This document is part of the Project “Machine Translation Enhanced Computer Assisted Translation (MateCat)”, funded by the 7th Framework Programme of the European Commission through grant agreement no.: 287688.

Machine Translation Enhanced
Computer Assisted Translation

D5.3 - First report on lab and field tests

Author(s): Nicola Bertoldi, Alessandro Cattelan, Marcello Federico
Dissemination level: Public
Date: December 5th, 2012
For copies of reports, updates on project activities and other MateCat-related information, contact:

FBK
MateCat
Marcello Federico federico@fbk.eu
Povo - Via Sommarive 18 Phone: +39 0461 314 552
I-38123 Trento, Italy Fax: +39 0461 314 591

Copies of reports and other material can also be accessed via http://www.matecat.com

© 2012, Nicola Bertoldi, Alessandro Cattelan, Marcello Federico
No part of this document may be reproduced or transmitted in any form, or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission from the copyright owner.
Executive Summary

This document reports on the first field test carried out with the technology developed in the MateCat project. The field test aims at evaluating the impact of self-tuning machine translation (MT) on user productivity. The field test employed the first version of the MateCat Tool developed by the industrial partner and the MT engines developed by the research partners. The field test was carried out with professional translators, under working conditions very similar to the usual practice. In addition to measuring the impact of self-tuning MT on the final user productivity, this report also provides a detailed evaluation of the performance of the MT component alone, that was carried out on the data collected during the field test. In particular, comparative results are provided in terms of three automatic metrics, BLEU, TER, and GTM.
Contents

1 Introduction 5

2 Field test 6
   2.1 Dry run ................................................................. 6
   2.2 Plan revision ......................................................... 6
   2.3 Protocol .............................................................. 6
   2.4 Translated documents .............................................. 7
   2.5 Translation environment .......................................... 8
   2.6 Data collection ..................................................... 9
   2.7 Data filtering ....................................................... 11
   2.8 Key performance indicators ...................................... 12
   2.9 Results ............................................................... 13

3 Lab tests 16
   3.1 Training Data ........................................................ 16
   3.2 Test Data ............................................................. 16
   3.3 Results ............................................................... 17

4 Conclusions and future work 20
1 Introduction

The most important evaluation activity of the MateCat project are the field tests, which are carried with annual cadence and whose main goal is to measure the impact of novel MT components on the final productivity of translators working with the MateCat Tool. A preliminary field test was already carried out at month 6, by employing a commercial CAT tool (SDL Trados) and a commercial MT engine (Google Translated). The goal of this preliminary test was to set a good reference baseline for the new MateCat Tool, especially in terms of achievable user productivity through the introduction of MT suggestions. Indeed, this experience was very useful to improve the design of the MateCat Tool itself, too.

This report documents the outcomes of the first field test, whose aim was to evaluate the impact of self-tuning MT on user productivity. Indirectly, we also evaluated the MateCat Tool itself, in particular the CAT client and the MT servers developed by the partners. According to the plan, the field test had to cover two translation directions, English-Italian and English-German, and two application domains, legal and information technology (IT) documents. The outcome of a preliminary dry run, however, limited to the IT domain, suggested to postpone the evaluation on the German to English direction, as the level of performance of the MT engine seemed not yet sufficient to provide suggestions worth to be post-edited.

The field test was organized over two days in which translators had to work on different portions of a document. Moreover, during the first day they received suggestions by an MT engine without self-tuning MT, while during the second day MT suggestions came from an MT with self-tuning MT. The impact of the self-tuning MT was measured by comparing productivity of translators during the first and the second day. Productivity was measured by two key performance indicators: average translation time for each word (time to edit) and average estimated number of edit operations applied on the suggestions (post-edit effort).

After the field test, the collected translations were used to evaluate the employed MT engines. In particular, the MT engines corresponding to the different training conditions were compared with three automatic MT metrics: BLEU [Papineni et al., 2002], TER [Snover et al., 2006], and GTM [Turian et al., 2003].

