Relevance Language Modeling for Speech Recognition and Related Applications

Berlin Chen (陳柏琳)
Department of Computer Science & Information Engineering
National Taiwan Normal University
2011/12/28
Outline

- Introduction
- Automatic Speech Recognition (ASR)
- Relevance Language modeling for ASR
- Related Tasks: Speech Retrieval and Summarization
- Conclusions
Information -> Knowledge -> Wisdom?

Information Overload

Creating content means more to manage.

The figure is adapted from the presentation slides of Prof. Ostendorf at Interspeech2009
Introduction (1/2)

- Communication and search are by far the most popular activities in our daily lives
  - **Human-Computer Interaction:** Speech is the most nature and convenient means of communication between humans, and between humans and machines
    - A spoken language interface could be more convenient than a visual interface on a small device
    - Provide "anytime" and "anywhere" access to information
  - **Multimedia Content Processing:** Already over half of the internet traffic consists of video data
    - Though visual cues are important for search, the associated spoken documents often provide a rich set of semantic descriptions (e.g., transcripts, speakers, emotions, and scenes) for the data
Introduction (1/2)

- Automatic speech recognition (ASR)
  - Transcribe the linguistic contents of speech utterances
  - Play a vital role in multimedia information retrieval, summarization and mining, as well as computer-assisted language learning (CALL), such as
    - Transcribing spoken queries and documents
    - Determine pronunciation accuracy and intelligibility
Automatic Speech Recognition (ASR)

- Decision Rule of ASR (Risk-Minimization Principle)

\[ W_{opt} = \arg \min_{W \in \mathcal{W}} \text{Risk} \left( W \mid O \right) \]

\[ = \arg \min_{W \in \mathcal{W}} \sum_{W' \in \mathcal{W}} \text{Loss} \left( W, W' \right) P(W' \mid O) \]

\[ \approx \arg \max_{W \in \mathcal{W}} P(W \mid O) \quad \text{Assumption of Using the “0-1” Loss Function} \]

\[ \approx \arg \max_{W \in \mathcal{W}} \frac{p(O \mid W) P(W)}{p(O)} \]

\[ = \arg \max_{W \in \mathcal{W}} p(O \mid W) P(W) \quad \text{Linguistic Decoding} \]

- The ASR problem is reduced to finding the most likely word sequence \(W\) in response to an input speech signal \(O\).
Speech Feature Extraction

- The raw speech waveform is passed through feature extraction to generate relatively compact feature vectors at a frame rate of around 100 Hz
  - Parameterization: an acoustic speech feature is a simple compact representation of speech and can be modeled by cepstral features such as the Mel-frequency cepstral coefficient (MFCC)

raw (perception-driven) features vs. discriminant (posterior) features
Acoustic Modeling: HMMs (1/2)

- An inventory of phonetic hidden Markov models (HMMs) can constitute any given word in the pronunciation lexicon with two assumptions
  - **First-order (Markov) assumption**: the state transition depends only on the origin and destination
  - **Output-independent assumption**: all observation frames are dependent on the state that generated them, not on neighboring observation frames

Steve Young et al. The HTK Book. Version 3.4, March 2006
Acoustic Modeling: HMMs (2/2)

- Three fundamental problems
  1. Computation of the probability (likelihood) of a sequence of observations given a specific HMM
     - Forward/backward algorithms for efficient computation
  2. Determination of a best sequence of model states
     - Viterbi algorithm for state alignment
  3. Adjustment of model parameters so as to best account for observed signals (or discrimination purposes)
     - Maximum Likelihood (ML), Maximum A Posteriori (MAP) and Discriminative Training (DT) criteria
     - DT considers not only the correct (or reference) transcript of a training utterance, but also the competing hypotheses for better model discrimination

Language Modeling: *n*-grams (1/2)

