# Chapter 13: Summarization and Information Presentation

# Overview

- Definitions
- Oldest approaches
- Summarization of a single text
- Summarization of multiple texts
- Text compaction for display on small screens
- Evaluation and results of TAC
- Information presentation

# Definitions

- Text summarization = generating the content of a text in a condensed form
  - highly valued in information retrieval:
    - **topic-general** summary with essential content
    - **user-specific** (**task/viewpoint-oriented, query-based**) summary

1) when selecting information when browsing document collections and retrieval results

2) when used as a surrogate when indexing documents

# Definitions

- **Indicative summary**: helps a reader to decide whether consulting the complete document will be worthwhile
- **Informative summary**: reports on the actual content of the document and presents as much as possible the information contained in it (cf. surrogate)
- **Extract:** is composed of pieces of text extracted verbatim from the original document <-> **Abstract**: is composed of text that is not verbatim extracted from the original document
- **Comparative summary**: summary description and evaluation of multiple document texts
- **Update summary:** summary of one or multiple documents given a initial summary, which reflects what a user already knows
- **Length of the summary**: variable, brevity is important, ...

# Oldest approaches

- End 1950s, 1960s, beginning 1970s
- **Extraction of important sentences**:
  - sentences with high-weighted (e.g., *tf* x *idf)* content terms [Luhn IBM 1958]:
    - often additional constraint: terms in close proximity
    - the oldest, but still one of the most successful approaches for summarizing news text [McKeown et al. SIGIR 2006]
    - importance of a term may depend on the number of repetitions, synonyms, coreferents or related terms that occur in the texts and that together form a **lexical chain**

# Oldest approaches

- sentences cued by indicator phases (often of rhetorical nature) that signal important content (e.g., “we conclude that”)
- sentences cued by their location (e.g., first sentence of a paragraph)

- **Problems:**
  - sentences containing high-weighted terms not always most informative
  - heuristic cues might work for 1 text type, subject domain or author, not for others
  - result = extract, not always coherent summary

# Oldest approaches: current variant

- Supervised learning given a **training corpus of source texts and their summaries**
- Feature selection: length of a sentence, sentences containing indicator phrases, the first, final or medium sentences, sentences with high-weighted content words, and sentences with proper names that occur more than once, ...
  - learning of: importance of the features for text type / subject domain / author of the training corpus or for the user-specific summary
  - extraction of sentences with important features
- **Techniques:** naive Bayes, learning of rules and trees, support vector machines

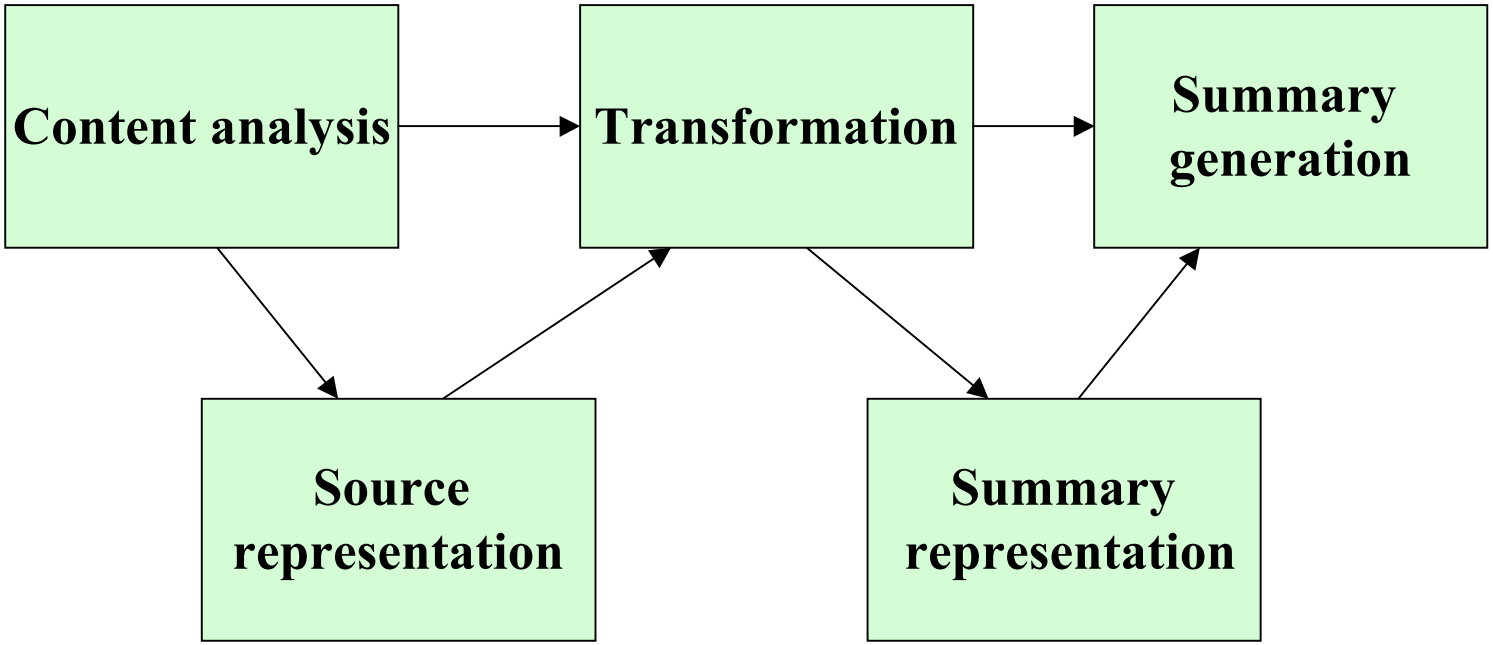
# Oldest approaches: current variant

- Earliest example: use of a naive Bayes classifier [Kupiec et al. SIGIR 1995]
- DUC-2002: good results for summarizing single news stories
- Problem: availability of sufficient training examples

