Generating Summary Keywords for Emails Using Topics

Hanna M. Wallach

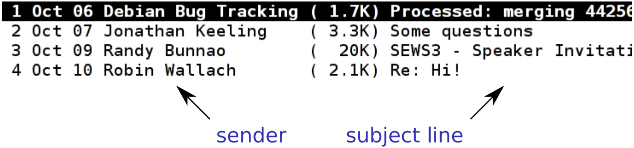
University of Cambridge/University of Massachusetts Amherst
hmw26@cam.ac.uk

October 17, 2007

(Joint work with M. Dredze, D. Puller, F. Pereira)

Email Triage

- Email triage: deciding how to handle incoming email
- User has a limited amount of information about each email:

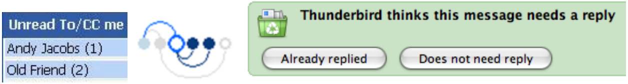


- Decisions made using available information
- Goal: provide user with additional concise information

Incorporating Information

Previous work:

- Social information, thread indicators, reply prediction

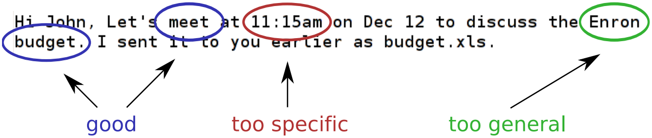


- Message snippets, full-sentence summaries, summary keywords

Our approach: generate a concise summary of the message's contents – summary keywords – in an unsupervised fashion

Good Summary Keywords

- Prepare user for message contents
- Cannot be too specific or too general



- Represent the gist of the email
- Should be associated with coherent user concepts

Our Approach

- Unsupervised framework for choosing summary keywords
  - No annotated training data required
- Use latent concept models to represent topics in user's mailbox
  - A good summary keyword relates the message to other topically similar messages in the user's mailbox
- Two ways of selecting keywords, analogous to:
  - Query-document similarity
  - Word association

Latent Concept Models

- Documents are assumed to have latent semantic structure
- Latent structure is inferred from word-document co-occurrences

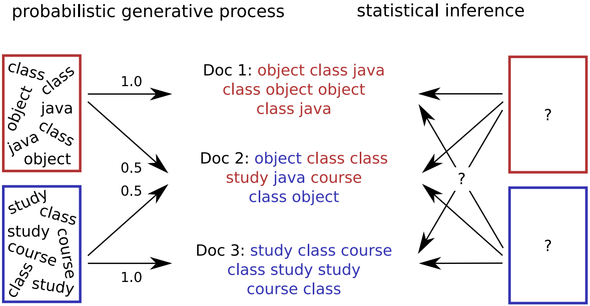
... Germany hosted the World Beard and Mustache Championships [1]

- Relates words to concepts and concepts to documents
- Used in information retrieval, classification, collaborative filtering

[1] `http://www.worldbeardchampionships.com`

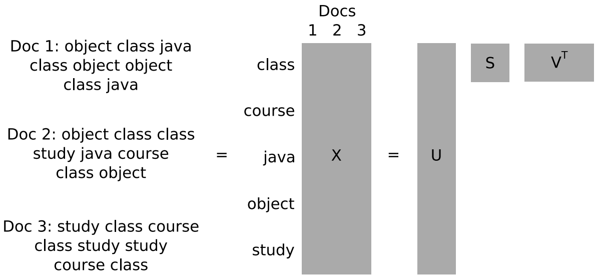
Latent Dirichlet Allocation (Blei et al., '03)

Models documents as mixtures of latent topics. Topics inferred from word correlations, independent of word order: "bag-of-words"



Latent Semantic Analysis (Deerwester et al., '90)

LSA decomposes the word-document co-occurrence count matrix into a set of orthogonal factors that represent latent concepts



Query-Document Similarity

Treat candidate keywords as one-word queries, compute similarity between each keyword and the email, choose those that are most similar

- Latent Dirichlet allocation:

$$P(\text{keyword k} \mid \text{email d}) = \sum_{\text{topics t}} P(\text{k} \mid \text{t}) P(\text{t} \mid \text{d})$$

- Latent semantic analysis:

$$\text{score(keyword k, email d)} = \mathrm{U_k} \cdot \mathrm{V_d}$$

Word Association

Compute the association between each candidate keyword and each of word in the email, choose those that are most closely associated

- Latent Dirichlet allocation:

$$P(\text{keyword k} \mid \text{email d}) = \prod_{\text{w in d}} \sum_{\text{topics t}} P(\text{k} \mid \text{t}) P(\text{t} \mid \text{w, d})$$

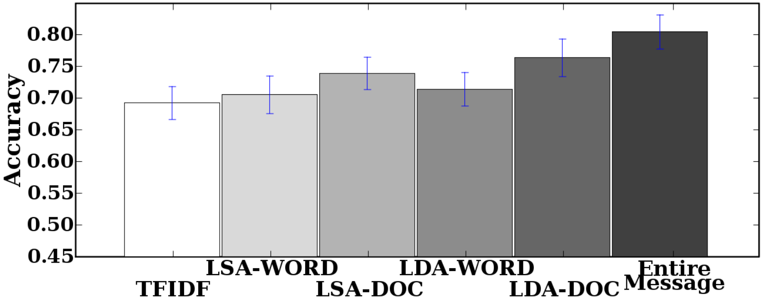
- Latent semantic analysis:

$$\text{score(keyword k, email d)} = \sum_{\text{w in d}} \sum_{\text{factors f}} \text{U}_{\text{k,f}} \, \text{U}_{\text{w,f}} \, \text{V}_{\text{f,d}}$$

