End-to-end Speech-to-Text Translation
An Annotated Bibliography

Satvik Sethia (ssethia2@illinois.edu)
University of Illinois at Urbana-Champaign

October 17, 2019

References

[1] A. Anastasopoulos and D. Chiang, "Tied multitask learning for neural speech translation," *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018.

This paper explores tied multi-task learning with sequence-to-sequence models, where the second task decoder receives input from both the encoder as well as the first task decoder. This information from a higher-level intermediate representation is intuitively useful, and is proved by the improved results. The authors also reason that transitivity and invertibility are intuitive principles, and add a regularization term to the model's attention mechanism so that it conforms to these principles.

[2] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, "Low-resource speech-to-text translation," *Interspeech 2018*, 2018.

This paper aims to investigate if the neural encoder-decoder model approach, that can directly translate speech to text in high-resource settings, will also work in low-resource settings (in terms of both data and compute power). The authors use word-level decoding instead of character-level to use fewer computational resources, which, along with the small amount of data, yields a lower BLEU score, but better precision and recall scores for word prediction.

[3] ——, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," *Proceedings of the 2019 Conference of the North*, 2019.

This paper investigates a method to improve direct speech-to-text translation (ST) for low resource languages. Since there is very sparse data available, instead of traditional ST (ASR and MT), the authors implement end-to-end ST by pre-training a model on a high resource language (English in this example) and fine-tuning the parameters for the low-resource language. BLEU Machine Translation scores are used as the metric to show improvement in results.

[4] A. Bérard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin, "End-to-end automatic speech translation of audiobooks," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

The authors of this paper investigate end-to-end speech translation in one single pass. The authors, however, make source language transcription available during training. An encoder-decoder model with attention is used on a corpus of audiobooks specifically augmented for this task. The authors propose their model as the baseline that performs almost as well as cascading two neural models for ASR and MT.

[5] D. Chen and B. Mak, "Multi-task learning of deep neural networks for low-resource speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015.

The authors of this paper propose a deep neural network based multi-task learning approach for low-resource ASR. They show that training grapheme models in parallel improves the performance of phone models for a single language. While training multiple languages, the authors explore the learning of a set of Universal Phones to improve the phone models of all the languages. Using this approach on three low-resource languages shows a significant improvement in word-recognition gains.

[6] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, and et al., "Recent advances in deep

learning for speech research at microsoft," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

This paper explores deep learning techniques for speech recognition and feature coding that are suitable replacements for earlier architecture that used Gaussians associated with HMM states. The authors apply these methods to various speech technology applications, and present improved results for many classic tasks. These techniques can be extended to other areas, end-to-end speech translation in particular.

[7] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, and et al., "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Transactions of the Association for Computational Linguistics*, vol. 5, 2017.

This paper proposes a standard neural MT system for multilingual translation. The encoder, decoder and attention module are shared across all languages, and the model accepts a token that indicates the target language. Using a shared word-piece vocabulary and no increase in parameters, this approach surpasses state-of-the-art on WMT'14 and WMT'15 benchmarks. The multilingual model allows for better translation, including transfer learning and zero-shot translation by implicit bridging between language pairs that were not seen during training.

[8] Y. Miao, F. Metze, and S. Rawat, "Deep maxout networks for low-resource speech recognition," *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013.

This paper applies deep maxout networks architecture on low-resource languages with limited text transcriptions. The success of this model lies in the reduced hidden activations, that shrink the size of the parameters, which make it very suitable for low-resource settings. This introduction of sparsity in the hidden activations also make these models suitable sparse feature extractors.

[9] S. Thomas, K. Audhkhasi, J. Cui, B. Kingsbury, and B. Ramabhadran, "Multilingual data selection for low resource speech recognition," *Interspeech 2016*, 2016.

This paper introduces a technique for data selection that discovers language groups from a set of training languages. The authors recognize that feature representations extracted from deep neural network based multilingual frontends show improved results for speech recognition in low-resource settings. They argue that their data selection model can be used to effectively reduce training data and time significantly, without affecting performance. This approach is extremely useful for low-resource settings with only a few hours of data available.

[10] S. Thomas, M. L. Seltzer, K. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

The authors of this paper explore a new technique for training deep neural networks for large vocabulary continuous speech recognition in low-resource settings. Transcribed multilingual data and semi-supervised training are used to tackle the lack of sufficient data for acoustic modeling. Both these approaches show a great improvement in results with a very small amount of data.

[11] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly translate foreign speech," *Interspeech 2017*, 2017.

The authors present an encoder-decoder deep recurrent neural network for direct speech-to-text translation, without transcription into text in source language. A single attention-based sequence-to-sequence model is used for end-to-end speech translation, which the authors argue is more powerful than independent ASR and MT models, as indicated by the improved BLEU score. It is mentioned that this method is viable for low-resource settings, and can be augmented if speech in a low-resource language has text transcriptions in a high-resource language.

[12] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

The authors of this paper propose a transfer-learning approach by training a model on a pair of high-resource languages, and then transfer those parameters to the low-resource pair. This paper tackles the limitations of traditional neural MT networks in low-resource settings, which perform poorly in comparison to other MT systems. It is also concluded that a model based on a pair of high-resource language outperforms that based on a single language, and even the choice of languages affects performance.