Machine Translation for Subtitling: A Large-Scale Evaluation

Thierry Etchegoyhen,[1] **Lindsay Bywood,**[2] **Mark Fishel,**[3] **Panayota Georgakopoulou,**[4] **Jie Jiang,**[5] **Gerard van Loenhout**,[6] **Arantza del Pozo**,[1] **Mirjam Sepesy Maučec**,[7] **Anja Turner**,[8] **Martin Volk**[3]

[1] Vicomtech-IK4, Donostia San Sebastián, Spain, [2] Voice & Script International, London, United Kingdom,
[3] Text Shuttle GmbH, Zurich, Switzerland, [4] Deluxe Media, London, United Kingdom,
[5] Capita TI, Greater Manchester, United Kingdom, [6] invision Ondertiteling, Amsterdam, The Netherlands,
[7] University of Maribor, Maribor, Slovenia, [8] Titelbild Subtitling and Translation, Berlin, Germany

[1]{tetchegoyhen, adelpozo}@vicomtech.org, [2]lindsay@vsi.tv, [3]{fishel, volk}@cl.uzh.ch, [4]yota.georgakopoulou@bydeluxe.com,
[5]jie.jiang@capita-ti.com, [6]gerard@ondertiteling.nl, [7]mirjam.sepesy@uni-mb.si, [8]Anja.Turner@titelbild.de

Abstract

This article describes a large-scale evaluation of the use of Statistical Machine Translation for professional subtitling. The work was carried out within the FP7 EU-funded project SUMAT and involved two rounds of evaluation: a quality evaluation and a measure of productivity gain/loss. We present the SMT systems built for the project and the corpora they were trained on, which combine professionally created and crowd-sourced data. Evaluation goals, methodology and results are presented for the eleven translation pairs that were evaluated by professional subtitlers. Overall, a majority of the machine translated subtitles received good quality ratings. The results were also positive in terms of productivity, with a global gain approaching 40%. We also evaluated the impact of applying quality estimation and filtering of poor MT output, which resulted in higher productivity gains for filtered files as opposed to fully machine-translated files. Finally, we present and discuss feedback from the subtitlers who participated in the evaluation, a key aspect for any eventual adoption of machine translation technology in professional subtitling.

Keywords: statistical machine translation, user evaluation, subtitling

1. Introduction

Thanks to the availability of large amounts of parallel and monolingual corpora, statistical machine translation (SMT) systems are being developed for a wide range of domains and real-world applications. Subtitling has been previously recognized as a domain which was likely to benefit from machine translation technology (Volk, 2009). Although the variety of genres and content covered in subtitling represents a challenge for MT technology, subtitles are short and meaningful units which can serve as adequate training material for SMT systems.

In this paper, we describe a large-scale evaluation of SMT technology for professional subtitling work and present results describing the quality and usefulness of SMT systems whose cores were built on professionally created subtitle corpora (Petukhova et al., 2012). Quality evaluation was undertaken by professional subtitlers, who post-edited machine translated output, rated individual subtitles in terms of their quality, and collected recurrent errors. Usefulness of the SMT systems in the domain is also assessed through a measure of productivity gain/loss, comparing timed post-editing of machine translated output to translation from source.

The work we describe is part of the SUMAT project (www.sumat-project.eu), funded through the EU ICT Policy Support Programme (2011-2014), and involving nine partners: four subtitle companies (Deluxe Media, InVision, Titelbild, Voice & Script International) and five technical partners (Athens Technology Center, CapitaTI, TextShuttle, University of Maribor and Vicomtech-IK4). The goal of the project is to explore the impact of machine translation on subtitle translation and develop an online subtitle translation service catering for nine European languages combined into 14 bidirectional language pairs: English-Dutch, English-French, English-German, English-Portuguese, English-Spanish, English-Swedish, and Serbian-Slovenian. A subset of the language pairs was used for the evaluation, selected in terms of market potential, with Serbian-Slovenian as a test-case of an under-resourced language pair. The selected translation pairs were: English into Dutch, French, German, Portuguese, Spanish & Swedish; French, German & Spanish into English; and Serbian-Slovenian in both directions.

We first present an overview of the systems developed for the project and the corpora used to build them, followed by a description of the quality evaluation design and results. We then describe the experimental design and results for the productivity evaluation round, and the feedback collected throughout the evaluation.

2. SUMAT: Corpora & Systems

At their core, the machine translation systems developed within the project are phrase-based SMT systems (Koehn et al., 2003), built with the Moses toolkit (Koehn et al., 2007) and trained on professional parallel corpora provided by the subtitle companies in the SUMAT consortium. More than 2.5 million parallel subtitles were added to the resources described in (Petukhova et al., 2012), resulting in an average of 1 million aligned parallel subtitles for our language pairs, and approximately 15 million monolingual subtitles overall which were used to train the language model components of the systems.

