COM4511 Speech Technology: Integrating with Others

Anton Ragni

March 2, 2020

Statistical Speech Recognition

- Given a parameterised audio sequence, infer the underlying latent representation
  - parameterised audio: sequences of feature vectors $\boldsymbol{O}_{1:T} = \boldsymbol{o}_1, \ldots, \boldsymbol{o}_T$
  - latent representation: sequences of words $\boldsymbol{w}_{1:L} = w_1, \ldots, w_L$
- Options for inference
  - maximum-a-posteriori

$$\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} \left\{ P(\boldsymbol{w}|\boldsymbol{O}_{1:T}) \right\}$$

    - yields most probable sequence of words (sentence-level)
  - minimum Bayes' risk

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \left\{ \sum_{\boldsymbol{w}'} P(\boldsymbol{w}'|\boldsymbol{O}_{1:T}) \mathcal{L}(\boldsymbol{w}, \boldsymbol{w}') \right\}$$

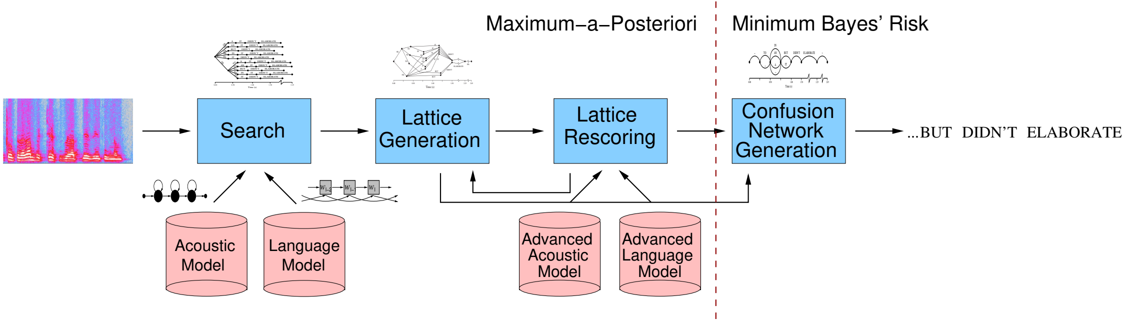
    - yields sequence of words with the smallest expected loss (word or character level)
- Need to know:
  - how to model the posterior and perform inference (search)

Why Integration?

- Speech technology used across large number of applications
  - back-end: dictation, subtitling, machine translation
  - front-end: spoken dialogue systems, information retrieval
- No machine learning solution (including speech recognition!) is free from mistakes
  - error mitigation critical for successful use of speech technology
- Error mitigation strategies
  - information preservation (rich representations)
  - uncertainty measures (confidence scores)

Anton Ragni

Inference (Revisited)



- Standard inference pipeline
  - multiple passes of gradual search space refinement
  - may additionally involve rounds of acoustic and language model adaptation

Anton Ragni

Why One Best? (or Why Not?)

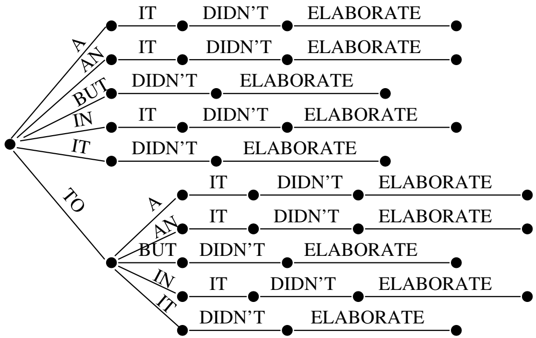
Start Time	Duration	Word	Confidence
303.95	0.12	ONE	0.26
304.07	0.06	OF	0.25
304.13	0.07	THE	0.34
304.20	0.22	HARRY	0.62
304.42	0.25	POTTER	0.29
304.67	0.21	BOOKS	0.23
304.88	0.17	IN	0.59
305.05	0.42	CHINESE	0.15

An excerpt from BBC4 The Book Quiz

- Advantages:
  - highly compact representation
  - may include addition information (start time, duration, confidence)
- Disadvantages:
  - significant loss of information

