Tutorial: End-to-End Speech Translation

Jan Niehues[1], **Elizabeth Salesky**[2], **Marco Turchi**[3] **and Matteo Negri**[3]
[1]Maastricht University
[2]Johns Hopkins University
[3]Fondazione Bruno Kessler
jan.niehues@maastrichtuniversity.nl esalesky@jhu.edu
{turchi,negri}@fbk.eu

Abstract

Speech translation is the translation of speech in one language typically to text in another, traditionally accomplished through a combination of automatic speech recognition and machine translation. Speech translation has attracted interest for many years, but the recent successful applications of deep learning to both individual tasks have enabled new opportunities through joint modeling, in what we today call 'end-to-end speech translation.'

In this tutorial we will introduce the techniques used in cutting-edge research on speech translation. Starting from the traditional cascaded approach, we will given an overview on data sources and model architectures to achieve state-of-the art performance with end-to-end speech translation for both high- and low-resource languages. In addition, we will discuss methods to evaluate analyze the proposed solutions, as well as the challenges faced when applying speech translation models for real-world applications.

1 Description

Machine translation (MT) and automatic speech recognition (ASR) have been mainstays of the speech and natural language processing communities for decades. Speech translation (ST), the combination of both tasks to translate from speech in one language typically to text in another, has existed for nearly as long as either of these (Waibel et al., 1991), attracting interest from both academia and industry. Until very recently, however, research in this area involved a cascade of separately trained speech recognition and machine translation models, with main questions pertaining to intermediate representations and processing steps to best connect these models.

The successful application of deep learning methods to speech and language processing has not only significantly improved the quality of models for both tasks (Sennrich et al., 2016; Hinton et al., 2012), but has also enabled new opportunities through joint modeling of speech and translation in what is today referred to as end-to-end speech translation (Bérard et al., 2016; Weiss et al., 2017). By integrating ideas from machine translation and speech recognition, this research topic is at the intersection of speech and language processing, traditionally two separate communities.

The paradigm switch to neural, end-to-end models has brought a significant increase in research interest and data resources for ST. The yearly evaluation campaign organized by IWSLT has seen large increases in participation in recent years (Ansari et al., 2020), and this year brought the creation of a joint special interest group (SIGSLT) spanning the ACL and ISCA communities. "Simpler" sequence-to-sequence architectures have lowered the barrier to entry; where previously researchers wishing to work in this area typically needed to either have significant knowledge of both ASR and MT or work in large collaborations, this is no longer the case. However, it remains the case that the best-performing models do draw on insights from both of these fields, and so we think that the time is ripe for a tutorial to better introduce the techniques to do cutting-edge research in ST.

This tutorial will summarize recent developments in end-to-end speech translation. We will start with discussion about the term 'end-to-end'[1] as well as a comparison to the traditional cascaded approach. In the subsequent sections, we will summarize ideas leveraged from automatic speech recognition (e.g. Chan et al. (2016)) and machine translation (e.g. Vaswani et al. (2017)) that are part of current state-of-the-art models, which are cur-

[1]For example, is use of pretrained models end-to-end? Is use of additional steps to create auxiliary target tasks like phoneme recognition? When do these distinctions matter?

Proceedings of EACL: Tutorials, pages 10–13
April 19 - 20, 2021.

rently demonstrated through evaluation campaigns like IWSLT. A particular focus point of the tutorial will be the current data landscape, as well as techniques to exploit different resources (Kano et al., 2020; Sperber et al., 2019) to enable speech translation not just for the few high-resource languages for which multi-parallel speech, transcripts, and translations exist.

After the survey of current state-of-the-art methods, we will present evaluation and analysis methods, and challenges when bringing these models from the lab to real-world environments. For example, one challenge of end-to-end models is their 'opaqueness'; with one joint system, it is more difficult to isolate causes of particular model behaviors and perhaps intervene, to avoid situations where key terms are translated in unexpected ways. Further, most training examples used fixed, pre-segmented input with parallel sentences, while in most practical applications the audio is not segmented. This brings additional challenges both in processing and also scoring. Finally, there are aspects of speech, such as speaker gender, accent, and prosody, which in cascaded systems the MT model did not have access to. We will touch on the impacts of some of these aspects, and provide greater detail about the specific example of gender bias mitigation (Bentivogli et al., 2020).

During the tutorial, we will highlight the present successes and challenges in end-to-end speech translation using examples from current state-of-the-art systems. Resources and teaching materials will be made available at `https://st-tutorial.github.io`.

2 Tutorial Type

This tutorial will cover cutting-edge research in the emerging field of end-to-end speech translation, and the aspects from speech and MT needed for this interdisciplinary research. The topic has not been previously covered in *CL tutorials.

3 Outline

- Introduction **(30 min)**
  - Task definition
  - Challenges/differences in translating speech rather than text
  - Traditional cascade approach to ST
- End-to-End **(45 min)**
  - Current state (high level overview)
  - Input representation
  - Architecture modifications
  - Output representation
- Data **(30 min)**
  - Available data for end-to-end ST
  - Different ways to leverage data sources:
    - Multi-task learning
    - Transfer learning and pretraining
    - Alternate data representations (e.g. phonemes)
- Evaluation/Analysis **(20 min)**
  - Automatic metrics
  - Utterance segmentation for automatic scoring
  - Mitigating errors due to speaker variation (gender, accent, etc.)
- Advanced topics **(30 min)**
  - Utterance segmentation
  - Making ST work for under-resourced languages
  - Multilingual ST
- From the lab to the real-world **(20 min)**
  - Automatic generation of subtitles
  - Simultaneous translation
  - Other Topics: system intervention, etc
- Conclusion **(5 min)**

4 Prerequisites

We would assume acquaintance with basic knowledge of machine learning and sequence-to-sequence models for machine translation, such as are covered in most introductory NLP courses. Any programming examples will be shown in Python.