To improve systems coverage and quality, various approaches have been explored over the course of the project (Etchegoyhen et al., 2013), from the inclusion of various linguistic features to domain adaptation through additional data incorporation and selection. The most successful approach, in terms of improvement in automated metrics and systems efficiency, was translation model domain adaptation (Sennrich, 2012). In this approach, separately trained translation models are combined into a joint model and their combination weights are optimized for a specific domain by reducing the perplexity of the resulting model on a domain-specific dataset. For our models, the systems were tuned on the SUMAT development sets.

We tested various combinations of models, built on separate data, eventually retaining the optimal combination which consisted of models trained on the SUMAT, Europarl and OpenSubs corpora.[1] Tables 1 and 2 provide an overview of the parallel corpora used to train the systems that were evaluated, and the systems' respective scores on the SUMAT test sets.[2] For each language pair, the development and test sets consisted of 2000 and 4000 subtitles respectively, randomly selected across genres and domains.

	SUMAT	Europarl	OpenSubs
EN-DE	1 488 341	3 763 616	4 631 974
EN-ES	978 705	1 011 054	31 456 400
EN-FR	1 326 616	977 225	19 006 604
EN-NL	1 397 810	3 762 663	21 260 772
EN-PT	762 490	4 223 816	20 128 490
EN-SV	786 783	1 862 234	7 302 603
SL-SR	167 717	n/a	1 921 087

Table 1: Parallel training data

3. Quality Evaluation

The first round of evaluation was designed to estimate the quality of the systems. Subtitles were assigned quality scores by subtitlers and we evaluated the correlation between these scores and automated metrics computed on post-edited files. We also asked subtitlers for general feedback on the post-editing experience and any additional comments they had regarding their perception of MT output quality. Furthermore, we collected recurrent MT errors in order to gradually improve the systems throughout the three phases of the evaluation, each phase consisting of MT output evaluation followed by systems improvement.

Each phase involved two subtitlers per translation pair, who were asked to post-edit to their usual translation quality standards and perform the task in their usual subtitling software environment. There were two input files for each of the first two phases, and one for the third, consisting of both scripted and unscripted material from different genres and domains (e.g. drama, documentaries, magazine programmes, corporate talk shows). Note that, to increase the overall amount of different subtitles to be annotated, the evaluators did not process the same files. There was thus no measure of inter-annotator agreement in this phase. Correlation measures between ratings and post-editing effort were however computed, and are discussed in Section 3.3. Overall, 27565 subtitles were post-edited, rated and annotated in this evaluation round. The main aspects and results of the evaluation are described hereafter.

	BLEU	TER	Equal	Lev5
EN to DE	19.7	66.3	6.02	10.65
EN to ES	32.3	51.3	3.92	9.88
EN to FR	28.2	59.4	2.80	8.62
EN to NL	24.3	58.8	4.51	9.57
EN to PT	25.8	56.5	2.92	8.85
EN to SV	33.0	50.5	11.9	20.8
DE to EN	23.2	60.0	6.25	12.16
ES to EN	36.2	47.5	5.12	12.93
FR to EN	29.2	54.9	3.23	9.03
NL to EN	28.0	55.2	5.13	10.76
PT to EN	33.1	48.1	5.61	10.90
SL to SR	17.8	66.1	4.0	11.6
SR to SL	19.1	65.0	4.8	12.3
SV to EN	34.3	47.3	11.6	20.6

Table 2: Systems evaluation on SUMAT test sets

3.1. Quality Rating

First, professional subtitlers evaluated the quality of machine translation output by assigning a score to each machine translated subtitle. The rating scale was the one established for the WMT 2012 Shared Task on MT quality estimation:[3] each subtitle was to be annotated on a 1 to 5 scale indicating the amount of post-editing effort, where subtitles rated 1 signal incomprehensible and unusable MT, and subtitles rated 5 denote perfectly clear and intelligible MT output, with little to no post-editing required. Figure 1 summarizes the results for our SMT systems, taking the average of all evaluated translation pairs. The results rise in percentage from poor to good MT, with a predominance of machine translated output that required little post-editing effort. Given the unrestricted nature of the input data, which covered various genres, domains and language registers, these results can be considered quite satisfactory. Table 3 summarizes the average rating assigned by the evaluators, and the average results on automated metrics using post-edited files as references, for all translation pairs in the experiment. With post-editing in mind, two results are worth noting: 1 in 5 machine translated subtitles required no post-editing at all and more than 1 in 3 required less than five character-level editing steps. These two measures indicate a substantial volume of unambiguously useful MT output, with only minor post-editing needed.

[1] For both Europarl and OpenSubs, we used the corpora available in the OPUS repository (Tiedemann, 2012) and experimented with various types of data selection in distinct language pairs (e.g., data selection through bilingual cross-entropy difference (Axelrod et al., 2011)).

