Using Unpaired Text in Encoder-Decoder Models for Speech Recognition

Shubham Toshniwal and Karen Livescu

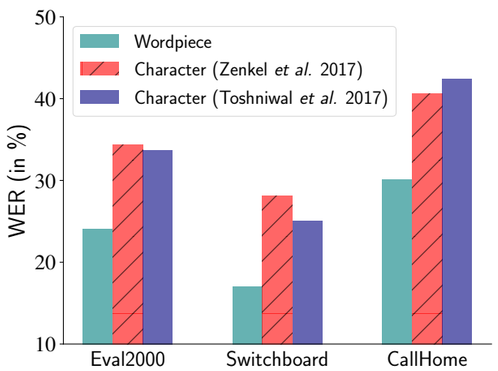
Toyota Technological Institute at Chicago

Abstract

- Attention-based encoder-decoder models present an elegant solution to the automatic speech recognition (ASR) problem.
- These models require only a parallel corpus of speech and text.
- However, unlike earlier models that combine separate acoustic and language models, it is not clear how to use additional (unpaired) text.
- In this study, we propose: (a) integrating a pretrained language model as a lower layer of the decoder, and (b) jointly training the whole model for the speech recognition task and the pretrained part for the language modeling task (Ramachandran *et al.* 2017).
- We find that:
  - Joint training improves performance over pretraining alone.
  - Sharing softmax layer parameters is crucial for good results.
  - The performance gain is comparable to that of shallow fusion
  - Combination of the proposed approach with shallow fusion further improves the performance.
- All of our models use word-pieces as output units, which improves performance over using characters as output units (Chiu *et al.* 2018).

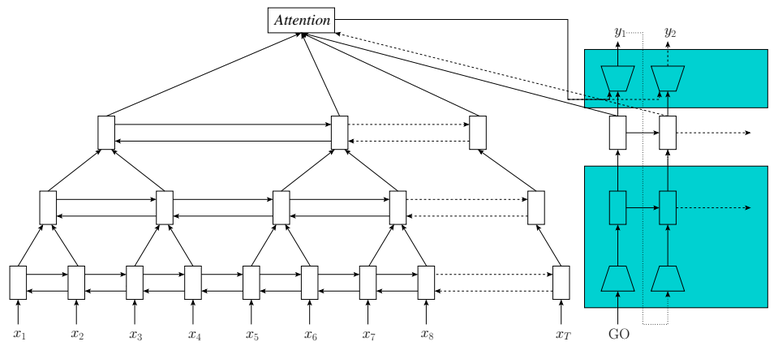
Word-Pieces as Output Units

- We use a byte pair encoding (BPE) compression algorithm to build a vocabulary of word-pieces (Sennrich *et al.* 2015).
- BPE starts with the symbol set of characters and iteratively adds the most frequent symbol n-gram to the set.
- Word-pieces provide a middle ground between:
  (a) Characters: Allow for flexible outputs with the disadvantage of computational inefficiency.
  (b) Words: Reduce the decoding time but have limited coverage.
- We use a vocabulary of size 1000, which includes all of the characters to avoid UNKs.
- Our baseline word-piece model improves over prior character-based models, as shown below:



Eval2000 results for character vs word-piece encoder-decoder models.

Model



Attention-based encoder-decoder model with a pyramidal bidirectional LSTM (we use 4 layers; 3 layers shown for simplicity). In our proposed model, the colored part of the decoder is both initialized with a pretrained LM and jointly trained for LM loss.

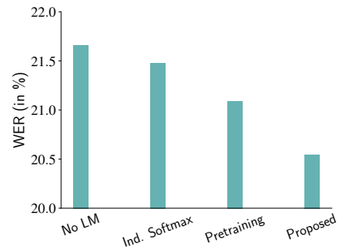
Model Variants

- No LM: Neither initialization nor joint training with LM.
- Pretraining: Only initialization with pretrained LM.
- Independent Softmax: Same as proposed model except separate softmax layer.

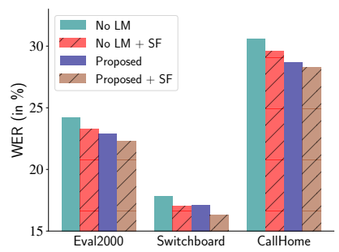
Experimental Setup

- We train all models using the 300 hr Switchboard conversational speech corpus.
- For LM training, we combine Switchboard and Fisher (also a conversational speech corpus).
- To avoid domain mismatch, we removed additional noise/hesitation markers from Fisher transcripts and filtered out utterances not covered by the word-piece set trained on Switchboard ($\sim$ 400K/2.2M Fisher utterances removed).
- We decode with beam size 16 with/without log-linear interpolation with the pretrained LM.
- We refer to beam search decoding with separate LM as *Shallow Fusion*.

Results



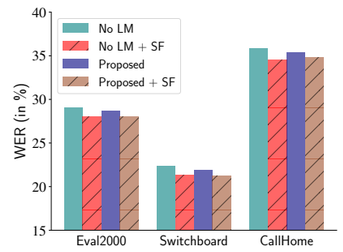
Dev set result for model variants



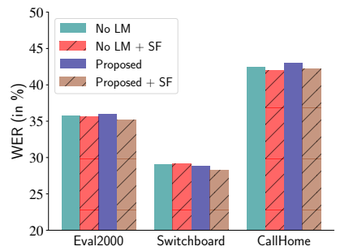
Eval2000 results with and without shallow fusion (SF)

- The dev set results show that:
  (a) Sharing the softmax layer is very important.
  (b) LM-based pretraining alone improves performance, but further joint training with both ASR and LM losses is better.
- The test set results show that:
  (a) The proposed LM integration outperforms shallow fusion.
  (b) The approach is complementary to shallow fusion: the combination achieves a relative $\sim$8% WER reduction over "No LM".

Reduced Training Data



Eval2000 results for $\sim$150 hrs data



Eval2000 results for $\sim$75 hrs data

Ongoing/Future Work

- Introduce additional terms such as word insertion penalty, in beam search.
- Further tuning in low-data regimes.
- Compare against other LM integration approaches.
- Error analysis.

References

- P. Ramachandran *et al.* Unsupervised pretraining for sequence to sequence learning. EMNLP 2017.
- C. Chiu *et al.* State-of-the art speech recognition with sequence-to-sequence models. ICASSP 2018.
- R. Sennrich *et al.* 2016. Neural machine translation of rare words with subword units. ACL 2016.
- T. Zenkel *et al.* Comparison of decoding strategies for CTC acoustic models. Interspeech 2017.
- S. Toshniwal *et al.* Multitask Learning with Low-Level Auxiliary Tasks for Encoder-Decoder Based Speech Recognition. Interspeech 2017.