Latent Topic Modeling of Word Co-Occurrence Information for Spoken Document Retrieval

Berlin Chen

Department of Computer Science & Information Engineering
National Taiwan Normal University, Taiwan

2009/04/23

Outline

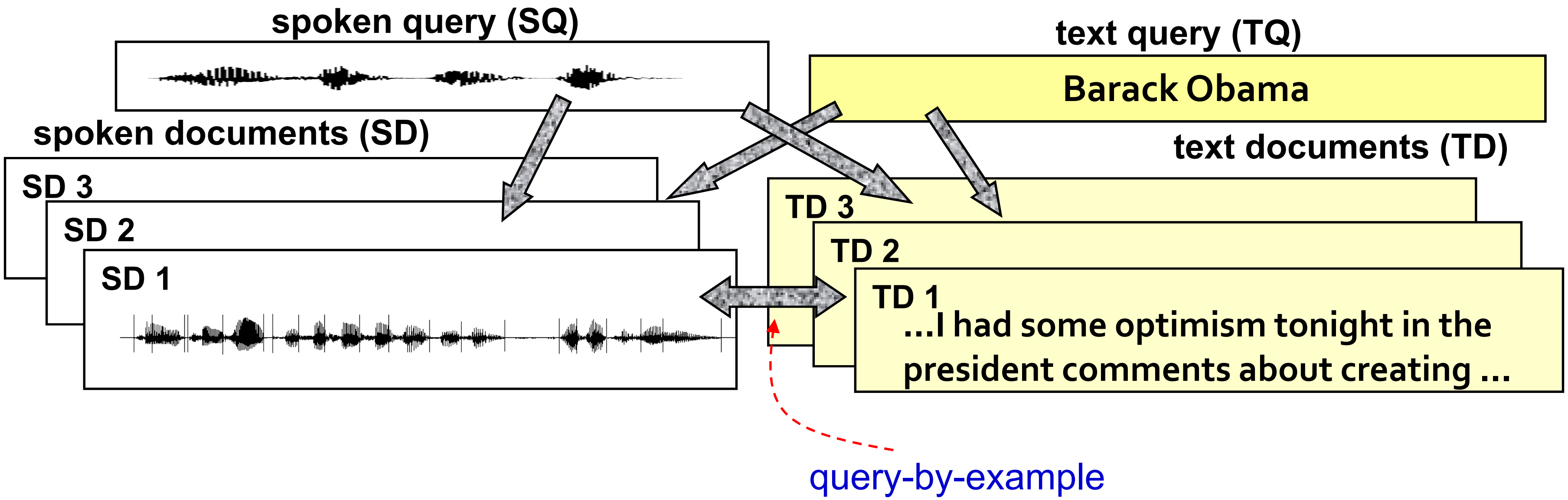
- Introduction
- Document Topic Models (DTM)
- Word Topic Model (WTM)
- Comparisons and Experiments on SDR
- Applications of WTM to Other Related Tasks
- Conclusions

Introduction

- Large volumes of multimedia associated with speech are now made available on the Internet
  - Voice search provides a natural way for multimedia access

- Task Definition for Voice Search
  - 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

Scenarios for Spoken Document Retrieval (SDR)



- SQ/SD is the most difficult
- TQ/SD is studied most of the time
  - This paper investigates using (Xinhua) text news to retrieve relevant (Voice of America) broadcast news
    - "query-by-example"
    - Useful for news monitoring and tracking

Language Modeling (LM) Approaches

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

$$P_{\mathrm{LM}}(D|Q) = \frac{P(Q|\mathrm{M}_{\mathrm{D}})P(D)}{P(Q)} \propto P(Q|\mathrm{M}_{\mathrm{D}})$$

  – A probabilistic framework for ranking documents given a query
  – Each document is viewed as a language model for generating the query
  – Those documents with higher query-likelihoods are more relevant to the query

The so-called query-likelihood methods !

LM for SDR: Two Matching Strategies

- Literal Term Matching: Each document offers a *n*-gram (usually unigram) distribution for observing a query word

$$P_{\text{Unigram}}(Q|\mathrm{M}_D) = \prod_{i=1}^{L} \left[\lambda \cdot P(w_i|\mathrm{M}_D) + (1-\lambda) \cdot P(w_i|\mathrm{M}_C)\right]$$

- Concept Matching: Each document as a whole consists of a set of shared latent topics with different weights -- A document topic model (DTM)
  - Each topic offers a unigram (multinomial) distribution for observing a query word

$$P_{\text{PLSA/LDA}}(Q \mid \mathrm{M}_D) = \prod_{i=1}^{L} \left[\sum_{k=1}^{K} P(w_i \mid T_k) P(T_k \mid \mathrm{M}_D)\right]$$

  - PLSA (Probabilistic Latent Semantic Analysis) and LDA (Latent Dirichlet Allocation) are the two good examples
    - Mainly differ in inference of model parameters (fixed & unknown vs. Dirichlet distributed)

Most of the popular LMs in IR/SDR are bag-of-words (unigram) modeling !

