ADAM MICKIEWICZ UNIVERSITY IN POZNAŃ

Recent Advances in Error Correction of ASR

2019-04-09

Tomasz Ziętkiewicz

Outline

A SPELLING CORRECTION MODEL FOR END-TO-END SPEECH RECOGNITION

Jinxi Guo[1*] , *Tara N. Sainath*[2], *Ron J. Weiss*[2]

[1]University of California, Los Angeles, USA
[2]Google Inc., USA

lennyguo@g.ucla.edu, {tsainath, ronw}@google.com

[GSW19]

IEEE International Conference on Acoustics, Speech and Signal Processing May 12-17, 2019

Motivation

- Popularity of end-to-end ASR models
- Acousting, pronunciation and language model combined in one neural network
- Problem: needs annotated audio data
- LM trained on small dataset compared with "traditionalapproach
- Worse performance on rare words

Motivation

- Popularity of end-to-end ASR models
- Acousting, pronunciation and language model combined in one neural network
- Problem: needs annotated audio data
- LM trained on small dataset compared with "traditionalapproach
- Worse performance on rare words

Motivation

- Popularity of end-to-end ASR models
- Acousting, pronunciation and language model combined in one neural network
- Problem: needs annotated audio data
- LM trained on small dataset compared with "traditionalapproach
- Worse performance on rare words

Motivation

- Popularity of end-to-end ASR models
- Acousting, pronunciation and language model combined in one neural network
- Problem: needs annotated audio data
- LM trained on small dataset compared with "traditionalapproach
- Worse performance on rare words

Motivation

- Popularity of end-to-end ASR models
- Acousting, pronunciation and language model combined in one neural network
- Problem: needs annotated audio data
- LM trained on small dataset compared with "traditionalapproach
- Worse performance on rare words

Possible solutions

- Incorporating external LM trained on text-only data
  - Rescoring n-best decoded hypothesis from end-to-end ASR:

$$y* = \underset{y}{\operatorname{argmax}}\, logP(y|x) + \lambda logP_{LM}(y)$$

  - Incorporate RNN-LM into first-pass beam search by shallow, cold or deep fusion
- Use TTS to generate audio-text pairs training data from text-only data
- Rare words and proper nouns are still problematic with this approach
- Why? Hypothesis: LM trained with other objective then correcting e2e model's errors

Possible solutions

- Incorporating external LM trained on text-only data
  - Rescoring n-best decoded hypothesis from end-to-end ASR:

$$y* = \underset{y}{\operatorname{argmax}}\, logP(y|x) + \lambda logP_{LM}(y)$$

  - Incorporate RNN-LM into first-pass beam search by shallow, cold or deep fusion
- Use TTS to generate audio-text pairs training data from text-only data
- Rare words and proper nouns are still problematic with this approach
- Why? Hypothesis: LM trained with other objective then correcting e2e model's errors

Possible solutions

- Incorporating external LM trained on text-only data
  - Rescoring n-best decoded hypothesis from end-to-end ASR:

$$y* = \underset{y}{\operatorname{argmax}}\, logP(y|x) + \lambda logP_{LM}(y)$$

  - Incorporate RNN-LM into first-pass beam search by shallow, cold or deep fusion
- Use TTS to generate audio-text pairs training data from text-only data
- Rare words and proper nouns are still problematic with this approach
- Why? Hypothesis: LM trained with other objective then correcting e2e model's errors

Possible solutions

- Incorporating external LM trained on text-only data
  - Rescoring n-best decoded hypothesis from end-to-end ASR:

$$y* = \underset{y}{\operatorname{argmax}}\, logP(y|x) + \lambda logP_{LM}(y)$$

  - Incorporate RNN-LM into first-pass beam search by shallow, cold or deep fusion
- Use TTS to generate audio-text pairs training data from text-only data
- Rare words and proper nouns are still problematic with this approach
- Why? Hypothesis: LM trained with other objective then correcting e2e model's errors

Possible solutions

- Incorporating external LM trained on text-only data
  - Rescoring n-best decoded hypothesis from end-to-end ASR:

$$y* = \underset{y}{\operatorname{argmax}}\, logP(y|x) + \lambda logP_{LM}(y)$$

  - Incorporate RNN-LM into first-pass beam search by shallow, cold or deep fusion
- Use TTS to generate audio-text pairs training data from text-only data
- Rare words and proper nouns are still problematic with this approach
- Why? Hypothesis: LM trained with other objective then correcting e2e model's errors

Possible solutions

- Incorporating external LM trained on text-only data
  - Rescoring n-best decoded hypothesis from end-to-end ASR:

$$y* = \underset{y}{\operatorname{argmax}}\, logP(y|x) + \lambda logP_{LM}(y)$$

  - Incorporate RNN-LM into first-pass beam search by shallow, cold or deep fusion
- Use TTS to generate audio-text pairs training data from text-only data
- Rare words and proper nouns are still problematic with this approach
- Why? Hypothesis: LM trained with other objective then correcting e2e model's errors

Solution

Proposed solution: spelling corrector model on text-to-text (hypothesis-to-refrerence) pairs.

