Extending the MuST-C Corpus for a Comparative Evaluation of Speech Translation Technology

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

Abstract

This project aimed at extending the test sets of the MuST-C speech translation (ST) corpus with new reference translations. The new references were collected from professional post-editors working on the output of different ST systems for three language directions: English–German/Italian/Spanish. In this paper, we describe how the data were collected and how they are distributed. As an evidence of their usefulness, we also summarize the findings of the first comparative evaluation of *cascade* and *direct* ST approaches, which was carried out relying on the collected data. The project was partially funded by the European Association for Machine Translation (EAMT) through its 2020 Sponsorship of Activities programme.

1 Project overview

In this project we created and released additional reference translations for the test sets of the MuST-C corpus (Cattoni et al., 2021). The new references were collected for three language directions, i.e. En–De/Es/It, and consist of professional post-edits of the output of two state-of-the-art systems that represent the main current ST approaches, namely a *cascade* and a *direct* system.

Data. Our evaluation data are drawn from MuST-C, which is the largest freely available multilingual corpus for ST. It is based on English TED talks and currently covers 14 language directions, with English audio segments automatically aligned with their corresponding manual transcripts and translations. In MuST-C, a *Common* Test Set includes segments from talks that are common in all directions, thus making it possible to evaluate and compare systems across languages. For the three language directions addressed in the project, this common section includes the same 27 TED talks, for a total of around 2,500 largely overlapping segments.[1] For all language directions, we selected from MuST-C *Common* the same English audio portions from each talk, in order to obtain representative groups of contiguous segments that are comparable across languages. Furthermore, to ensure high data quality, we manually checked the selected samples and kept only those segments for which the *audio-transcript-translation* alignment was correct. Each of the 3 resulting post-editing test sets – henceforth *PE sets* – contains 550 segments, corresponding to ∼10,000 English source words. Then, we translated the PE sets with two ST systems. One represents the traditional *cascade* approach, in which the task is performed by means of a pipeline of separate automatic speech recognition (ASR) and machine translation (MT) components. The other adopts the more recent *direct* approach, which relies on a single encoder–decoder architecture that directly translates the source audio signal bypassing intermediate representations.

Post-editing. To prepare the data for the two post-editing (PE) tasks, we followed the main criteria adopted in the IWSLT PE-based evaluation campaigns (Cettolo et al., 2013). To guarantee high-quality data, we relied on two professional translators with experience in subtitling and

[1]Note, however, that due to automatic segmentation and alignment of the talks, segments can vary across languages.

post-editing, who were hired through a language service provider (Translated.com). Furthermore, in order to cope with translators' variability (i.e. one translator could systematically correct more than the other), the outputs of the two ST systems were randomly assigned to them, ensuring that each translator worked on all the 550 segments, equally post-editing both systems (cascade and direct). Another aspect inherent to our ST framework, which differentiates it from the traditional MT PE scenario, is the nature of the input (speech vs text). Since ST systems take spoken utterances as input, the traditional bilingual MT PE task, where translators are required to post-edit the system output according to the source text, is not feasible. For this reason, while the PE task was run using the MateCat tool (Federico et al., 2014), which displays the transcript together with the ST output to be edited, we also provided translators with the audio file of each segment, and asked them to post-edit according to it. The complete *ad hoc* guidelines given to the translators are available at: https://bit.ly/3gXEQin.

Final release. The project resulted in a significant extension of the MuST-C En–De/Es/It test sets. Specifically, for each of the 550 segments in the corresponding PE sets, two new reference translations were added. The data release includes, for each segment: *i)* the audio file, *ii)* the original reference transcript, *iii)* the original reference translation, *iv)* two ST outputs (from the cascade and direct systems), and *v)* the professional post-edits of the two ST outputs. The resource is distributed under a CC BY-NC-ND 4.0 license and is downloadable at: https://ict.fbk.eu/mustc-post-edits/.

2 Experiments with the released data

The collected high-quality post-edits can be exploited for different purposes, not limited to the standard one of computing more reliable multi-reference automatic evaluations. In a recent study (Bentivogli et al., 2021), we used them to analyse the relation between systems performance and specific characteristics of the input audio, and to investigate possible differences between the systems in terms of lexical, morphological and word ordering errors. We also explored whether the output of cascade and direct systems can be distinguished by humans or by automatic classifiers. Our investigation showed that the performance gap between the two technologies is now substantially closed. Subtle differences in their behavior exist: overall performance being equal, the cascade still seems to have an edge in terms of morphology, word ordering and lexical diversity, which is balanced by the advantages of direct models in audio understanding and capturing prosody. However, these differences do not seem sufficient to make the output of the two approaches easily distinguishable by humans.

3 Conclusion

In this project we released new high-quality reference translations which extend the En–De/Es/It test sets of MuST-C. These additional references consist of professional post-edits of the output of two state-of-the-art ST systems. The collected data are distributed as a special release of MuST-C, thus providing the community with a valuable resource to foster additional research in the ST field. Along this direction, we employed this resource to carry out a multi-faceted analysis that resulted in a timely contribution towards taking stock of the situation of ST technology advancements.

Acknowledgements

This project was partially funded by the European Association for Machine Translation (EAMT) through its 2020 Sponsorship of Activities programme.

References

Bentivogli, Luisa, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus direct speech translation: Do the differences still make a difference? In *Proceedings of ACL 2021 (Volume 1: Long Papers)*, pages 2873–2887, Online, August.

Cattoni, Roldano, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. MuST-C: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, 66:101155.

Cettolo, Mauro, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2013. Report on the 10th IWSLT Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany.

Federico, Marcello, Nicola Bertoldi, Mauro Cettolo, et al. 2014. The MateCat tool. In *Proceedings of COLING 2014 (System Demonstrations)*, pages 129–132, Dublin, Ireland, August.