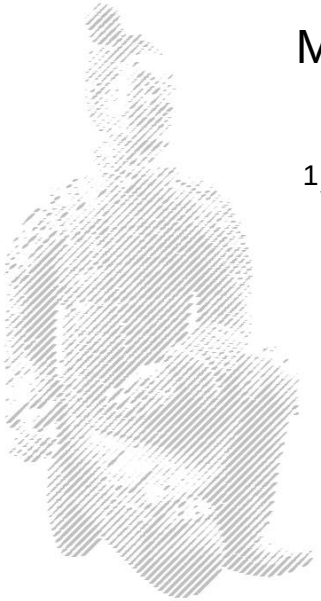
Interspeech 2011, Florence, Italy

Probabilistic Latent Semantic Analysis for Broadcast News Story Segmentation



Mimi LU[1,2], Cheung-Chi LEUNG[2], Lei XIE[1], Bin MA[2] and Haizhou LI[2]

[1]Shaanxi Provincial Key Lab of Speech and Image Information Processing, Northwestern Polytechnical University, China

[2]Institute for Infocomm Research, A*STAR, Singapore

Broadcast news story segmentation

- The task of dividing broadcast news (BN) programs into homogeneous units each addressing a main topic

- A key precursor to various tasks, such as spoken document retrieval and summarization
- Three categories of cues for story segmentation:
  - Lexical, acoustic and visual cues



Motivation

- Lexical cohesion based methods
  - Words in a story hang together by semantic relations
  - Different stories deploy different set of words
  - Usually measured by rigid word counts
- Literal matching on individual terms is unreliable:
  - Synonym: “car”, “automobile”; …
  - Polysemy: “china” can refer to a nation or porcelain; “apple” can refer to Apply Computer Inc or fruit; …
- Conceptual matching is introduced:
  - E.g. Latent semantic analysis (LSA), Probabilistic latent semantic analysis (PLSA)



Motivation

- Spoken document segmentation is different from text segmentation:
  - Task performs on LVCSR outputs where erroneous words exist, thus breaking the lexical cohesion
  - Many recognition errors are from Out-Of-Vocabulary (OOV) words, which are typically name entities that are key to topics
- Phoneme *n*-gram: partial matching
  - Incorrectly recognized words may contain subword units correctly recognized



Contributions

- We use PLSA for story segmentation for broadcast news
- We use phoneme *n*-gram as the basic unit for lexical cohesion measure to handle erroneous LVCSR transcripts
- Cross entropy based approach is introduced for lexical cohesion measure and it is compared with cosine similarity
- We compare dynamic programming (DP) with TextTiling for story boundary identification



