Uncertainty and confidence scores in sequence data

Alexandros Kastanos, Mark Gales

Department of Engineering, University of Cambridge

UNIVERSITY OF CAMBRIDGE

Introduction

- The aim of this research to improve confidence scores in the context of speech processing.
- Automatic speech recognition aims to generate a transcription for a given speech recording.
- A good confidence score is able to predict errors in the generated transcription.
- This information is useful for applications such as speaker adaptation [1] and error detection [2].

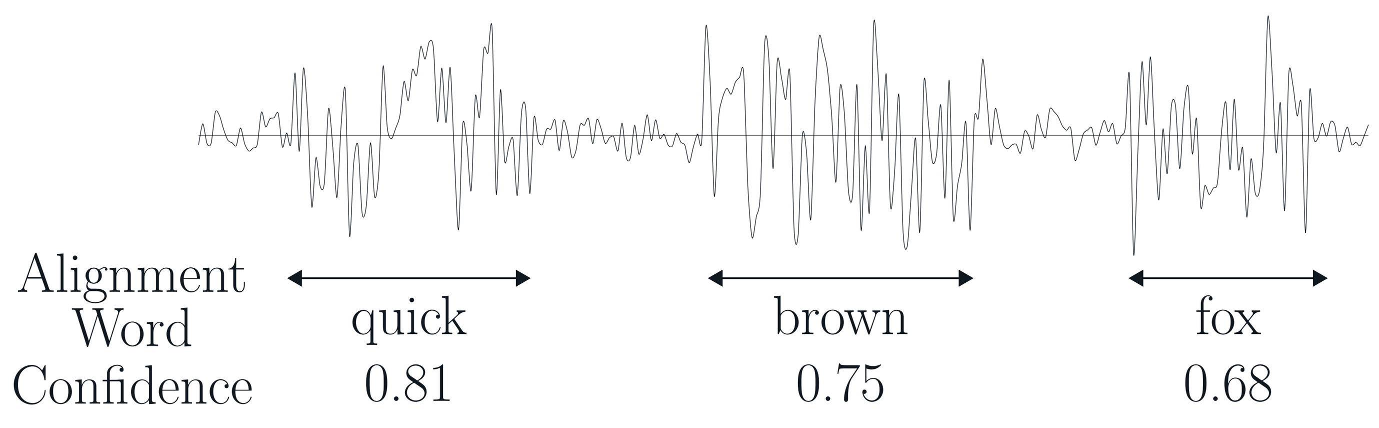

Fig. 1: Example audio recording of the phrase "quick brown fox" with word and confidence score predictions.

Confidence scores for 1-best sequences

- Traditionally confidence scores use 1-best hypotheses.
- Propagate information through the sequence in one direction to generate confidence scores.

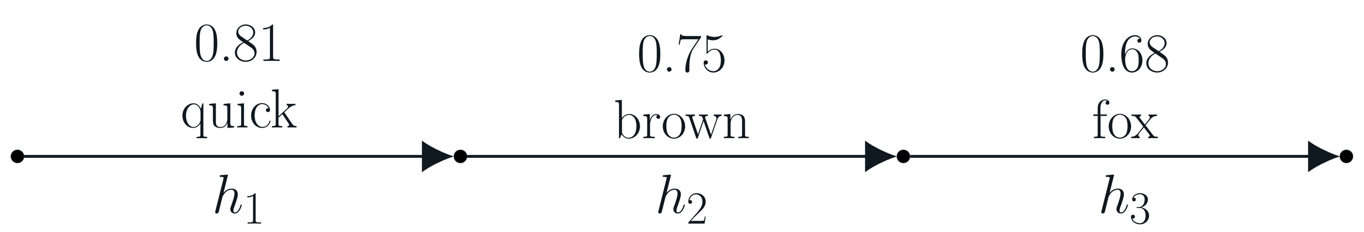

Fig. 2: A 1-best hypothesis with the predicted word, confidence score, and hidden state h_i

Lattice representation

- Represents N-best lists in an efficient and compact form.
- Additional information from each confusion is considered.

