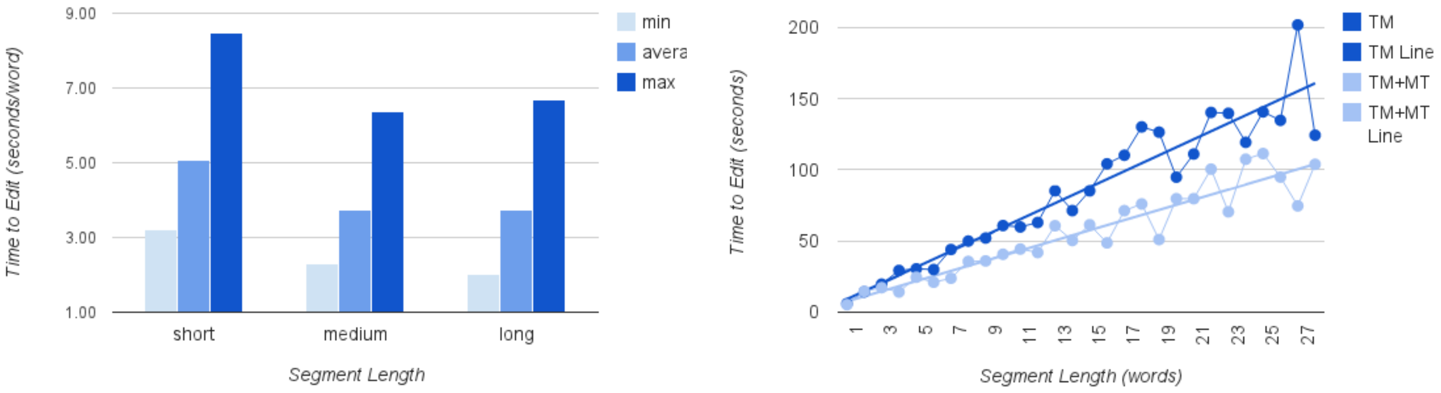

Figure 4: Left. minimum, average, and maximum time-to-edit by translators on short segments (1–10 words), medium-length segments (11–20), and long segments (>20 words). Average post-edit time versus sentence length. Regression lines fitting TM and TM+MT are respectively: $y = 5.615x + 3.586$ and $y = 3.571x + 4.024$.

a commercial MT engine. Results reported productivity gains in terms of post-editing speed by all translators. For 10 out of 12 translators, the corresponding time gains were also statistically significant at level $\alpha = 0.01$. As a difference with previous studies, we analysed productivity gains coming from MT suggestions in a real world setting. Professional translators were in fact asked to work with their preferred and full fledged CAT tool, on real translation projects, and at their usual workplace. Translators received minimal instructions about how to perform their task so as to maximize the outcome of the experiment. Carrying out the evaluation with such a weak supervision introduced however unreliable or useless measurements which had to be filtered out from the log-files. This trial was indeed propaedeutic for future field tests that will be carried out with a newly developed web-based CAT tool powered with a Moses-based engine featuring novel functionalities. After this experience, we plan to introduce some improvements in order limit the percentage of useless measurements, by still continuing to evaluate translators' productivity in the "real worl". In particular, we will try to cope with non-sequential translation patterns, which seem to be relatively frequent with some translators. We are designing our new CAT tool in such a way to allow tracing all the time spent on each segment also through multiple passes.

Acknowledgments

This work partially funded by the European Commission under the FP7 project MateCat, Grant 287688. The authors wish to thank Reviewer #1 for her/his many useful suggestions and comments.

References

Jeffrey Allen. 2003. Post-editing. In Harold Somers, editor, *Computers and Translation*, chapter 16, pages 297–317. John Benjamins.

Ergun Biçici and Marc Dymetman. 2008. Dynamic translation memory: using statistical machine translation to improve translation memory fuzzy matches. In *Proceedings of the 9th international conference on Computational linguistics and intelligent text processing*, CICLing'08, pages 454–465, Berlin, Heidelberg. Springer-Verlag.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *Proceedings of Coling 2004*, pages 315–321, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Jorge Civera, Elsa Cubel, Antonio L. Lagarda, David Picó, Jorge González, Enrique Vidal, Francisco Casacuberta, Juan M. Vilar, and Sergio Barrachina. 2004. From machine translation to computer assisted translation using finite-state models. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 349–356, Barcelona, Spain, July. Association for Computational Linguistics.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. In *Proceedings of Interspeech*, pages 1618–1621, Melbourne, Australia.

Ana Guerberof. 2009. Productivity and quality in MT post-editing. In *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII)*, Beyond Translation Memories: New Tools for Translators Workshop. International Association for Machine Translation.

Yifan He, Yanjun Ma, Josef van Genabith, and Andy Way. 2010. Bridging smt and tm with translation recommendation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 622–630, Uppsala, Sweden, July. Association for Computational Linguistics.

Philipp Koehn and Barry Haddow. 2009. Interactive assistance to human translators using statistical machine translation methods. In *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII)*. International Association for Machine Translation.

Philipp Koehn and Jean Senellart. 2010. Convergence of translation memory and statistical machine translation. In *Proceedings of the Second Joint EM+CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry"*, pages 21–31, Denver, CO.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

Philippe Langlais, George Foster, and Guy Lapalme. 2000. Transtype: a computer-aided translation typing system. In *Proceedings of the ANLP-NAACL 2000 Workshop on Embedded Machine Translation Systems*, pages 46–51, Seattle, WA.

Sharon O'Brien. 2011. Towards predicting post-editing productivity. *Machine Translation*, 25:197–215.

Franz Josef Och, Richard Zens, and Hermann Ney. 2003. Efficient search for interactive statistical machine translation. In *Proceedings of Meeting of the European Chapter of the Association of Computational Linguistics (EACL)*, pages 387–394, Budapest, Hungary.

Ian Oliver. 1993. *Programming classics: implementing the world's best algorithms*. Prentice Hall.

Mirko Plitt 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.

Chris Quirk. 2004. Training a sentence-level machine translation confidence measure. In *Proceedings of the 4th Conference of Language Resources and Evaluation (LREC)*, pages 825–828, Lisbon, Portugal.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Michel Simard and Pierre Isabelle. 2009. Phrase-based machine translation in a computer-assisted translation environment. In *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII)*. International Association for Machine Translation.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the Fifth Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231, Boston, MA.

Lucia Specia and Atefeh Farzindar. 2010. Estimating Machine Translation Post-Editing Effort with HTER. In *Proceedings of the Second Joint EM+CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry"*, pages 33–41, Denver, CO.