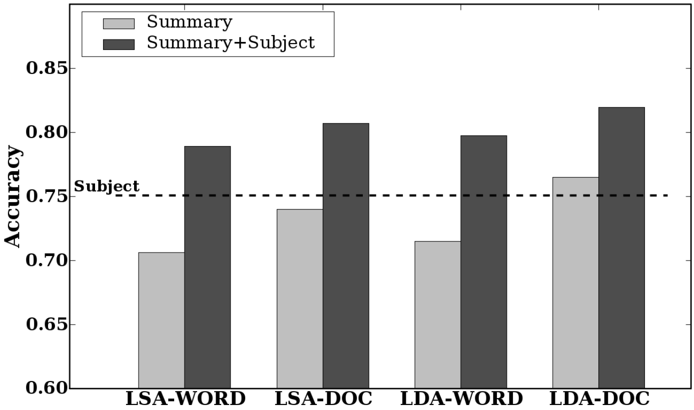
Evaluation

- Summaries evaluated using two proxy tasks
  - Automated foldering
  - Recipient prediction
- Compared on users from the Enron data set
- Length of each summary was set to nine keywords
- Two baselines:
  - Term frequency-inverse document frequency (TF-IDF) keywords
  - Full message contents

Automated Foldering: Prediction Accuracy



Automated Foldering: Improvement Over Subject



Summary Keywords

Sally -
Attached are the hypertiles from the final report out at yesterday's ASE Studio Workshop. The CD is finished and on its way to Houston. The files are organized by team:
Hammer - Sales and Marketing, Vision Stmt, Mission Stmt, Target Market, How to Approach, Pricing, SLA
Pliers - Producst and Services - Consulting Based
Saw - Infrastructure Transition Plan
Wrench - Producst and Services - Basic Outsourcing
I hope these help with your meeting tomorrow. Let me know if there is anything else I can do to help.
Lisa P

- TF-IDF: producst pliers stmt hammer wrench
- LDA-doc: team meeting services lisa ase

Findings and Future Work

Key finding:

- Summary keywords generated using topic models are a good approximation of message content and provide additional information over the message subject line

Future work:

- Other latent concept models, *e.g.*, topical *n*-gram model
- Incorporating person-specific information
- Research on incorporation of keywords into user interfaces

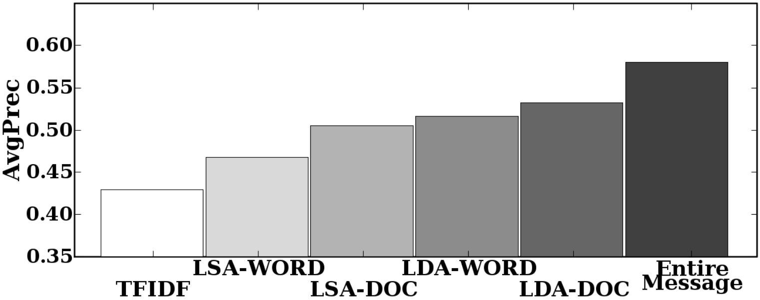
Questions?

hmw26@cam.ac.uk

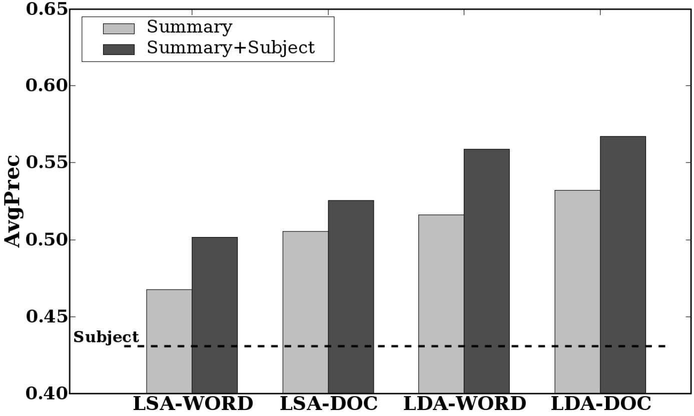
http://www.inference.phy.cam.ac.uk/hmw26/

Dredze was funded by a NDSEG fellowship. Puller was funded by a University of Pennsylvania Provost's Undergraduate Research Mentoring Program Fellowship. Work conducted in part at the University of Pennsylvania, supported by DARPA contract NBCHD030010.

Recipient Prediction: Average Precision



Recipient Prediction: Improvement Over Subject



User Interface Design

1. Display keywords with subject and sender entries in mailbox listing
2. Separate visualization, such as a tag cloud:

Each word is scaled according to its relevance as a keyword