Towards a methodology for evaluating automatic subtitling

Alina Karakanta[1,2], Luisa Bentivogli[1], Mauro Cettolo[1], Matteo Negri[1], Marco Turchi[1]
[1]Fondazione Bruno Kessler
[2]University of Trento
{akarakanta,bentivo,cettolo,negri,turchi}@fbk.eu

Abstract

In response to the growing interest towards automatic subtitling, the 2021 EAMT-funded project "Towards a methodology for evaluating automatic subtitling" aimed at collecting subtitle post-editing data in a real use case scenario where professional subtitlers edit automatically generated subtitles. The post-editing setting includes, for the first time, automatic generation of timestamps and segmentation, and focuses on the effect of timing and segmentation edits on the post-editing process. The collected data will serve as the basis for investigating how subtitlers interact with automatic subtitling and for devising evaluation methods geared to the multimodal nature and formal requirements of subtitling.

1 Project overview

Automatic subtitling is the task of generating target language subtitles for a given video without any intermediate human transcription and timing of the source speech. The source speech in the video is automatically transcribed, translated and segmented into subtitles, which are synchronised with the speech – a process called automatic spotting (or auto-spotting). Automatic subtitling is becoming a task of increasing interest for the MT community, practitioners and the audiovisual industry. Despite the technological advancements, the evaluation of automatic subtitling still represents a significant research gap. Popular MT evaluation metrics consider only content-related parameters (translation quality), but not form-related parameters, such as format (length and segmentation) and timing (synchronisation with speech, reading speed), which are important features for high-quality subtitles (Carroll and Ivarsson, 1998). Moreover, the way subtitlers interact with automatically generated subtitles has not been yet explored, since the majority of works which conducted human evaluations of the post-editing effort in MT for subtitling have focused on edits in the textual content (Volk et al., 2010; Bywood et al., 2017; Matusov et al., 2019; Koponen et al., 2020).

This project seeks to investigate automatic subtitling, the factors contributing to post-editing effort and their relation to the quality of the output. This is achieved through the collection of rich, product- and process-based subtitling data in a real use case scenario where professional subtitlers edit automatically translated, spotted and segmented subtitles in a dedicated subtitling environment. The richness of the data collected during this one-year project is ideal for understanding the operations performed by subtitlers while they interact with automatic subtitling in their professional environment and for applying mixed methods approaches to:

- Investigate the correlation between amount of text editing, adjustments in auto-spotting and post-editing temporal/technical effort
- Explore the effect of auto-spotting edits on the total post-editing process
- Investigate the variability in subtitle segmentation decisions among subtitlers
- Propose tentative metrics for auto-spotting quality and subtitle segmentation

2 Data collection

Three professional subtitlers with experience in post-editing tasks (two subtitlers en→it, one

en→de) were asked to post-edit 9 single-speaker TED talks from the MuST-Cinema test set,[1] the only publicly available speech subtitling corpus (Karakanta et al., 2020), amounting to one hour of video (10,000 source words) in total. The post-editing task was performed in a novel PE subtitling tool, Matesub,[2] which features automatic speech recognition, machine translation, automatic generation of timestamps and automatic segmentation of the translations into subtitles.

For each subtitler, we collected the following data: 1) original automatically-generated subtitle files and the corresponding final human post-edited subtitle files in SubRip `.srt` format; 2) process logs from the Matesub tool, which records the original and final subtitle, original and final timestamps and total time spent on the subtitle; 3) keystrokes, using InputLog[3] (Leijten and Van Waes, 2013). Screen recordings were also collected to trace the translation and segmentation decisions of the subtitlers and identify possible outliers. At the end of the task, the subtitlers completed a questionnaire giving feedback on their user experience with automatic subtitling, particular problems faced, and their general impressions on automatic subtitling.

For en→it, we collected in total 1,199 subtitles from the first subtitler (it1) and 1,208 subtitles from the second subtitler (it2), while for en→de 1,198 subtitles. Based on the process logs we can define the status of each subtitle: *new* – a new subtitle is added by the subtitler; *deleted* – an automatically generated subtitle is discarded by the subtitler; or *edited* – any subtitle that is not new or deleted, regardless of whether it was confirmed exactly as generated by the system or changed. Table 1 shows the distribution of subtitles based on their status, with *edited* being the majority.

Subtitler	Edited	New	Deleted
it1	1,015 (84,7%)	59 (4.9%)	125 (10.4%)
it2	953 (78.9%)	68 (5.7%)	187 (15.4%)
de	1,051 (87.7%)	59 (4.9%)	88 (7.4%)

Table 1: Distribution of subtitles based on their status.

3 Final remarks

This project focuses on automatic subtitling and the challenges in its evaluation due to the multimodal nature of the source medium (video, audio) and the formal requirements of the target (format and timing of subtitles). The data collected constitute the basis for future multi-faceted analyses to explore correlations between translation quality, spotting quality, and post-editing effort, possibly leading to new metrics for automatic subtitling. The subtitling data collected will be publicly released to promote research in automatic subtitling.

[1] `https://ict.fbk.eu/must-cinema/`
[2] `https://matesub.com/`
[3] `https://www.inputlog.net/`

Acknowledgements

This project has been partially funded by the EAMT programme "2021 Sponsorship of Activities - Students' edition". We kindly thank the subtitlers Giulia Donati, Paolo Pilati and Anastassia Friedrich for their participation in the PE task.

References

Bywood, Lindsay, Panayota Georgakopoulou, and Thierry Etchegoyhen. 2017. Embracing the threat: machine translation as a solution for subtitling. *Perspectives*, 25(3):492–508.

Carroll, Mary and Jan Ivarsson. 1998. *Code of Good Subtitling Practice*. Simrishamn: TransEdit.

Karakanta, Alina, Matteo Negri, and Marco Turchi. 2020. MuST-Cinema: a Speech-to-Subtitles corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3727–3734, Marseille, France. ELRA.

Koponen, Maarit, Umut Sulubacak, Kaisa Vitikainen, and Jörg Tiedemann. 2020. MT for subtitling: User evaluation of post-editing productivity. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 115–124, Lisboa, Portugal, November. European Association for Machine Translation.

Leijten, Mariëlle and Luuk Van Waes. 2013. Keystroke logging in writing research: Using inputlog to analyze writing processes. *Written Communication*, 30:358–392.

Matusov, Evgeny, Patrick Wilken, and Yota Georgakopoulou. 2019. Customizing Neural Machine Translation for Subtitling. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 82–93, Florence, Italy, August. Association for Computational Linguistics.

Volk, Martin, Rico Sennrich, Christian Hardmeier, and Frida Tidström. 2010. Machine Translation of TV Subtitles for Large Scale Production. In Zhechev, Ventsislav, editor, *Proceedings of the Second Joint EM+/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC'10)*, pages 53–62, Denver.