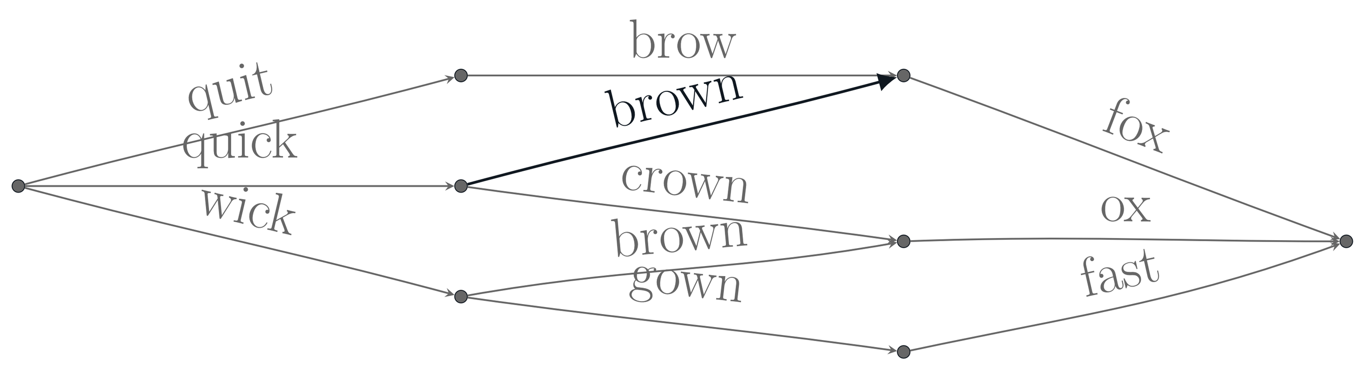

Fig. 3: A simple word-marked lattice

Lattice enrichment

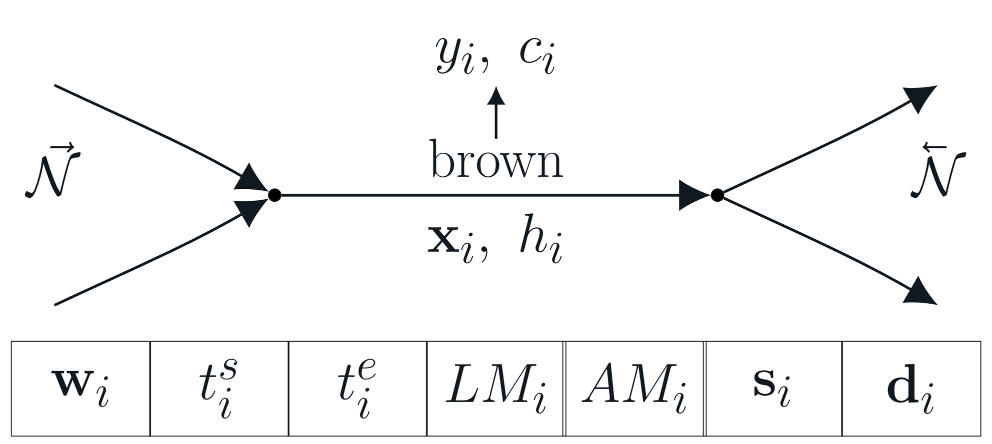

Fig. 4: Edge e_i with enriched features

$$h_i = f(h_{\mathcal{N}}) = f(h_{\vec{\mathcal{N}}}, h_{\overleftarrow{\mathcal{N}}}) \quad (1)$$

$$c_i = f(h_i, \mathbf{x}_i) \quad y_i = f(h_i, \mathbf{x}_i) \quad (2)$$

- Using grapheme-marked lattices allows subword level features to be embedded in the arcs.
- These features include the language model score (LM_i), acoustic model score (AM_i), and a fixed length representation of the grapheme information such as the duration ($\mathbf{d}_i$).
- Arc combination and aggregation of grapheme information is achieved through attention.

LatticeRNN

- Bidirectional recurrent architecture which considers the forward and backward probabilities distinctly.
- The existing arc combination procedure can be improved by applying attention over arc neighbourhoods $\vec{\mathcal{N}}$ and $\overleftarrow{\mathcal{N}}$ rather than just the incoming and outgoing arcs.