Word Topic Models (WTM)

- Each word of language is treated as a word topic model (WTM) for predicting the occurrences of other words

$$P_{\mathrm{WTM}}\left(w_i \mid \mathrm{M}_{w_j}\right) = \sum_{k=1}^{K} P\left(w_i \mid T_k\right) P\left(T_k \mid \mathrm{M}_{w_j}\right)$$

- The relevance measure between a query and a document can be expressed by

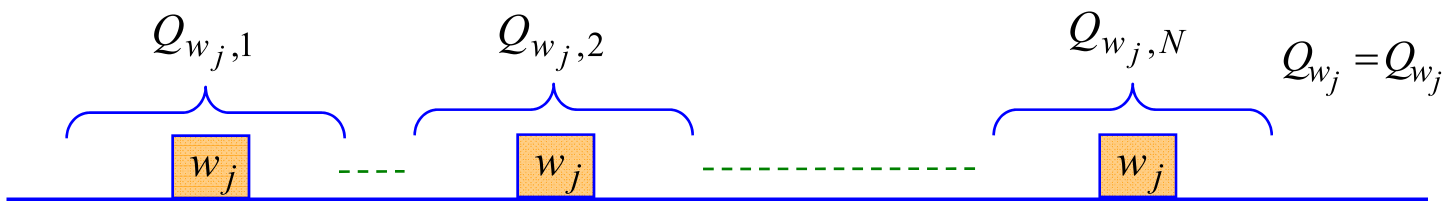
$$P_{\mathrm{WTM}}\left(Q \middle| D\right) = \prod_{i=1}^{L} \left[\sum_{w_j \in D} P_{\mathrm{WTM}}\left(w_i \middle| \mathrm{M}_{w_j}\right) P\left(w_j \middle| D\right) \right]$$

  - A spoken document can be viewed as a composite WTM
  - WTM is a kind of LM for translating words in the document to words in the query
  - $P\left(w_j \middle| D\right)$ is estimated according to the frequency of w_j in D

Can we model topic al information using other units beyond "documents" ?

Unsupervised Training of WTM

- The WTM $P_{\mathrm{WTM}}\left(w_i \mid \mathrm{M}_{w_j}\right)$ of each word can be trained with maximum likelihood estimation (MLE)
  - By concatenating those words occurring within a context window around each occurrence of the word, which are assumed to be relevant to the word, to form the training observation



$$Q_{w_j} = Q_{w_j,1}, Q_{w_j,2}, \cdots, Q_{w_j,N}$$

$$\log L_{\mathbf{W}} = \sum_{w_j \in \mathbf{W}} \log P_{\mathrm{WTM}}\left(Q_{w_j} \middle| \mathrm{M}_{w_j}\right) = \sum_{w_j \in \mathbf{W}} \sum_{w_i \in Q_{w_j}} c\left(w_i, Q_{w_j}\right) \log P_{\mathrm{WTM}}\left(w_i \middle| \mathrm{M}_{w_j}\right)$$

  - $\mathbf{W}$: the set of words in the language

  - WTM was trained to optimize its prediction power over the observation

Comparison Between WTM and DTM
-- Probabilistic Matrix Decompositions

PLSA/LDA

documents / words: $\mathbf{A}$ $\approx$ topics / words: $\mathbf{G}$; documents / topics: $\mathbf{H}^{\mathrm{T}}$

normalized “word-document” co-occurrence matrix

mixture components

mixture weights

$$P_{\text{PLSA/LDA}}\left(w_i \mid \mathrm{M}_D\right) = \sum_{k=1}^{K} P\left(w_i \mid T_k\right) P\left(T_k \mid \mathrm{M}_D\right)$$

WTM

vicinities of words / words: $\mathbf{B}$ $\approx$ topics / words: $\mathbf{Q}$; vicinities of words / topics: $\mathbf{Q}'^{\mathrm{T}}$

normalized “word-word” co-occurrence matrix

mixture components

mixture weights

$$P_{\text{WTM}}\left(w_i \mid \mathrm{M}_{w_j}\right) = \sum_{k=1}^{K} P\left(w_i \mid T_k\right) P\left(T_k \mid \mathrm{M}_{w_j}\right)$$

Unsupervised training for PLSA/LDA and WTM!