In the following two sections, the field test of the MateCat Tool and lab tests on the MT engines are reported in detail. Finally, the achieved results are compared with the initial goals set in the work plan and general remarks and requirements about future developments of the MateCat project are set.
2 Field test

2.1 Dry run

Before starting the field test, a few dry run tests were performed to verify the level of performance of the MT engines under conditions comparable to the field test. The reason was to check if accuracy of MT was adequate for post-editing. The dry run tests were performed on the information technology (IT) only, which is potentially the most problematic given the smaller amount of available training data and the lower achieved automatic MT scores. The tests were performed with the baseline SMT engines developed by the partners and with two translators for each translation direction. Each translator was asked to translate 75 segments for a total of around 1,000 words. For the Italian direction positive results were achieved: post-edit efforts were respectively 22% and 26%. For the German direction the same figure were significantly higher: 40% and 45%, which means that translators were more often translating from scratch rather than editing the MT suggestions. The same experiment run with Google Translated reported post-editing effort figures only slightly higher than those for Italian, that is 22% and 31%, respectively.

2.2 Plan revision

After analyzing the outcome of the dry-run test we decided to postpone the field test for English-German on the IT domain. The reason was simply to avoid wasting project resources on a test that would not provide any useful information. Moreover, as the consortium has already reached the capability to run field tests very efficiently, it was also decided to limit the first field test to 2,500 words per language/domain, and to run additional intermediate field tests between the official field tests planned by the project, either to gain additional information (e.g. statistical significance of improvements) or to field test different translation directions. In fact, during the next months it will be decided if the intermediate field tests will be run also on English-French. The decision will also take into account progress made on English to German in the lab tests.

2.3 Protocol

The field test was organized by running two separate experiments with the MateCat Tool. In the first experiment (Day 1), each user translated the first half of a document by receiving translation hints from the translation memory (TM) and a baseline MT engine. In the second, contrastive experiment (Day 2) the same users translated the second half of the document by receiving this time translation hints from the TM and an adapted MT engine. The baseline MT engines were trained on data from the application domain (legal or information technology),
while the adapted engines were trained by also taking advantage of the source part of the whole document as well as its target side produced during Day 1. For each language and domain, four translators were involved. For each source segment to be translated, the MateCat Tool provided three ranked suggestions coming from the TM and the MT engine.

The field test was carried out in two non-consecutive days (Day 1 and Day 2), from 17 to 22 October 2012. The test document was split into two parts of similar size, translated in Day 1 and in Day 2, respectively. Two MT systems (SYS1 and SYS2) were sequentially developed and interfaced with the Matecat Tool, to provide suggestions to the translators at Day 1 and Day 2, respectively. Although the source side of the document was released before the field test started, SYS1 did not exploit any part of it. The interval between the Day 1 and Day 2 was employed for finalizing the project adaptation, and building SYS2. To the purpose, the source side of the whole document and the post-edits of Day 1 were exploited.

For each language pair and domain, four professional translators performed the translation for both Day 1 and Day 2. Hence, a total of 12 translators took part in the field test.

### 2.4 Translated documents

For the IT domain, the field test was carried out for the English-Italian language pair. The text to be translated was taken from a user manual of a specific customer of the industrial partner. For the legal domain, the test has been conducted for both language pairs. The text translated during the two days was taken from a recent motion for a European Parliament resolution published on the EUR-Lex platform.

The choice of the documents was driven by different practical considerations, depending on the domain. For the IT domain, we could choose among documents (mainly software manuals) from different IT companies, which however resulted in rather different MT accuracies. The main reason is that the available training data does cover well all the available sources, which in addition seem rather different from each other, especially for what concerns the translation of technical terms. Hence, the choice went for some not yet translated manuals of a company that is well represented in the training data, so that accuracy was enough good to encourage human translators to post-edit MT suggestions. For the legal domain, the situation was instead much simpler. The training data (JRC Aquis corpus) are indeed quite homogeneous and allow for a rather good MT accuracy. Hence, we selected documents from the European Parliament that were not yet translated (or better whose translation was not yet put on-line) and that addressed a specific topic rather than generic one. This to avoid finding matches of the document segments in the translation memory (MyMemory) and to slightly lower MT accuracy in order to provide more potential for adaptation between day one to day two.

