Speech Recognition with Segmental Conditional Random Fields

JHU Summer workshop
June 18, 2010
Patrick Nguyen
(panguyen@microsoft.com)

Agenda

• What is speech recognition?

• State-of-the-art
  – Markov Random Fields
  – Hidden Markov Models
  – (Hidden) Conditional Random Fields
  – Semi-Markov Random Fields

• Segmental Conditional Random Fields
  – Model
  – Parameter estimation

Agenda (2)

• Feature design with Segmental Conditional Random Fields
  – Problem Statement
  – Model-based features
  – Detection-Based features
    • Template-based features
    • Existence
    • Expectation
    • Levenshtein

Speech recognition

• Sometimes called Automatic Speech Recognition (ASR) or Speech-To-Text (STT)

• Identify what is being said, by machines

Applications of ASR

• Query by voice (web search)
• Medical Transcriptions
• Audio Indexing
• Voicemail transcription
• Downstream processing
  – Speech-to-speech translation
  – Topic detection and tracking

Speech Recognition

• A sequence transduction problem:
• Sequence of Audio

\[ X = \]

• Sequence of Words

\[ Y = “Nineteenth century” \]

http://en.wikipedia.org/wiki/Spectrogram
Warping and segmentation

- Audio can have different length
- There can be multiple words
- Introduce a new random variable: segmentation
- A segmentation assigns one audio “frame” to one word (without specifying identity)

Segmentation assigns audio to words

- Input is $x = \{x_1, x_2, \ldots, x_T\}$
- Segmentation is $q = \{q_1, q_2, \ldots, q_T\}$ = \{word 1, word 2, \ldots\}
- Output is $y = \{y_1, y_2, \ldots, y_T\}$ = \{19th, 19th, cent., cent.\}

Markov Random Fields

- A log-linear model in the sense defined in the first part of the lecture
- We are given segmented audio $(x, q)$
- We use a log-linear generative model:
  $$\log p(x, q, y) \propto \pi(x, q, y)$$
- We apply a Markov assumption, at each time $t$
  $$\log p(x, q, y) \propto \sum_t \pi(x_t, q_t, y_t)$$

Linear Chain MRF

- Graph structure for “dynamic” networks (ie time sequences)
- Unrolled: \(y_1\) \(y_2\)

Hidden Markov Models

- We don’t know the segmentation $q$
  $$\log p(x, q, y) \propto \pi(x, q, y)$$
- We marginalize over it
  $$p(x, y) = \sum_q p(x, q, y)$$
- This is akin to a mixture model
  $$p(x, y) = \frac{\sum_q \exp \pi(x, q, y)}{\int dx \sum_{y'} \exp \pi(x', q', y')}$$

Exercise

- Laplace distribution:
  $$\log p(z) \propto \beta |z|, z \in \mathbb{R}$$
- Is this a log-linear? Features?
- Computing the partition function
  $$Z(z) = \int dx \exp[\beta |z|]$$
- Add square features. Partition?
Conditional Random Fields

- Partition function must be computed analytically. (Remember the debate about direct vs generative?)
- Take the Markov Random Field, but use a conditional model instead
  \[ \log p(y, q|x) \propto \pi(y, q|x) \]
- Notice that segmentation becomes an output variable
- Decompose, per time, as previously

Hidden Conditional Random Fields

- The segmentation is hidden (unobserved)
  \[ \log p(y, q|x) \propto \pi(y, q|x) \]
- Marginalize over it
  \[ p(y|x) = \sum_q p(y, q|x) \]
- Notice how the partition becomes easier
  \[ p(y|x) = \frac{\sum_q \exp[\pi(y, q|x)]}{\sum_{y',q'} \exp[\pi(y', q'|x)]} \]

Exercise

- What are the differences between Markov random fields and conditional random fields?
  - Computationally
  - Analytically
- Can you think of applications where one would prefer using Markov vs conditional random fields?

