State of the Art Summary of Low-Resource End-to-end Speech Translation

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

October 2019

Speech translation is the process of translating spoken words of a source language to spoken words of a target language. Traditionally, speech translation has been broken down into three distinct steps – Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-To-Speech (TTS) conversion. These three steps implemented in that order form the complete speech translation pipeline. In recent years, however, a new approach to speech translation has been adopted, that combines the first two steps for an end-to-end speech translation. This end-to-end approach converts speech from the source language to text in the target language directly, skipping the intermediate source language text state.

Much success has been achieved by using neural network models for speech translation. An RNN based model is necessary for this task, since the input of speech translation, spoken language or speech, is sequential, and thus requires some sort of history to have been kept for effective translation. Particularly, deep learning techniques have been used in speech recognition and research for quite a few years, especially in low-resource settings. There are several advantages of using these methods over more conventional methods. Deep neural networks extract higher level features, that are closer to what humans perceive, as opposed to low-level feature extraction in neural networks with a fewer number of layers. This is more suited to speech research tasks, where data needs an abstract representation. An end-to-end approach particularly needs higher-level feature extraction, specifically since source language text is not an intermediate state in the process. The RNN based sequence-to-sequence models use encoders that represent the input in a new way extracting high-level features, which are then decoded by a decoder in the output format. Further, an attention mechanism can be added to these sequence-to-sequence models. Attention is exactly what it sounds like, it enables the network to focus on more important aspects of the data, instead of everything encoded by the network encoder, just like how humans focus on some details over others. For all these reasons, it makes sense to use deep learning models for an end-to-end approach, which is more abstract than the two steps (ASR and MT) that were previously used for this task. Pre-

vious research has shown that deep learning models are indeed an improvement over and are increasingly replacing traditional models in speech research [6, 8].

Then there is the added complication of a low-resource setting, which is the environment for this summary and research project. Attaining good results in a low-resource setting requires many more features and improvements to baseline models of high-resource settings. A lot of techniques have been experimented with, to try and compensate for the sparsity of data, which is extremely difficult to work around. Even so, developing models that can directly translate just a few hours of speech from a low-resource language to speech or text in a high-resource language is a critical task for documenting and preserving those languages. Deep learning neural networks are the preferred choice in this specialized sub-domain too, and have been extensively used to obtain promising results [8, 10, 5, 9]. This document summarizes the state-of-the-art models for end-to-end speech translation in low-resource settings and the techniques and methods proposed to implement the model to improve results.

The biggest drawback of the traditional ASR followed by MT approach lies in combining two different models for two different tasks. These models might yield good results for their specific tasks, but in composing them together, errors of the ASR system intensify the errors of the MT system which has not been trained to handle the errors of the ASR system in its input. This is easily addressed in the end-to-end model, by tweaking the parameters for the ultimate goal. From Weiss et al. [11], it can be concluded that end-to-end speech translation is a more effective method than the traditional ASR and MT system. A recurrent encoder-decoder deep neural network model with attention is used that specifically does not require source language text supervision nor explicitly transcribes the speech to text in the source language. This end-to-end method shows significant improvements in results, and is also particularly suited for low-resource settings, where transcriptions are often available only in the target language, which can be used to train the model, and also in the case of spoken languages that have no written form. Thus, this method not only yields better results over the traditional approach, but is also more suited for low-resource settings. Hence, variations in implementations of this approach form the state-of-the-art in this domain.

A particular example of the case where source language text transcription is available, only during training, is explored in Bérard et al. [4], where the source language speech is translated to target language text in a single pass. Results of four different models, with attention, for speech translation – cascaded ASR and MT, end-to-end with no source transcriptions, pre-trained with ASR and MT models as encoder-decoder, and multitask – are presented. For the high-resource corpus used in this paper, the cascaded model performs the best, but the end-to-end model is not very far behind. It must also be noted that ASR and MT models for the languages used – English and French – are extremely sophisticated, due to a lot of experiments conducted on two extremely high-resource languages. Sufficient cascading techniques have also been developed for this pair, explaining the performance of that particular model.