PLSA model

- Probabilistic latent semantic analysis

$$P(d,w) = P(d)P(w|d) \qquad P(w|d) = \sum_{z \in Z} P(w|z)P(z|d)$$

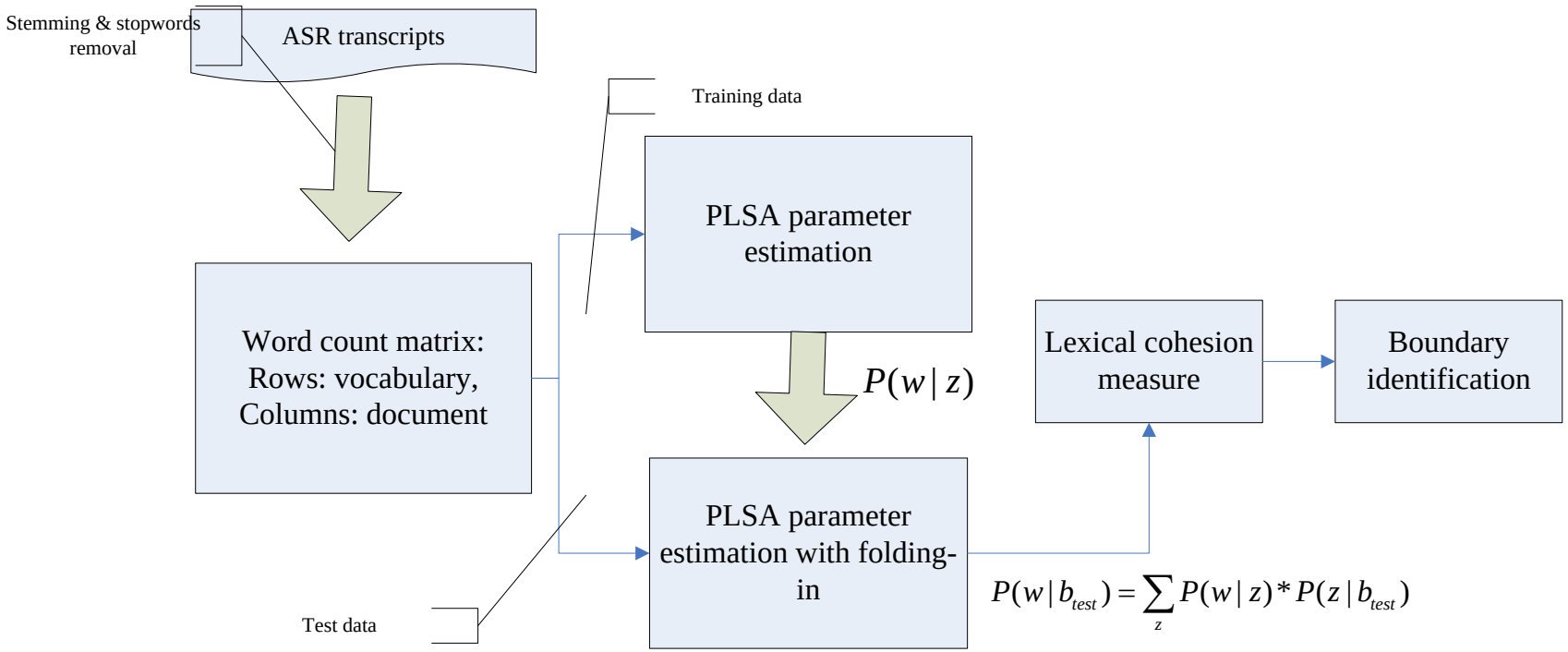
d: document, w: word, z: topic

- Maximum Likelihood Estimation
  - Maximize log-likelihood of co-occurrence pairs $L = \sum_d \sum_w n(d,w) \log P(d,w)$
  - E-step $P(z|d,w) = \frac{P(w|z)P(z|d)}{\sum_z P(w|z)P(z|d)}$
  - M-step $P(w|z) = \frac{\sum_d n(d,w)P(z|d,w)}{\sum_w \sum_d n(d,w)P(z|d,w)}$ $P(z|d) = \frac{\sum_w n(d,w)P(z|d,w)}{\sum_z \sum_w n(d,w)P(z|d,w)}$

- Folding-in process for unseen test data: keep P(w|z) fixed



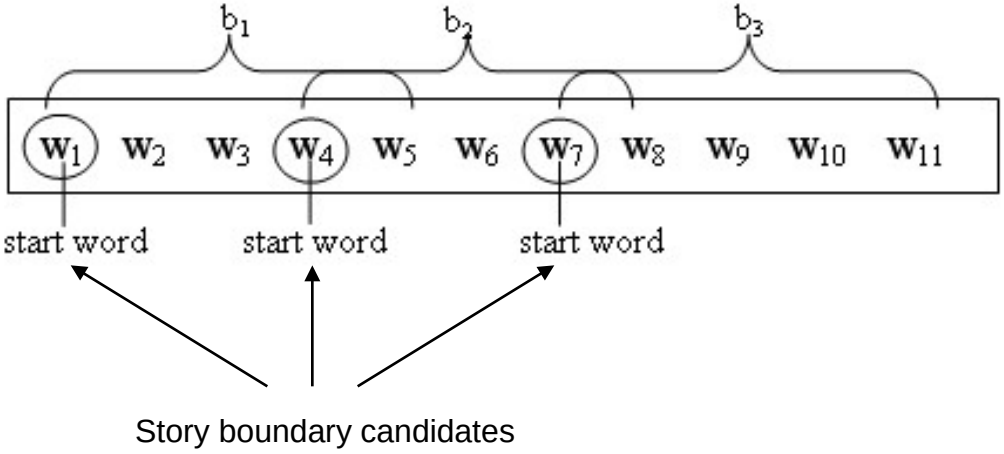
System overview





Sentence Construction

- Sentence delimiters are not available in LVCSR transcripts
- Pseudo-sentence: each text block with a fixed number of consecutive words is formed