A second look at the Markov product

- It is too strong
  \[ \log p(x, q, y) \propto \sum_t \pi(x_t, q_t, y_t) \]
- Problem topology dictates longer-term effects
- One can distinguish “trajectories”

Agenda

- What is speech recognition?
- State-of-the-art
  - Markov Random Fields
  - Hidden Markov Models
  - (Hidden) Conditional Random Fields
  - Semi-Markov Random Fields
- Segmental Conditional Random Fields
  - Model
  - Parameter estimation

Relax the Markov assumption

- Instead of segmenting audio time, segment by word
  \[ \log p(x, q, y) \propto \sum_t \pi(x_t, q_t, y_t) \]
- Time-based:
  \[ \log p(y, q|x) \propto \sum_{j=1,2} \pi(y_j, \{x_i\}_{\text{start}(q_j)}) \]
- Segmental:
  \[ \log p(y, q|x) \propto \sum_{j=1,2} \pi(y_j, \{x_i\}_{\text{end}(q_j)}) \]
Segmental Models

Exercise

• What is the implied duration model in an HMM?

\[ Y_1 = \begin{array}{ccc} 1 & 2 & 3 \end{array} \]

– Duration model: probability of staying at, say, state 2

• How to implement arbitrary duration models?

Time-Synchronous Markov Models

Segmental Conditional Random Fields

Exercise

• Flat direct models
• How about just using the following?
  \[ \log p(y|x) \propto \pi(y|x) \]
• Can we do it?
• What does it buy us?

From Exponential to Segmental Conditional Random Fields

• Step 0: start with the log-linear model
  \[ \log p(x, y) \propto \pi(x, y) \]
• Step 1: define a segmentation among words
  \[ \log p(y, q, x) \propto \sum_{j=1,2} \pi(y_j, x_{\text{end}(q_j)}) \]
• Step 2: condition on the input (direct model)
  \[ \log p(y, q|x) \propto \pi(y, q|x) \]
• Step 3: sum over all possible segmentations
  \[ p(y|x) = \sum_q p(y, q|x) \]
Segmental Conditional Random Fields

- States represent whole words (not phonemes)
- Observations blocked into groups corresponding to words. Observations typically detection events.
- Log-linear model relates words to observations

\[ p(y|x) = \frac{\sum_y \prod_{j} \pi \left( y_j, q_j \left| x_{\text{end}(q_j)}, x_{\text{start}(q_j)} \right. \right)}{\sum_{y'} \prod_{j} \pi \left( y_j', q_j \left| x_{\text{end}(q_j)}, x_{\text{start}(q_j)} \right. \right)} \]

Parameter estimation

- We are given a database of labeled examples \((x, y)\)
- We need to draw a good hyperplane
- We optimize the log-likelihood w.r.t. the parameter vector

Exercise

- Write down the gradient for the unstructured log-linear case and put it in proper form
- Write down the gradient for the semi-Markov conditional random field and put it in proper form
  - Can you get to this result faster?
- Write down the gradient for the hidden case and put it in proper form

So far

- We have shown how Markov Random Fields, Hidden Markov Models, (Hidden) Conditional Random Fields, and Segmental conditional random fields relate to each other
- Again, the magic is in the features
  - More than frame similarity
  - Long-term effects can be modeled
  - Agnostic feature combination

The features: agenda

- Problem Statement
- Model-based features
  - Fisher Trick
- Detection-based recognition
  - Template-based features
  - Existence Features
  - Expectation Features
  - Edit Distance Features

Problem Statement

- Given a word identity and a chunk of audio
  \((w = \text{“nineteenth”}, \{x\} = \text{image})\)
- Find a fixed-dimension vector which will be consistent
- The best feature: 1 if \(\{x\}\) sounds like nineteenth, -1 otherwise
Exercise

• Can you define a feature based on x only, without reference to the word?

• Discuss why it would make a difference or not

• How about a feature on Y only? Why?

Model-based scores

• A researcher has a measure of goodness

\[ s(w | \{x\}_1^n; \theta) \in \mathbb{R} \]

• The model family \( s() \) is parameterized by \( \theta \).

