VOCABULARY INDEPENDENT SPOKEN TERM DETECTION

Author: Jonathan Mamou, Bhuvana Ramabhadran, Olivier Siohan
Presenter: Chiahua
OUTLINE

1. INTRODUCTION
2. AUTOMATIC SPEECH RECOGNITION SYSTEM
3. RETRIEVAL MODEL
4. EXPERIMENTS
5. RELATED WORK
6. CONCLUSIONS
INTRODUCTION

• the effectiveness of retrieval mostly depends on the accuracy of automatic speech recognition (ASR) systems of the transcripts.

• In many applications the OOV rate may get worse over time unless the recognizer’s vocabulary is periodically updated.

• Another approach consists of converting the speech to phonetic transcripts and representing the query as a sequence of phones.

• The retrieval is based on searching the sequence of phones representing the query in the phonetic transcripts. The main drawback of this approach is the inherent high error rate of the transcripts.

• Therefore, such approach cannot be an alternative to word transcripts, especially for in-vocabulary (IV) query terms that are part of the vocabulary of the ASR system.
A solution would be to combine the two different approaches presented above:

1. we index both word transcripts and phonetic transcripts; during query processing, the information is retrieved from the word index for IV terms and from the phonetic index for OOV terms.

2. We would like to be able to process also hybrid queries, i.e., queries that include both IV and OOV terms. Consequently, we need to merge pieces of information retrieved from word index and phonetic index.
AUTOMATIC SPEECH RECOGNITION SYSTEM

• We use an ASR system for transcribing speech data. It works in speaker-independent model.

• For best recognition results, a speaker-independent acoustic model and a language model are trained in advance on data with similar characteristics.

• ASR generates lattices that can be considered as directed acyclic graphs.

• Each vertex in a lattice is associated with a timestamp and each edge \((u, v)\) is labeled with a word or phone hypothesis and its prior probability, which is the probability of the signal delimited by the timestamps of the vertices \(u\) and \(v\), given the hypothesis.

• The 1-best path transcript is obtained from the lattice using dynamic programming techniques.
AUTOMATIC SPEECH RECOGNITION SYSTEM

- Mangu et al. [18] and Hakkani-Tur et al. [13] propose a compact representation of a word lattice called word confusion network (WCN).

- As explained in [13], the three main steps for building a WCN from a word lattice are as follows:
  1. Compute the posterior probabilities for all edges in the word lattice.
  2. Extract a path from the word lattice (which can be the 1-best, the longest or any random path), and call it the pivot path of the alignment.
  3. Traverse the word lattice, and align all the transitions with the pivot, merging the transitions that correspond to the same word (or label) and occur in the same time interval by summing their posterior probabilities.

- The 1-best path of a WCN is obtained from the path containing the best hypotheses.
AUTOMATIC SPEECH RECOGNITION SYSTEM
The main problem with retrieving information from spoken data is the low accuracy of the transcription particularly on terms of interest such as named entities and content words.

Generally, the accuracy of a word transcript is characterized by its word error rate (WER).

There are three kinds of errors that can occur in a transcript:
1. Substitution
2. Deletion
3. Insertion
Spoken document detection task

- As stated in the STD 2006 evaluation plan [2], the task consists in finding all the exact matches of a specific query in a given corpus of speech data. A query is a phrase containing several words. The queries are text and not speech.
- found automatically by searching the manual transcripts using the following rule: the gap between adjacent words in a query must be less than 0.5 seconds in the corresponding speech.
Indexing

• We have used the same indexing process for WCN and phonetic transcripts. Each occurrence of a unit of indexing (word or phone) \( u \) in a transcript \( D \) is indexed with the following information:
  • the begin time \( t \) of the occurrence of \( u \),
  • the duration (為期) \( d \) of the occurrence of \( u \).

• In addition, for WCN indexing, we store
  • the confidence level of the occurrence of \( u \) at the time \( t \) that is evaluated by its posterior probability \( Pr(u|t, D) \),
  • the rank of the occurrence of \( u \) among the other hypotheses beginning at the same time \( t \), \( rank(u|t, D) \).