Before starting their task, translators were instructed not to use search engines to look up
terms while translating. This limitation was imposed so as to avoid their using pre-translated documents on the web. They were allowed to perform terminology search on MyMemory [MyMemory].

Statistics on the test documents translated during the field test are reported in Table 1. Figures on the source side (English) refer to the texts the users are requested to translate; for the Legal domain, they are shared among the English-Italian and English-German language pairs. Figures on the target side (Italian) refer to the suggestions given by either the TM or the MT engine and chosen for post-editing by one translator and the post-edits provided by one translator. Although the documents of the Legal domain are the same for both English-Italian and English-German, the processing step resulted in a different amount of segments, but not in a different amount of words. All figures refer to the tokenized texts.

<table>
<thead>
<tr>
<th>field test</th>
<th>test set</th>
<th>#segments</th>
<th>#source words</th>
<th>#suggestion words</th>
<th>#target words</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-It, IT</td>
<td>Day 1</td>
<td>177</td>
<td>3,332</td>
<td>3.488</td>
<td>3,544</td>
</tr>
<tr>
<td></td>
<td>Day 2</td>
<td>176</td>
<td>3,066</td>
<td>3,168</td>
<td>3,336</td>
</tr>
<tr>
<td>En-It, Legal</td>
<td>Day 1</td>
<td>91</td>
<td>2,960</td>
<td>3.056</td>
<td>3,202</td>
</tr>
<tr>
<td></td>
<td>Day 2</td>
<td>90</td>
<td>3,007</td>
<td>3,153</td>
<td>3,421</td>
</tr>
<tr>
<td>En-De, Legal</td>
<td>Day 1</td>
<td>86</td>
<td>2,960</td>
<td>2.741</td>
<td>3,060</td>
</tr>
<tr>
<td></td>
<td>Day 2</td>
<td>89</td>
<td>3,007</td>
<td>2,931</td>
<td>3,080</td>
</tr>
</tbody>
</table>

Table 1: Statistics on test sets used in Day 1 and Day 2 of the field tests. All figures refer to tokenized texts.

2.5 Translation environment

The translation environment used to conduct the field test was the first version of the MateCat tool developed in the first year of the project.

Each test translation project was created using the MateCat tool connected to a specific machine translation engine. The web interface shown in Figure 1 allows to upload the document to translate and then to specify the language pair and MT engine to use.

The translation environment shown in Figure 2 was developed so as to be as fast and easy to use as possible for professional translators. One of the key goals was to minimize the learning curve so that translators could be as efficient with the MateCat tool as they would be with their standard CAT tool without extensive training and/or experience with the tool. Hence, the text is presented in a tabular view where the document is broken down into minimal units (segments). For the active segment (i.e. the segment the translator is editing), three translation matches are presented with their relative quality match and origin (TM, MT). The GUI was carefully
Machine Translation Enhanced Computer Assisted Translation
D5.3: First report on lab and field tests

Figure 1: Project set-up and MT engine selection in MateCat Tool.

designed to allow translators to focus their attention on the active segment and on the supplied suggestions.

While translators usually work on the text segment by segment, the MateCat Tool allows them to also move across segments, edit or proofread their output more times, without any restriction. For each interaction with a segment, the cumulative time needed to elaborate the final version of the translation is collected (see Figure 3). The time to edit is shown on the right-hand side of each segment and then collected in the editing log (see Figure 4).

2.6 Data collection

The MateCat tool provides functionalities to collect information on the translation activities of each translator. For each translated segment we collect the following information:

- matches provided by the TM server (if any) with their relative quality match,
- matches provided by the MT engine (if any) with their relative quality match,
- target segments edited by the translator,
- view of the differences between the first match provided and the final translation,
- time taken to edit each segment (measured by adding the time used to perform multiple edits on the same segment),
Figure 2: MateCat Tool translation environment.