• For instance, an HMM score

\[ s(w | \{x\}_1^n; \theta) = \log p_{hmm}(\{x\}_1^n|w; \theta) \]

• This allows us to turn a variable-length input into a single scalar

The Fisher trick

• How to get more than a single number?

• The parameters are fixed-dimension!

\[ \nabla_\theta \log p_{hmm}(\{x\}_1^n | w; \theta) \in \mathbb{R}^{\text{dim}(\theta)} \]

• Variant: can use the mode of the distribution

\[ \text{argmax}_\theta \log p_{hmm}(\{x\}_1^n | w; \theta) \in \mathbb{R}^{\text{dim}(\theta)} \]

• Obviates a need for joint training of \((\theta, \lambda)\)

Generative models and likelihood ratios

• Generative story

\[ q(x, y) \]

• Direct Model

\[ f(y|x) = \log q(x,y) \]

• Bayes Rule

\[ q(y|x) = \frac{q(x,y)}{\sum_y q(x,y')} \]

• Partition

\[ p(y|x) = \frac{\exp[f(x,y)]}{\sum_y \exp[f(x,y)']} \]

• Likelihood Ratio test

\[ \frac{q(y|x)}{q(y'|x)} > 1? \]

• Decision Boundary

\[ \pi(y|x) > \pi(y'|x) \]

The features: agenda

• Problem Statement

• Model-based features

  – Fisher Trick

• Detection-based recognition

  – Template-based features

  – Existence Features

  – Expectation Features

  – Edit Distance Features

Detection-based ASR

• Audio events occur in the utterances

• It is possible to “spot” them, the simple instance is a word spotter

Observations blocked into groups corresponding to words. Observations typically detection events.
That rings a bell...

- Template features are exemplar-based
- Nearest-neighbor

<table>
<thead>
<tr>
<th>Training DB</th>
<th>Test sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>nineteenth</td>
<td>nineteenth</td>
</tr>
<tr>
<td>twentieth</td>
<td>nineteenth</td>
</tr>
</tbody>
</table>

Templates

- Embedding defined by $d(x_{train}, x_{test})$

Exercise

- Features based on templates $v$ for $(w, \{x\})$

<table>
<thead>
<tr>
<th>Description</th>
<th>Cardinality (# features)</th>
<th>Density (#features per example)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w=$nineteenth and closest word is $w=$twentieth</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closest example of $w$ in the database is $x$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closest example $v=w$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ratio of distance between closest word $v$ when $v=w$ and $v\neq w$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closest example $w=v$, and number of letters in $w$ is $K$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Breaking down the problem

- Word-level is too sparse
- Audio-level is continuous and unwieldy

Forget about the audio!

- Forget about what happens at the sentence level

A tale of two Markovs

- Hidden Markov Model vs syllable Markov

<table>
<thead>
<tr>
<th>Word level</th>
<th>Syllable level</th>
<th>Audio level</th>
</tr>
</thead>
<tbody>
<tr>
<td>nineteenth</td>
<td>nineteenth</td>
<td>time</td>
</tr>
<tr>
<td>century</td>
<td>century</td>
<td></td>
</tr>
</tbody>
</table>
Subword detections

- We have syllable-like detections. Now what?

existence features

- Also informally called Pavlov features
- Do not take order into account ("bag" model)
- Do not generalize across words
- No prior knowledge, learn any pronunciation

expectation features

- We know what to expect for "century"
- Order not taken into account
- Generalize across words

Levenshtein features

- Take order into account

Features: conclusion

- Score-based, Fisher kernel trick
- Detection-based
  - Template-based
  - Break down the problem (syllable Markov)
    - Knowledge-free model (existence)
    - Bag-to-bag (expectation)
    - Sequence transduction (Levenshtein)
- It is left to your imagination...

Recap: A taxonomy of models
Conclusion

• Segmental conditional random fields
  – Direct model
  – Get rid of frame-level Markov assumption
  – Retain Markov at the word level
• Features
  – Score-based
  – Detection-based
    • Template-based
    • Existence, expectation, and Levenshtein