# Text summarization today

- Very large interest (for text and multimedia) due to need of content reduction and abstraction:
  - for display on small screens (e.g., of smartphones)
  - for display in exploratory search

- But, many research questions still to be solved




[Sparck Jones 2007]

# Content analysis step

- Focus upon techniques that consider the discourse structure:
  - **Schematic structure**:
    - text-type dependent, informative segments (e.g., lead paragraph in news stories)
  - **Rhetorical structure**:
    - Rhetorical Structure Theory (see course Natural Language Processing - Marcu 2000)
  - **Thematic structure**:
    - text block, sentence and paragraph clustering
    - latent semantic topic models

# Text block, sentence and paragraph clustering

- = text blocks of fixed length, sentence and paragraphs of a text are clustered:
  - features used in the clustering: mostly content terms, but also stems, named entities, words of specific syntactic class, ...
  - vector representations of the blocks, sentences and paragraphs
- **Topic segmentation:**
  - adjacent blocks are clustered to form a topic segment
- **Topic detection:**
  - blocks (usually adjacent) are clustered to identify the main (sub)topics of the text

# Text block, sentence and paragraph clustering

- **Text summarization**:
  - extraction of topical sentences in reading order or key terms from the clusters to form the summary:
    - key terms: e.g., most highly weighted terms of the cluster centroid

# Linear topic segmentation

- = linearly segment the text in homogeneous units that are topically coherent
  - (1) unsupervised: use of lexical features that signal lexical cohesion: term repetition, synonyms, related terms, ... => form **lexical chains**
    - chains might be weighted (e.g., proportional with their length)
  - (2) supervised: use of cues that are indicative of topic shifts: paragraph boundaries, referential noun phrases, ...

[Galley et al. 2003]

# Linear topic segmentation

- Example: TextTiling [Hearst CL 1997]

Given:

*n* text blocks where a block = sequence of e.g., 20 words

*n*-1 gaps where a gap = point between two adjacent blocks

compute the similarity or cohesion score $s_i$ of each gap $g_i$ (= similarity between the two blocks) where $i = 1,\ldots,n-1$:

e.g., cosine of term vectors, or of lexical chain vectors

resulting sequence of cohesion scores is placed in a gap by cohesion graph

smooth the cohesion scores:

e.g., smoothed cohesion score $s_i = (s_{i-1} + s_i + s_{i+1})/3$

other heuristics are possible

translate each cohesion score into a depth score:

$d_i = (s_{i-1} - s_i) + (s_{i+1} - s_i)$

(in gap by cohesion graph: summing the heights of the two sides of the valley the gap is located in)

valleys in the graph identify ruptures in the topic structure:

compute the average $\mu$ of the $n$-1 depth scores and the standard deviation $\sigma$

select a gap $g_i$ as a topic boundary when e.g., $d_i > \mu - c\sigma$ where $c$ is a constant (e.g., $c = 0.5$ or $c = 1.0$)

