Automatic Speech Recognition (CS753)
Lecture 18: Search & Decoding (Part I)

Instructor: Preethi Jyothi
Mar 23, 2017
Recall ASR Decoding

\[
W^* = \arg \max_W \Pr(O_A|W) \Pr(W)
\]

\[
W^* = \arg \max_{W_1^N, N} \left\{ \prod_{n=1}^{N} \Pr(w_n|w_{n-1}^{n-1}) \left[ \sum_{q_1^T, w_1^N} \prod_{t=1}^{T} \Pr(O_t|q_t, w_1^N) \Pr(q_t|q_{t-1}, w_1^N) \right] \right\}
\]

Viterbi

\[
\approx \arg \max_{W_1^N, N} \left\{ \prod_{n=1}^{N} \Pr(w_n|w_{n-1}^{n-1}) \left[ \max_{q_1^T, w_1^N} \prod_{t=1}^{T} \Pr(O_t|q_t, w_1^N) \Pr(q_t|q_{t-1}, w_1^N) \right] \right\}
\]

- Viterbi approximation divides the above optimisation problem into sub-problems that allows the efficient application of dynamic programming
- An exact search using Viterbi is infeasible for large vocabulary tasks!
Recall Viterbi search

- Viterbi search finds the most probable path through a trellis of time on the X-axis and states on the Y-axis.

Viterbi algorithm: Only needs to maintain information about the most probable path at each state.

9.5 HMM Training: The Forward-Backward Algorithm

Image from [JM]: Jurafsky & Martin, 3rd edition, Chapter 9
the boy is walking

Network of HMM states

Network of phones

Network of words

0

the birds are is

boy
Time-state trellis

Time, $t \rightarrow$
Viterbi search over the large trellis

- Exact search is infeasible for large vocabulary tasks
  - Unknown word boundaries
  - N-gram language models greatly increase the search space

- Solutions
  - Compactly represent the search space using WFST-based optimisations
  - Beam search: Prune away parts of the search space that aren’t promising
Viterbi search over the large trellis

- Exact search is infeasible for large vocabulary tasks
  - Unknown word boundaries
  - Ngram language models greatly increase the search space

- Solutions
  - Compactly represent the search space using WFST-based optimisations
    - Beam search: Prune away parts of the search space that aren’t promising
Two main WFST Optimizations

- Use determinization to reduce/eliminate redundancy

Recall not all weighted transducers are determinizable

To ensure determinizability of $L \circ G$, introduce disambiguation symbols in $L$ to deal with homophones in the lexicon

- `read : r eh d #0`
- `red   : r eh d #1`

Propagate the disambiguation symbols as self-loops back to $\hat{C}$ and $\hat{H}$. Resulting machines are $\hat{H}$, $\hat{C}$, $\hat{L}$
Two main WFST Optimizations

- Use determinization to reduce/eliminate redundancy
- Use minimization to reduce space requirements

Minimization ensures that the final composed machine has minimum number of states

Final optimization cascade:

$$N = \pi_\varepsilon(\min(\det(\tilde{H} \circ \det(\tilde{C} \circ \det(\tilde{L} \circ G))))))$$

Replaces disambiguation symbols in input alphabet of $\tilde{H}$ with $\varepsilon$
Example G

Diagram:

- States: 0, 1, 2
- Transitions:
  - 0 to 1: bob:bob, bond:bond, rob:rob
  - 1 to 2: slept:slept, read:read, ate:ate

- Arrows indicate the transitions between states.
Compact language models (G)

- Use Backoff Ngram language models for G
Example G
Example $\tilde{L}$: Lexicon with disambig symbols
\( \det(\tilde{L} \circ G) \)

\( \min(\det(\tilde{L} \circ G)) \)
Viterbi search over the large trellis

- Exact search is infeasible for large vocabulary tasks
  - Unknown word boundaries
  - Ngram language models greatly increase the search space

Solutions

- Compactly represent the search space using WFST-based optimisations
- Beam search: Prune away parts of the search space that aren’t promising
Beam pruning

- At each time-step $t$, only retain those nodes in the time-state trellis that are within a fixed threshold $\delta$ (beam width) of the best path

- Given active nodes from the last time-step:
  - Examine nodes in the current time-step ...
  - ... that are reachable from active nodes in the previous time-step
  - Get active nodes for the current time-step by only retaining nodes with hypotheses that score close to the score of the best hypothesis
Beam search

- Beam search at each node keeps only hypotheses with scores that fall within a threshold of the current best hypothesis.