- For a word sequence $W$, $P(W)$ can be decomposed into a product of conditional probabilities

$$P(W) = P(w_1, w_2, ..., w_m)$$

$$= P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)...P(w_m|w_1, w_2, ..., w_{m-1})$$

$$= P(w_1)\prod_{i=2}^{m} P(w_i|w_1, w_2, ..., w_{i-1})$$

- *n*-gram modeling: the history is put into $V^{n-1}$ equivalence classes, where $V$ is the vocabulary size

$$P(w_i|w_1, w_2, ..., w_{i-1}) \approx P(w_i|w_{i-n+1}, w_{i-n+2}, ..., w_{i-1})$$

  - Bigram ($n=2$) and trigram ($n=3$) are the most prevalent

$$P(w_i|w_1, w_2, ..., w_{i-1}) \approx P(w_i|w_{i-2}, w_{i-1}) \text{ or } P(w_i|w_{i-3}, w_{i-2}, w_{i-1})$$

R. Rosenfeld, “Two Decades of Statistical Language Modeling: Where Do We Go from Here?,” Proceedings of IEEE, 2000
Language Modeling: $n$-grams (2/2)

- **Known Weakness of $n$-grams**
  - Sensitive to changes in the style or topic of the text on which they are trained
  - Assume the probability of next word in a sentence depends only on the identity of last $n-1$ words
    - Capture only local contextual information or lexical regularity of a language

- **F. Jelinek said “put language back into language modeling”**
  - Structure and topic models and language models have been proposed to harness extra information cues complementary to $n$-grams; e.g., a typical topic model

$$P_{\text{Topic}}(w_i \mid \text{History}) = \sum_{k=1}^{K} P(w_i \mid T_k) \cdot P(T_k \mid \text{History})$$

---

Linguistic Decoding (1/2)

- Find the most likely word sequence on top of the acoustic and language models and through
  - A dynamically-built word network: tree-copy search
  - A statically-built word network: finite state transducer, FST
- Efficient search algorithms and pruning techniques are highly demanded
  - Breadth-first search (BFS) with path pruning (beam search)
  - A* search (or stack decoding) with heuristics/evaluation functions
- Need to strike the balance between time and space requirements

Aubert, X. L., "An Overview of Decoding Techniques for Large Vocabulary Continuous Speech Recognition," Computer Speech and Language, 2002
Linguistic Decoding (2/2)

- E.g., tree-copy search with $n$-gram (bigram) models

  - The pronunciation lexicon is structured as a tree
  - Due to the constraints of $n$-gram language modeling, a word’s occurrence is dependent on the previous $n-1$ words
  - We have to search through all possible tree copies from the start time to the end time of the utterance to find a best sequence of word hypotheses
ASR Robustness is Crucial

- The difficulty of ASR is further exacerbated by the speaker and environment variability

Variability caused by the environment
Variability caused by the context
Intra-speaker variability
Inter-speaker variability
Linguistic variability

Pronunciation Variation

Speaker-independency
Speaker-adaptation
Speaker-dependency

Context-Dependent
Acoustic Modeling

Robustness
Enhancement

C.-H. Lee et al. (eds.) Automatic Speech and Speaker Recognition: Advanced Topics, 1996.
State-of-the-art ASR Performance

- Word error rate (WER) performance over time for a range of DARPA large-vocabulary speech recognition tasks

Applications of ASR

- Multimedia (spoken document) retrieval and organization
  - Speech-driven Interface and multimedia content processing
  - Work in association with information retrieval techniques
  - A wild variety of potential applications (to be introduced later)

- Computer-Aided Language Learning (CALL)
  - Speech-driven Interface and multimedia content processing, in conjunction with natural language processing techniques
    - Synchronization of audio/video learning materials
    - Automatic pronunciation assessment/scoring
    - Automated reading tutor

- Among many others
Prototype and Deployed Systems

- **Informedia** System at Carnegie Mellon Univ.
- **Rough’n’Ready** System at BBN Technologies
- SpeechBot Audio/Video Search System at HP Labs
- IBM Speech Search for Call-Center Conversations & Call-Routing, Voicemails, Monitoring Global Video and Web News Sources (*TALES*)
- Google’s **411 Voice Search**
- MIT Lecture Browser
- Apple’s **Siri**

