EAMT-funded Project
"Extending the MuST-C Corpus for a Comparative Evaluation of Speech Translation Technology"

Final Report

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Matteo Negri, Marco Turchi
Fondazione Bruno Kessler
Trento, Italy

1 Introduction

This project is aimed at the creation and release of additional reference translations to extend the test sets of MuST-C, a publicly released multilingual Speech Translation (ST) corpus based on English TED Talks (Di Gangi et al., 2019c). The additional references are collected for three language directions, i.e. English-Italian/German/Spanish, 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 system and a direct system. The collected post-edits allow us to carry out and share a fine-grained comparative and cross-lingual analysis of the two ST solutions, aimed at shedding light on the strengths and limitations of the rapidly advancing direct technology with respect to the traditional cascaded methodology.

In this report we describe the methodology devised to collect the post-edits (Section 2), the features of the state-of-the-art ST systems we developed for English-German, English-Italian and English-Spanish (Section 3), our evaluation methodology based on post-editing (Section 4), and finally the results of the comparative evaluation carried out exploiting the collected post-edits (Sections 5 and 6).

2 Data Collection

Our evaluation data are drawn from the MuST-C corpus (Cattoni et al., 2020). MuST-C 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 transcriptions and translations. MuST-C *Common* Test Set includes segments from talks that are common to all directions, thus making it possible to evaluate and compare systems across languages. The *Common* Test Sets of the three language directions addressed in the project are composed of the same 27 TED talks, for a total of around 2,500 largely overlapping segments,[1] and include one reference translation manually created from scratch.

For all language pairs, we selected from MuST-C *Common* the same English audio portions from each talk, so as to obtain representative groups of contiguous segments that are comparable across languages. Furthermore, to ensure high data quality, we carried out a preliminary manual check and included only those segments *i)* containing only speech and *ii)* for which *audio-transcript-translation* alignment is correct. Each of the three resulting test sets – henceforth *PE-sets* – is composed of 550 segments, corresponding to about 10,000 English source words.

Our cascade and direct systems (see Section 3) were then run on the PE-sets be post-edited. To prepare the data for the two post-editing (PE) tasks, we followed the main criteria adopted in the IWSLT PE-based evaluations campaigns (Cettolo et al., 2013). To guarantee high quality data, we relied on two professional translators with experience in subtitling and 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.

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

Since ST systems take an audio signal as input, the traditional bilingual MT PE task, where translators are required to post-edit the system output directly according to the input source text, is not appropriate. In ST PE, the audio must be the primary source of information. This is even more important in our study since we specifically aim to understand if direct approaches leverage the audio input in a different way with respect to ASR+MT cascaded approaches.

For this reason, while the post-editing task was run using the MateCat tool, 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. We also prepared ad hoc guidelines where we highlighted all the specific characteristics of the task. The complete guidelines given to translators are available at: `https://bit.ly/3gXEQin`.

The resulting collected data for each of the three languages consist of two new reference translations for each of the 550 segments of the PE-set. The complete data release includes:

- audio files (from MuST-C)
- manual transcriptions (from MuST-C)
- manual translations (from MuST-C)
- Cascade and Direct systems' outputs
- Post-Edits of the Cascade and Direct systems' outputs

and can be found here: `https://bit.ly/3pQ6Zw1`

3 ST Systems

To maximize the cross-language comparability of our analyses, cascade and direct ST systems for en–de/es/it were built with the same core technology, based on Transformer. Their good quality is attested by the comparison with the winning system at the IWSLT-20 offline ST task,[2] which consists of an ensemble of two cascade models scoring 28.8 BLEU on the en-de portion of MuST-C *Common* test set (Bahar et al., 2020). On the same data, our cascade and direct models achieve similar scores, respectively 28.9 and 29.1. On en-es and en-it, identical architectures perform similarly or better (up to 32.9 on en-es). Although BLEU scores are not strictly comparable across languages, we can safely consider all our models as state-of-the-art.

In the following, we present the architectures of the two approaches.

