Leveraging Relevance Cues for Improved Spoken Document Retrieval

Pei-Ning Chen¹, Kuan-Yu Chen² and Berlin Chen¹

National Taiwan Normal University, Taiwan¹
Institute of Information Science, Academia Sinica, Taiwan²

2011/08/30 @ Interspeech 2011
Outline

• Introduction

• Language Modeling for Information Retrieval (IR)

• Various Relevance Models Proposed in this paper

• Experimental Results

• Conclusions and Future Work
Introduction

• Large volumes of multimedia associated with speech are now made available on the Internet
  – *Speech retrieval* provides a natural way for multimedia access
  – It has been extensively studied and practiced in the speech processing community
Introduction (cont.)

• Task Definition for speech retrieval
  – Robustly Index spoken documents with speech recognition techniques
  – Retrieve relevant spoken documents in response to a user query
    • Spoken Term Detection (STD)
      — Find “literally matched” spoken documents where all/most query terms should be present (much like Web search)
    • Spoken Document Retrieval (SDR)
      — Find spoken documents that are “topically related” to a given query
The fundamental problems facing SDR are generally three-fold

- First, a query is often only a vague expression of an underlying information need

- Word usage mismatch between a query and a spoken document even if they are topically related to each other

- The imperfect speech recognition transcript carries wrong information and thus deviates somewhat from representing the true theme of a spoken document
Language Modeling for IR

- LM approaches have been introduced to IR (and SDR), and demonstrated with good success

\[ P_{LM}(D|Q) = \frac{P(Q|D)P(D)}{P(Q)} \propto P(Q|D) \]

- The Kullback-Leibler (KL)-Divergence measure is another basic formulation of LM for IR

\[ KL(Q \parallel D) = \sum_{w \in V} P(w|Q) \log \frac{P(w|Q)}{P(w|D)} \propto - \sum_{w \in V} P(w|Q) \log P(w|D) \]

- A query is treated as a probabilistic model rather than simply an observation

- KL-divergence supports us to improve not only the document model but also the query model for better document ranking
Relevance Modeling (RM)

• In the conventional relevance modeling
  – Each query $Q$ is assumed to be associated with an unknown relevance class $R_Q$
  – Documents that are relevant to the information need expressed in the query are samples drawn from $R_Q$

• The document ranking problem can be reduced to determine the probability $P_{RM}(w \mid Q)$
  – The relevance model can be defined as the probability of the word selected from relevance documents

$$P_{RM}(w \mid Q) \propto \sum_{m=1}^{M} P(D_m)P(q_1,\ldots,q_L,w \mid D_m)$$
$$= \sum_{m=1}^{M} P(D_m)P(w \mid D_m)\prod_{l=1}^{L} P(q_l \mid D_m)$$
Incorporating Topical Information in RM

• Topic-based relevance model (TRM) makes a step forward by incorporating latent topic information into RM
  – As conventional topic models, the probability that a word occurs is estimated from a set of latent topics

\[
P_{TRM}(w | Q) \propto \sum_{m=1}^{M} \sum_{k=1}^{K} P(D_m) P(T_k | D_m) P(w | T_k) \prod_{l=1}^{L} P(q_l | T_k)
\]

• TRM has some assumptions and properties:
  – Relevant documents are assumed to share a set of pre-defined latent topic variables \( \{T_1, \ldots, T_K\} \)
  – When given a latent topic, words and documents are independent of each other
  – TRM assumes that the additional cues of how words are distributed across a set of latent topics can carry useful global topic structure for relevance modeling
Modeling Pairwise Word Association in RM

- RM and TRM be used to model the association between an entire query \( Q \) and a word \( w \).

- We propose Pairwise-based RM (PRM) to render the pairwise word association between a word in the query and any word.
  - It can be regarded as a kind of LM for translating words \( q_l \) in the query to \( w \):
    \[
    P_{PRM}(w | q_l) \propto \sum_{m=1}^{M} P(D_m)P(w | D_m)P(q_l | D_m)
    \]
    \[
    P_{PRM}(w | Q) = \frac{1}{L} \sum_{l=1}^{L} P_{PRM}(q_l, w | Q)
    \]
  - Again, a set of latent topics is introduced into PRM (denoted by TPRM) to describe the word-word co-occurrence relationships:
    \[
    P_{TPRM}(q_l, w | Q) \propto \sum_{m=1}^{M} \sum_{k=1}^{K} P(D_m)P(T_k | D_m)P(q_l | T_k)P(w | T_k)
    \]
In practice, the relevant documents are unknown in advance

- First-round retrieval with the “query-likelihood” LM approach is applied to obtain a set of top-ranked (pseudo-relevant) documents to approximate the relevance class
- Second-run retrieval with the “KL-divergence” measure is used to re-rank the spoken documents
Different Granularities of Index Features

- Word-level index features possess more semantic information than subword-level ones
  - Enhance the precision

- Subword-level index features are more robust against the open vocabulary problem and speech recognition errors
  - Enhance the recall

- Distinct syllable pairs occurring in the spoken document collection were then identified to form a vocabulary of syllable pairs for indexing
Incorporating Non-Relevance Information

• Further, in addition to using the relevance information, we also hypothesize that the non-relevant (low-ranked) documents can provide extra useful cues.

• For this idea to work, we attempt to estimate a non-relevance model $P(w | NR_Q)$ for each test query $Q$.
  – The non-relevance model can be estimated simply based on the ML criterion or be further optimized with the EM algorithm.
  – E-step:
    $$P(NR_Q | w) = \frac{\lambda \cdot P(w | NR_Q)}{\lambda \cdot P(w | NR_Q) + (1 - \lambda) \cdot P(w | BG)}$$
  – M-step:
    $$P(w \mid NR_Q) = \frac{\sum_{D' \in D_{Low}} c(w, D') \cdot P(NR_Q \mid w)}{\sum_w \sum_{D' \in D_{Low}} c(w, D') \cdot P(NR_Q \mid w)}$$
Incorporating Non-Relevance Information

• The similarity measure between query \( Q \) and any document \( D \) thus can be computed as follows:

\[
SIM(Q, D) = SIM_{RM}(Q, D) + \alpha \cdot KL(NR_Q \parallel D)
\]

- Relevance Information
- Penalty Factor
- Non-Relevance Information

• Note also that
  – Here we adopt an unsupervised way to estimate the non-relevance model
  – We intend to explore whether the relevance and non-relevance cues of a test query can conspire to enhance the SDR performance
Conclusions

• In this paper, we have investigated a relevance language modeling framework for SDR
  – The utility of the methods have also been validated by extensively comparisons with several widely used retrieval methods
  – The experimental results indeed demonstrate the applicability of our methods

• As to future work, we envisage two directions:
  – One is utilizing speech summarization techniques to help better estimate the query and document models
  – The other is training the query and document models in a lightly-supervised manner through the exploration of users’ click-through data
Thank you!

Questions?