[2] *Equal* indicates the percentage of MT output identical to the reference and *Lev5* is a Levenshtein-distance metric measuring the percentage of MT output that can reach a reference translation in less than five character editing steps (Volk, 2009).

[3] http://www.statmt.org/wmt12/quality-estimation-task.html

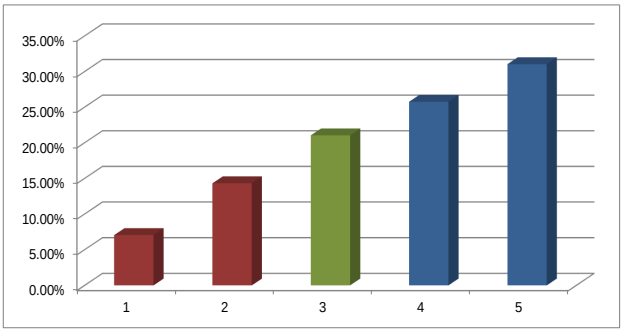

Figure 1: Global rating averages

	Averages
Rating	3.60
BLEU	39.69
TER	44.88
Equal	20.1
Lev5	35.69

Table 3: Average metrics on post-edited files

3.2. Translation Pair Comparison

The previous results were based on global averages for automated metrics and ratings. Figure 2 presents a comparative view, where the following elements were measured for each translation pair: i) the BLEU scores on the SUMAT test sets, ii) the average hBLEU scores on the post-edited files, and iii) the average rating, ported to a [0-50] scale for easier visualization.[4]

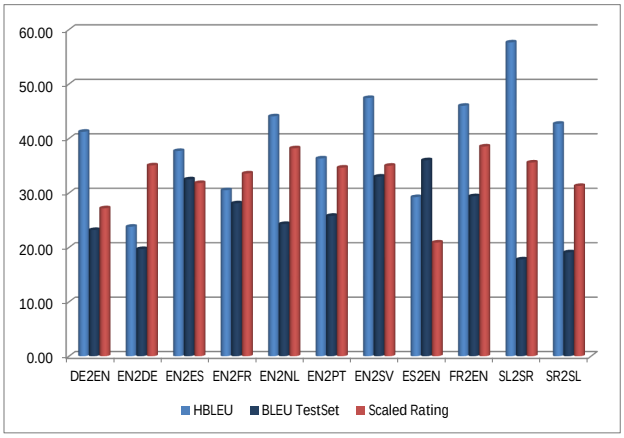

Figure 2: Language pairs comparative results

As hBLEU scores are measured on post-edited files, they are expected to be higher than the BLEU scores on test sets, as there should be a higher amount of common n-grams in a transformed (i.e., post-edited) reference text than in an independently translated reference. As can be seen in the figure, this has been the case for all but one translation pairs, namely, Spanish to English. This is one of the surprising results in this evaluation round, given that this translation pair is the highest scoring one on the SUMAT test sets. The hBLEU and hTER results were consistent with the manual quality rating, where Spanish to English was the only translation pair where the volume of MT output rated as poor was larger than the one rated as good. A manual examination of a subset of the annotation showed that a noticeable amount of MT output rated 1 or 2 was actually grammatically correct, but fully discarded by post-editors as they had offered a different translation alternative. Although it could be argued that the letter of the evaluation guidelines was partially respected here, with low scores given for fully discarded MT output, this translation pair stands isolated with respect to the way in which grammatically correct MT output was considered. Finally, the subtitlers working on this language pair noted that several of the source files were difficult to use, with audio and template issues that rendered the post-editing task all the more difficult.

Another notable result is the very positive evaluation scores obtained for Serbian and Slovenian, which scored the lowest on the SUMAT test sets but gave the highest ratings and best metrics on post-edited files. Previous manual examination of the test sets had shown them to contain large volumes of difficult and unusual text, and the results from this evaluation round seem to confirm that the quality of the SMT systems for this language pair is undervalued by current test set scores.

For the other language pairs, the differential between metrics is quite uniform, with hBLEU scores consistently higher than test set BLEU scores, and quality ratings seemingly correlating with the automated metrics. A finer-grained analysis of correlation aspects is presented in the next section.

3.3. Correlation Measures

To estimate the degree to which rating was correlated to the actual post-editing effort, we computed the Pearson correlation coefficient between average ratings and automated metrics for each post-edited file. As can be seen in Table 4, when estimated on all translation pairs, the results ranged from moderate correlation for BLEU to strong for TER (both above statistical significance). As expected, the correlation between the percentage of subtitles rated 5 and Lev5 was strong.

