Learning Word-Level Confidence for Subword End-to-End ASR

Google

David Qiu Qiujia Li[1] Yanzhang He Yu Zhang Bo Li Liangliang Cao Rohit Prabhavalkar Deepti Bhatia Wei Li Ke Hu Tara N. Sainath Ian McGraw

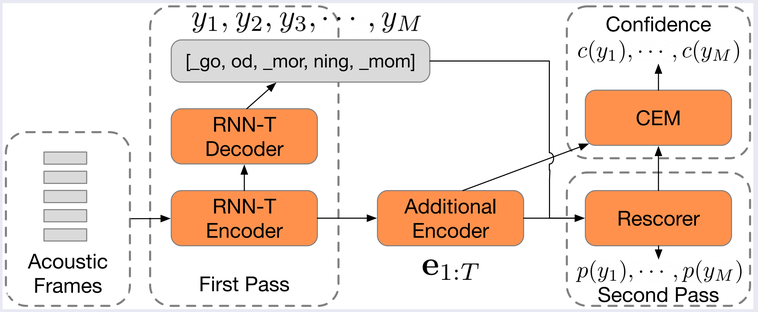
Google, LLC, USA [1]University of Cambridge, UK

{qdavid,qiujia,yanzhanghe}@google.com

1. Summary

- We present a dedicated confidence estimation module (CEM) that:
  - learns from the word-level Levenshtein edit distance tags over N-best hypotheses
  - uses transformer self-attention to eliminate the need for hand-designed subword to word confidence aggregation functions
  - uses deliberation to attend to consensus from multiple hypotheses
  - is better matched to downstream WER-related applications (e.g., system combination)
- Empirically evaluate the improvement with each model enhancement
- Show that the proposed model is better calibrated and improves long-tail WER when combined with a conventional ASR

2. Confidence Estimation Module for Two-Pass Subword ASR



We base the CEM on the state of the art two-pass ASR in [2].

1. Word-piece (WP) as the vocabulary
2. Streaming first-pass RNN-T generates N-best hypotheses; label-synchronous second-pass transformer rescores hypotheses
3. CEM outputs a scalar value $\in [0, 1]$ for each output WP

3. Multiple Tokenization Problem

Hyp:	_go	od	_mor	ning	_mom
Ref:	_go	od	_morn	ing	
WP edit:	*cor*	*cor*	*sub*	*sub*	*ins*
Word edit:	–	*cor*	–	*cor*	*ins*
$d(w_j)$:	–	1	–	1	0
$m(y_i)$:	0	1	0	1	1
$\mathcal{L}(w_j)$:	–	$\log c_w(w_1)$	–	$\log c_w(w_2)$	$\log(1 - c_w(w_3))$

- ASR outputs WP sequence while reference is word sequence
- Possible to tokenize the reference and run edit distance on the WPs
  - WP tokenization is not unique, and can cause incorrect WP tags when the word is correct
- Instead, we only retain the final CEM output per word and train it with cross-entropy loss to predict whether the word-level edit distance tag is "cor"

4. Subword Deletion Problem

Hyp:	_go		_mor	ning
Ref:	_go	od	_morn	ing
WP edit:	*cor*	*del*	*sub*	*sub*
Word edit:	*sub*	–	–	*cor*
$d(w_j)$:	0	–	–	1
$m(y_i)$:	1	0	0	1
$\mathcal{L}(w_j)$:	$\log(1 - c_w(w_1))$	–	–	$\log c_w(w_2)$

- Words where some constituent WPs are correct and some are deletions should be tagged as word-level substitution errors
- Deletions are not modeled – training on WP tags results in overconfidence (all "cor" labels)

5. Word-Level Confidence Using Transformer / Deliberation

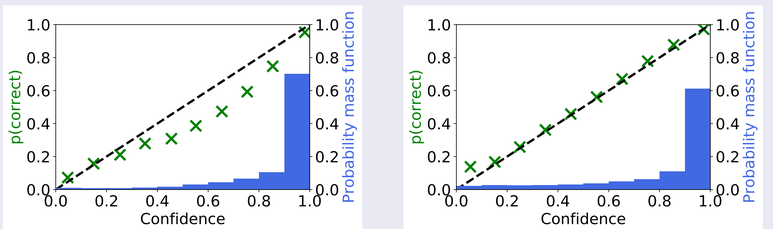
- ϕ: rescorer's penultimate layer activations; Emb: output token embedding; top-K: K most likely output tokens' rescorer log probabilities; log($p(\cdot)$): output token's log probability
- Deliberation attends to multi-hypotheses consensus from the concatenated BLSTM encodings of the top 1 or 8 hypotheses

6. Experiments

Confidence Models	NCE	AUC ROC	AUC PR	WCR RMSE	(1 – WER) RMSE
ASR Softmax	0.241	0.873	0.280	0.140	0.244
WP MLP [1]	0.269	0.885	0.329	0.138	0.233
WP Xformer	0.280	0.885	0.347	0.137	0.231
E2E Xformer	0.367	0.928	0.466	0.130	0.221
+Delib 1-Hyp	0.361	0.923	0.474	0.128	0.206
+Delib 8-Hyp	**0.425**	**0.941**	**0.508**	**0.127**	**0.204**

- Higher is better for NCE, AUC-ROC, AUC-PR; lower is better for WCR RMSE, (1 – WER) RMSE

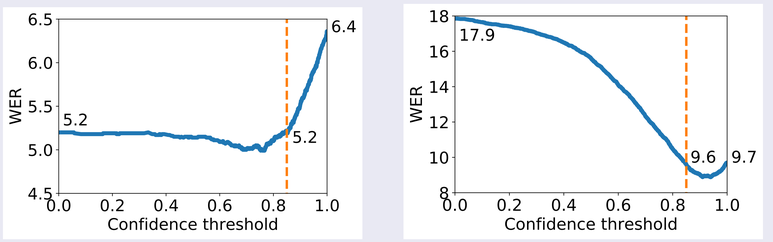
7. Confidence Calibration Curves



(a) ASR Softmax (b) E2E Xformer

Figure: The black and green curves show the ideal and actual calibration curves, respectively. The blue bar plot shows the probability mass in each bin.

8. Application: System Combination



(a) Voice Search (b) Long-tail Maps

Figure: Overall WER at different operating points for system combination. When confidence threshold is 0, all utterances are processed on-device. When it is 1, all utterances are processed on the server.

- On-device is more accurate on Voice Search, while server is more accurate on long-tail (due to language model, etc.)
- System combination logic: trust on-device recognition if confidence is larger than a threshold, use server recognition otherwise
- Combined system achieves the lower WER between the two systems on Voice Search and long-tail Maps

9. References

[1] Q. Li, D. Qiu, Y. Zhang, B. Li, Y. He, et al. Confidence estimation for attention-based sequence-to-sequence models for speech recognition. In *ICASSP*, 2021.

[2] W. Li, J. Qin, C.-C. Chiu, R. Pang, and Y. He. Parallel rescoring with transformer for streaming on-device speech recognition. In *Interspeech*, 2020.