- Hypotheses with \( Q(t, s) < \delta \cdot \max Q(t, s') \) are pruned.

  Here, \( \delta \) controls the beam width.

- Search errors could occur if the most probable hypothesis gets pruned.

- Trade-off between balancing search errors and speeding up decoding.
Static and dynamic networks

• What we’ve seen so far: Static decoding graph
  • $H \circ C \circ L \circ G$
  • Determinize/minimize to make this graph more compact

• Another approach: Dynamic graph expansion
  • Dynamically build the graph with active states on the fly
  • Do on-the-fly composition with the language model $G$
  • $(H \circ C \circ L) \circ G$
Multi-pass search

- Some models are too expensive to implement in first-pass decoding (e.g. RNN-based LMs)

- First-pass decoding: Use simpler model (e.g. Ngram LMs)
  - to find most probable word sequences
  - and represent as a word lattice or an N-best list

- Rescore first-pass hypotheses using complex model to find the best word sequence
Multi-pass decoding with N-best lists

- Simple algorithm: Modify the Viterbi algorithm to return the N-best word sequences for a given speech input.
Multi-pass decoding with N-best lists

- Simple algorithm: Modify the Viterbi algorithm to return the N-best word sequences for a given speech input

<table>
<thead>
<tr>
<th>Rank</th>
<th>Path</th>
<th>AM logprob</th>
<th>LM logprob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>it’s an area that’s naturally sort of mysterious</td>
<td>-7193.53</td>
<td>-20.25</td>
</tr>
<tr>
<td>2.</td>
<td>that’s an area that’s naturally sort of mysterious</td>
<td>-7192.28</td>
<td>-21.11</td>
</tr>
<tr>
<td>3.</td>
<td>it’s an area that’s not really sort of mysterious</td>
<td>-7221.68</td>
<td>-18.91</td>
</tr>
<tr>
<td>4.</td>
<td>that scenario that’s naturally sort of mysterious</td>
<td>-7189.19</td>
<td>-22.08</td>
</tr>
<tr>
<td>5.</td>
<td>there’s an area that’s naturally sort of mysterious</td>
<td>-7198.35</td>
<td>-21.34</td>
</tr>
<tr>
<td>6.</td>
<td>that’s an area that’s not really sort of mysterious</td>
<td>-7220.44</td>
<td>-19.77</td>
</tr>
<tr>
<td>7.</td>
<td>the scenario that’s naturally sort of mysterious</td>
<td>-7205.42</td>
<td>-21.50</td>
</tr>
<tr>
<td>8.</td>
<td>so it’s an area that’s naturally sort of mysterious</td>
<td>-7195.92</td>
<td>-21.71</td>
</tr>
<tr>
<td>9.</td>
<td>that scenario that’s not really sort of mysterious</td>
<td>-7217.34</td>
<td>-20.70</td>
</tr>
<tr>
<td>10.</td>
<td>there’s an area that’s not really sort of mysterious</td>
<td>-7226.51</td>
<td>-20.01</td>
</tr>
</tbody>
</table>

- N-best lists aren’t as diverse as we’d like. And, not enough information in N-best lists to effectively use other knowledge sources

---

Image from [JM]: Jurafsky & Martin, SLP 2nd edition, Chapter 10
Multi-pass decoding with lattices

- ASR lattice: Weighted automata/directed graph representing alternate word hypotheses from an ASR system

so, it’s

it’s

there’s

that’s

that

an area that’s naturally sort of mysterious

the scenario

not really
Multi-pass decoding with lattices

- *Confusion networks/sausages*: Lattices that show competing/confusable words and can be used to compute posterior probabilities at the word level.