Figure 3: Time to edit for each segment on the right-hand side of the translation editor window.
- post-edit effort measured by the word edit distance between the first match provided and the final translation.

### Job 1000 - Editing Log

*Slowest 5000 segments by time-to-edit*

#### Summary

<table>
<thead>
<tr>
<th>Words</th>
<th>Avg Secs per Word</th>
<th>% of MT</th>
<th>% of TM</th>
<th>Total Time-to-edit</th>
<th>Avg PE Effort %</th>
<th>% of words in too SLOW edits</th>
<th>% of words in too FAST edits</th>
</tr>
</thead>
<tbody>
<tr>
<td>2735</td>
<td>5</td>
<td>99%</td>
<td>1%</td>
<td>03h:47m:32s</td>
<td>31%</td>
<td>0%</td>
<td>1%</td>
</tr>
</tbody>
</table>

#### Editing Details

<table>
<thead>
<tr>
<th>Secs/Word</th>
<th>Job ID</th>
<th>Segment ID</th>
<th>Words</th>
<th>Suggestion source</th>
<th>Match percentage</th>
<th>Time-to-edit</th>
<th>Post-editing effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>344.5</td>
<td>1000</td>
<td>581590</td>
<td>6</td>
<td>Machine Translation</td>
<td>36%</td>
<td>14m:27s</td>
<td>47%</td>
</tr>
</tbody>
</table>

![Editing log](image)

Figure 4: Editing log - Real time report on time to edit and post-editing effort.

The information are displayed in real time on a web interface (see Figure 4) and are also available as a CSV file, which allows for in-depth analysis of the results of each field test.

### 2.7 Data filtering

In order to remove unreliable measurements, we assumed that translation times shorter or longer than a given threshold were not related to the translation work flow but were more likely dependent on errors or on the translators behavior – e.g. translators who stopped translating for whatever reason without saving the segment they were editing.

Two time-based thresholds were applied to filter reliable segments:

- **< 30 seconds per word**: Translation times over 30 sec/word are assumed to be dependent on factors unrelated to the complexity of the source text and more likely dependent on translators behavior – e.g. pauses, distractions, etc.
- **> 0.5 seconds per word**: Translation times below 0.5 sec/word are assumed to be unrealistic for most segments and result of an accidental interaction with the software – e.g. saving a segment without reading or editing it.
2.8 **Key performance indicators**

We used two key performance indicators to measure the effectiveness of the adaptation of the MT server and of the MateCat platform:

- **Time to edit (TTE)**, which is the average translation drafting speed by the translators. TTE aims at measuring the average productivity of translators. In particular, we measure the average time taken by the translator to complete a segment in seconds per word.

- **Post-editing effort (PEE)**, which is the average percentage of word changes applied by the translators on the suggestions provided by the CAT tool. PEE aims at defining the quality of the matches provided by MT engine. We measured the percentage of words edited in a segment by comparing the match provided by the system and the edited segment submitted by the translator. A proprietary function was used which compares two segments and assigns a match percentage based on factors such as same words in the two segments and word order.
2.9 Results

Figures 5–7 and Table 2 report results of key performance indicators (TTE and PEE) for all field tests and for all translators. For the English-Italian direction, significant TTE and PEE
improvements can be observed between Day 1 and Day 2. In particular, on the IT domain, two translators of four improved significantly both figures (t1 and t4), while on the legal domain this was the case of three of four (t6-t8). All observed TTE reductions were statistically significant, while the same hold only for three of the observed PEE variations. By looking at the average productivity gains, on the IT domain we observed a 11.2% improvement in TTE and a 6.5% in PEE, while on the Legal domain we observed a 22.3% gain in TTE and a 10.8% in PEE. Finally, the good correlation observed between PEE and TTE under the different conditions show that very likely the translators were able to took advantage of MT suggestions, and that the adapted MT engine suggestions were in general better. In fact, better PEE effort was observed for 7 translators of 8.