- Identify likely errors in ASR output
- Propose alternatives
- Combine with LM-rescoring

Solution

Proposed solution: spelling corrector model on text-to-text (hypothesis-to-refrerence) pairs.

- Identify likely errors in ASR output
- Propose alternatives
- Combine with LM-rescoring

Solution

Proposed solution: spelling corrector model on text-to-text (hypothesis-to-refrerence) pairs.

- Identify likely errors in ASR output
- Propose alternatives
- Combine with LM-rescoring

ASR train/eval Dataset

- LibriSpeech [PCPK15]
- Large-scale (1000 hours) corpus of read English speech
- audiobooks from the LibriVox project
- carefully segmented and aligned
- http://www.openslr.org/12/
- License: CC BY 4.0

ASR train/eval Dataset

- LibriSpeech [PCPK15]
- Large-scale (1000 hours) corpus of read English speech
- audiobooks from the LibriVox project
- carefully segmented and aligned
- http://www.openslr.org/12/
- License: CC BY 4.0

ASR train/eval Dataset

- LibriSpeech [PCPK15]
- Large-scale (1000 hours) corpus of read English speech
- audiobooks from the LibriVox project
- carefully segmented and aligned
- http://www.openslr.org/12/
- License: CC BY 4.0

ASR train/eval Dataset

- LibriSpeech [PCPK15]
- Large-scale (1000 hours) corpus of read English speech
- audiobooks from the LibriVox project
- carefully segmented and aligned
- http://www.openslr.org/12/
- License: CC BY 4.0

ASR train/eval Dataset

- LibriSpeech [PCPK15]
- Large-scale (1000 hours) corpus of read English speech
- audiobooks from the LibriVox project
- carefully segmented and aligned
- http://www.openslr.org/12/
- License: CC BY 4.0

ASR train/eval Dataset

- LibriSpeech [PCPK15]
- Large-scale (1000 hours) corpus of read English speech
- audiobooks from the LibriVox project
- carefully segmented and aligned
- http://www.openslr.org/12/
- License: CC BY 4.0

Text-only dataset

- Spelling correction needs parallel corpus: ASR hypothesis + ground truth text
- 800M word LibriSpeech language modeling corpus
- Selected 40M sentences not overlapping with test set
- Generated audio using TTS (WaveNet [vdOLB+17])
- Added noise and reverbation to get additional 40M utterances
- Decode using pretrained ASR model
- From each TTS utterance ASR produces 8 hypotheses
- All of them used to form hypothesis-reference pairs: 640M hypothesis-reference pairs
- also added to ASR trainset

Text-only dataset

- Spelling correction needs parallel corpus: ASR hypothesis + ground truth text
- 800M word LibriSpeech language modeling corpus
- Selected 40M sentences not overlapping with test set
- Generated audio using TTS (WaveNet [vdOLB+17])
- Added noise and reverbation to get additional 40M utterances
- Decode using pretrained ASR model
- From each TTS utterance ASR produces 8 hypotheses
- All of them used to form hypothesis-reference pairs: 640M hypothesis-reference pairs
- also added to ASR trainset

Text-only dataset

- Spelling correction needs parallel corpus: ASR hypothesis + ground truth text
- 800M word LibriSpeech language modeling corpus
- Selected 40M sentences not overlapping with test set
- Generated audio using TTS (WaveNet [vdOLB+17])
- Added noise and reverbation to get additional 40M utterances
- Decode using pretrained ASR model
- From each TTS utterance ASR produces 8 hypotheses
- All of them used to form hypothesis-reference pairs: 640M hypothesis-reference pairs
- also added to ASR trainset

Text-only dataset

- Spelling correction needs parallel corpus: ASR hypothesis + ground truth text
- 800M word LibriSpeech language modeling corpus
- Selected 40M sentences not overlapping with test set
- Generated audio using TTS (WaveNet [vdOLB+17])
- Added noise and reverbation to get additional 40M utterances
- Decode using pretrained ASR model
- From each TTS utterance ASR produces 8 hypotheses
- All of them used to form hypothesis-reference pairs: 640M hypothesis-reference pairs
- also added to ASR trainset