_We are witnessing the golden age of ASR!_

Investigate a novel use of relevance information cues to dynamically complement (or adapt) the conventional n-gram models, assuming that:

- During ASR, a search history $H = h_1, h_2, \ldots, h_L$ is a sample from a relevance class $R$ describing some semantic content.
- A probable word $w$ that immediately succeeds the $H$ is a sample from $R$.

$$P(w|H)$$

Relevance language modeling for ASR (1/4)

Relevance Language Modeling for ASR (2/4)

• Leverage the top-$M$ relevant documents of the search history to approximate the relevance class $R$
  ◦ Take $H$ as a query to retrieve relevant documents
  ◦ **Relevance Model**: Multinomial view (**bag-of-words modeling**) of $R$

\[
P_{RM}(w|H) = \frac{P_{RM}(H,w)}{P_{RM}(H)}
= \frac{\sum_{m=1}^{M} P(D_m)P(H,w|D_m)}{\sum_{m=1}^{M} P(D_m)P(H|D_m)}
= \frac{\sum_{m=1}^{M} P(D_m)P(w|D_m)\prod_{l=1}^{L} P(h_l|D_m)}{\sum_{m=1}^{M} P(D_m)\prod_{l=1}^{L} P(h_l|D_m)}
\]

• **Combined with n-gram models**

\[
P_{Adapt}(w|H) = \lambda \cdot P_{RM}(w|H) + (1 - \lambda) \cdot P_{BG}(w|h_{L-1}, h_L)
\]
Relevance Language Modeling for ASR (3/4)

- Further incorporation of latent topic information
  - A shared set of latent topic variables \( \{T_1, T_2, \ldots, T_K\} \) is used to describe “word-document” co-occurrence characteristics

\[
P(w \mid D_m) = \sum_{k=1}^{K} P(w \mid T_k)P(T_k \mid D_m)
\]

\[
P_{\text{TRM}}(H, w) = \sum_{m=1}^{M} \sum_{k=1}^{K} P(D_m)P(T_k \mid D_m)P(w \mid T_k)\prod_{l=1}^{L} P(h_l \mid T_k)
\]

- Alternative modeling of pairwise word associations

\[
P_{\text{PRM}}(h_l, w) = \sum_{m=1}^{M} P(D_m)P(h_l \mid D_m)P(w \mid D_m)
\]

\[
P_{\text{PRM}}(w \mid H) = \sum_{l=1}^{L} \alpha_l \cdot P_{\text{PRM}}(w \mid h_l)
\]

\[
P_{\text{TPRM}}(h_l, w) = \sum_{m=1}^{M} \sum_{k=1}^{K} P(D_m)P(T_k \mid D_m)P(h_l \mid T_k)P(w \mid T_k)
\]
Tested on a large vocabulary broadcast new recognition task
  ◦ Character error rate (CER) results (the lower the better)

<table>
<thead>
<tr>
<th>n-gram</th>
<th>RM</th>
<th>TRM</th>
<th>PRM</th>
<th>TPRM</th>
<th>PLSA</th>
<th>LDA</th>
<th>Cache</th>
<th>TBLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.08</td>
<td>19.29</td>
<td>19.08</td>
<td>19.23</td>
<td>19.09</td>
<td>19.15</td>
<td>19.15</td>
<td>19.86</td>
<td>20.02</td>
</tr>
</tbody>
</table>

  ◦ The various RM models achieve results compared to PLSA and LDA (topic models) and are considerably better than Cache and TBLM (trigger-based language model)

  ◦ The various RM models are more efficient than PLSA and LDA
    • The various RM probabilities can be easily composed on the basis of the component probability distributions that were trained beforehand, without recourse to any complex inference procedure during the recognition (or rescoring) process
    • Computationally tractable and feasible for ASR
Speech Retrieval