A closer examination made apparent that three of the eleven language pairs, namely German to English, English to Spanish and English to Portuguese, showed weak inverse correlation below statistical significance. Excluding these three pairs resulted in the figures shown in the third and fourth lines of Table 4, with stronger correlation for all metrics. These results indicate that rating was strongly correlated with the actual post-editing effort, except in a minority of cases where a larger number of subtitlers would have been needed to balance individual rating to post-editing effort disparities.

[4] We present results in terms of BLEU scores here, rather than TER, as it makes it easier to compare them with average ratings, an increase in both being positive. The BLEU and TER metrics were very strongly correlated, with a Pearson correlation coefficient of 0.96 $\pm$ 0.002, and either one can thus be safely used to present the results.

	Rating-TER	Rating-BLEU	Rating-Lev5
r (all pairs)	-0.626	0.574	0.715
p-value (all pairs)	0.030	0.039	0.019
r (8 pairs)	-0.734	0.746	0.822
p-value (8 pairs)	0.024	0.023	0.014

Table 4: Rating-Metric correlations

3.4. Error Collection

As mentioned above, we also collected recurrent MT errors for possible correction by the technical partners in the project. For this purpose, we provided evaluators with an error taxonomy and asked them to indicate the errors for subtitles rated 3 or higher only, since we assumed that lower rated subtitles would contain too many errors to properly distinguish them. The taxonomy included: *agr* for grammatical agreement errors; *miss(ing)* for content words/segments that were lost in the translation process; *order* for grammatical ordering errors in the target language; *phrase* for any multiword expression wrongly treated as separate words, or any separate words wrongly translated as a unit; *cap* for capitalization errors; *punc* for punctuation errors; *spell(ing)* for any spelling mistake; *length* for any machine translated output deemed too long given constraints on subtitle length; and *trans(lation)* for mistranslations, a large category that includes any lexical or phrasal mistranslation.

The results are given in Figure 3.[5] Overall, the distribution shows a dominance of mistranslations, followed by agreement errors and segments lost in the translation process. This is not unexpected for phrase-based SMT systems, with no access to linguistic information to handle grammatical errors like agreement, for instance. Over the three phases, the systems were improved for other more manageable categories, e.g. punctuation, capitalization and multi-word units. Given the amount of named entities in the overall subtitling domain, improving the systems in this regard was strongly requested by post-editors and led to the systems being retrained with truecasing. Finally, the results on the subtitle-specific category length are also worth noting; further research would be necessary to tune the statistical translation engine towards producing output adjusted to subtitle length constraints in the target language (see (Aziz et al., 2012) for an approach along those lines).

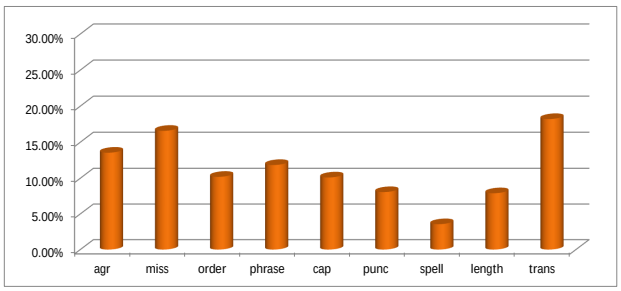

Figure 3: Global distribution of errors

4. Productivity Measurement

The second major phase of the evaluation focused on measuring productivity gain/loss by comparing the time needed to translate a subtitle file from source vs. post-editing machine translated output. We hypothesized that this type of evaluation could be a strong additional indicator of the usefulness of machine translation for pamongrofessional subtitling. A pilot study was executed in 2012 for the English-Swedish language pair, as described in (Bywood et al., 2012). There were large variations in the results, which showed both increases and decreases in productivity for subtitlers post-editing MT output.

4.1. Experimental Design

The experimental design involved the same translation pairs used for the quality evaluation round, with two subtitlers per pair. Productivity was measured in terms of subtitles per minute, comparing speed of post-editing to translation from source.

In this round, an additional scenario was implemented, with automatic quality estimation and filtering of MT output.[6] In this configuration, poor machine translated subtitles were removed from the MT output files, thus providing post-editors with empty MT subtitles to be translated from the source; good quality MT went through the filters unmodified, to be post-edited. The main driver for adding this third use-case came from general feedback provided by subtitlers in the quality evaluation round. Although the feedback included comments regarding the surprisingly good MT quality for some translation pairs, with post-editing becoming easier after some practice, it also included repeated men-

[5]For the distribution of errors shown here, the *agr* category has been weighted, to account for a change in the error typology which was effected in phases 2 and 3. In the first phase, the *trans* category was omitted, as this class of errors is difficult to correct in SMT systems and no technical fixes were envisioned. However, subtitlers requested the inclusion of this error category, as they frequently felt the need to indicate such translation errors. During Phase 1, mistranslations were eventually marked as *agr* errors, thus over-representing this category. The above figure provides for a more representative view of the distribution of errors, using the ratio of *trans* and *agr* errors that were found in phases 2 and 3.