Text-only dataset

- Spelling correction needs parallel corpus: ASR hypothesis + ground truth text
- 800M word LibriSpeech language modeling corpus
- Selected 40M sentences not overlapping with test set
- Generated audio using TTS (WaveNet [vdOLB+17])
- Added noise and reverbation to get additional 40M utterances
- Decode using pretrained ASR model
- From each TTS utterance ASR produces 8 hypotheses
- All of them used to form hypothesis-reference pairs: 640M hypothesis-reference pairs
- also added to ASR trainset

Text-only dataset

- Spelling correction needs parallel corpus: ASR hypothesis + ground truth text
- 800M word LibriSpeech language modeling corpus
- Selected 40M sentences not overlapping with test set
- Generated audio using TTS (WaveNet [vdOLB+17])
- Added noise and reverbation to get additional 40M utterances
- Decode using pretrained ASR model
- From each TTS utterance ASR produces 8 hypotheses
- All of them used to form hypothesis-reference pairs: 640M hypothesis-reference pairs
- also added to ASR trainset

Text-only dataset

- Spelling correction needs parallel corpus: ASR hypothesis + ground truth text
- 800M word LibriSpeech language modeling corpus
- Selected 40M sentences not overlapping with test set
- Generated audio using TTS (WaveNet [vdOLB+17])
- Added noise and reverbation to get additional 40M utterances
- Decode using pretrained ASR model
- From each TTS utterance ASR produces 8 hypotheses
- All of them used to form hypothesis-reference pairs: 640M hypothesis-reference pairs
- also added to ASR trainset

Text-only dataset

- Spelling correction needs parallel corpus: ASR hypothesis + ground truth text
- 800M word LibriSpeech language modeling corpus
- Selected 40M sentences not overlapping with test set
- Generated audio using TTS (WaveNet [vdOLB+17])
- Added noise and reverbation to get additional 40M utterances
- Decode using pretrained ASR model
- From each TTS utterance ASR produces 8 hypotheses
- All of them used to form hypothesis-reference pairs: 640M hypothesis-reference pairs
- also added to ASR trainset

Text-only dataset

- Spelling correction needs parallel corpus: ASR hypothesis + ground truth text
- 800M word LibriSpeech language modeling corpus
- Selected 40M sentences not overlapping with test set
- Generated audio using TTS (WaveNet [vdOLB+17])
- Added noise and reverbation to get additional 40M utterances
- Decode using pretrained ASR model
- From each TTS utterance ASR produces 8 hypotheses
- All of them used to form hypothesis-reference pairs: 640M hypothesis-reference pairs
- also added to ASR trainset

Baseline ASR model

- LAS - Listen, Attend and Spell [CJLV16]
- Encoder-decoder with attention
- encoder: 2 convolutional layers, 3 bidirectional LSTM layers
- decoder: single undirectional LSTM layer

Spelling correction model

- attention-based encoder-decoder sequence-to-sequence
- similar to Neural Machine Translation model from [CFB+18]
- Encoder: 3 bi-directional LSTM layers
- Decoder: 3 unidirectional LSTM layers