• Note that since the task is to find exact matches of the phrase queries, we have not filtered stopwords and the corpus is not stemmed before indexing.
Search

- we present our approach for accomplishing the STD task using the indices described above. The terms are extracted from the query. The vocabulary of the ASR system building word transcripts is given. Terms that are part of this vocabulary are IV terms;
- the other terms are OOV. For an IV query term, the posting list is extracted from the word index. For an OOV query term, the term is converted to a sequence of phones using a joint maximum entropy N-gram model [10]. For example, the term prosody.
Search

• The next step consists of merging the different posting lists according to the timestamp of the occurrences in order to create results matching the query.

1. we check that the words and phones appear in the right order according to their begin times.
2. we check that the gap in time between adjacent words and phones is “reasonable”.
   ➢ Conforming to the requirements of the STD evaluation, the distance in time between two adjacent query terms must be less than 0.5 seconds.
   ➢ For OOV search, we check that the distance in time between two adjacent phones of a query term is less that 0.2 seconds; this value has been determined empirically.
Search

- our indexing model allows to search for different types of queries:
  1. queries containing only IV terms using the word index.
  2. queries containing only OOV terms using the phonetic index.
  3. keyword queries containing both IV and OOV terms using the word index for IV terms and the phonetic index for OOV terms;
  4. phrase queries containing both IV and OOV terms;
Ranking

• Since IV terms and OOV terms are retrieved from two different indices, we propose two different functions for scoring an occurrence of a term; afterward, an aggregate score is assigned to the query based on the scores of the query terms.

• Because the task is term detection, we do not use a document frequency criterion for ranking the occurrences.
Ranking

• In vocabulary term ranking

\[ \text{score}(k, t, D) = Brank(k|t, D) \times \text{Pr}(k|t, D) \]

\[ 0 \leq \text{score}(k, t, D) \leq 1. \]

• Out of vocabulary term ranking

\[ \text{score}(k, t^{i}_0, D) = 1 - \frac{\sum_{i=1}^{l} 5 \times (t^{k}_i - (t^{k}_{i-1} + d^{k}_{i-1}))}{l} \]

• Combination

\[ \text{score}(Q, t_0, D) = \prod_{i=0}^{n} \text{score}(k, t_i, D)^{\gamma_n} \]
Experimental setup

- Our corpus consists of the evaluation set provided by NIST for the STD 2006 evaluation [1]. It includes three different source types in US English: three hours of broadcast news (BNEWS), three hours of conversational telephony speech (CTS) and two hours of conference room meetings (CONFMTG).
- We have processed the query set provided by NIST that includes 1100 queries.
- Each query is a phrase containing between one to five terms, common and rare terms, terms that are in the manual transcripts and those that are not.
- Testing and determination of empirical values have been achieved on another set of speech data and queries, the development set, also provided by NIST.
Experimental setup

- We have used the IBM research prototype ASR system, described in [26], for transcribing speech data.
- We have produced WCNs for the three different source types.
- 1-best phonetic transcripts were generated only for BNEWS and CTS, since CONFMTG phonetic transcripts have too low accuracy. We have adapted Juru [7], a full-text search library written in Java, to index the transcripts and to store the timestamps of the words and phones;
NIST for the 2006 STD evaluation

- term-weighted value (TWV)
  \[ TWV(\theta) = 1 - \text{average}_q\{P_{\text{miss}}(q, \theta) + \beta \times P_{FA}(q, \theta)\} \]
  \[ \beta = \frac{c}{V}(Pr_q^{-1} - 1). \]
  - \( \theta \) is the detection threshold.
  - \( \frac{c}{V} = 0.1 \)
  - \( Pr_q^{-1} = 10^{-4} \)
  - \( \beta = 999.9 \)

- Actual Term-Weighted Value (ATWV)
- Maximum Term-Weighted Value (MTWV).
NIST for the 2006 STD evaluation

• term-weighted value (TWV)

\[ TWV(\theta) = 1 - \text{average}_q \{ P_{\text{miss}}(q, \theta) + \beta \times P_{\text{FA}}(q, \theta) \} \]

\[ P_{\text{miss}}(q, \theta) = 1 - \frac{N_{\text{correct}}(q, \theta)}{N_{\text{true}}(q)} \]

\[ P_{\text{FA}}(q, \theta) = 1 - \frac{N_{\text{spurious}}(q, \theta)}{N_{\text{NT}}(q)} \]

