New Developments in Spoken Query Transcription

Jonathan Mamou, IBM
Bhuvana Ramabhadran, IBM
Abhinav Sethy, IBM

Afeka Speech Processing Conference
June 2012
Spoken queries are a natural medium for searching the Mobile Web
   – Especially where typing on a keyboard is not practical or inconvenient

The relevant hits should contain the query terms, but not necessarily adjacent and ordered as in the query

If the terms in an ASR hypothesis co-occur frequently in documents of the training data, the hypothesis is more likely to be correct
Challenges in Language Modeling for Voice Search

- Modeling the distribution of words in spoken queries offers different challenges compared to the more conventional speech applications like Voicemail to text or dictation.

- Most queries are short
  - 1-6 words, with median length of 2 words
  - “Sheen’s korner”, “My Meebas”

- Weak word proximity and syntax dependencies between the query terms
  - Most queries are “google style” conjunction of terms
  - “traverse city Michigan events”, “address morimoto nyc”

- About 71% of search queries contain named entities
Language Modeling for Voice Search

- A Language Model (LM) is trained using **textual query logs** and **web documents**
  - Not enough query logs to build LM
  - We can crawl the web but queries do not follow the same syntax and grammar as natural language (web documents)

- Explore techniques to leverage the potentially infinite document data that can be crawled from the web as a source to augment limited sources of query data
N-Best Rescoring

- Offline collection of **co-occurrence** information from different data sources
- **Named-entity extraction** from web textual data
  - N-gram LM
  - Generate permutations in order to capture co-occurrence information
- Given n-best hypotheses of a voice query, lookup co-occurrence statistics
  - Different co-occurrence levels and scoring functions
  - Interpolation of co-occurrence based scores with acoustic and n-gram language model scores
Scoring

- **Document frequency**
  The number of documents matching the hypothesis in the corpus

\[
DF(h, D) = \frac{|\{d : h \in d\}|}{|D|}
\]

- **Term Frequency Inverse Document Frequency**
  The sum of the TFIDF of the hypothesis terms over the top n matching documents

\[
hypN\text{orm}(h) \times \sum_{\{d : h \in d\}_n} \text{coord}(h, d) \\
\times \text{norm}(d) \times \sum_{i=0}^{k} \left( \sqrt{tf(t_i, d) \times idf(t_i)^2} \right)
\]

Coord is the ratio of the number of hypothesis terms matched in d over the total number of hypothesis terms.
Experiments

- Lucene based implementation
  - Full indexing of the training data
  - Not necessary to do a full search, retrieve only co-occurrence statistics

- **Test data**
  - 43K utterances
  - 160K words

- **Data Collection** for training co-occurrence models:
  - Street address corpus
  - Directory assistance data
  - Other in-house data
  - Stock names
  - Unsupervised transcripts for spoken web queries
  - List of common URLs
  - Data collected from the web

<table>
<thead>
<tr>
<th>corpus</th>
<th>corpus size (G)</th>
<th>index size (G)</th>
<th>posting list length</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADDRESS</td>
<td>6</td>
<td>20</td>
<td>15378</td>
</tr>
<tr>
<td>DA</td>
<td>0.5</td>
<td>2.5</td>
<td>1689</td>
</tr>
<tr>
<td>OTHER</td>
<td>0.05</td>
<td>0.23</td>
<td>219</td>
</tr>
<tr>
<td>STOCK</td>
<td>0.05</td>
<td>0.25</td>
<td>3902</td>
</tr>
<tr>
<td>UNSUP</td>
<td>0.01</td>
<td>0.07</td>
<td>96</td>
</tr>
<tr>
<td>URL</td>
<td>0.1</td>
<td>0.61</td>
<td>157</td>
</tr>
<tr>
<td>WEBDT</td>
<td>3.3</td>
<td>16.8</td>
<td>287</td>
</tr>
</tbody>
</table>
Co-Occurrence Based Scoring

- Weights of the Interpolation of co-occurrence based scores with acoustic and n-gram language model scores are determined on a heldout set optimizing for word error rate (WER)

<table>
<thead>
<tr>
<th>method</th>
<th>WER</th>
<th>relative reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>unpruned LM</td>
<td>18.9</td>
<td>0</td>
</tr>
<tr>
<td>co-occurrence</td>
<td>18.76</td>
<td>0.7</td>
</tr>
<tr>
<td>n-gram NE</td>
<td>18.58</td>
<td>1.7</td>
</tr>
<tr>
<td>co-occurrence NE</td>
<td>18.53</td>
<td>2</td>
</tr>
<tr>
<td>combination</td>
<td>18.44</td>
<td>2.4</td>
</tr>
</tbody>
</table>
Analysis and Conclusion

- The gain is focused on short queries
  - These queries are difficult to model using more conventional language modeling techniques which rely on the structure of the word sequence
- Co-occurrence based approach and Named-Entity recognition used in order to improve the performance of the voice search system
- The co-occurrence models give a 2.4% relative accuracy improvement over a state of the art system