Quite different figures were observed for the English to German translation direction. While all translators improved their translation speed (TTE) from Day 1 to Day 2, only two of them (t9 and t12) indeed showed to take advantage of better MT suggestions. Moreover, PEE figures of all users are in fact all very high, above 41%, and even increase between Day 1 and Day 2 on the average (3.4%). The simple explanation for this is that users disregarded most of the time the provided MT suggestions and basically translated the segments from scratch, because for this language pair the MT quality is not yet suited for post-editing. The significant reductions in TTE (-11.3% on the average) observed from Day 1 to Day 2 are less easy to explain, but maybe a possible reason is that translators simply get more acquainted with the task and the MateCat Tool. In the future, the consortium will try to improve MT quality on English-German, in
order to make post-editing actually useful for gaining productivity. In case of negative result, the consortium will consider running the future field tests on one of the other two investigated translation directions, namely English-French.

The problem of comparing translation productivity over two different documents or two parts of the same document poses in fact issues that will need to be faced in the future.
3 Lab tests

This section analyses the performance of the MT engines employed in the field test by means of standard automatic metrics. The evaluation was carried out on the same test data collected during the field test. Before introducing the results, details about the training, adaptation and testing conditions are reported.

3.1 Training Data

The data available for training the baseline systems are briefly summarized in the following.

- **English-Italian, Information Technology domain**: Translation Memory of IT domain, corpora (KDE4, KDE4-GB, KDeoc, PHP) from the Opus collection, 6 software manuals for a total of about 2M segments, 28M English words and 29M Italian words

- **English-Italian, Legal domain**: JRC-Acquis corpus, for a total of 1.5M segments, 48M English words and 49M Italian words

- **English-German, Legal domain**: Translation Memory of Legal domain, JRC-Acquis corpus, a corpus (ECB) from the Opus collection, News Commentary (v7) and EuroParl (v7), for a total of 5.1M segments, 154M English words and 125M Italian words

These data were exploited for building both the baseline and the adapted systems. The adaptation of SYS2 exploited the whole source side of the test data and the post-edits of the Day 1 of all translators.

Deliverable 1.1 provides a detailed description of the training data and methods followed to train the SMT engines.

3.2 Test Data

Table 3 shows perplexity (PP) and out-of-vocabulary rate (OOV) of the source side of field test documents with respect to 6-gram LMs estimated on the (English) training data used for building the baseline MT engine.

It is worth noticing that, in both domains, the documents proposed in the second day are slightly more difficult than those of the first day: for the IT domain, there is a 26% increase in terms of PP, 42% in terms of OOV rate; for the Legal domain, the PP is higher by 8%, while the OOV rate almost doubles.
### Results

Tables 4–6 report the performance of the baseline and adapted MT systems, respectively SYS1 and SYS2, on the test data. Performance are given in terms of standard automatic metrics: BLEU, TER, and GTM; for BLEU and TER, standard deviation $\sigma$ is reported as well, computed via bootstrap resampling [Clark et al., 2011]. Figures are reported separately for each translator and for each task.

First of all, it is worth noticing that scores of Day 1 and Day 2 are not directly comparable, since documents translated in the two days are different. Nevertheless, even considering that Day 2 documents seem to be more difficult than Day 1 documents (see Table 3), the trend is clear: the adapted systems employed in Day 2 outperform the baseline systems employed in Day 1 with respect to all automatic metrics.

Let us first consider the English-Italian field test (see Tables 4 and 5). Interestingly, we observed that in English-Italian field tests two translators (t3 for the IT domain and t8 for the Legal domain) seemed particularly unsatisfied of MT output and/or inclined to an extensive...
Table 5: Legal domain, English-Italian: BLEU, TER, and GTM scores and corresponding standard deviations (σ) of MT outputs with respect to post-edited texts in Day 1 and Day 2.