Architecture

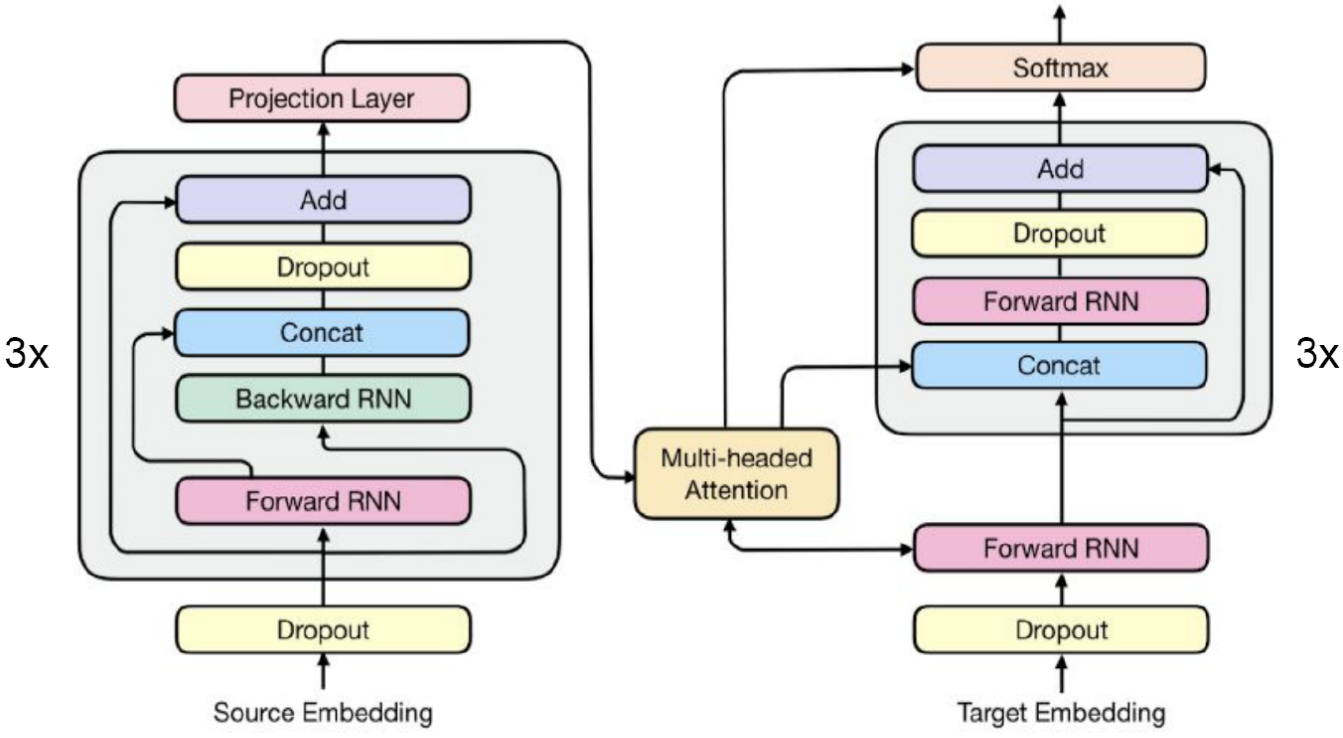

Fig. 1. Spelling Correction model architecture.

Language model

- 2 unidirectional LSTM layers
- used to rescore n-best list generated by ASR

Inference

- ASR produces N-best list of hypotheses with log prob scores (p_i)
- Spelling correction produces M-best list for each ASR hypothesis with scores (q_{ij})
- LM rescroing of each of $M \times N$ hypotheses with r_{ij} score
- Most likely hypothesis:

$$A^* = \underset{A}{\operatorname{argmax}} \lambda_{LAS} * p_i + \lambda_{SC} * q_{ij} + \lambda_{LM} * r_{ij}$$

Results

System	Dev-clean	Test-clean
LAS	5.80	6.03
LAS $\rightarrow$ LM (8)	4.56	4.72
LAS-TTS	5.68	5.85
LAS-TTS $\rightarrow$ LM (8)	**4.45**	**4.52**
LAS $\rightarrow$ SC (1)	5.04	5.08
LAS $\rightarrow$ SC (8) $\rightarrow$ LM (64)	**4.20**	**4.33**
LAS $\rightarrow$ SC-MTR (1)	4.87	4.91
LAS $\rightarrow$ SC-MTR (8) $\rightarrow$ LM (64)	**4.12**	**4.28**

Table 1. Word error rates (WERs) on LibriSpeech "clean" sets comparing different techniques for incorporating text-only training data. Numbers in parentheses indicate the number of input hypotheses considered by the corresponding model.

Results

System	Dev-clean	Test-clean
LAS	3.11	3.28
LAS $\rightarrow$ SC (1)	3.01	3.02
LAS $\rightarrow$ SC (8)	**1.63**	**1.68**

Table 2. Oracle WER before and after applying the SC model.

Results

System	Dev-clean	Dev-TTS
LAS baseline	5.80	5.26
LAS $\rightarrow$ SC (1)	5.04	**3.45**
LAS $\rightarrow$ SC (8) $\rightarrow$ LM (64)	4.20	**3.11**

Table 3. WER comparison on a real audio and TTS dev sets.

Thank you

Thank you for your attention!

References I

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Niki Parmar, Mike Schuster, Zhifeng Chen, Yonghui Wu, and Macduff Hughes, The best of both worlds: Combining recent advances in neural machine translation, CoRR abs/1804.09849 (2018).

William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals, Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, ICASSP, 2016.

Jinxi Guo, Tara N Sainath, and Ron J Weiss, A spelling correction model for end-to-end speech recognition, arXiv preprint arXiv:1902.07178 (2019).

References II

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, Librispeech: An asr corpus based on public domain audio books, 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 2015, pp. 5206–5210.

Aäron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis C. Cobo, Florian Stimberg, Norman Casagrande, Dominik Grewe, Seb Noury, Sander Dieleman, Erich Elsen, Nal Kalchbrenner, Heiga Zen, Alex Graves, Helen King, Tom Walters, Dan Belov, and Demis Hassabis, Parallel wavenet: Fast high-fidelity speech synthesis, CoRR abs/1711.10433 (2017).