[6]Quality estimation was performed with the QuEst toolkit (Specia et al., 2013). Space limitations prevent us from providing the complete experimental design and results here. Summarizing the approach, ROC curves were constructed to choose between different binary classification schemes, and further heuristic rules were applied to avoid the over-discarding issue. Experimental results showed that the binary classification equal error rate varied from 22.07% to 42.87% across different language pairs, while the overall discarding rate was kept under 25%, to match the amount of poor MT output observed during the first evaluation round.

tions of the additional cognitive effort required to work with poor MT output. Introducing a mixed-case scenario with integrated quality estimation and filtering was an attempt to evaluate a possible solution for this important issue.

For each translation pair, two subtitlers handled the same 6 files each: 2 machine translated files, to be post-edited; 2 source files, to be translated directly ; and 2 files where machine translated subtitles classified as below required quality had been removed. In this latter scenario, subtitlers thus performed both post-editing and translation from source. Overall, 114 files and 37104 subtitles were processed.[7]

Input files were also selected to reflect the scripted/unscripted dichotomy, as it was hypothesized that this particular characteristic might affect machine translation quality.[8] For the English and Spanish source files, 3 scripted and 3 unscripted files were chosen, each subgroup composed of 1 file used as benchmark, 1 file fully machine translated and a third file with MT content partially filtered, as described above.

Finally, although post-editing was timed in this evaluation round to measure productivity differences, subtitlers were instructed to work at their normal rhythm, using their usual subtitling software environment, and to post-edit or translate to their usual quality standards. The two source files which were subtitled directly served as benchmarks for productivity gain/loss measurement.

4.2. Productivity Gain/Loss

The global results are shown in Figure 4, with productivity gain/loss expressed in percentage of speed increase/decrease over the benchmark source files that were translated directly from the source.[9]

Taking the average productivity for each translation pair, and considering all machine translated files, filtered and unfiltered, the gain in productivity reached 38.2%. This gain can be viewed as a significant positive result, considering that machine translation was applied in the open subtitling domain, which covers highly varying language across the board.

The scripted/unscripted split gave surprising results for the 7 translation pairs where it was introduced, with unscripted files giving markedly better results than scripted files. It would seem premature to conclude from these results that the distinction is not worth maintaining, given the respective properties of these two categories; further evaluations of this dichotomy will most likely be necessary. However, these results do show the high dependence of MT quality and usefulness on the contents of a specific source file.

[7]Note that for Serbian-Slovenian the number of processed files was reduced in this round, as one of the post-editors had to retire from the evaluation process for independent reasons. Three files, all scripted, were handled by one post-editor, the 3 files being split between source, full MT and filtered, as designed for the other language pairs.

[8]Scripted material typically contains more controlled, or predictable, language, whereas unscripted shows tend to contain spontaneous speech, interrupted dialog, ellipsis, and other properties which are difficult for natural language processing in general.

[9]In the figure, MT indicates files that contained fully machine translated content, and FILT denotes machine translated files with partially filtered content, as previously described.

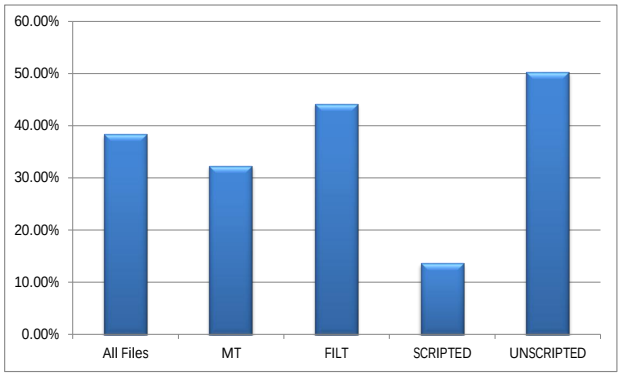

Figure 4: Global productivity gain/loss (in %)

Among the unscripted files used in this evaluation round, one English source file in particular, a subtitled talk show, gave productivity gains that overshadowed all other results. As this file was among the filtered ones, it could also be the case that the filtering process produced an optimal combination of MT output and empty subtitles to be translated directly.

The experiment in using quality estimation and filtering to provide a mixed post-editing scenario also gave positive results. Filtered files showed a 36.7% gain over files containing the complete MT output, a trend that was unaffected by the scripted/unscripted distinction. A more thorough investigation and exploitation of the filtering approach for post-editing open domain material seems to be a promising path for future research and exploitation of MT technology.