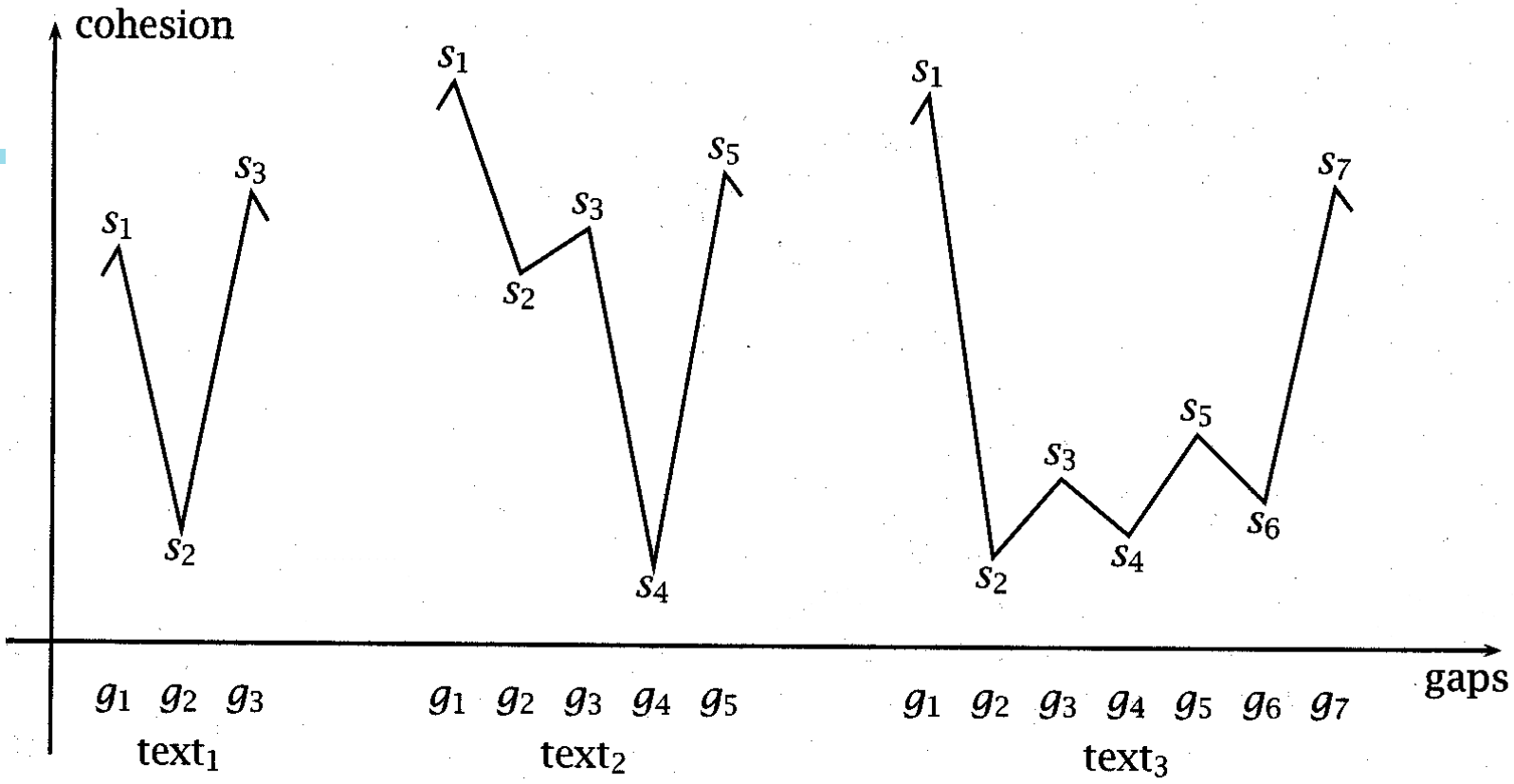



**Figure 15.12** Three constellations of cohesion scores in topic boundary identification.

[Manning & Schütze 1999]

# pLSA in topic segmentation

- Cf. TextTiling: document is divided into text blocks (e.g., sentence, block of predefined number of words)
- pLSA provides a representation for each text block $b_j$ in terms of topic and word distributions
- The estimated distribution for each word $w_i$ for each block $b_j$ is then computed as:

$$P(w_i|b_j) = \sum_{k=1}^{K} P(w_i|z_k)P(z_k|b_j)$$

# pLSA in topic segmentation

- The distribution of words $w_i$ in adjacent text blocks is compared using a (dis)similarity metric (e.g., Kullback-Leibler divergence between two probability distributions)

# Link analysis in summarization

- Computing the importance of sentences, entities, etc. based on mutual links:
  - TextRank: variation on the PageRank algorithm where words are linked based on their co-occurrence in a sentence
  - LexRank: variation on the PageRank algorithm where sentences are linked by shared content terms
  - Aboutness detection: variation on the HITS algorithm where entities are linked with their mutual coreferent en referent relations
- PageRank and HITS: see **Web information retrieval**

[Mihalcea & Tarau 2004] [Erkan & Radev 2004] [Moens et al., 2006]

# Transformation step

1) **Selection of content**:

- topic-general summary: selection of the main content
- task-specific summary: selection of the main content that is relevant for the task
- many practical systems combine text analysis and information selection: e.g., summarization = extraction of sentences that have a high relevancy score

# Transformation step

2) **Generalization of selected content**:

- = condensing the information to a more abstract form
  - e.g., replacing specific concepts by more general ones (use of thesaurus, ontology, … )
  - e.g., script-based inference
- very important task in text summarization
- very difficult task: fusion of concepts often requires knowledge about the world or domain which is seldom included in the text explicitly and understanding of content

- Potential of hierarchical Latent Dirichlet Allocation models for concept learning (e.g., from a pure and noise free, domain-specific corpus)

Example of the difficulty in generalization:

Text:

John and Bill wanted money. They bought ski-masks and guns and stole an old car from a neighbor. Wearing ski-masks and waving their guns, the two enter the bank, and within minutes left the bank with several bags of $100 bills. They drove away happy, throwing away their ski-masks and guns in a trash can. They were never caught.

Word counting would indicate that the story is about ski-masks and guns, both of which are mentioned more than any other word count. Clearly, the story is about robbery and the summary must mention this fact.