- Robustly Index spoken documents with speech recognition techniques
  - Explore better ways to represent the recognition hypotheses of spoken documents beyond the top scoring ones
  - Hybrid of words and subwords (phone/syllable/character n-grams) for indexing

- Retrieve relevant spoken documents in response to a user query
  - *Spoken Document Retrieval* (SDR)
    - Find spoken documents that are “topically related” to a given query
  - *Spoken Term Detection* (STD)
    - Find “literally matched” spoken documents where all/most query terms should be present (much like Web search)
Scenarios of Spoken Document Retrieval (SDR)

- Scenarios
  - spoken query (SQ)
  - text query (TQ)
  - spoken documents (SD)
  - text documents (TD)

- SQ/SD is the most difficult
- TQ/SD is studied most of the time
  - "query-by-example": e.g., use text news documents to retrieve relevant broadcast news documents
  - Useful for news monitoring and tracking

B. Chen et al., "Speech Retrieval of Mandarin Broadcast News via Mobile Devices," *Interspeech 2005*
Representations of Spoken Queries and Documents

- Lattice/confusion network structures for retaining multiple recognition hypotheses

Retrieval Models for SDR

- Information retrieval (IR) models, for example, can be characterized by two different matching strategies
  - Literal term matching
    - Match queries and documents in an index term space
  - Concept matching
    - Match queries and documents in a latent semantic space

Relevance Language Modeling for SDR

- Schematic illustration

![Diagram showing the relationship between Document, Document Model, Query, Query Model, and Information Need. The diagram includes arrows indicating the flow of information, with labels for Document Likelihood, Query Likelihood, and Information Need (Relevance Class R)].

How to estimate RM?

MAP Evaluated on the TDT collection (the higher the better)

<table>
<thead>
<tr>
<th></th>
<th>ULM</th>
<th>RM</th>
<th>TRM</th>
<th>RM+NR</th>
<th>TRM+NR</th>
<th>PLSA</th>
<th>LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0.323</td>
<td>0.364</td>
<td>0.394</td>
<td>0.392</td>
<td>0.402</td>
<td>0.345</td>
<td>0.341</td>
</tr>
</tbody>
</table>

B. Chen et al., "Query modeling for spoken document retrieval," ASRU2011.
Extractive Speech Summarization

Pre-processing
- Speech Detection
- Speaker Identification
- Speech Recognition
- Spontaneous Effect Removal
- Sentence Boundary Detection
- Summarization Unit Selection

Feature Extraction
- Structural Info. Extraction
- Prosodic Info. Extraction
- Lexical Info. Extraction
- Acoustic Info. Extraction
- Discourse Info. Extraction

Summarization
- Summarization Algorithms

Post-processing
- Compaction
- Representation
- Evaluation

Relevance Language Modeling for Summarization

• Schematic illustration

spoken document $D$ be summarized

\[
S^* = \arg \min_{S_n} \lambda \text{KL}(D\|S_n) - (1 - \lambda) \text{KL}(S\|S_n)
\]

○ Iteratively select important sentences $S_n$ that have a small model distance to $D$ but have a large distance to the set $S$ of already selected sentences

○ Leverage sentence-specific relevance model (RM) and non-relevance model (NR) to enhance each sentence model
NTNU Lecture/News Browsing System

Spoken Document Browser
Spoken Language Processing Laboratory, NTNU
Conclusions

- Multimedia information access (over the Web) using speech will be very promising in the near future
  - Speech is the key for multimedia understanding and organization
  - Several task domains still remain challenging

- Speech retrieval and summarization provide good assistance for companies, for instance, in
  - Contact (Call)-center conversations: monitor agent conduct and customer satisfaction, increase service efficiency
  - Content-providing services: such as MOD (Multimedia on Demand): provide a better way to retrieve and browse described program contents

- Speech processing technologies are expected to play an essential role in computer-aided (language) learning