4.3. Translation Pair Comparison

Productivity gain/losses results per translation pair are shown in Figure 5. At the two extremes are the same translation pairs found to be successful or problematic in the first round of evaluation: Serbian-Slovenian, two closely related languages, gave the best results, whereas Spanish to English and English to German showed a slight loss and no gain, respectively. The latter translation pair is notorious for being difficult for SMT in general, which makes this particular result unsurprising. Spanish to English results, however, were unexpected, matching the poor results observed in Round 1.

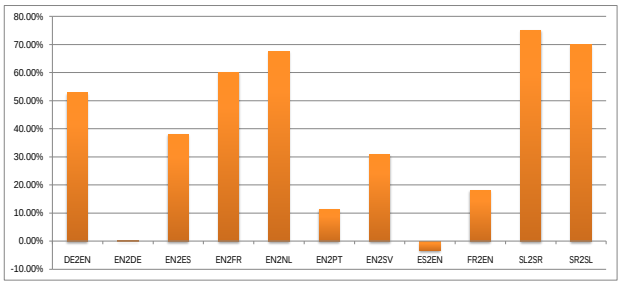

Figure 5: Productivity gain/loss per translation pair (in %)

The first and most likely explanation for the latter results is the impact of the specific source files that were used. As discussed in the previous section, MT quality will vary

depending on the specific content of a source file (vocabulary, simplicity of syntactic constructs, etc.), sometimes to a large degree, and post-editors signaled, in this round as well, that the source files were particularly difficult to work with. Another possible explanation is the perception of the task by the subtitlers: as will be discussed in section 5., post-editing is not always well-perceived among professional subtitlers, and this can at times translate into larger amounts of discarded MT output.

For the other translation pairs, English as a source language gave better results, which was also somewhat unexpected. As a matter of fact, the English language models are the largest ones, being trained with English data from the target side of all translation pairs. Translation into English was thus expected to perform better on average, on a par with results on the project's test sets. Overall, metrics results on post-edited files were very similar in both rounds, as shown in Figure 6.[10]

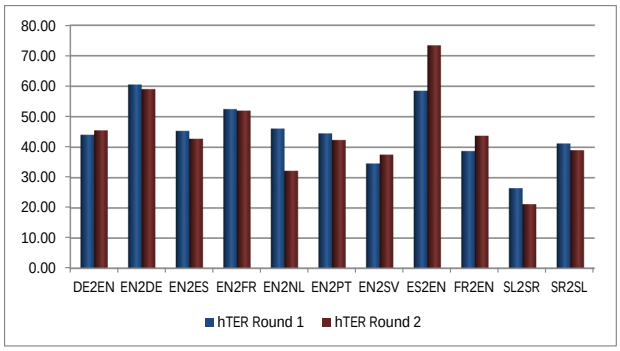

Figure 6: hTER metric results

The standard deviation from the mean on productivity grounds was 50.50 for the translation pairs in this evaluation round. This large variance is unsurprising considering the multiple parameters that impact the post-editing process: source file quality and type; data on which the SMT systems were trained and tuned for each translation pair; perception of the task by post-editors; varying degrees of practice in post-editing; among other significant variables. Given this, the most notable result was the fact that 9 of the 11 translation pairs evaluated in this round showed significant productivity gain, with only minor loss and a neutral result respectively for the remaining two translation pairs.

4.4. Correlation and Variation

We measured the correlation between productivity gain/loss and automated metrics computed on post-edited files, for all translation pairs. For the TER metric on all translation pairs, the Pearson correlation coefficient was -0.34, a weak correlation, with a p-value of 0.09, below statistical significance. The main conclusion that can be drawn from this result is that the post-editing effort, i.e. the amount of transformations performed on MT output, is not the most impactful indicator of productivity gain/loss. This seems to indicate that post-editing practice, or smoother integration of post-editing in the subtitlers own translation flow, has a higher impact: a subtitler may actually use more of the MT output, post-editing rather than deleting the output and translating from scratch, and perform this task faster than another subtitler who post-edited similar or lower amounts of MT output. Deciding to post-edit a machine translated subtitle, vs. deleting it and translating directly from the source, is in itself a crucial part of post-editing, on a par with efficiency in transforming MT output. Variation on these grounds is expected between subtitlers, especially considering that the vast majority of the subtitlers who participated in the evaluation rounds had no previous experience in post-editing.

To evaluate the variation between subtitlers, we computed the absolute differences between productivity gain/loss for each of the two subtitlers working on the same translation pairs and files, as well as the absolute difference in terms of metrics on post-edited files. The results are shown in Figure 7.