# Transformation step

- **Synthesis** of selected information (cf. multi-document summarization)
  - Needs **coreferent resolution**:
    - noun phrases
    - temporal expressions
    - spatial expressions

# Summary generation step

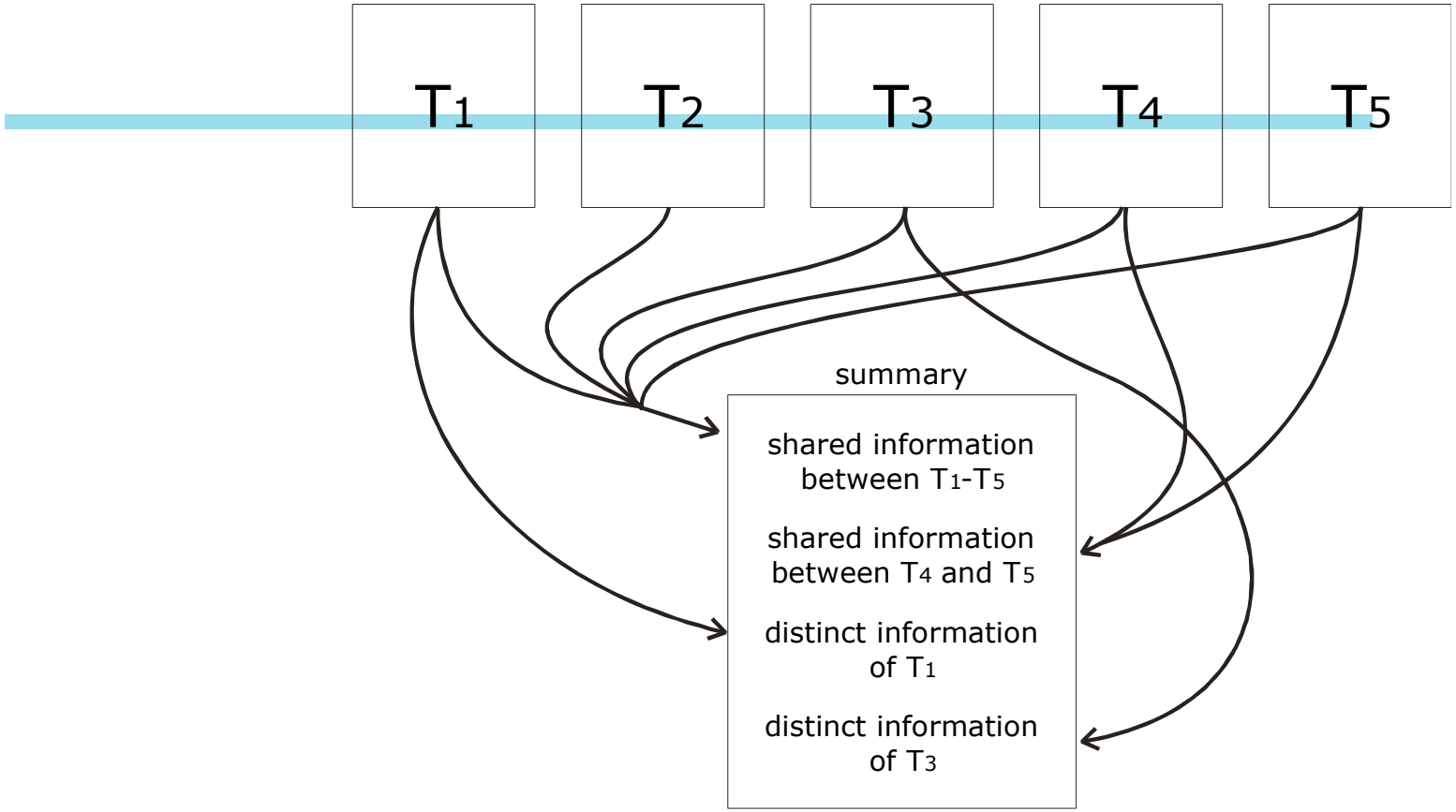
- **Text generation**:
  - NLP task
  - content planning: organization of the content of the summary
  - generation of well-formed sentences from the summary representation
  - constraint: summaries must convey maximal information in a minimum amount of space:
    - use of complex sentence structures, including multiple modifiers of a noun or verb, conjunction (e.g., “and”) and ellipsis (i.e., the deletion of repetitions across conjoined phrases)

# Summary generation step

- In practice: often “polishing” of the extracted sentences:
  - deletion of sentences with anaphoric references or the addition of a preceding sentence to the one that contains the anaphor
  - deletion of rhetorical connectives (e.g., "However" at the start of a sentence)
  - merging sentences with repetitious structure and coordinate sentences with conjunctions

# Summarization of multiple texts

- = summarizing the similarities and differences in information content between the texts
  - often referred to as multi-document summarization
  - useful:
    - e.g., when summarizing retrieval results (e.g., cluster of relevant documents)
    - e.g., when information in texts changes over time: merging and generalizing the stable information and identifying the most recent dynamic information
  - important: to find similarities between text blocks, sentences and paragraphs across the texts and to eliminate redundancy




Summarization of multiple texts ($T_{1, \ldots, T_5}$).

