Unsupervised Word Discovery from Speech using Bayesian Hierarchical Models

Oliver Walter and Reinhold Häb-Umbach

12.09.2016
Modelling of Speech

- Sequential
- Hierarchical
  - Audio
  - Phonemes
  - Words
- Increasing abstraction with each layer
Unsupervised Learning

- Labeled speech data needed to train recognition system
  - Transcription in terms of the spoken words
  - Lexicon to map words to phoneme sequences

- Labeled data not always available
  - Costly to obtain
  - So called low resource languages
  - Zero resource
    - Languages without written form
    - Endangered languages
  - Speech changes constantly

- Learn a language like a child
Speech recognition

- Components of traditional speech recognition system
  - Acoustic model
  - Lexicon
  - Language model
Generative model of speech

- Language model generates sequence of words
- Lexicon maps words to phoneme sequences
- Acoustic model emits speech features
• Language model generates sequence of words
• Lexicon maps words to phoneme sequences
• Acoustic model emits speech features
• For the remainder of the presentation we assume an already trained acoustic model
  - Represent input as graph of possible phoneme sequences
  - Learn language model and lexicon
Unsupervised Word Discovery

- **Goal**: Segment the Phone lattice into words
- **Approach**:
  - Exploit that the sequence of units is more predictable within words than at word boundaries
  - Iteratively alternate between word segmentation and language model estimation
  - Use Nested Hierarchical Pitman-Yor Language Model

Phone Lattice → Word Discovery → Word sequence

"I am"
HPYLM

- Bayesian interpolated n-gram Language Model with back off:

\[
Pr(w|u) = \frac{c_{uw} - d_u t_{uw}}{\theta_{|u|} + c_{u.}} + \frac{\theta_{|u|} + d_u t_{u.}}{\theta_{|u|} + c_{u.}} Pr(w|\pi(u))
\]

- Hierarchy of Pitman-Yor processes
- Base measure \(G_0^{(w)}\) of unigram: zero-gram (uniform distribution)
- Inference performed with Gibbs Sampling
Idea: Words consist of units (Characters, Phones, Acoustic Units)

Model unit n-gram probabilities with another HPYLM

Replace word LM zerogram with resp. unit sequence likelihood

Result: Able to learn words from unit sequences
Finite State Machine based Implementation

- Input sequence:
  \[
  0 \xrightarrow{I, I} 1 \xrightarrow{a, a} 2 \xrightarrow{m, m} 3 \xrightarrow{EOS, EOS} 4
  \]

- Lexicon:
Segmentations and Language Model
Weighted Segmentation

- Weighted possible Paths

- Learning is done using Gibs Sampling
  - Forward filtering / Backward sampling

- Iterate between sampling of possible Path and Language model Estimation
Experimental Results

• Use English Recognizer to decode Xitsonga speech
  ▪ ZeroSpeech 2015 challenge Data
  ▪ Type (Lexicon) F-Score
  ▪ Token (Segmentation) F-Score

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Xitsonga</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Type</td>
<td>Token</td>
</tr>
<tr>
<td>Type</td>
<td>24.0</td>
<td>5.1</td>
</tr>
<tr>
<td>Token</td>
<td>25.0</td>
<td>3.7</td>
</tr>
</tbody>
</table>

Unsupervised Word Discovery

O. Walter
• Learning from untranscribed data is possible
  ▪ Learn: Language model and Lexicon
  ▪ But: Noisy input deteriorates the results

• Still open problems:
  ▪ How to deal with recognition errors
  ▪ Consider pronunciation variants
  ▪ Integrate Acoustic model learning
Vielen Dank für ihre Aufmerksamkeit

Fragen?

Oliver Walter
Universität Paderborn
Fachgebiet Nachrichtentechnik

walter@nt.uni-paderborn.de
nt.uni-paderborn.de