Anton Ragni

Prefix Trees



Available information

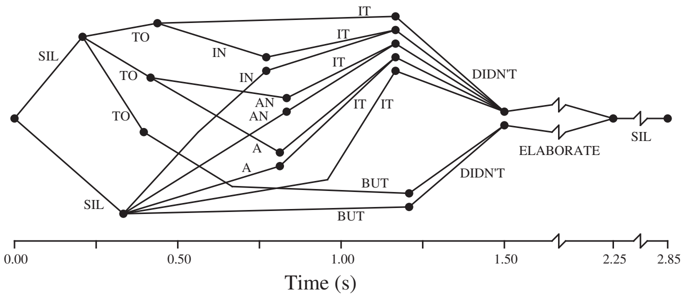
- nodes:
  - word (n-gram)
- arcs:
  - language probability

Other structures possible

- Number of possible word sequences grows exponentially with sequence length
  - possible to filter out unlikely sequences using language model
- Example:
  - how many 10-word sequences generated by language model with perplexity of 60?
  - how many paths fit into 1GB (4-byte word, language probability)?

Anton Ragni

Lattices



Available information

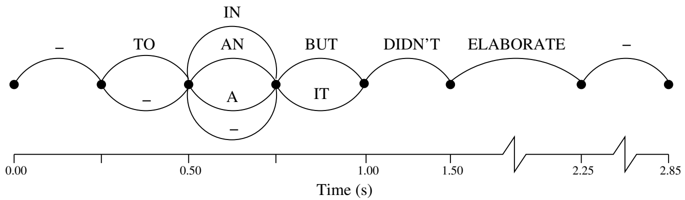
- nodes:
  - word
  - end time
- arcs:
  - acoustic probability
  - language probability
  - pronunciation probability

Other structures possible

- Apply *n*-gram approximation during prefix tree generation
  - merge any paths where past $n-1$ words are identical
- Yields variant of directed acyclic graph or lattice offering multiple advantages
  - highly compact representation of numerous word sequences
  - possible to use graph algorithms (determinisation, weight pushing, shortest path)
- Example:
  - compute cost of storing 3.2×10^{44} paths using 8823 nodes and 119975 arcs

Anton Ragni

Confusion Networks



Available information

- nodes:
  - approximate end time
- arcs:
  - word
  - posterior probability

- Simplify general acyclic graph structure
  - cluster nodes in time and aggregate posterior probabilities
- Yields linear graph structure offering multiple advantages
  - even more compact form yet contains all seen and many unseen word sequences
  - enables simple minimum Bayes' risk inference

Confidence Scores

Reference	—	AN	ELABORATE	MEAL
Hypothesis	BUT	DIDN'T	ELABORATE	—
Posterior	0.3	0.9	0.7	—
Error	Ins	Sub	—	Del

- Useful to know whether hypothesised transcriptions are correct or not
  - three error classes: substitution, insertion, deletion
- Uncertainty measure provides a principled approach
  - BUT hard to derive for standard sequence models
  - alternatively could use surrogate quantities — confidence scores
- Arc posterior probabilities offer simplest form of confidence score
  - cannot handle deleted words
  - a form of self-assessment (bias)
  - over-estimates confidence due to limited lattice paths

Evaluation

- Standard measure of confidence score accuracy is normalised cross-entropy
  - information gain from predicting confidence rather than using average reference value

$$\bar{\mathcal{H}}(\boldsymbol{c}^*_{1:L}, \boldsymbol{c}_{1:L}) = \frac{\mathcal{H}(\bar{\boldsymbol{c}}^*_{1:L}, \boldsymbol{c}^*_{1:L}) - \mathcal{H}(\boldsymbol{c}^*_{1:L}, \boldsymbol{c}_{1:L})}{\mathcal{H}(\bar{\boldsymbol{c}}^*_{1:L}, \boldsymbol{c}^*_{1:L})}$$