# Summarization of multiple texts

- Techniques: cf. single-document summarization applied across documents with extra attention to:
  - **Transformation: selection and synthesis**:
    - elimination of redundant content or novelty detection: e.g., clustering of similar sentences, MMR (**Maximum Marginal Relevance)**, which compares a sentence's term vector with the term vectors of sentences already present in the summary based on cosine similarity
    - cross-document coreferent resolution of noun phrases, temporal and spatial expressions, etc.
    - fusion of, reasoning with the information
  - **Summary generation**
    - logical (e.g., temporal, causal) ordering of the information in the summary

# Summarization of multiple texts

Example of comparative summaries: SUMMONS system [McKeown & Radev SIGIR 1995]

- summaries of multiple news articles on the same event
- made from templates generated with the MUC-4 information extraction techniques
- text fragments are compared for change of perspective, contradictory statements, gaps in the information, additions, confirmations, and refinements of the information

# Example: SUMMONS system

| | |
|---|---|
| MESSAGE: ID | TST-COL-001 |
| SECSOURCE: SOURCE | Reuters |
| SECSOURCE: DATE | 26 FEB 93<br>EARLY AFTERNOON |
| INCIDENT: DATE | 26 FEB 93 |
| INCIDENT: LOCATION | WORLD TRADE CENTER |
| INCIDENT: TYPE | BOMBING |
| HUM TGT: NUMBER | AT LEAST 5 |

Figure 5: Template for newswire article 1.

| | |
|---|---|
| MESSAGE: ID | TST-COL-002 |
| SECSOURCE: SOURCE | Associated Press |
| SECSOURCE: DATE | 26 FEB 93 19:00 |
| INCIDENT: DATE | 26 FEB 93 |
| INCIDENT: LOCATION | WORLD TRADE CENTER |
| INCIDENT: TYPE | BOMBING |
| HUM TGT: NUMBER | 5 |

Figure 6: Template for newswire article 2. **[McKeown & Radev SIGIR 1995]**

```
((#TEMPLATES == 2) && (T[1].INCIDENT.LOCATION ==
T[2].INCIDENT.LOCATION) &&
(T[1].INCIDENT.TIME < T[2]. INCIDENT.TIME) &&  ….
(T[1].SECSOURCE.SOURCE != T[2].SECSOURCE.SOURCE)) ==>
(apply("contradiction", "with – new – account", T[1], T[2]))
```

Given two templates, if INCIDENT.LOCATION is the same, the time of the first report is before the time of the second report, the report sources are different, and at least one other slot differs in value (this rule is not shown), apply the contradiction operator to combine the templates.

## Rules for the contradiction operator

**[McKeown & Radev SIGIR 1995]**

## Alternative to rule set: machine learning of the contradiction patterns

**[de Marneffe et al. ACL 2008]**

Example summary produced by the SUMMONS system:

In the afternoon of February 26, 1993, Reuters reported that a suspected bomb killed at least 5 people in the World Trade Center. However, Associated Press announced that exactly five people were killed in the blast.

[McKeown & Radev 1995]

# Probabilistic topic model

- We draw from a background vocabulary distribution $\phi_B$ from Dirichlet ($V$, $\lambda_B$) ($V$= vocabulary, $\lambda$ = pseudo-count) shared across document collections: background distribution of vocabulary words models the stop words: BACKGROUND topic
- For each document set $\mathcal{D}$ we draw from a content distribution $\phi_C$ from Dirichlet($V$, $\lambda_c$) representing the significant content of $\mathcal{D}$ that we want to summarize: CONTENT topic
- For each document D in set $\mathcal{D}$ we draw from a document specific vocabulary distribution from Dirichlet($V$, $\lambda_D$) representing words which are local to a single document, but do not appear across several documents: DOCSPECIFIC topic
- For each sentence $S$ of each document $D$, we draw from a distribution $\psi_T$(CONTENT, DOCSPECIFIC, BACKGROUND) from a Dirichlet prior with pseudo-counts (1.0, 5.0, 10.0)
- For each word position in the sentence, we draw from a topic Z from $\psi_T$ and a word $W$ from the topic distribution that $Z$ indicates

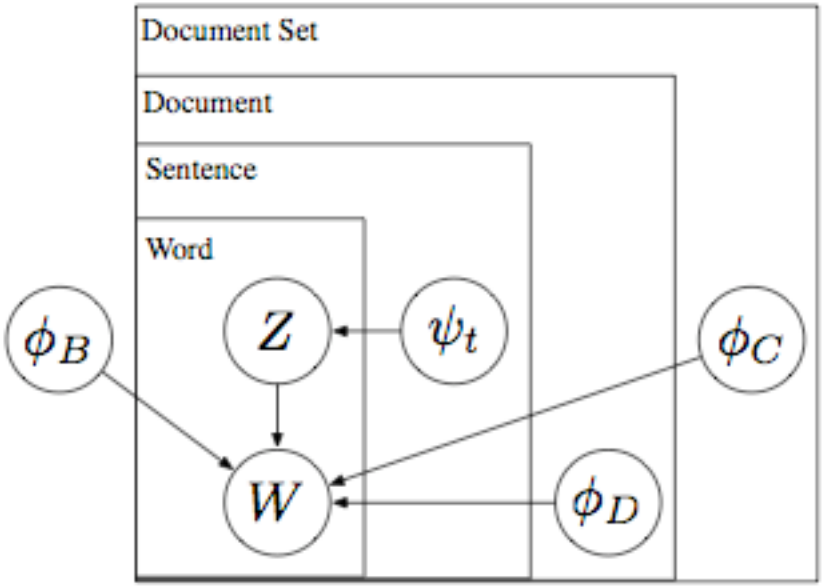



Figure 1: Graphical model depiction of TOPICSUM model (see section 3.3). Note that many hyperparameter dependencies are omitted for compactness.

