Confidence Estimation and Deletion Prediction using Bi-Directional Recurrent Neural Networks

Anton Ragni, Qiujia Li, Mark Gales, Yu Wang
{ar527,ql264,mjfg,yw396}@eng.cam.ac.uk

Department of Engineering / University of Cambridge

UNIVERSITY OF CAMBRIDGE

1. Introduction

Important role of confidence scores

Up-stream applications: adaptation, semi-supervised training

Down-stream applications: information retrieval, language assessment

Issues with current confidence estimation approaches

Use limited context: current word, past words

Cannot predict deletions: only prior work with CRFs

This work — bi-directonal recurrent neural networks

- Modified to predict confidence scores and deletions
- Examined in both matched and highly mismatched domains

2. Confidence Scores and Deletions

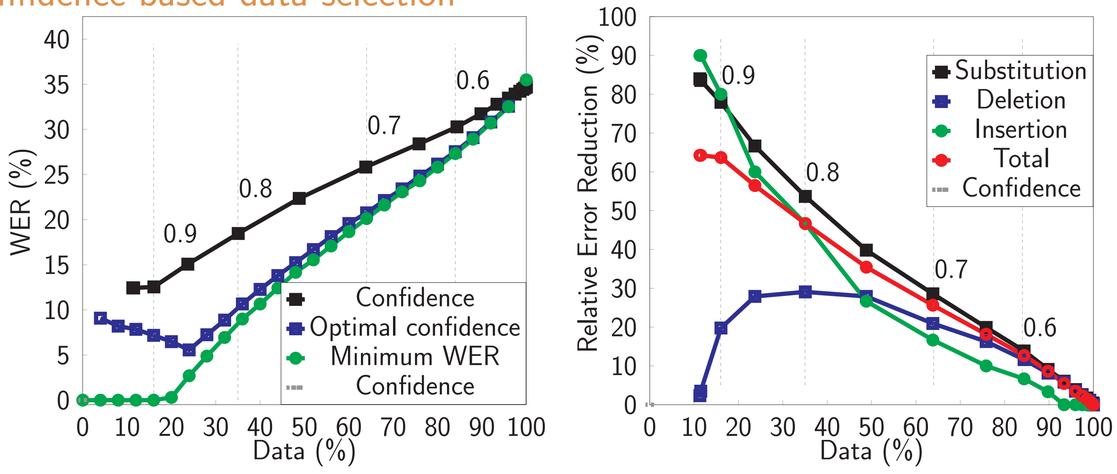
Confidence score estimation

- standard approach uses word posterior probabilities, $c_t = P(w_t|\mathbf{O})$
- map confidence scores using decision trees

Utterance/waveform confidence estimate

$$C = \sum_{t=1}^{T} \lambda_t c_t \Big/ \sum_{t=1}^{T} \lambda_t$$

Confidence-based data selection



- Impact of deletions can be moderated given accurate confidences
- Distribution of errors varies with confidence

Impact of domain mismatch

Type	Band		Error (%)			
	Model	Data	Sub	Del	Ins	Tot
Sup	Narrow	Narrow	24.1	8.3	3.1	35.5
	Narrow	Wide	23.4	19.0	1.3	43.7
Unsup	Wide	Wide	15.6	25.5	0.9	42.0

- Domain mismatch causes large increase in deletions
- Cannot be handled well by standard ASR approaches

3. Bi-Directional Recurrent Neural Networks

(a) confidence prediction (b) +deletion prediction

Use standard network topology to predict confidence scores

- Hidden state dynamics (forward and backward directions similar)

$$\overrightarrow{\boldsymbol{h}}_{t+1} = \boldsymbol{\sigma}\left(\boldsymbol{W}^{(\overrightarrow{h})}\overrightarrow{\boldsymbol{h}}_t + \boldsymbol{W}^{(x)}\boldsymbol{x}_{t+1}\right) \qquad \boldsymbol{h}_t = \left[\overrightarrow{\boldsymbol{h}}_t^{\mathsf{T}}\ \overleftarrow{\boldsymbol{h}}_t^{\mathsf{T}}\right]^{\mathsf{T}}$$

- Probability of current word being correct - confidence score

$$c_t = \sigma\left(\boldsymbol{w}^{(c)\mathsf{T}}\boldsymbol{h}_t + b^{(c)}\right)$$

Modify standard topology to predict deletions

- Probability that deletion occurs before the first, s_1, and next, d_t, word

$$s_1 = \sigma\left(\boldsymbol{w}^{(s)\mathsf{T}}\boldsymbol{h}_1 + b^{(s)}\right), \qquad d_t = \sigma\left(\boldsymbol{w}^{(d)\mathsf{T}}\boldsymbol{h}_t + b^{(d)}\right)$$