3.1 Cascade approach

The Cascade system is composed of a pipeline of automatic speech recognition (ASR) and machine translation (MT) models.

The **ASR** component of our cascade systems is a slightly revised version of S-Transformer (Di Gangi et al., 2019b). It was trained on 1.25M (*audio*, *transcript*) pairs, containing 22M English words, in a multi-task setting with an additional CTC loss (Graves et al., 2006) on the encoder output.

The **MT** component is built on the Transformer implementation provided by the ModernMT framework.[3] We trained a *base* model for en-it, *big* for en-es and en-de. Training data were automatically selected from corpora publicly available in the OPUS repository.[4] After data selection, the amount of data used for training is: 68M pairs ($\sim$800M En words) for En-It, 19M pairs ($\sim$330M En words) for En-Es, and 17M pairs ($\sim$260M En words) for en-de. To mitigate error propagation and make the MT system more robust to ASR errors, similarly to (Di Gangi et al., 2019a), each MT model was fine tuned on the concatenation of human and automatic transcripts of MuST-C, both paired with manual translations.

[2] In the pre-segmented data condition (Ansari et al., 2020).

[3] `https://www.modernmt.com/`

[4] `http://opus.nlpl.eu`

3.2 Direct approach

Our direct model uses the same architecture of the ASR component of our cascade system but it has 11 Transformer encoder layers and 4 Transformer decoder layers. It is trained on 300K audio-translation pairs, augmented by generating 1.1M synthetic samples with the translation of ASR transcripts with an NMT model, and with SpecAugment (Park et al., 2019) and time stretch (Nguyen et al., 2020). The encoder is initialized with the encoder of the English ASR model and the model is optimized distilling knowledge from an NMT model (Liu et al., 2019) trained on the OPUS datasets (Tiedemann, 2016), before fine-tuning on label-smoothed cross entropy (Szegedy et al., 2016). Finally, we distinguish synthetic and real data providing the model with an apposite *token*.

4 Evaluation Methodology

Besides making new ST test sets available to the community, this project aims at sharing the results of a cross-lingual comparative evaluation of cascade and direct approaches.

The evaluation is based on post-editing, which is one of the most prominent methodologies used for the human evaluation of translation quality (Bentivogli et al., 2018b). PE-based evaluation was also chosen as the official evaluation in the IWSLT campaigns from 2013 to 2017.

All the analyses conducted in this study are based on the *Translation Edit Rate* (TER) metric (Snover et al., 2006).[5] Depending on which of the available references are used (2 post-edits and the official MuST-C reference translation), we rely on different variants of TER: *(i)* standard TER, which is computed against the MuSt-C reference, *(ii) Human-targeted TER* (HTER), which is computed between the automatic translation and its post-edited version; *(iii) Multiple reference TER* (mTER), which is computed against all the three available references. For comparison purposes, we also report sacreBLEU scores (Post, 2018).[6]

Besides presenting systems' overall performance, we also automatically detect and classify translation errors, exploiting the methodology and tools used in (Bentivogli et al., 2018a). The procedure is based on HTER computation under the assumption that, since the post-edit is generated by correcting the ST output, it directly points to translation errors. This type of analysis has proved able to provide useful insights on what linguistic phenomena are best modeled by systems while pointing out other aspects that remain to be improved. The tool – downloadable through the WIT^3 repository (Cettolo et al., 2012)[7] – is a modified version of the *tercom* script requiring the lemmatized versions of both systems' outputs and post-edits. To lemmatize the data we used the *TreeTagger*.[8]

5 Overall Systems' Performance

Table 1 presents overall systems' performance results, computed both on the PE-sets and on the MuST-C *Common* test sets. Our primary evaluation (grey background columns) is based on the collected post-edits. In addition to HTER, we also report mTER (two post-edits and the official reference from MuST-C), since – being computed on 3 references – better accounts for post-editors' variability, making the evaluation more reliable and informative. For the sake of completeness, we also report TER and Sacre-BLEU scores computed only on the official MuST-C references.