Lexical cohesion measure

- Cosine similarity
  - Measure the closeness between two vectors, usually calculated on term frequencies
  - Apply with PLSA statistics:

$$Sim(i, j) = \frac{\sum_{w} P(w|b_i)P(w|b_j)}{\sqrt{\sum_{w} P(w|b_i)^2 \sum_{w} P(w|b_j)^2}}$$



Lexical cohesion measure

- Cross entropy
  - A divergence measure to depict how different two distributions are

$$H(p,q) = -\sum_{x} p(x) \log q(x)$$

  - Minimum obtained when p = q



Lexical cohesion measure

- Cross entropy
  - Apply with PLSA statistics:

$$CrossEnt(i, j) = -\sum_{w} P(w \mid b_i) \log P(w \mid b_j)$$

  - Normalization:

$$Dissim(i, j) = \frac{CrossEnt(i, j) - CrossEnt(i, i)}{CrossEnt(i, j)}$$



Boundary identification

- Local comparison
  - Compute lexical scores between adjacent blocks
  - Locate valleys (similarity) or peaks (dissimilarity)
  - E.g. TextTiling

When salient topic change occurs



Boundary identification

- Global optimization

When topic transitions are smooth

  - Minimize the cost of a specific segmentation

$$C(S) = \sum_{k=1}^{K} Cost(s_k)$$

$$Cost(s_k) = \frac{\sum_i \sum_j Dissim(i, j)}{N(len(s_k))}$$

Normalization factor

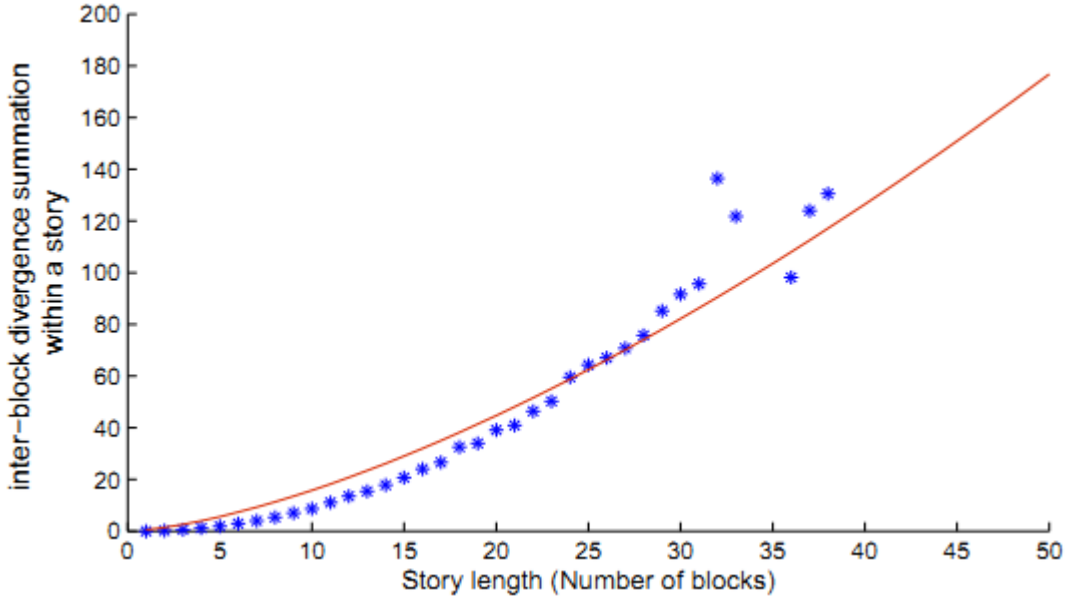
$$\hat{S} = \arg\min_S C(S)$$

  - S = {s_1, …, s_k, …, s_K} is a segmentation of document D
  - Implementation: dynamic programming (DP)