One of the few deterring factors behind using a neural network is the amount of computation it requires. Neural networks require significantly more compute power, and adding deep learning techniques with multiple layers and features like attention further increase computations. A novel study to investigate whether end-to-end models can be used in low-resource settings, in terms of both data and compute power, to tackle the issue raised above, is presented in Bansal et al. [2]. Citing the common problem of cascading ASR and MT systems for low-resource settings, the authors borrow from the experiments of [11] to train an end-to-end model, but make a few adjustments to the design of the models, so that it trains significantly faster on a single GPU. The biggest change is using a word-level decoding instead of the character-level models used in most previous experiments. This speeds training up by many orders of magnitude. Using fewer computational resources with a substantial amount of data, this approach produces comparable precision and recall. The trade-off for an increase in speed-up due to word-level decoding is errors in translating uncommon words. Subsequently, the amount of training data is incrementally reduced, and while the model performs worse, precision and recall values are good enough to suggest that this approach may find application in critical situations with acutely under-resourced languages and limited computing power.

Transfer learning is a promising approach to improve neural MT models for high-resource settings under low-resource settings. Zoph et al. [12] presents a transfer learning method by first training on a high-resource language pair, the parent model, and then transferring some of the learned parameters to initialize as well as constrain training for the low-resource pair, the child model. Once the child model is initialized with the transferred parameters, the rest of the parameters are fine-tuned. Training time on the child model is reduced as well, since the model is already initialized. This method produces results better than state-of-the-art neural MT results for low-resource languages, according to the BLEU metric. Experiments also show that the choice of high-resource language pair affects performance. This approach holds a lot of promise and could improve end-to-end speech translation.

A comparative approach to the above method is using multilingual data for training. Two works are discussed, Thomas et al. [9] that introduces a data selection technique to find language clusters that are (phonemically) similar, and Johnson et al. [7] that applies neural MT on multilingual training data. In [9], the motivation behind searching for languages that use a common phoneme set is to improve speech recognition systems. A data-driven technique analysing confusion matrices outputs similarity of similar languages. Generally, using languages belonging to the same family should give good enough results, although this method is directly useful for the task at hand, because it can find acutely under-resourced languages from different families that share features important for speech recognition. In [7], a state-of-the-art neural MT model is selected, and slightly tweaked to accept a token that indicates the target language. With this addition, a variety of different translations are possible – one or many source languages to one or many target languages. For the task at hand, consider the

case of many source languages to one target language, with variable amounts of data in the source languages. Training time for multiple languages is shown to be significantly faster, and results show an improvement over baselines, especially for the case where parallel data for the source languages and target language is available.

A strong argument highlighting the advantage of deep learning and the reason behind choosing them is presented in Anastasopoulos et al. [1], the fact that higher-level intermediate representations contain useful information. The model used by the authors aims to perform two tasks, low-resource speech transcription and translation. It uses a shared encoder to output sequences, but separate decoders. The multitasking approach is characterized by using the output of the encoders as well as the first decoder as input for the second decoder. This model is shown to improve results against all baselines, and also gives transcription in addition to translation.

The most recent work in low-resource end-to-end speech translation, Bansal et al. [3], builds on the previous approach in [2] of using fewer computational resources and less data. Inspired by [12], a model is pre-trained on a high-resource ASR task, and then its parameters are fine-tuned for speech translation. High-resource ASR tasks are, as mentioned previously, excellent. Thus, pre-training on those models, where the high-resource language is similar to the low-resource source language, shows a very high improvement against baselines. Experiments with 20 hours of data and less consistently showed improvements of 5-10 points according to the BLEU metric. In a truly low-resource setting with about 4 hours of available data, the BLEU score was doubled with this method.

This document has reviewed many techniques for a lot of different translation tasks, with a focus on low-resource settings. State-of-the-art end-to-end speech translation efforts have also been discussed. Combining these approaches have been shown to produce improvements, with different methods working for different languages under various scenarios. A subset of these methods can be chosen based on the task at hand, depending on the amount of data and computational resources available, deciding on including certain techniques for agglutinative versus fusional languages, tonal languages (an important feature to consider for speech recognition and translation), etc.

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.