  - average sample (binary) cross-entropy

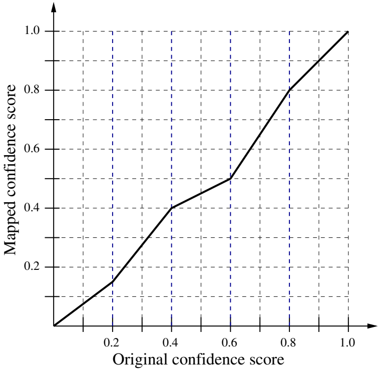
$$\mathcal{H}(\boldsymbol{c}^*_{1:L}, \boldsymbol{c}_{1:T}) = -\frac{1}{L}\sum_{l=1}^{L} c^*_l \log(c_l) + (1 - c^*_l)\log(1 - c_l)$$

  - reference $\boldsymbol{c}^*_{1:L}$ and predicted $\boldsymbol{c}_{1:L}$ confidences, average reference confidence $\bar{c}^* = \frac{1}{L}\sum_{l=1}^{L} c^*_l$
- Often interested in a simple threshold rather than perfect confidence predictions
  - area under the curve (precision-recall or ROC) type metric more appropriate

SPandH

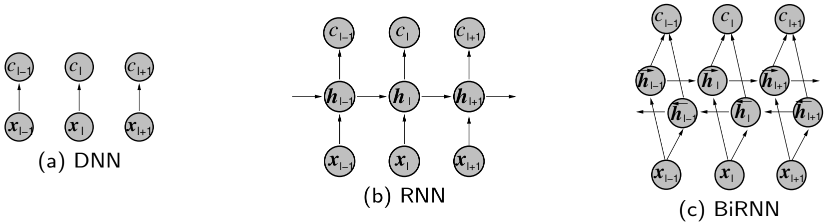
Anton Ragni

Decision Trees



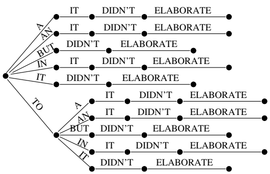
- Calibrate confidence scores using piece-wise linear mapping
  - partition scores into non-overlapping confidence ranges
  - estimate linear correction by fitting held-out confidence scores
- Simple yet effective confidence score calibration approach

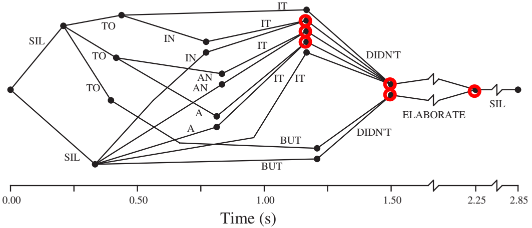
Neural Networks



- Alternatively use any suitable form of neural network
  - wide range of features and architectures possible
- Issues with machine learning approaches for confidence estimation
  - hard to find large quantities of labelled held-out data
  - cannot use recurrent neural networks with graph structures

Lattice Embeddings





- Recurrent unit introduces dependency on the complete word history
  - lattices do not maintain unique word histories (red circles) unlike prefix trees
- Merge available word histories using an attention mechanism

$$\boldsymbol{h}_i^{(n)} = \sum_{j \in \overrightarrow{\mathcal{A}}_i} \alpha_j \boldsymbol{h}_j^{(a)}$$

  - attention weights reflect relevance of word histories
- Alternatively, it is possible to cluster similar word histories
  - *n*-gram approximation, distance measure
  - standard approach to lattice rescoring with RNN language models
- Multiple options for lattice embeddings available
  - final history vector, attention over all history vectors

Anton Ragni

Audio Embeddings

- Lattice embeddings rely on external lattice generation mechanism
  - normally optimised to fit a different objective function
  - computationally expensive and not always available
- Alternatively, it is possible to learn embeddings directly from audio
  - BUT need to know how to define "good" embedding for any word sequence

Summary

- This lecture examined integration with down-stream applications
  - possible speech representations
  - uncertainty measures
- Focus on graph representations and confidence scores
  - lattices and confusion networks
  - confidence score calibration and evaluation
- Next lectures will examine advanced topics
  - adaptation, diarisation