Boundary identification

- Normalization factor $N(len(s_k))$
  - To make long and short segments comparable
  - Inter-block disparity distribution:





Experimental setup

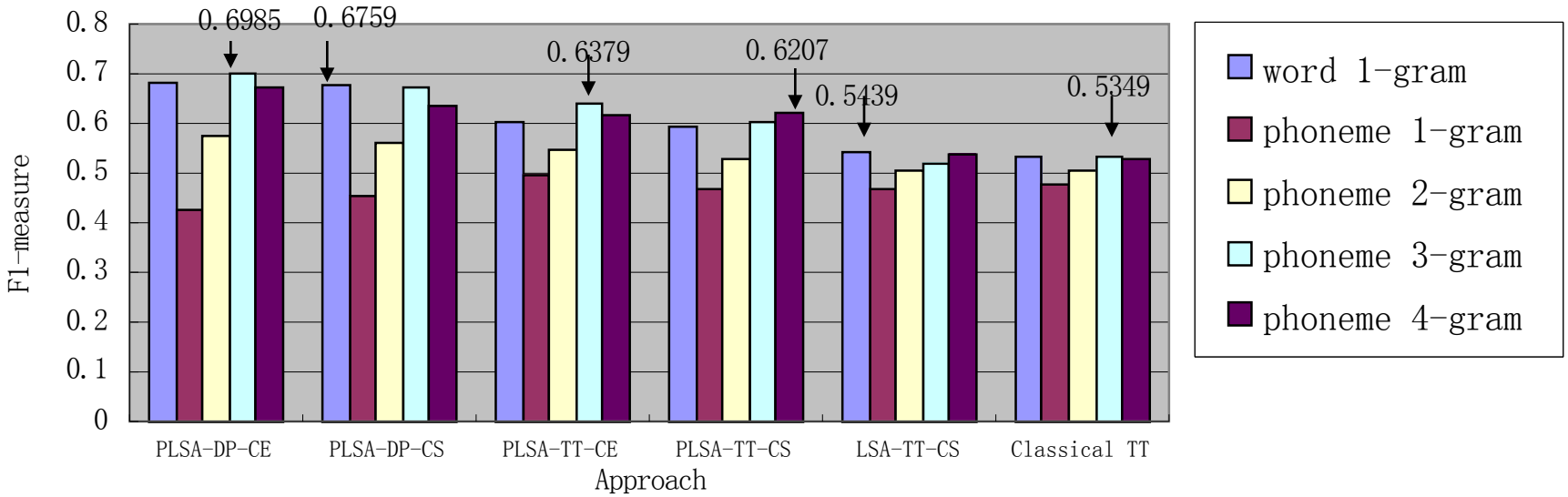
- Corpus:
  - LVCSR transcripts of TDT2 VOA English broadcast news
  - Data used (Number of programs):

    training = 56, development = 27, test = 28
- Tuning parameters:
  - TextTiling: block length, sliding window shift, lexical score threshold
  - DP: block length, alpha in normalization factor
- Phoneme *n*-gram sequences generated from word transcripts using the CMU dictionary
- Evaluation criterion: F1-measure

$$F1-measure = \frac{2 * recall * precision}{recall + precision}$$



Experimental results



DP: dynamic programming; TT: TextTiling

CE: cross entropy; CS: cosine similarity



Conclusions

- We investigate the use of PLSA for BN story segmentation
  - Phoneme subwords are adopted to address problems from LVCSR errors
  - Cross entropy and cosine similarity for lexical cohesion measure, and DP and TextTiling for story boundary identification are compared respectively
- Experimental results suggest:
  - PLSA can effectively boost story segmentation performance
  - Cross entropy shows advantages for describing distributional variation
  - DP provides better performance for story boundary identification
  - Performance gain using phoneme *n*-gram shows its ability to handle erroneous LVCSR transcripts



Thanks for your attention!