- \(N_{\text{correct}}(q, \theta)\) is the number of correct detections of the query q with a score greater than or equal to \(\theta\).
- \(N_{\text{spurious}}(q, \theta)\) is the number of spurious detections of the query q with a score greater than or equal to \(\theta\).
- \(N_{\text{true}}(q)\) is the number of true occurrences of the query q in the corpus.
- \(N_{\text{NT}}(q)\) is the number of opportunities for incorrect detection of the query q in the corpus; it is the ”Non-Target” query trials.
  - \(N_{\text{NT}}(q) = T_{\text{speech}} - N_{\text{true}}(q)\)
• term-weighted value (TWV)
  \( \text{TWV}(\theta) = 1 - \text{average}_q\{P_{\text{miss}}(q, \theta) + \beta \times PFA(q, \theta)\} \)

• Actual Term-Weighted Value (ATWV)
  • ATWV is the "actual term-weighted value";
  • it is the detection value attained by the system as a result of the system output and the binary decision output for each putative occurrence.
  • It ranges from \(-\infty\) to +1.

• Maximum Term-Weighted Value (MTWV).
  • MTWV is the "maximum term-weighted value" over the range of all possible values of \(\theta\). It ranges from 0 to +1.
WER analysis

• WER is defined as follows:
  \[ \frac{S + D + I}{N} \times 100 \]

• substitution error rate (SUBR) is defined by
  \[ \frac{S}{S + D + I} \times 100 \]

<table>
<thead>
<tr>
<th>corpus</th>
<th>WER(%)</th>
<th>SUBR(%)</th>
<th>DELR(%)</th>
<th>INSR(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNEWS WCN</td>
<td>12.7</td>
<td>49</td>
<td>42</td>
<td>9</td>
</tr>
<tr>
<td>CTS WCN</td>
<td>19.6</td>
<td>51</td>
<td>38</td>
<td>11</td>
</tr>
<tr>
<td>CONFMTG WCN</td>
<td>47.4</td>
<td>47</td>
<td>49</td>
<td>3</td>
</tr>
</tbody>
</table>
Theta threshold

- We have determined empirically a detection threshold $\theta$ per source type and the hard decision of the occurrences having a score less than $\theta$ is set to false; false occurrences returned by the system are not considered as retrieved and therefore, are not used for computing ATWV, precision and recall.

<table>
<thead>
<tr>
<th></th>
<th>BNEWS</th>
<th>CTS</th>
<th>CONFMTG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.4</td>
<td>0.61</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Processing resource profile

- We report in Table 3 the processing resource profile. Concerning the index size, note that our index is compressed using IR index compression techniques. The indexing time includes both audio processing (generation of word and phonetic transcripts) and building of the searchable indices.

<table>
<thead>
<tr>
<th>Index size</th>
<th>0.3267 MB/HS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indexing time</td>
<td>7.5627 HP/HS</td>
</tr>
<tr>
<td>Index Memory Usage</td>
<td>1653.4297 MB</td>
</tr>
<tr>
<td>Search speed</td>
<td>0.0041 sec.P/HS</td>
</tr>
<tr>
<td>Search Memory Usage</td>
<td>269.1250 MB</td>
</tr>
</tbody>
</table>
### Retrieval measures

<table>
<thead>
<tr>
<th></th>
<th>measure</th>
<th>BNEWS</th>
<th>CTS</th>
<th>CONFMTG</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WCN Phonetic</strong></td>
<td>ATWV</td>
<td>0.8485</td>
<td>0.7392</td>
<td>0.2365</td>
</tr>
<tr>
<td></td>
<td>MTWV</td>
<td>0.8532</td>
<td>0.7408</td>
<td>0.2508</td>
</tr>
<tr>
<td></td>
<td>precision</td>
<td>0.94</td>
<td>0.90</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>recall</td>
<td>0.89</td>
<td>0.81</td>
<td>0.37</td>
</tr>
<tr>
<td><strong>1-best-WCN Phonetic</strong></td>
<td>ATWV</td>
<td>0.8279</td>
<td>0.7102</td>
<td>0.2381</td>
</tr>
<tr>
<td></td>
<td>MTWV</td>
<td>0.8319</td>
<td>0.7117</td>
<td>0.2512</td>
</tr>
<tr>
<td></td>
<td>precision</td>
<td>0.95</td>
<td>0.91</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>recall</td>
<td>0.84</td>
<td>0.75</td>
<td>0.37</td>
</tr>
</tbody>
</table>
Figure 2: DET curve for WCN phonetic approach.
Influence of the duration of the query on the retrieval performance