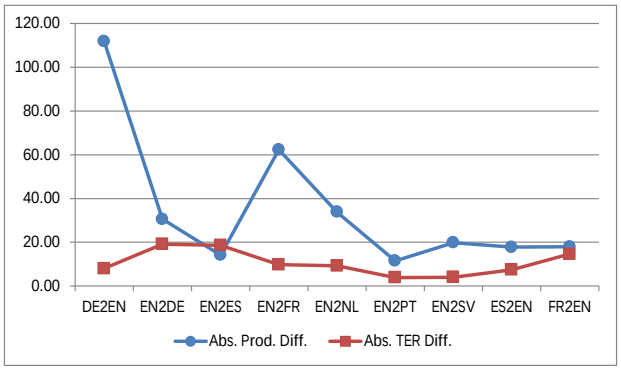

Figure 7: Absolute differences: productivity and TER

Although absolute productivity and TER differences should not be directly compared, their respective distributions show an interesting trend, with large variations between subtitlers in terms of productivity and less variation in terms of post-editing effort, as measured on the TER metric. Average absolute difference and standard deviation from the mean, are given in Table 5, where the averages and deviations were computed using productivity and TER results from all translation pairs.

	Productivity	TER
Average	35.50	10.51
Stdev	30.66	5.39

Table 5: Absolute differences: averages and deviations

The results show medium to low dispersion in terms of post-editing translation error rate, but a larger variation in terms of productivity. This can be viewed as a confirmation of the impact on productivity of variables other than post-editing effort, as discussed above.

[10]For clarity of presentation, we only present results on the TER metric; the other metrics showed comparable results between evaluation rounds as well.

5. User Feedback

For both evaluation rounds, subtitlers were asked to provide feedback on the post-editing process and the overall experience of using MT for professional subtitling. Objective results, such as productivity gains or MT quality evaluation results, might be promising, but of equal importance is the perception of the task by professional subtitlers, for whom post-editing is not part of the usual workflow. Below are excerpts from the free-form feedback provided, both positive and negative:

- EN2PT (Round 1): 'Once I got it going, it was quite easy.'
- EN2ES (Round 1): 'With shorter and simpler sentences like the ones in this episode, I think having the translation there saves quite some time.'
- EN2SV (Round 1): 'Hugely improved since last year! I have many 4 and 5 and am really quite amazed. There's still a long way to go, but it's usable already now.'
- FR2EN (Round 1): 'Overall pretty good. Simple sentences were usually perfect, but the machine has problems when the sentence is complicated []'
- ES2EN (Round 2): 'Generally speaking, it was only in very rare instances that the level of translation generated was such that it needed little or no editing at all. Frequently, it was just easier to get rid of everything and start from scratch.'
- EN2FR (Round 2): 'I was quite surprised by the quality of the translation of [File 1], which was pretty good, given the difficulty of the translation. Most of the terms and expressions were correctly translated, and this was really time-gaining. But for [File 2], I had to delete everything and translate from scratch. It was full of mistakes (even spelling mistakes), false sense, mistranslations and the translation was way too long.'
- EN2PT (Round 2): 'Sometimes the MT subtitles get it right and its limited vocabulary is often passable. But if I were to deliver a finished job by usual standards, I wouldn't use more than 30% of it. Sometimes it's just an order issue, but the general quality of it is not good. We can certainly use lots of words, but only rarely whole sentences.'
- EN2ES (Round 1): '[...] I guess all in all everything depends on the type of show being subtitled.'

Feedback from both rounds included both positive and negative comments, although the general trend was more negative in the second round than in the first. This might be due to the specific files being machine translated in each case, although the post-editing metrics were very similar, which would seem to discard this explanation. Another reason might come from the nature of the productivity evaluation task, which conveyed time-measurement pressure and negative connotations regarding the evolution of translation work conditions in the subtitling industry. The feedback was nonetheless precise and extensive in both positive and negative cases, and will need to be taken into account for any integration of MT in the subtitling workflow.

To quantify the perception of the tasks, a questionnaire was provided to participants in Round 2, asking them to rate various aspects of the post-editing process. We provided a 1 to 5 scale, from poor to excellent in this order, with 3 denoting neutral appreciation. The average results from Round 2 are given below:

- *How did you find the post-editing process?* 2.37
- *How did you find the mixed task?* 2.49
- *What is your overall perception of the quality of machine translated subtitles?* 2.18

As shown above, the average results were on the negative side of the scale, despite the generally good results on objective metrics. This correlates with the overall tone of the free-form feedback in the second round, and clearly marks that the task was perceived rather negatively overall. Improving on these aspects is of paramount importance, in order for machine translation technology to be a successful part of the translation process in subtitling, on a par with translation memory use in text translation.