A bird's-eye view of the results shows that, in more than half of the cases, performance differences between cascade and direct systems are not statistically significant. When they are, the raw count of wins for the two approaches is the same (4), attesting their substantial parity.

Looking at our primary metrics (HTER and mTER – grey background columns), systems are on par on en-it and en-de, while for en-es the direct approach significantly outperforms the cascade one. This difference, however, does not emerge with the other metrics. Indeed, BLEU and TER scores computed against the official references are less coherent across metrics and test sets. For instance, in terms of

[5]We used the *tercom* implementation of TER available at `www.cs.umd.edu/~snover/tercom9`

[6]`https://github.com/mjpost/sacrebleu/` Version signature: BLEU+c.mixed+#.1+s.exp+tok.13a+v.1.4.3

[7]`wit3.fbk.eu/2016-02`

[8]`www.cis.uni-muenchen.de/~schmid/tools/TreeTagger`

		PE Set				M. Common	
		HTER 1 PE	mTER 2 PE + 1 Ref	BLEU 1 Ref	TER 1 Ref	BLEU 1 Ref	TER 1 Ref
de	C	28.65	24.41	28.96	53.23	28.86	53.93
	D	30.22	25.60	28.46	52.56	29.05	**52.77***
es	C	29.96	25.30	**34.05***	50.75	**32.93***	**53.21***
	D	**28.19***	**24.02***	32.17	51.08	31.98	54.00
it	C	25.69	23.29	**30.04***	54.01	28.56	56.29
	D	26.14	23.26	28.81	54.06	28.56	**55.35***

Table 1: Performance of (C)ascade and (D)irect systems on the PE-sets and MuST-C *Common* test sets. Statistically significant differences (*) are computed with Paired Bootstrap Resampling (Koehn, 2004).

BLEU score the cascade system significantly outperforms the direct one on the en-it PE-set, while TER shows the opposite on MuST-C *Common*.

Interestingly, the scores obtained using independent references can also disagree with those computed with post-edits. This is the case of en-es, where significant HTER and mTER reductions attest the superiority of the direct system, while most BLEU and TER scores are still in favor of the cascade.

On the one hand, primary evaluation scores suggest that the rapidly advancing direct technology has eventually reached the traditional cascaded approach. On the other, the highlighted incongruities confirm widespread concerns about the reliability of fully automatic metrics – based on independent references – to properly evaluate neural systems (Way, 2018). This calls for a deeper analysis, which we carry out by investigating linguistic errors made by the systems.

6 Linguistic Analysis of Translation Errors

In this section we presents the results obtained by the tool that exploits manual post-edits and HTER-based computations to detect and classify translation errors according to three linguistic categories: lexicon, morphology and word order. Table 2 shows their distribution for each approach. As expected from the HTER scores reported in Table 1, results vary across language pairs. On en-it, systems show pretty much the same number of errors, with a slight percentage gain (+1.1) in favor of the cascade. For the other two pairs, differences are more marked and opposite, with an overall error reduction for the direct system on en-es (-6.7) and in favor of the cascade on en-de (+6.7).

	en-de			en-es			en-it		
	C	D	Δ%	C	D	Δ%	C	D	Δ%
L	2481	2560	+3.2	2674	2497	-6.6	2264	2264	0.0
M	468	536	+14.5	535	494	-7.7	433	470	+8.6
R	398	476	+19.6	308	290	-5.8	230	226	-1.7
	3347	3572	+6.7	3517	3281	-6.7	2927	2960	+1.1

Table 2: Distribution of (L)exical, (M)orphological and (R)eordering errors. Absolute numbers are presented together with the percentage of reduction/increase of the (D)irect system with respect to the (C)ascade (Δ%).

Looking at the distribution of errors across categories, while for en-es the direct system is always better and the percentage reduction is homogeneously distributed, for en-de the better performance of the cascade system is concentrated in the morphology and word order categories. Since English and German are the most different languages in terms of morphology and word order, this result suggests that cascade systems still have an edge on the direct ones in their ability to handle morphology and word reordering. This is further supported by en-it: the only difference, in favor of the cascade, is indeed observed in the morphology category.

