Enhancing Multilingual Graphemic RNN Based ASR Systems Using Phone Information

Markus Müller, Sebastian Stüker, Alex Waibel

Lately, systems based on recurrent neural networks (RNNs), trained using the connectionist temporal classification (CTC) loss function, have gained substantial research interest. Due to the recurrent nature of the network used, such systems are able to capture temporal dependencies implicitly, without the need to explicitly model context by, e.g., context-dependent phone states. In this work, we focus on multilingual systems, based on graphemes as acoustic modelling units. We aim at training a single, multilingual model that is able to recognize speech from multiple languages simultaneously.

In the past, we presented an approach to train RNN/CTC systems jointly on multiple languages using a global set of either graphemes or phonemes as acoustic units[1,2]. While using graphemes as acoustic units can be a challenging task for languages with complex pronunciation rules (e.g., English), training a system on a combination of languages is even more difficult: The network has to infer the pronunciation rules for multiple instead of a single language. In order to adapt networks to multiple languages, we proposed Language Feature Vectors (LFVs). Extracted via a neural network, they encode language specific peculiarities as low dimensional vector. We evaluated different methods for integration into the neural network architecture. Similar to feed-forward networks, appending the LFVs to the input features increases the performance, but we determined in a series of experiments, that modulating the layers of RNNs does result in better performance[3]. Akin to dropout, the modulation emphasizes or attenuates the outputs of neurons, based on encoded language properties. This forces the network to learn features based on language features which improves the performance in a multilingual setting.

In this work, we extended our approach in two ways: a) integrating phonetic information into grapheme based systems and b) by optimizing the network architecture. Based on Baidu's Deepspeech2, the RNN features 2 convolutional layers, 4 bidirectional LSTM layers and a feed-forward output layer. We trained the network jointly on 4 languages (English, French, German, Turkish), using graphemes as targets with an additional symbol to indicate word boundaries. We used greedy decoding with a grapheme based language model, trained on the transcripts of the training utterances only. Being character based, the resulting system does not have a fixed vocabulary.

In order to add phonetic information, we choose to train the network in a two step approach: First, we pre-trained the network using phones as targets. Next, we replaced the output layer trained on phones and fine-tuned the network again, using graphemes as targets. Pre-training with phonetic targets forced the network to extract features to discriminate phones. The fine-tuning process using graphemes then allowed for the network to learn a mapping from the phonetic features to graphemes. This improved the system performance from 29.9% WER (12.4% Token Error Rate, TER) to 28.3% WER (11.7% TER).

In addition, we optimized the network architecture. Since we are using bidirectional LSTM layers, the output of each layer has twice the size of the number of LSTM cells. As input to the next layer, these outputs could be passed along unchanged, or the outputs for each direction could be combined pairwise. In a series of experiments, we compared the performance of addition, multiplication or taking the maximum of the outputs (inspired by maxout networks) to simply passing the outputs along unchanged. Multiplying or taking the maximum improves the performance, with using the maximum resulting in the best performance: Starting with a baseline of 28.3% WER (11.7% TER), multiplication results in 26.9% WER (11.1% TER) and "maxout" in 26.7% WER (11.0% TER).

Future work includes the evaluation of additional network configurations and training techniques to further improve the performance by maximizing the language adaptation using LFVs.

---

[1] Markus Müller, Sebastian Stüker and Alex Waibel. "Language Adaptive Multilingual CTC Speech Recognition." In *International Conference on Speech and Computer*, pp. 473-482. Springer, Cham, 2017.

[2] Markus Müller, Sebastian Stüker and Alex Waibel. "Phonemic and Graphemic Multilingual CTC Based Speech Recognition." *arXiv preprint arXiv:1711.04564* (2017).

[3] Markus Müller, Sebastian Stüker and Alex Waibel. "Multilingual Adaptation of RNN Based ASR Systems." *arXiv preprint arXiv:1711.04569* (2017).