[Haghighi & Vanderwende 2009]

# Probabilistic topic model: variant

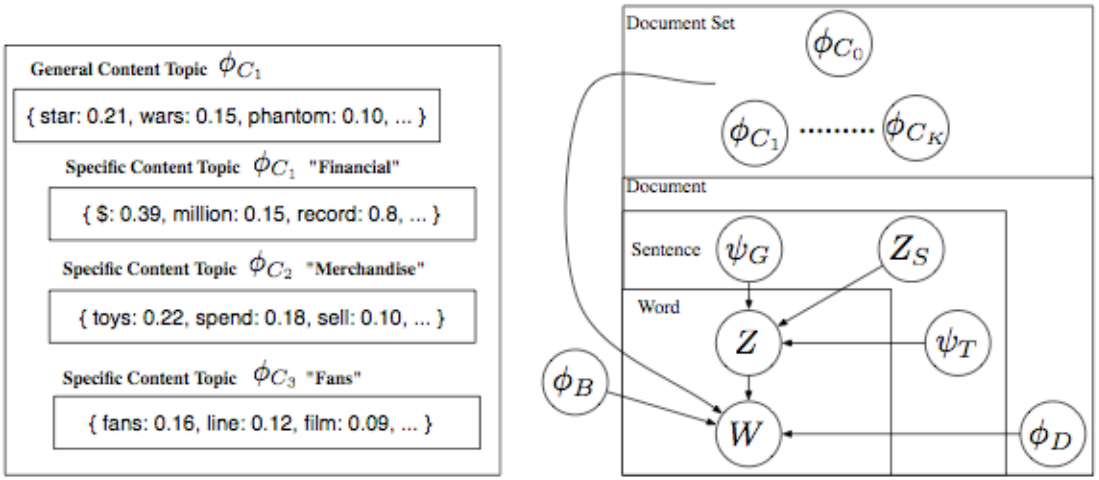
- We draw a general content distribution $\phi_{C0}$ from Dirichlet($V$, $\lambda_G$) representing the general content of the document set
- We draw specific content distributions $\phi_{Ci}$, for $i = 1, \ldots, K$ representing the specific sub-stories
- For each sentence *S* of each document *D*, we draw a binomial distribution $\psi_G$ determining whether a content word in the sentence will be drawn from the general or specific topic distribution
  - Reflecting the intuition that earlier sentences in a document describe the general content of a story, we bias $\psi_G$ BETA(5,2), preferring general content words, and every later sentence BETA(1,2)
  - BETA(a,b) represents the beta prior over binomial random variables with *a* and *b* being pseudo-counts for the first and second outcomes respectively

# Probabilistic topic model: variant

- If in the previous step a content specific content word is decided, we assume that each sentence draws a single specific topic $Z_s$ for every content word in the sentence
- Concretely, $Z_s$ for the first sentence in $D$ is drawn uniformly and each subsequence sentence's $Z_s$ will be identical to the previous sentence with probability $\sigma$ and with probability (1- $\sigma$ ) we select a successor topic from a learned transition distribution amongst 1, ..., K.

# Probabilistic topic model

- Distributions can be estimated with Gibbs sampling
- Content models allow to characterize the sentences of a document or set of documents based on which sentences that make up the summaries can be chosen (general for set, substory-specific, document-specific, etc.)




(a) Content Distributions (b) HIERSUM Graphical Model

Figure 2: (a): Examples of general versus specific content distributions utilized by HIERSUM (see section 3.4). The general content distribution $\phi_{C_0}$ will be used throughout a document collection and represents core concepts in a story. The specific content distributions represent topical 'sub-stories' with vocabulary tightly clustered together but consistently used across documents. Quoted names of specific topics are given manually to facilitate interpretation. (b) Graphical model depiction of the HIERSUM model (see section 3.4). Similar to the TOPICSUM model (see section 3.3) except for adding complexity in the content hierarchy as well as sentence-specific prior distributions between general and specific content topics (early sentences should have more general content words). Several dependencies are missing from this depiction; crucially, each sentence's specific topic $Z_S$ depends on the last sentence's $Z_S$.

[Haghighi & Vanderwende 2009]

# Text compaction for display on very small screens

- To show on smartphones, personal digital assistants (PDA), pocket PCs, …

  e.g., compaction of web pages

- Techniques:
  - often strong compression needed: e.g., words, sentences
  - keywords and single sentence summaries
  - often summaries expandable with different levels of detail


[Otterbacher et al. IP & M 2008]

# Word compression

- Removal of certain characters based on information theory: characters are predictable given context and therefore can be omitted
- Use of abbreviations
- Experiment with e-mails in different languages: humans could successfully decode

**Compacted output (English):**

PrblmOfAutmtcSmmrztnPssVrtyOfTghChllngsInBthNL Undrstndng&Gnrtn.

**Source text:**

The problem of automatic summarization poses a variety of tough challenges in both NL understanding and generation.