Taking the general feedback into account, the core aspects below will need to see an improvement in order to compensate for the more frustrating parts of subtitle post-editing:

1. *Improving the quality of machine translation*: although several subtitlers expressed their surprise at the quality of MT output for some files, even for quite colloquial or fragmented text, overall a new leap forward in MT quality would help reduce the cognitive effort in post-editing machine translated text in the open subtitling domains.
2. *Improving quality estimation and filtering*: the preliminary results on QE and filtering gave quite good outcomes in terms of productivity, although the task was still perceived negatively. In some instances, it was not clear to the subtitlers why a given subtitle, which seemed easy to machine translate, had been automatically filtered. Filtering also necessitates going back to the original source in between post-editing steps, a process which is unusual and would require further practice. Finally, improving quality estimation would help reduce the cognitive effort needed for a human estimation of the machine translated output, and allow subtitlers to focus on the less frustrating parts of the post-editing task.
3. *Augmenting post-editing user-interfaces*: with the varying domains and genres found in subtitling, most of the typical errors made by SMT engines can be predicted to occur on a regular basis. As mentioned by several evaluators, improved user interfaces, with integrated short-cuts to enable the efficient correction of the most typical MT errors (e.g. adjacent word reordering), would greatly improve the post-editing experience.

In general, it appears that more thorough communication is necessary between researchers in machine translation and professional subtitlers. On the one hand, it will be necessary to better describe the current and projected limitations of MT technology, as post-editors were often frustrated by MT errors that seemed unproblematic to a trained human translator.[11] A better description and understanding of MT limitations will help decrease the frustrating aspects of post-editing. On the other hand, a more precise understanding of the cognitive and practical efforts involved in post-editing subtitles from open-domains, including the most typical correction processes, will enable the development of SMT systems and post-editing interfaces that address the actual needs of professional users.

6. Conclusions

In this paper, we described a large-scale evaluation of machine translation for subtitling. The MT systems that were used make full use of both professionally-created and crowd-sourced corpora, aiming to achieve an optimal balance between the use of large language resources and system adaptation for the many domains and genres found in subtitling.

The quality evaluation round yielded positive results, with a consistent distribution of MT output rising from lower percentages of poor quality output to higher amounts of good quality machine translated subtitles. Measures of productivity gain/loss were also positive, with an overall increase of nearly 40% in terms of subtitles per minute. The second evaluation round included a measure of the impact of MT quality estimation, coupled with filtering of poor MT output, which also yielded positive results.

On the negative side, the cognitive effort in assessing poor MT output, before proceeding with either significant post-editing or re-translation, is an aspect that clearly needs to be taken into account for a useful integration of MT technology. Further improvements in terms of machine translation quality, combined with better quality estimation, would help reduce the frustrating aspects of post-editing, as would a better communication of MT limitations and adapted user-interfaces for post-editing.

Overall, the SMT systems developed within the SUMAT project, which were trained and tuned on subtitles, have shown promising results in terms of quality and usefulness for a professional use applied to the variety of domains and genres found in subtitling.

7. Acknowledgements

The authors wish to thank all the subtitlers who took part in the evaluation, for their time, effort, feedback and expertise, which was central to the work described in this article. We also wish to thank the three anonymous LREC reviewers for their insightful comments; opinions, remaining errors and omissions are our own. SUMAT is funded through the EU ICT Policy Support Programme (2011-2014), under grant agreement 270919.

[11] Contextual errors, e.g. registry changes in languages that distinguish between polite and familiar forms, are a typical example, mentioned across the evaluation rounds.

8. References

Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 355–362. Association for Computational Linguistics.

Aziz, W., de Sousa, S. C., and Specia, L. (2012). Cross-lingual sentence compression for subtitles. In *proceedings of the 16th Annual Conference of the European Association for Machine Translation, Trento, Italy*.

Bywood, L., Georgakopoulou, P., Volk, M., and Fishel, M. (2012). What is the productivity gain in machine translation of subtitles? In *Proceedings of the 9th International Conference on Language Transfer in Audiovisual Media, Berlin, Germany*.

Etchegoyhen, T., Fishel, M., Jiang, J., and Sepesy Maučec, M. (2013). SMT experiments for commercial translation of subtitles. In *Proceedings of MT Summit XIV, User Track, Nice, France*.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics-Volume 1*, pages 48–54. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

Petukhova, V., Agerri, R., Fishel, M., Penkale, S., del Pozo, A., Maucec, M. S., Way, A., Georgakopoulou, P., and Volk, M. (2012). SUMAT: Data collection and parallel corpus compilation for machine translation of subtitles. In *Proceedings of the 8th Language Resources and Evaluation Conference*, pages 21–28.

Sennrich, R. (2012). Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 539–549. Association for Computational Linguistics.

Specia, L., Shah, K., de Souza, J. G., Cohn, T., and Kessler, F. B. (2013). QuEst–a translation quality estimation framework. *Proceedings of the 51st ACL: System Demonstrations*, pages 79–84.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th Language Resources and Evaluation Conference*, pages 2214–2218.

Volk, M. (2009). The automatic translation of film subtitles. A machine translation success story? *Journal for Language Technology and Computational Linguistics*, 24(3):115–128.