Discriminative Rescoring using Landmarks

Katrin Kirchhoff
Rationale

- WS04 approach: lattice/N-best list rescoring instead of first-pass recognition
- baseline system already provides high-quality hypotheses
  - 1-best error rate from N-best lists: 24.4% (RT-03 dev set)
  - oracle error rate: 16.2%
- ⇒ use landmark detection only where necessary, to correct errors made by baseline recognition system
Ref: that cannot be that hard to sneak onto an airplane
Hyp: they can be a that hard to speak on an airplane

- Identify word confusions
- Determine most important acoustic-phonetic features that distinguish confusable words
- Use high-accuracy landmark detectors to determine probability of those features
- Use resulting output for rescoring
Identifying Confusable Hypotheses

- Use existing alignment algorithms for converting lattices into confusion networks (Mangu, Brill & Stolcke 2000)

- Hypotheses ranked by posterior probability

- Generated from n-best lists without 4-gram or pronunciation model scores (⇒ higher WER compared to lattices)

- Multi-words (“I_don’t_know”) were split prior to generating confusion networks
Identifying Confusable Hypotheses

- How much can be gained from fixing confusions?
- Baseline error rate: 25.8%
- Oracle error rates when selecting correct word from confusion set:

<table>
<thead>
<tr>
<th># hypotheses to select from</th>
<th>Including homophones</th>
<th>Not including homophones</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>23.9%</td>
<td>23.9%</td>
</tr>
<tr>
<td>3</td>
<td>23.0%</td>
<td>23.0%</td>
</tr>
<tr>
<td>4</td>
<td>22.4%</td>
<td>22.5%</td>
</tr>
<tr>
<td>5</td>
<td>22.0%</td>
<td>22.1%</td>
</tr>
</tbody>
</table>
Selecting relevant landmarks

- Not all landmarks are equally relevant for distinguishing between competing word hypotheses (e.g. vowel features irrelevant for sneak vs. speak)
- Using all available landmarks might deteriorate performance when irrelevant landmarks have weak scores (but: redundancy might be useful)

Automatic selection algorithm
- Should optimally distinguish set of confusable words (discriminative)
- Should rank landmark features according to their relevance for distinguishing words (i.e. output should be interpretable in phonetic terms)
- Should be extendable to features beyond landmarks
Selecting relevant landmarks

- Words are associated with variable-length sequences of landmarks

- Options for selection:
  - Use a discriminative sequence model: Conditional Random Fields
  - Convert words to fixed-length representation and use standard discriminative classifier, e.g. maximum-entropy model, MLP, SVM
  - Related work (e.g. by Byrne, Gales): Fisher score spaces + SVMs
  - Here: phonetic vector space + maxent model (interpretable)
Maximum-Entropy Landmark Selection

- Convert each word in confusion set into fixed-length landmark-based representation using idea from information retrieval:
- Vector space consisting of binary relations between two landmarks
  - Manner landmarks: precedence, e.g. $V < \text{Son. Cons.}$
  - Manner & place features: overlap, e.g. $V o +\text{high}$
  - preserves basic temporal information
- Words represented as frequency entries in feature vector
- Not all possible relations are used (phonotactict constraints, place features detected dependent on manner landmarks)
- Dimensionality of feature space: 40 - 60
- Word entries derived from phone representation plus pronunciation rules
## Vector-space word representation

<table>
<thead>
<tr>
<th></th>
<th>Start &lt; Fric</th>
<th>Fric &lt; Stop</th>
<th>Fric &lt; Son</th>
<th>Fric &lt; Vowel</th>
<th>Stop &lt; Vowel</th>
<th>Vowel o high</th>
<th>Vowel o front</th>
<th>Fric o strident</th>
</tr>
</thead>
<tbody>
<tr>
<td>speak</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>sneak</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>seek</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>he</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>she</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>steak</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>....</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Maximum-entropy discrimination

- Use maxent classifier

\[ P(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_k \lambda_k f_k(x, y)\right) \]

- Here: \( y \) = words, \( x \) = acoustics, \( f \) = landmark relationships

- Why maxent classifier?
  - Discriminative classifier
  - Possibly large set of confusable words
  - Later addition of non-binary features

- Training: ideally on real landmark detection output
- Here: on entries from lexicon (includes pronunciation variants)
Maximum-entropy discrimination