<table>
<thead>
<tr>
<th>user</th>
<th>metric</th>
<th>Day 1</th>
<th>Day 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>t5</td>
<td>BLEU/σ</td>
<td>54.59/2.22</td>
<td>56.66/2.13</td>
</tr>
<tr>
<td></td>
<td>TER/σ</td>
<td>30.28/1.69</td>
<td>28.09/1.63</td>
</tr>
<tr>
<td></td>
<td>GTM</td>
<td>0.7735</td>
<td>0.7986</td>
</tr>
<tr>
<td>t6</td>
<td>BLEU/σ</td>
<td>51.59/2.21</td>
<td>55.28/2.46</td>
</tr>
<tr>
<td></td>
<td>TER/σ</td>
<td>33.86/1.74</td>
<td>30.39/1.98</td>
</tr>
<tr>
<td></td>
<td>GTM</td>
<td>0.7491</td>
<td>0.7797</td>
</tr>
<tr>
<td>t7</td>
<td>BLEU/σ</td>
<td>58.22/2.13</td>
<td>66.19/2.16</td>
</tr>
<tr>
<td></td>
<td>TER/σ</td>
<td>26.91/1.55</td>
<td>21.16/1.54</td>
</tr>
<tr>
<td></td>
<td>GTM</td>
<td>0.8029</td>
<td>0.8406</td>
</tr>
<tr>
<td>t8</td>
<td>BLEU/σ</td>
<td>36.47/1.48</td>
<td>37.55/1.40</td>
</tr>
<tr>
<td></td>
<td>TER/σ</td>
<td>40.38/1.38</td>
<td>38.89/1.32</td>
</tr>
<tr>
<td></td>
<td>GTM</td>
<td>0.6804</td>
<td>0.6825</td>
</tr>
</tbody>
</table>

post-editing. For these two translators, the effectiveness of self-tuning is lower than for the others, or even negligible.

The gain achieved in the IT domain is much larger than in the Legal domain on the average. Two reasons can explain this behavior; (i) performance in the Legal domain are already very high so that improving them is hard; (ii) the genre and the lexicon of the IT documents are very specific (to a specific customer), and difficult to be modeled from heterogeneous training data.

As concerns the English-German direction the automatic metrics show that the self-tuning MT module is very effective, although this behavior is not confirmed by the KPIs (see Figure 7) observed in the field test. Further investigation will be carried out to understand the reason of this mismatch through a deeper analysis of the translators’ behavior.

Table 7 shows performance of the baseline (SYS1) and adapted (SYS2) systems on the same test set, namely the Day 2 document. For English-Italian, the scores are computed comparing the MT outputs versus the three post editions not used for project adaptation, so that to reduce biasing, while for English-German all four post editions are exploited. The adapted systems consistently outperform the baseline systems in both domain, clearly showing that the self-tuning modules implemented by partners are very effective.

We are aware that this comparison slightly rewards the adapted systems with respect to the baseline systems, because the reference exploited are post-edits of their actual suggestions. Nevertheless, this bias can only partially explain the big improvement observed; hence, we think that the self-tuning module is really effective in improving MT performance.
### Table 6: Legal domain, English-German: BLEU, TER, and GTM scores and corresponding standard deviations (\(\sigma\)) of MT outputs with respect to post-edited texts in Day 1 and Day 2.