Comparison Between WTM and DTM -- Spoken Document Retrieval

- Experiments were conducted on the TDT-2 spoken document collection (~50h broadcast news stories, 16 test queries)
  - Results were measured by Mean Average Precision (*m*AP)

PLSA		LDA		WTM		WTM-L	
TD	SD	TD	SD	TD	SD	TD	SD
0.627	0.568	0.641	0.570	0.636	0.573	0.644	0.574

  - PLSA, LDA and WTM (8 topics) are all trained without supervision (without using additional query-document relevance information)
    - PLSA or LDA maximizes the collection likelihood
    - WTM maximizes the likelihood of words in each word's vicinity
  - WTM-L: Further assume the parameters of WTM follow Dirichlet distributions

–
$$\hat{P}_{\text{DTM/WTM}}(w_i \mid \mathrm{M}_D) = \rho_1 \cdot P_{\text{DTM/WTM}}(w_i \mid \mathrm{M}_D) + \rho_2 \cdot P(w_i \mid \mathrm{M}_D) + (1 - \rho_1 - \rho_2) \cdot P(w_i \mid \mathrm{M}_C)$$

Supervised Training of WTM

- Maximum Likelihood Estimation (MLE)
  - Maximize the log-likelihood of an outside training set of (~800) query exemplars generated by their relevant documents

$$\log L_{\mathbf{Q}_{TrainSet}} = \sum_{Q \in \mathbf{Q}_{TrainSet}} \sum_{D_r \in \mathbf{D}_{R \text{ to } Q}} \log P_{\mathrm{WTM}}\left(Q \middle| \mathrm{M}_{D_r}\right)$$

- Minimum Classification Error Training (MCE)
  - Given a training query exemplar, we can instead minimize the following error function

relevant document

irrelevant document

$$E(Q, D_r, D_{irr}) = \frac{1}{|Q|}\left[-\log P_{\mathrm{WTM}}\left(Q \middle| \mathrm{M}_{D_r}\right) + \max_{D_{irr}} \log P_{\mathrm{WTM}}\left(Q \middle| \mathrm{M}_{D_{irr}}\right)\right]$$

Other irrelevant documents for the training query can be into consideration

  - Further converted to a loss function with a Sigmoid operator
  - Corresponding parameters of WTM then are updated with a generalized probabilistic descent (GPD) procedure

Associate documents with queries even if they do not share any of the query words!

Results of Supervised Training

	WTM				PLSA				Unigram	
	MIX-8		MIX-32		MIX-8		MIX-32			
	TD	SD	TD	SD	TD	SD	TD	SD	TD	SD
MLE	0.689	0.617	0.735	0.686	0.675	0.592	0.683	0.626	0.633	0.566
MCE	0.700	0.631	0.760	0.710	0.679	0.608	0.685	0.628	0.646	0.581

– For WTM, if training query-relevant document pairs were available, significantly better results could be achieved by either MLE or MCE
– PLSA and Unigram LM (i.e., the simple literal term matching model) can also be trained with supervision
– Notice also that, MCE seems to provide additional performance gains over MLE

Results of Various Vector Space Approaches

- Here we also list the results of retrieval using three popular vector space approaches

VSM		LSA		SVM	
TD	SD	TD	SD	TD	SD
0.555	0.512	0.551	0.531	0.580	0.532

  – SVM (Support Vector Machine) treats IR as a classification problem
    - A set of 11 heterogeneous features is used to represent each spoken document given an input query
    - SVM was trained by leveraging the relevance information of the outside training query exemplars

  – All LM-based retrieval approaches are significantly better than these vector space approaches

WTM Applied to Other Related Tasks

- Language Modeling in Speech Recognition

$$P\left(w_i \middle| H_{w_i}\right) = \sum_{j=1}^{i-1} P_{\mathrm{WTM}}\left(w_i \middle| \mathrm{M}_{w_j}\right) P\left(w_j \middle| H_{w_i}\right)$$
$$= \sum_{j=1}^{i-1} P\left(w_j \middle| H_{w_i}\right) \sum_{k=1}^{K} P\left(w_i \middle| T_k\right) P\left(T_k \middle| \mathrm{M}_{w_j}\right)$$

- Extractive Spoken Document Summarization

$$P\left(D \middle| S\right) = \prod_{i=1}^{L} \left[\sum_{w_j \in S} P_{\mathrm{WTM}}\left(w_i \middle| \mathrm{M}_{w_j}\right) P\left(w_j \middle| S\right) \right]$$
$$= \prod_{i=1}^{L} \left[\sum_{w_j \in S} P\left(w_j \middle| S\right) \sum_{k=1}^{K} P\left(w_i \middle| T_k\right) P\left(T_k \middle| \mathrm{M}_{w_j}\right) \right]$$

- For both tasks, WTM has preliminarily demonstrated good results as compared to existing approaches

Conclusions

- This paper presented a word topic modeling (WTM) approach for spoken document retrieval
  - Simple and easy to implement
- Various model inference techniques were studied for WTM and other document topic models (DTMs)
  - Given an outside training set of query exemplars with relevance labels, the LM-based retrieval models can be steadily improved
- Future work on WTM: integration with more elaborate indexing mechanisms for large-scale SDR
  - Compared to more other retrieval models