Neural Speech Translation: from Neural Machine Translation to Direct Speech translation

Mattia A. Di Gangi*
Fondazione Bruno Kessler (FBK)
ICT Doctoral School - University of Trento
via Sommarive, Povo, Trento, Italy
mattia.digangi@unitn.it

Speech-to-text translation, or simply speech translation (ST), is the task of translating automatically a spoken speech. The problem has classically been tackled by combining the technologies of automatic speech recognition (ASR) and machine translation (MT) with different degrees of coupling (Takezawa et al., 1998; Waibel et al., 1991). The most popular approach is to cascade ASR and MT systems, as it can make use of the state of the art in such mature fields (Black et al., 2002). The goal of this thesis was to develop the so-called approach of direct speech translation, which translates audio without intermediate transcription (Duong et al., 2016; Bérard et al., 2016; Weiss et al., 2017). Direct speech translation (DST) is based on the sequence-to-sequence learning technology that allowed the spectacular advances of the field of neural MT (NMT) but introducing its own challenges (Sutskever et al., 2014; Bahdanau et al., 2015).

We started with a study about the effects of NMT in cascaded ST, where we analyzed the translation errors of NMT and phrase-based MT (PBMT) for automatically transcribed input text. Our results showed that NMT achieves an overall higher quality also in this setting, but its ability to model a theoretically-unlimited context can introduce subtle errors. Indeed, we found that in PBMT the errors are localized in correspondence to the source error, whereas NMT can introduce errors far from the source-side error position.

Motivated by application needs, in a following work we studied how to use a single NMT system to translate effectively clean source text and automatic transcripts. We found that a simple training algorithm that fine-tunes the model on both kinds of inputs improves the translation quality of corrupted input without any degradation on clean input.

In a parallel research line, we were interested in making the training of RNN-based NMT more efficient, as it required at the time long training time also for relatively small datasets. For this, we proposed simple-recurrent NMT (SR-NMT), an encoder-decoder architecture that requires a fraction of parameters and computing power than LSTM-based NMT. It is built on top of simple recurrent units (Lei et al., 2017), which are faster to train but achieve a lower translation quality than LSTMs, particularly because they do not benefit from the addition of computation layers. On the other side, SR-NMT has been designed to be trained as a deep network and our results show how the performance improves significantly up to 8 layers in the encoder and in the decoder.

Our two research lines converge in our work on DST. We start with a participation in IWSLT 2018, which introduced a separate evaluation for direct models in order to encourage participants to explore this new and promising technology. From this participation we learn that training such kind of models is really difficult, findings confirmed by the very low results of all but the winning model. We hypothesize that such difficulty is due also to the low availability of training data for the task, which in fact requires source audio matched with its translation. It is much easier to find transcribed audio data and separate translated text.

In a first effort to overcome this data paucity, we propose MuST-C, a Multilingual Speech Translation Corpus (Di Gangi et al., 2019a). It is obtained from TED talks and provides the audio (in English) segmented into sentences matched with the cor-

*Now at AppTek GmbH

responding audio transcripts and translations to 8 languages. MuST-C provides audio data ranging from 385 to 504 hours, according to the target language, filtered for achieving a high quality of parallel data.

With MuST-C available, we focused on deep learning methods for DST and proposed S-Transformer, an adaptation of Transformer to the task (Di Gangi et al., 2019b). The problems that S-Transformer aims to solve are the high resource burden in terms of computing power and training time of LSTM-based DST, and the difficulty of self-attention to model audio-like sequences, characterized by a very high number of time steps and low information density per step. The first problem is tackled effectively by the use of Transformer, which trains faster and scales better than LSTMs, while for modeling we used 2D CNNs, 2D self-attention, and time-biased self-attention, which help with both convergence time and translation quality.

Finally, we applied S-Transformer in a one-to-many multilingual fashion to make better use of the MusT-C data, as well as comparing character-level against BPE-level segmentation of the target sentence. Our results showed that the BPE-segmentation is generally better and achieves larger improvement also in the multilingual scenario. Moreover, we participated in the DST evaluation at IWSLT 2019 and 2020, where MuST-C became the main in-task training corpus, and our submissions' results were competitive with the ones of teams from the industry. The results and products of this thesis contributed to the fast development of the technology of DST and lowered the barrier of entry into the field by making data[1] and code[2] publicly available.

Acknowledgments

The author would like to thank his Ph.D. supervisors: Marcello Federico, Marco Turchi, and Matteo Negri; his thesis examiners: Evgeny Matusov, Jan Nieheus, and Loïc Barrault, as well as all the HLT-MT group at FBK. The author was financially supported by a Ph.D. scholarship from FBK. This thesis was partly financially supported by an Amazon AWS ML Grant.

[1] https://ict.fbk.eu/must-c/

[2] https://github.com/mattiadg/FBK-Fairseq-ST

References

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of ICLR 2015*.

Bérard, Alexandre, Olivier Pietquin, Laurent Besacier, and Christophe Servan. 2016. Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*.

Black, Alan W, Ralf D Brown, Robert Frederking, Kevin Lenzo, John Moody, Alexander I Rudnicky, Rita Singh, and Eric Steinbrecher. 2002. Rapid Development of Speech-to-Speech Translation Systems. In *Seventh International Conference on Spoken Language Processing*.

Di Gangi, Mattia A., Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019a. Must-c: a Multilingual Speech Translation Corpus. In *Proceedings of NAACL 2019*, pages 2012–2017, Minneapolis, MN, USA.

Di Gangi, Mattia A., Matteo Negri, and Marco Turchi. 2019b. Adapting Transformer to End-to-End Spoken Language Translation. In *Proceedings of Interspeech 2019*, pages 1133–1137, Graz, Austria.

Duong, Long, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. An attentional Model for Speech Translation Without Transcription. In *Proceedings of NAACL 2016*, pages 949–959.

Lei, Tao, Yu Zhang, and Yoav Artzi. 2017. Training RNNs as Fast as CNNs. *arXiv preprint arXiv:1709.02755*.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of NIPS 2014*.

Takezawa, Toshiyuki, Tsuyoshi Morimoto, Yoshinori Sagisaka, Nick Campbell, Hitoshi Iida, Fumiaki Sugaya, Akio Yokoo, and Seiichi Yamamoto. 1998. A Japanese-to-English speech translation system: ATR-MATRIX. In *Fifth International Conference on Spoken Language Processing*.

Waibel, Alex, Ajay N Jain, Arthur E McNair, Hiroaki Saito, Alexander G Hauptmann, and Joe Tebelskis. 1991. JANUS: a Speech-to-Speech Translation System Using Connectionist and Symbolic Processing Strategies. In *Proceedings of the ICASSP 1991*, pages 793–796.

Weiss, Ron J., Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-Sequence Models Can Directly Translate Foreign Speech. In *Proceedings of Interspeech 2017*, Stockholm, Sweden, August.