7 Conclusion

In this project we created and released additional reference translations which extend the test sets of MuST-C, the largest freely available multilingual corpus for ST, which is becoming a reference benchmark in the research community. The additional references are collected for three language directions, i.e.English-Italian/German/Spanish, 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 system and a direct system. All the collected data are freely distributed as a special release of MuST-C, thus providing the community with a valuable resource to be re-used for additional research in the ST field.

The high-quality post-edits have been exploited to analyse systems' behavior from different perspectives. We calculated overall systems' performance and investigated if the two approaches exhibit differences in terms of lexical, morphological and word ordering errors. The results suggest that i) overall the cascade and direct approaches now perform substantially on par, and ii) subtle differences can observed in their behavior, but are not sufficiently evident to draw clear conclusions. Thus, these results advocate for further, finer-grained, manual analyses, in an effort to answer important questions about ST technology that are arising within the community.

References

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. 2020. Findings of the IWSLT 2020 Evaluation Campaign. In *Proceedings of the International Conference on Spoken Language Translation (IWSLT)*, Virtual Event, July.

Parnia Bahar, Patrick Wilken, Tamer Alkhouli, Andreas Guta, Pavel Golik, Evgeny Matusov, and Christian Herold. 2020. Start-Before-End and End-to-End: Neural Speech Translation by AppTek and RWTH Aachen University. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Virtual Event.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2018a. Neural versus phrase-based MT quality: an in-depth analysis on English–German and English–French. *Computer Speech and Language"*, 49:52 – 70.

Luisa Bentivogli, Mauro Cettolo, Marcello Federico, and Christian Federmann. 2018b. Machine Translation Human Evaluation: an investigation of evaluation based on Post-Editing and its relation with Direct Assessment. In *Proceedings of the International Conference on Spoken Language Translation (IWSLT)*, Bruges, Belgium.

Roldano Cattoni, Mattia A. Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2020. MuST-C: A Multilingual Corpus for end-to-end Speech Translation. Computer Speech & Language Journal. Doi: https://doi.org/10.1016/j.csl.2020.101155.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web inventory of transcribed and translated talks. In *Proceedings of the Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May.

Mauro Cettolo, 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.

Mattia A. Di Gangi, Robert Enyedi, Alessandra Brusadin, and Marcello Federico. 2019a. Robust Neural Machine Translation for Clean and Noisy Speech Transcripts. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong.

Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. 2019b. Adapting Transformer to End-to-end Spoken Language Translation. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*, Graz, Austria, September.

Mattia Antonino Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019c. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, page 2012–2017, Minneapolis, Minnesota, June.

Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. 2006. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of the 23rd international conference on Machine learning (ICML)*, pages 369–376, Pittsburgh, Pennsylvania.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, Spain, July.

Yuchen Liu, Hao Xiong, Jiajun Zhang, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-End Speech Translation with Knowledge Distillation. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1128–1132, Graz, Austria, sep.

Thai-Son Nguyen, Sebastian Stueker, Jan Niehues, and Alex Waibel. 2020. Improving Sequence-to-sequence Speech Recognition Training with On-the-fly Data Augmentation. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Barcelona, Spain, may.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2613–2617, Graz, Austria, sep.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Conference on Machine Translation (WMT)*, pages 186–191, Brussels, Belgium, October.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the Conference of the Association for Machine Translation of the Americas (AMTA)*, pages 223–231, Cambridge, US-MA.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, Las Vegas, Nevada, United States, jun.

Jörg Tiedemann. 2016. Opus – parallel corpora for everyone. *Baltic Journal of Modern Computing*, page 384. Special Issue: Proceedings of the 19th Annual Conference of the European Association of Machine Translation (EAMT).

Andy Way. 2018. Quality Expectations of Machine Translation. In S. Castilho, J. Moorkens, F. Gaspari, and S. Doherty, editors, *Translation quality assessment: From Principles to Practice*, pages 159–178. Springer.