- Example: sneak vs. speak

<table>
<thead>
<tr>
<th>sneak</th>
<th>speak</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC ○ +blade  2.47</td>
<td>SC ○ +blade -2.47</td>
</tr>
<tr>
<td>FR &lt; SC      2.47</td>
<td>FR &lt; SC      -2.47</td>
</tr>
<tr>
<td>FR &lt; SIL     -2.11</td>
<td>FR &lt; SIL     2.11</td>
</tr>
<tr>
<td>SIL &lt; ST     -1.75</td>
<td>SIL &lt; ST     1.75</td>
</tr>
</tbody>
</table>

- Different model is trained for each confusion set ⇒ landmarks can have different weights in different contexts
**Landmark queries**

- Select N landmarks with highest weights
- Could scan bottom-up landmark detection output for presence of relevant landmarks
- Better: use knowledge of relevant landmarks in top-down fashion (suggestion by Jim)
- Ask landmark detection module to produce scores for selected landmarks within word boundaries given by baseline system
- Example:
  
  **Confusion networks**
  
  sneak 1.70 1.99 SC ○ +blade ?
  
  sneak 1.70 1.99 SC ○ +blade 0.75 0.56
  
  **Landmark detectors**
Rescoring

- Landmark detection scores: weighted combination of manner and place probabilities
- Normalization across words confusion set & combination (weighted sum or product) with original probability distribution given by baseline system
- Or: use as additional features in a maxent model for rescoring confusion networks (more on this in Kemal’s talk)
- Only applied to confusion sets that contain phonetically distinguishable hypotheses (e.g. not by - buy, to-two-too…)
- Only applied to sets where words do not compete with
Experiments

- Varied number of landmark scores to use (1, 2, ... all)
- Top 2 vs. 3 vs. all hypotheses in confusion network
- Use of entire word time interval vs. restricting time intervals to approximate location of landmarks
- Changes in feature-space representation of lexicon
- Various score combination methods for rescoring
- Initial experiments on learning lexicon representation from data (for most frequent words)
Results

RT-03 dev set, 35497 Words, 2930 Segments, 36 Speakers
(Switchboard and Fisher data)

<table>
<thead>
<tr>
<th></th>
<th>WER</th>
<th>Insertions</th>
<th>Deletions</th>
<th>Substitutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>25.8%</td>
<td>2.6% (982)</td>
<td>9.2% (3526)</td>
<td>14.1% (5417)</td>
</tr>
<tr>
<td>Rescored</td>
<td>25.8%</td>
<td>2.6% (984)</td>
<td>9.2% (3524)</td>
<td>14.1% (5408)</td>
</tr>
</tbody>
</table>

Rescored: product combination of old and new prob. distributions, weights 0.8 (old), 0.2 (new)

- Correct/incorrect decision changed in about 8% of all cases
- Slightly higher number of fixed errors vs. new errors
Analysis

- **When does it work?**
  - Detectors give high probability for correct distinguishing feature
    - *mean* (correct) vs. *me* (false)  \( V < +\text{nasal} \ 0.76 \)

- **When does it not work?**
  - Problems in lexicon representation
    - *once* (correct) vs. *what* (false):  \( \text{Sil} \circ +\text{blade} \ 0.87 \)
    - *can’t* [kæt] (correct) vs *cat* (false):  \( \text{SC} \circ +\text{nasal} \ 0.26 \)
  - Landmark detectors are confident but wrong
    - *like* (correct) vs. *liked* (false):  \( \text{Sil} \circ +\text{blade} \ 0.95 \)
Analysis

- Incorrect landmark scores often due to word boundary effects, e.g.:

- Word boundaries given by baseline system may exclude relevant landmarks or include parts of neighbouring words
Conclusions

- Positive trend but not strong enough yet to decrease word error rate
- Method can be used with classifiers other than landmark detectors (e.g. high-accuracy triphone classifiers)
- Can serve as diagnostic tool (statistics of score queries ⇒ relevance of phonetic distinctions for improving word error rate on given corpus)
- Provides information about which detector outputs are likely to help vs. likely to cause errors ⇒ feedback for developing landmark classifiers
- Advantage: little computational effort, fast
Future Directions

- Improve landmark detectors (e.g. specialized detectors for word endings)
- Select landmarks that are not only discriminative but can also be detected robustly
- Learn lexical representation from data (takes into account errors made by detectors)
- Change lexical representation to include more temporal constraints
- Try approach with other classifiers
- Allow flexible word segmentation