Confidence scores for sequence data

Alexandros Kastanos, Mark Gales, Anton Ragni
Department of Engineering, University of Cambridge

Introduction

- Automatic speech recognition aims to generate a transcription for a given speech recording.

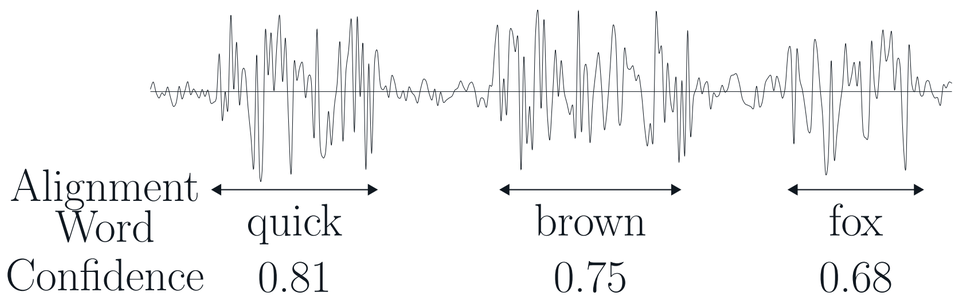

Fig. 1: Transcription and confidence score prediction for the phrase: *quick brown fox*.

- A good confidence score is able to predict errors in a hypothesis transcription.
- The aim of this research is to improve confidence scores in the context of speech processing.

LatticeRNN for Confidence

- Generalisation of BiLSTM to graph-like structures
- Incoming arcs are merged using attention

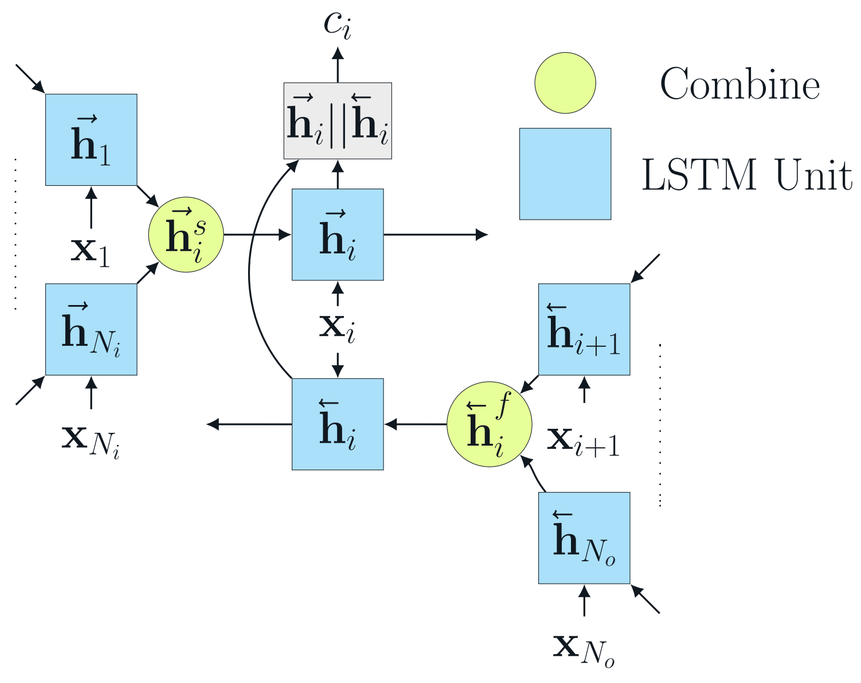

Fig. 2: Bi-Directional LatticeRNN for confidence estimation.

Results

Table 1: Improving confidence score estimates for one-best sequences

Model	NCE	AUC
Unmapped word posteriors	-0.1978	0.9081
Decision tree mapped word posteriors	0.2755	0.9081
BiLSTM	0.2911	0.9121
BiLSTM + grapheme encoder	0.2978	0.9139

Table 2: Improving confidence score estimates for confusion networks

	One-best arcs		All arcs	
Features	**NCE**	**AUC**	**NCE**	**AUC**
Word-based	0.2934	0.9201	0.4959	0.8406
+ grapheme features	0.2998	0.9228	0.4993	0.8432
+ LM and AM	0.3004	0.9231	0.5013	0.8444

Representing Hypotheses

- Simplest output from an ASR is a 1-best hypothesis.
- Confusion networks and lattices represent the N-best competing hypotheses efficiently.

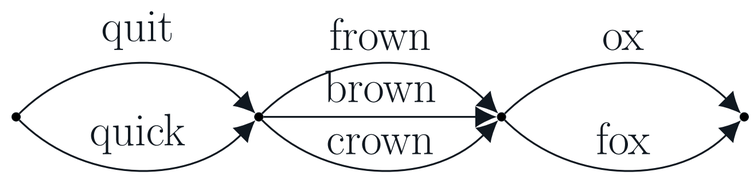

Fig. 3: Confusion network

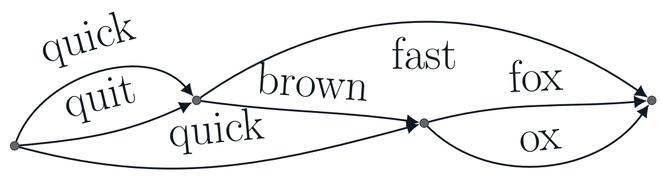

Fig. 4: Word lattice

Sub-word Level Features

- Typically, word-level features such as the word duration (t_i^d), the language model score (LM_i), acoustic model score (AM_i), and word posterior (p_i) are used.
- This research investigates methods for incorporating sub-word level features - specifically grapheme-level features.

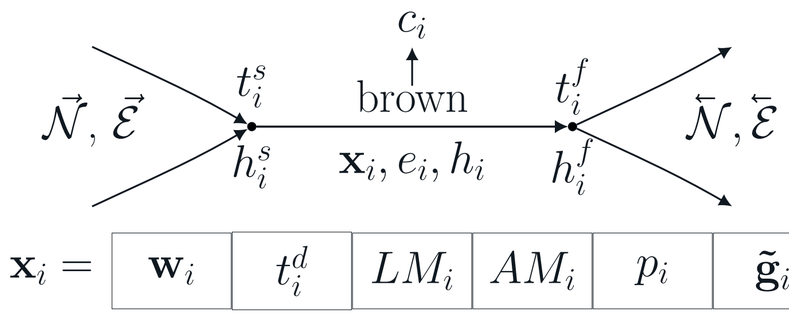

Fig. 5: Lattice edge with word and sub-word level features.

- Fixed representation of grapheme-level features is required.
- A BiGRU with additive attention operates as a grapheme encoder to produce $\tilde{\mathbf{g}}_i$.

Conclusion

- Sub-word level features can be incorporated into existing LatticeRNN models using a grapheme encoder.
- The grapheme embedding and grapheme duration are complimentary features for confidence estimation.

References

[1] Qiujia Li et al. "Bi-Directional Lattice Recurrent Neural Networks for Confidence Estimation". In: *arXiv preprint arXiv:1810.13024* (2018).

[2] Anton Ragni et al. "Confidence Estimation and Deletion Prediction Using Bidirectional Recurrent Neural Networks". In: *SLT*. 2018.