# Sentence compression

- The parse tree of a sentence allows deleting information from the sentence without harming grammatical correctness
- Approach: from parsed training examples of source sentences and their parsed summary sentences: learn reduction rules [Knight & Marcu AI 2002]:
  - features: reductions of elements of the parse tree and possibly features that indicate salience of words in the discourse

# Sentence compression

**ORIGINAL SENTENCE:**

```
(S1 (S (NP (NNP Lady) (NNP Hera))
    (VP (VBD was)
     (NP (NP (DT a)
             (ADJP (JJ jealous)
              (, ,)
              (JJ ambitious)
              (CC and)
              (JJ powerful))
             (NN woman))
      (SBAR (WHNP (WP who))
        (S (VP (VBD was)
          (ADVP (RB continually))
          (VP (VB irated)
           (PP (IN over)
            (NP (NP (NP (NNP Zeus) (POS ')) (NN pursuit))
             (PP (IN of)
                (NP (JJ mortal) (CC and) (JJ immortal) (NN woman)))))))))
    (. .)))
```

**REDUCED SENTENCE:**

**Lady Hera was a jealous, ambitio us and powerful woman .**

**RULE:**

**Removal of the deepest SBAR.**

# Text summarization results

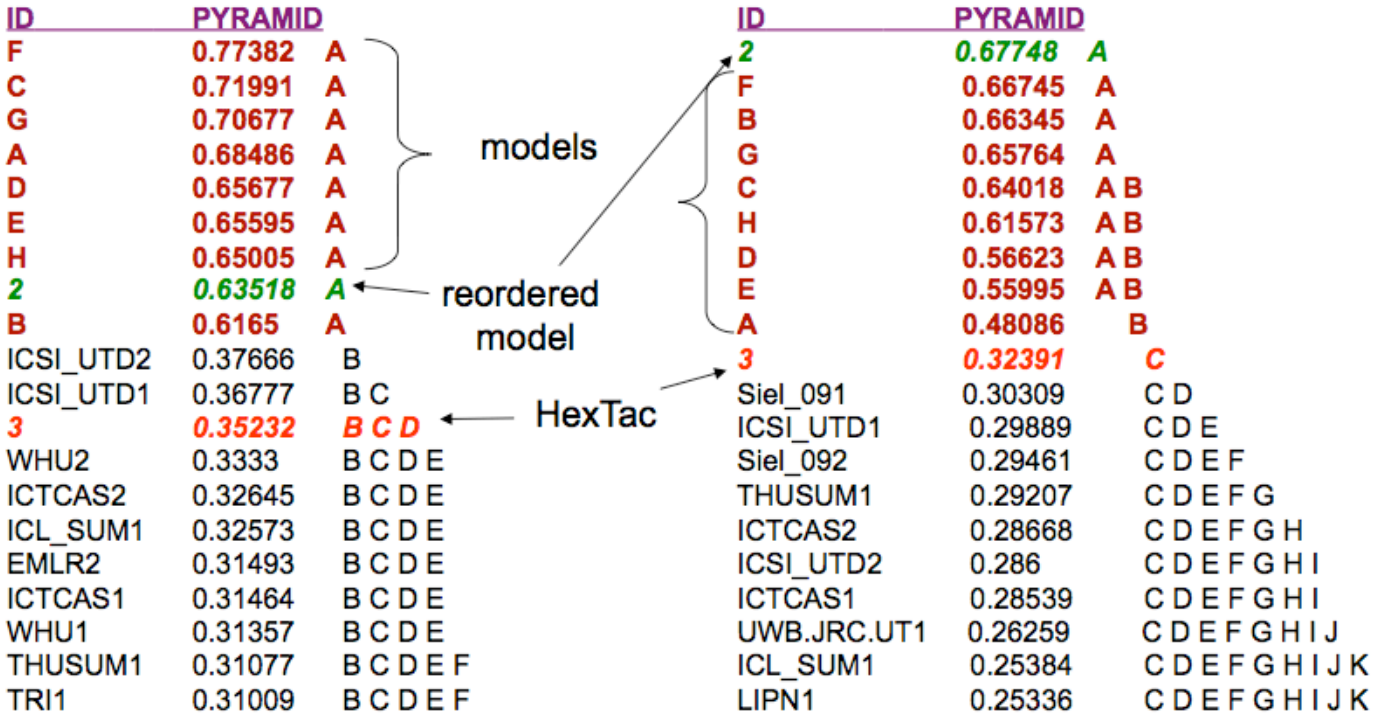
- Since 2008: international competition: TAC (Text Analysis Conference) organized by NIST, USA

- Evaluation of results of TAC
  - Content: by means of the pyramid method
  - Readability: judgments for grammaticality, coherence, referential clarity and organization
  - Responsiveness: judgments on how well the summary responds to the information need contained in the topic statement or query

# Results of summarization of TAC 2009

## Evaluation - Pyramid

[Dang & Owczarzak 2009]




**Initial summaries**

**Update summaries**

# Information presentation

- Summaries are often used in information presentation or for visualization of information:
  - E.g., text summarization results are transformed into visual summaries
  - In example below:
    - Latent Dirichlet Allocation extract the topic distributions of a text collection at different time moments (each topic is represented by a colored layer with keyword clouds)
    - Correlated topics are represented close to each other