<table>
<thead>
<tr>
<th>quantile</th>
<th>0-33</th>
<th>33-66</th>
<th>66-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNEWS</td>
<td>ATWV</td>
<td>0.7655</td>
<td>0.8794</td>
</tr>
<tr>
<td></td>
<td>MTWV</td>
<td>0.7819</td>
<td>0.8914</td>
</tr>
<tr>
<td>CTS</td>
<td>ATWV</td>
<td>0.6545</td>
<td>0.8308</td>
</tr>
<tr>
<td></td>
<td>MTWV</td>
<td>0.6551</td>
<td>0.8727</td>
</tr>
<tr>
<td>CONFMTG</td>
<td>ATWV</td>
<td>0.1677</td>
<td>0.3493</td>
</tr>
<tr>
<td></td>
<td>MTWV</td>
<td>0.1955</td>
<td>0.4109</td>
</tr>
</tbody>
</table>

Table 5: ATWV, MTWV according to the duration of the query occurrences per source type.
# OOV vs. IV query processing

<table>
<thead>
<tr>
<th>index</th>
<th>word</th>
<th>phonetic</th>
<th>word and phonetic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>precision</td>
<td>recall</td>
<td>precision</td>
</tr>
<tr>
<td>IV queries</td>
<td>0.8</td>
<td>0.96</td>
<td>0.11</td>
</tr>
<tr>
<td>OOV queries</td>
<td>0</td>
<td>0</td>
<td>0.13</td>
</tr>
<tr>
<td>hybrid queries</td>
<td>0</td>
<td>0</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Table 6: Comparison of word and phonetic approach on IV and OOV queries
RELATED WORK

• In the past decade, the research efforts on spoken data retrieval have focused on extending classical IR techniques to spoken documents.

• An LVCSR system is used to transcribe the speech into 1-best path word transcripts. The transcripts are indexed as clean text: for each occurrence, its document, its word offset and additional information are stored in the index.

• A generic IR system over the text is used for word spotting and search as described by Brown et al. [6] and James [14]
RELATED WORK

• Singhal et al. [24, 25] propose to add some terms to the transcript in order to alleviate the retrieval failures due to ASR errors.

• From an IR perspective, a classical way to bring new terms is document expansion using a similar corpus.

• Their approach consists in using word lattices in order to determine which words returned by a document expansion algorithm should be added to the original transcript.

• The necessity to use a document expansion algorithm was justified by the fact that the word lattices they worked with, lack information about word probabilities.
RELATED WORK

• Mamou et al. [17] propose a model for spoken document retrieval using WCNs in order to improve the recall and the MAP of the search.

• However, the problem of queries containing OOV terms is not addressed.

• Popular approaches to deal with OOV queries are based on sub-words transcripts, where the sub-words are typically phones, syllables or word fragments (sequences of phones).

• The classical approach consists of using phonetic transcripts. The transcripts are indexed in the same manner as words in using classical text retrieval techniques.
Saraclar and Sproat in [22] show improvement in word spotting accuracy for both IV and OOV queries, using phonetic and word lattices, where a confidence measure of a word or a phone can be derived.

They propose three different retrieval strategies:

1. search both the word and the phonetic indices and unify the two different sets of results;
2. search the word index for IV queries, search the phonetic index for OOV queries;
3. search the word index and if no result is returned, search the phonetic index.

However, no strategy is proposed to deal with phrase queries containing both IV and OOV terms.
CONCLUSIONS

• we have presented a vocabulary independent model of indexing and search that combines both the approaches.
• The system can deal with all kinds of queries although the phrases that need to combine for the retrieval, information extracted from two different indices, a word index and a phonetic index.
• The scoring of OOV terms is based on the proximity (in time) between the different phones.
• The scoring of IV terms is based on information provided by the WCNs.
CONCLUSIONS

• We have shown an improvement in the retrieval performance when using all the WCN and not only the 1-best path and when using phonetic index for search of OOV query terms. This approach always outperforms the other approaches using only word index or phonetic index.

• As a future work, we will compare our model for OOV search on phonetic transcripts with a retrieval model based on the edit distance.