Input features play crucial role - only simple features in this work

- Word posterior probability and duration
- Word embedding, n-gram probability, back-off order, # characters
- Time lag between previous and current word, following and current word

4. Deletion-aware Data Selection

Incorporate confidence **and** deletion predictions

- Modify the standard selection criterion *or* propose a new approach

Confidence discounting

$$\hat{c}_t = c_t - \theta_d d_t$$
$$\hat{c}_1 = c_1 - \theta_d d_1 - \theta_s s$$

Confidence thresholding

Estimate of WER for current utterance

$$\text{WER}(\boldsymbol{c}, \boldsymbol{d}; \boldsymbol{\theta}) = \frac{\text{Inc}(\boldsymbol{c}; \theta_c) + \text{Del}(\boldsymbol{d}; \theta_s, \theta_d)}{\theta_p \text{Inc}(\boldsymbol{c}; \theta_c) + \text{Cor}(\boldsymbol{c}; \theta_c)}$$

- number of correctly recognised words

$$\text{Cor}(\boldsymbol{c}; \theta_c) = \sum_{t=1}^{T} \delta(c_t \geq \theta_c)$$

- number of incorrectly recognised words

$$\text{Inc}(\boldsymbol{c}; \theta_c) = \sum_{t=1}^{T} \delta(c_t < \theta_c)$$

- number of deleted words

$$\text{Del}(\boldsymbol{d}; \theta_s, \theta_d) = \delta(s \geq \theta_s) + \sum_{t=1}^{T} \delta(d_t \geq \theta_d)$$

Thresholds set by minimising MMSE between WER and WER$(\boldsymbol{c}, \boldsymbol{d}; \boldsymbol{\theta})$

5. Experiments

Experimental setup

Narrow-band data: IARPA Babel Swahili 60 hrs pack

Wide-band training data: untranscibed Voice of America(Swahili) 100 hrs

Wide-band test data: IARPA MATERIAL Swahili analysis 10 hrs pack

ASR: lattice-free TDNN-LSTM and n-gram language model

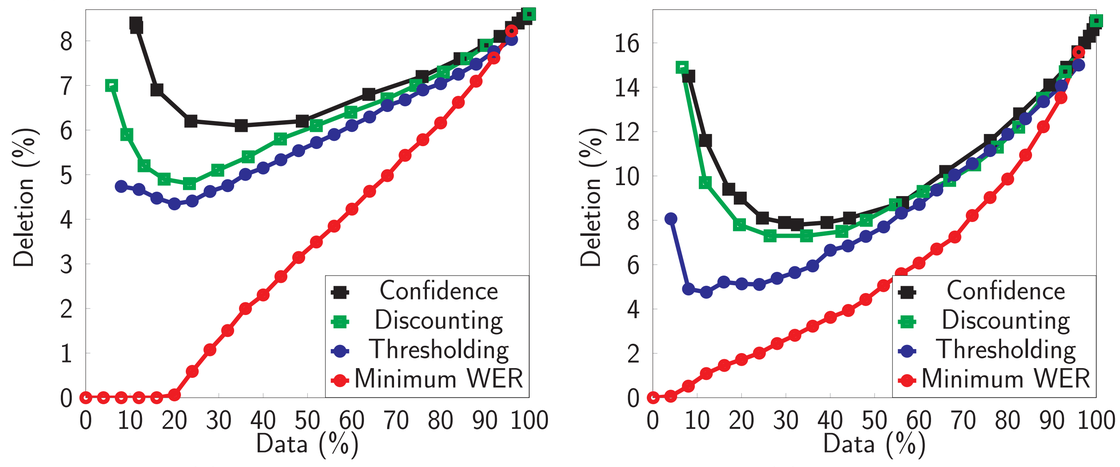
Accuracy of confidence and deletion prediction in matched domain

Predictor	Prediction	AUC (ROC) Posterior	AUC (ROC) +BiRNN
Cor/Inc	Cor/Inc	0.832	0.846
+Del	Cor/Inc	0.832	0.846
	Next Del	—	0.742
	Start Del	—	0.746

- Deletion prediction accuracy clearly above chance level

Deletion-aware data selection schemes in (mis)matched domains

Generalisation on narrow and wide-band data



(a) narrow-band (b) wide-band

- Different behaviour in low, mid and high confidence regions

Data selection scheme performance on wide-band data

- Select 50% data using each scheme

Selection	Error (%) Sub	**Del**	Ins	**Tot**
narrow-band	23.4	**19.0**	1.3	**43.7**
confidence	15.6	**25.5**	0.9	**42.0**
discount	16.9	**23.7**	1.1	**41.7**
threshold	16.5	**25.2**	1.1	**42.8**

- Discounting yields small drop in deletion and total error

6. Conclusions

- Proposed BiRNN for confidence and deletion prediction
- Proposed data selection schemes for semi-supervised training
- Assessed in highly challenging mismatched domain scenario