5 Reading list

- Survey paper (Sperber and Paulik, 2020)
- The first papers on end-to-end ST (Bérard et al., 2016; Weiss et al., 2017)
- Data for end-to-end ST (Di Gangi et al., 2019b)
- Integrating additional data (Bansal et al., 2019; Jia et al., 2019; Sperber et al., 2019)

- Data representation (Salesky and Black, 2020)
- Adapting the Transformer for ST (Di Gangi et al., 2019a)
- Multilingual models (Inaguma et al., 2019)

6 Presenters

Jan Niehues, *Maastricht University*
Email: jan.niehues@maastrichtuniversity.nl
Website: https://dke.maastrichtuniversity.nl/jan.niehues/

Jan Niehues is an assistant professor at Maastricht University. He received his doctoral degree from Karlsruhe Institute of Technology in 2014 on the topic of "Domain Adaptation in Machine Translation." He has conducted research at Carnegie Mellon University and LIMSI/CNRS, Paris. His research has covered different aspects of Machine Translation and Spoken Language Translation. He has been involved in several international projects on spoken language translation e.g. the German-French Project Quaero, the H2020 EU project QT21 EU-Bridge and ELITR. Currently, he is one of the main organizers of the spoken language track in the IWSLT shared task.

Elizabeth Salesky, *Johns Hopkins University*
Email: esalesky@jhu.edu
Website: https://esalesky.github.io

Elizabeth Salesky is a PhD student at Johns Hopkins University. She has previously studied at Carnegie Mellon University, been a research assistant at Karlsruhe Institute of Technology, and worked at MIT Lincoln Laboratory, focused on speech and text translation. Her research focuses on speech translation for real-world and low-resource scenarios; e.g. phoneme features to reduce data dependence, disfluency removal in translating conversational speech, and learning robust representations. She has organized shared tasks on speech translation at IWSLT.

Marco Turchi, *Fondazione Bruno Kessler*
Email: turchi@fbk.eu
Website: http://ict.fbk.eu/people/detail/marco-turchi/

Marco Turchi is the head of the machine translation unit at Fondazione Bruno Kessler (FBK). He received his PhD degree in Computer Science from the U. of Siena, Italy in 2006. Before joining FBK in 2012, he worked at the European Commission, at the University of Bristol, at the Xerox Research Centre Europe, and at Yahoo Research Lab. His research activities focus on various aspects of sequence-to-sequence modelling applied to machine translation, speech translation and automatic post-editing. He is the co-organizer of the Conference of Machine Translation, the Spoken Language Translation Workshop and the automatic post-editing evaluation campaigns. He has been involved in several EU projects such as SMART, Matecat, ModernMT and QT21. He was the recipient of the Amazon AWS ML Research Awards on the topic of end-to-end spoken language translation in rich data conditions. He is the secretary of the ISCA SIGSLT interest group.

Matteo Negri *Fondazione Bruno Kessler*
Email: negri@fbk.eu
Website: https://ict.fbk.eu/people/detail/matteo-negri

Matteo Negri is a senior researcher in the Machine Translation unit at Fondazione Bruno Kessler. He received his degree in Philosophy of Language from the University of Turin, Italy in 2000. His research interests are in the field of computational linguistics, particularly machine translation, spoken language translation, textual entailment and question answering. He worked in several EU projects (QT21, CRACKER, MMT, MateCat, CoSyne, QALL-ME) and co-organised conferences, workshops and evaluation campaigns in NLP and MT-related areas (including the Conference on Machine Translation, the International Workshop on Spoken Language Translation and SemEval shared tasks). Together with Marco Turchi, he was the recipient of an Amazon AWS ML Research Award on "End-to-end Spoken Language Translation in Rich Data Conditions."

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 EVALU-

ATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online. Association for Computational Linguistics.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 58–68, Minneapolis, Minnesota. Association for Computational Linguistics.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 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*, Barcelona, Spain.

W. Chan, N. Jaitly, Q. Le, and O. Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964.

Mattia A Di Gangi, Matteo Negri, and Marco Turchi. 2019a. Adapting transformer to end-to-end spoken language translation. In *INTERSPEECH 2019*, pages 1133–1137. International Speech Communication Association (ISCA).

Mattia Antonino Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019b. 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.

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2019. Multilingual end-to-end speech translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 570–577. IEEE.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2019. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7180–7184. IEEE.

T. Kano, S. Sakti, and S. Nakamura. 2020. End-to-end speech translation with transcoding by multi-task learning for distant language pairs. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1342–1355.

Elizabeth Salesky and Alan W Black. 2020. Phone features improve speech translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2388–2397, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. 2019. Attention-Passing Models for Robust and Data-Efficient End-to-End Speech Translation. *Transactions of the Association for Computational Linguistics (TACL)*.

Matthias Sperber and Matthias Paulik. 2020. Speech Translation and the End-to-End Promise: Taking Stock of Where We Are. In *Association for Computational Linguistic (ACL)*, Seattle, USA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

A. Waibel, A. N. Jain, A. E. McNair, H. Saito, A. G. Hauptmann, and J. Tebelskis. 1991. Janus: A speech-to-speech translation system using connectionist and symbolic processing strategies. In *Proceedings of the Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference*, ICASSP '91, page 793–796, USA. IEEE Computer Society.

Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. *Proc. Interspeech 2017*, pages 2625–2629.