Machine Translation for Subtitling: A Large-Scale Evaluation

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

[1]{tetchegoyhen, adelpozo}@vicomtech.org - Vicomtech-IK4, San Sebastián, Spain
[2]lindsay@vsi.tv - Voice & Script International, London, UK
[3]yota.georgakopoulou@bydeluxe.com - Deluxe Media, London, UK
[4]mirjam.sepesy@uni-mb.si - University of Maribor, Maribor, Slovenia
[5]{fishel, volk}@cl.uzh.ch - Text Shuttle GmbH, Zurich, Switzerland
[6]jie.jiang@capita-ti.com - Capita TI, London, United Kingdom
[7]gerard@ondertiteling.nl - Invision Ondertiteling, Amsterdam, The Netherlands
[8]Anja.Turner@titelbild.de - Titelbild Subtitling and Translation, Berlin, Germany

1. Introduction

As a result of 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. In this paper, we describe a large-scale evaluation of SMT technology for professional subtitling work and present results on the quality and usefulness of SMT systems whose core was built on professionally created subtitle corpora [6]. Quality evaluation was undertaken by professional subtitlers, who post-edited machine translated output, ranked 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 they were built on, followed by a description of the quality evaluation design and results. Finally, we will describe the goals and design of the productivity measurement evaluation round, which is under way, with final results expected in February 2014.

2. SUMAT: Corpora & Systems

At their core, the machine translation systems developed within the project are phrase-based SMT systems [5], built with the Moses toolkit [4] 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 [6], 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 [3], 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 grounds, has been translation model domain adaptation [7]. 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

	SUMAT	Europarl	OpenSubs
EN-DE	1 488 341	3 763 616	4 631 974
EN-ES	978 705	1 011 054	20 000 000
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

	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

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 consisting 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]

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). Overall, 27 565 subtitles were post-edited, ranked and annotated in this evaluation round. The main aspects and results of the evaluation are described hereafter.

3.1. Quality Ranking

First, professional subtitlers evaluated the quality of machine translation output by assigning a score to each machine translated subtitle. The ranking 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 ranked 1 signal incomprehensible and unusable MT, and subtitles ranked 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 follow a staircase distribution, rising 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 ranking 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 [9] and experimented with various types of data selection in distinct language pairs (e.g., data selection through bilingual cross-entropy difference [1]).

[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 [10].

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

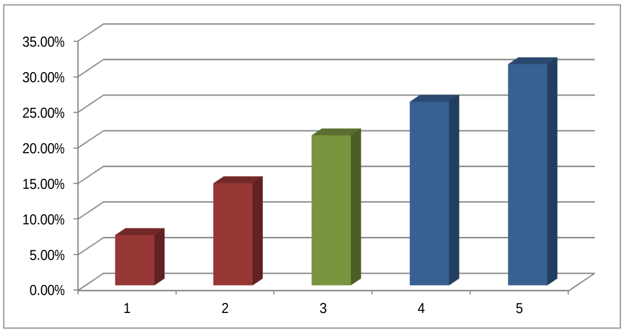

Figure 1: Global ranking averages

	Averages
Rank	3.60
BLEU	39.69
TER	44.88
Equal	20.1
Lev5	35.69

Table 3: Average metrics on post-edited files

3.2. Correlation Measures

To estimate the degree to which ranking was correlated to the actual post-editing effort, we computed the Pearson correlation coefficient between average rankings 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 ranked 5 and Lev5 was strong. A closer examination showed that three of the eleven language pairs, namely English to French, English to Spanish and English to Portuguese, showed weak 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 ranking was strongly correlated with the actual post-editing effort, modulo a minority of cases where a larger number of subtitlers would have been needed to balance individual ranking to post-editing effort disparities.

3.3. Error Collection

As mentioned above, we also collected recurrent MT errors which might be corrected 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 ranked 3 or higher only, since we assumed that lower ranked 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.

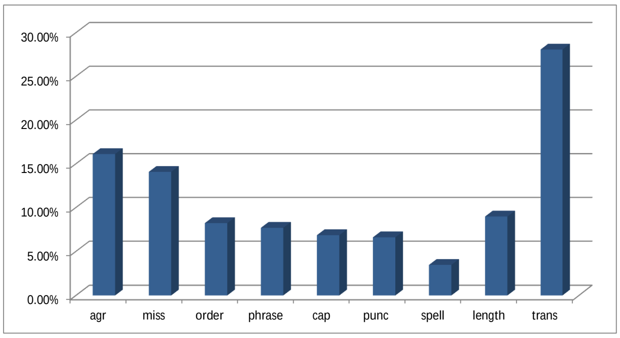

Figure 2: Global Errors

The results are given in Figure 2. 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 subtitling across domains, 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 [2] for an approach along those lines).

4. Productivity Measurement

The second major phase of the evaluation focuses on measuring productivity gain/loss by comparing the time needed to translate a source subtitle file from source vs. post-editing machine translated output. This round of evaluation is scheduled to start in October 2013, with results available in February 2014. In addition to the two aforementioned use cases, post-editing vs. direct translation, a third scenario is also considered: a mixed case with automatic quality es-

	Rank-TER	Rank-BLEU	Rank5-Lev5
r (all pairs)	-0.626	0.574	0.715
p-value (all pairs)	0.030	0.039	0.019
r (8 pairs)	-0.763	0.750	0.847
p-value (8 pairs)	0.020	0.022	0.012

Table 4: Ranking-Metric correlations

timation and filtering of MT output.[4] In this configuration, poor machine translated subtitles are removed from the MT output file, thus providing post-editors with empty MT subtitles to be translated from the source; good quality MT goes through the filters unmodified, to be post-edited. The main reason for adding this third use-case comes 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 mentions of the additional cognitive effort required to work with poor MT output. Introducing a mixed-case scenario with integrated quality estimation and filtering aims at evaluating a possible solution for this important issue. The experimental design involves the same translation pairs used for the quality evaluation round. Two subtitlers are involved for each translation pair and handle 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 will have been removed: in this scenario, subtitlers will thus perform both post-editing and translation from source.

Though post-editing is timed in this evaluation round to measure productivity differences, subtitlers are instructed to translate 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 to be subtitled directly will serve as benchmarks for productivity gain/loss measurement and the final results will include measures of inter-annotator agreement. We hypothesize that this type of evaluation will be a strong additional indicator of the usefulness of machine translation for professional subtitling.

5. 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 subtitling domain. 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. 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. The impact of MT output filtering before post-editing is being evaluated in the second evaluation round, where measuring productivity gain/loss will constitute a complementary assessment of the usefulness of MT technology for professional subtitling.

References

[1] A. Axelrod, X. He, and J. Gao. 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, 2011.

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

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

[4] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. 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, 2007.

[5] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proceedings of the*

[4] Quality estimation is performed with QuEst [8].

2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, pages 48–54. Association for Computational Linguistics, 2003.

[6] V. Petukhova, R. Agerri, M. Fishel, S. Penkale, A. del Pozo, M. S. Maucec, A. Way, P. Georgakopoulou, and M. Volk. 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, 2012.

[7] R. Sennrich. 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, 2012.

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

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

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