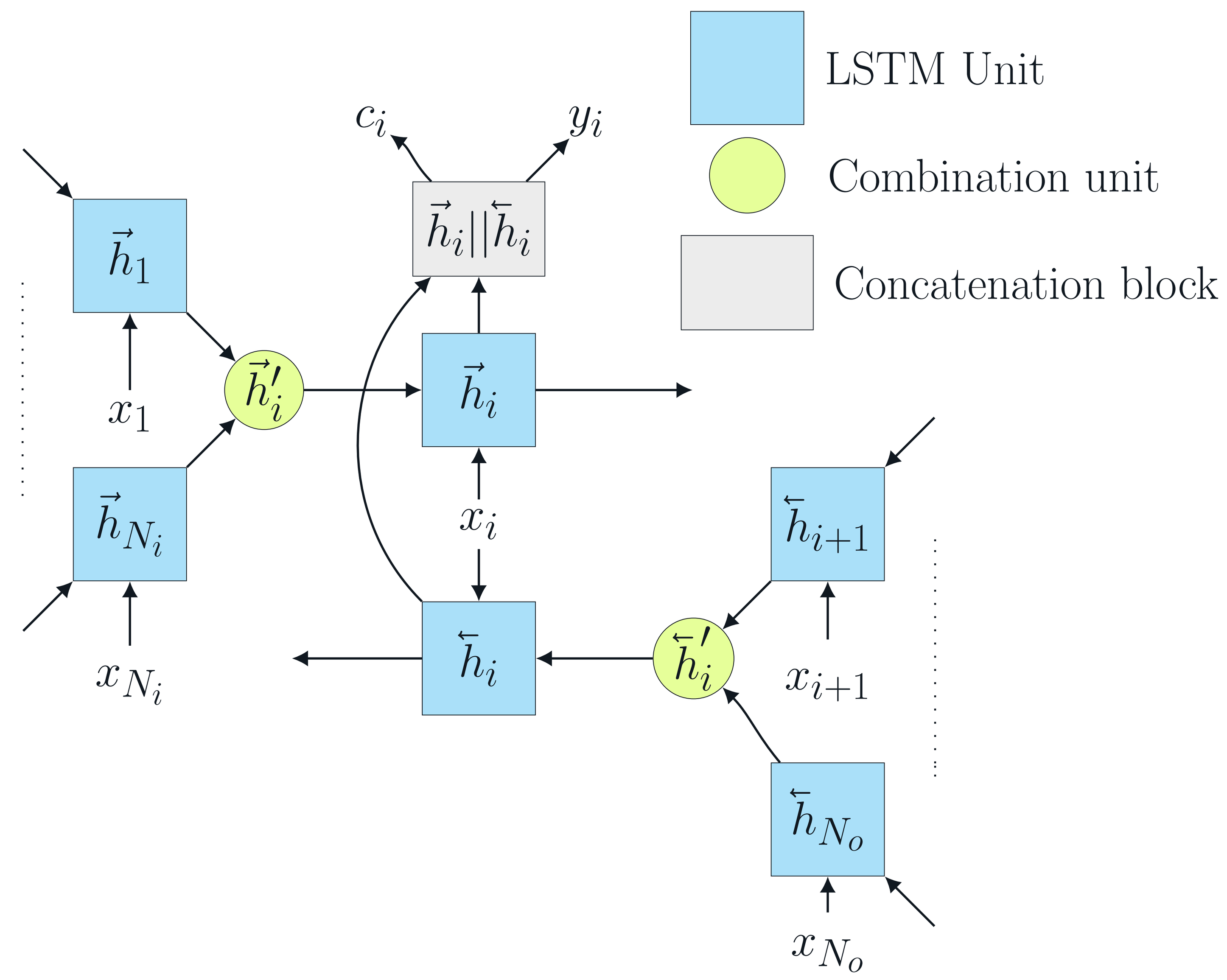

Fig. 5: LatticeRNN model for confidence score prediction [3].

References

[1] Luís Felipe Uebel and Philip C. Woodland. "Speaker adaptation using lattice-based MLLR". In: 2001.

[2] Hui Jiang. "Confidence measures for speech recognition: A survey". In: *Speech communication* 45.4 (2005), pp. 455–470.

[3] Qiujia Li et al. "Bi-directional Lattice Recurrent Neural Networks for Confidence Estimation". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 6755–6759.