<table>
<thead>
<tr>
<th>user metric</th>
<th>Day 1</th>
<th>Day 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU/(\sigma)</td>
<td>42.80/2.00</td>
<td>48.92/2.33</td>
</tr>
<tr>
<td>TER/(\sigma)</td>
<td>44.57/2.05</td>
<td>38.08/1.94</td>
</tr>
<tr>
<td>GTM</td>
<td>0.7051</td>
<td>0.7487</td>
</tr>
<tr>
<td>BLEU/(\sigma)</td>
<td>35.46/2.66</td>
<td>52.80/3.45</td>
</tr>
<tr>
<td>TER/(\sigma)</td>
<td>57.90/3.24</td>
<td>36.86/3.06</td>
</tr>
<tr>
<td>GTM</td>
<td>0.6140</td>
<td>0.7479</td>
</tr>
<tr>
<td>BLEU/(\sigma)</td>
<td>39.08/2.00</td>
<td>45.31/2.32</td>
</tr>
<tr>
<td>TER/(\sigma)</td>
<td>52.46/2.47</td>
<td>46.91/2.45</td>
</tr>
<tr>
<td>GTM</td>
<td>0.6706</td>
<td>0.7116</td>
</tr>
<tr>
<td>BLEU/(\sigma)</td>
<td>29.73/2.30</td>
<td>37.77/2.37</td>
</tr>
<tr>
<td>TER/(\sigma)</td>
<td>59.95/2.72</td>
<td>48.62/2.07</td>
</tr>
<tr>
<td>GTM</td>
<td>0.5900</td>
<td>0.6507</td>
</tr>
</tbody>
</table>

### Table 7: BLEU, TER and GTM scores, and corresponding standard deviations (\(\sigma\)), of domain- and project-adapted SMT systems on IT and Legal documents of Day 2. See text for more details.

<table>
<thead>
<tr>
<th>metric</th>
<th>En-It IT SYS1</th>
<th>En-It IT SYS2</th>
<th>En-It Legal SYS1</th>
<th>En-It Legal SYS2</th>
<th>En-De Legal SYS1</th>
<th>En-De Legal SYS2</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU/(\sigma)</td>
<td>42.27/1.53</td>
<td>61.09/2.05</td>
<td>59.06/1.62</td>
<td>69.27/1.69</td>
<td>51.46/2.17</td>
<td>67.91/2.33</td>
</tr>
<tr>
<td>TER/(\sigma)</td>
<td>41.93/1.20</td>
<td>28.09/1.36</td>
<td>27.04/1.47</td>
<td>19.79/1.44</td>
<td>44.33/2.20</td>
<td>30.31/2.25</td>
</tr>
<tr>
<td>GTM</td>
<td>0.7946</td>
<td>0.8633</td>
<td>0.8530</td>
<td>0.8971</td>
<td>0.8296</td>
<td>0.9059</td>
</tr>
</tbody>
</table>
4 Conclusions and future work

The first field test was run as planned by using the new MateCat Tool and MT engines developed by the consortium. The overall organization of the field test was rather smooth and efficient, although some improvements could be introduced in the protocol, in order to reduce secondary effects in the contrastive tests. For instance, running two different MT engines on two different portions even of the same document may right comparability issues, as was observed in this field test. Hence, for the future it might be better to interleave, randomly, the use of the two engines among the segments of the same document.

In the following we compare the outcomes of the first field test with the goals set at the begin of the project. The success criteria of the intermediate and final field tests were set in terms of relative improvement both in MT quality (by subjective ranking) and user productivity, against a baseline not using the new MT functionalities developed in the project. For the first field test the target relative improvements are set to 5%, while for the last field test to 15%.

For what concerns the English-Italian direction, we were able to fully reach the target on user productivity. Indeed, the average improvement on both key performance indicators is well above 5%. Concerning MT quality, given the improvements obtained with all automatic metrics (including the TER based on the references), we are rather confident that also this goal has been met. In fact, we will leave a subjective evaluation for next year.

For the English-German direction, instead, consistent MT quality improvements were measured with self-tuning on both tasks [Schwenk et al., 2012]; again we are rather confident that we met the goal relate to MT performance. Unfortunately, for English-German the reached MT quality does not yet meet the requirements by human translators. Both a preliminary dry run on the IT domain and the field test on the legal domain have shown that translators basically were not able to take advantage of MT suggestions. In the future, the consortium will try to get more training data for the English-German information technology task and to improve MT quality on English-German; it will possibly run an intermediate field test before the second official field test on month 24. In case of negative result, the consortium will consider running the future field tests on one of the other two investigated translation directions, namely English-French. In this way, progress in MT will still be pursues on all four language pairs, but only English-Italian and English-French will be field tested.
References

MyMemory. URL http://mymemory.translated.net/.