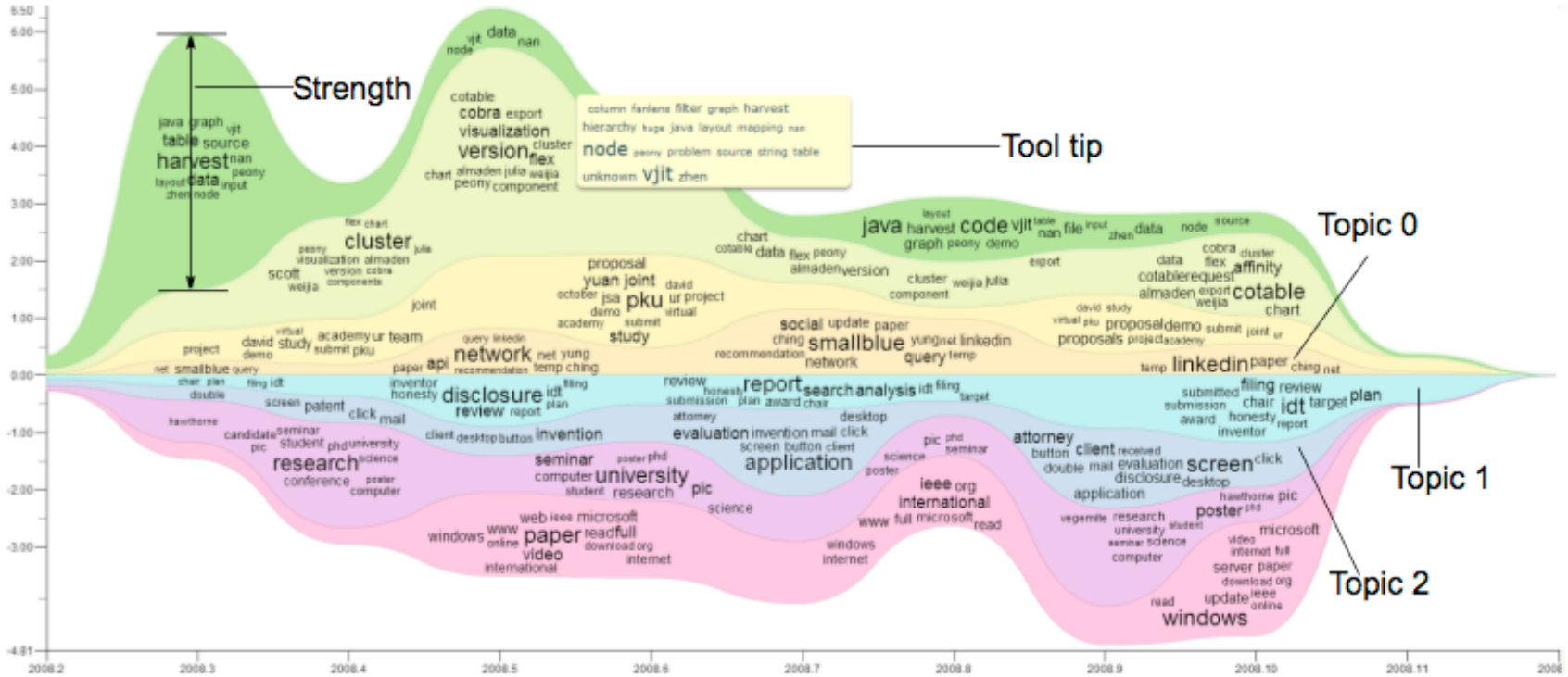



**Figure 1. Annotated TIARA-created visual summary of 10,000 emails in the year of 2008. Here, the x-axis encodes the time dimension, the y-axis encodes the importance of each topic. Each layer represents a topic, which is described by a set of keywords. These topic keywords are distributed along the time, summarizing the topic content and the content evolution over time. The tool tip shows the aggregated content of the top-most topic (green one).**

[Liu et al. 2009]

# What have we learned?

- Feature dependent extractive summarization
- Topic segmentation
- Probabilistic topic models for summarization
- Sentence compression

# Research questions to be solved

- Abstractive summarization both at a sentence and discourse level
- Sentence compression, paraphrasing and generation
- Cross-document, cross-lingual, cross-media synthesis of information
- Ordering of content when fusing information from different sources

## Further reading


Carbonell, J. & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* New York: ACM.

Chen, F. & Tsochantarides, I. (2002). Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of the 26th Annual International ACM Conference on Information and Knowledge Management.*

de Marneffe, M.-C., Rafferty, A.R. & Manning, C.D. (2008). Finding contradictions in text. In *Proceedings of ACL-08*: *HLT* (pp. 1039-1047). *Association for Computational Linguistics.*

Erkan, G. & Radev, D.R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research,* 22: 457-479.

Galley, M., McKeown, K., Fosler-Lussier, E. & Jing H. (2003). Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. ACL.

Hang, H. T. & Owczarzak, K. (2009). Overview of TAC 2009 summarization track. NIST.

Haghighi, A. & Vanderwende, L. (2009). Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual conference of the North American Chapter of the ACL* (pp. 362-370). East Stroudsburgh: ACL.

Hearst, M.A. (1997). TextTiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23 (1), 33-64.

Knight, K. & Marcu, D. (2002). Summarization beyond sentence extraction. A probabilistic approach to sentence compression. *Artificial Intelligence*, 139 (1).

Kupiec, J., Pedersen, J. & Chen, F. (1995). A trainable document summarizer. In E.A. Fox, P. Ingwersen, & R. Fidel (Eds.), *Proceedings of the 18th SIGIR Conference* (pp. 68-73). New York: ACM.

Liu, S., Zhou, M.X., Pan, S., Qian, W., Cai, W. & Lian, X. (2009). Interactive, topic-based visual text summarization and analysis. In *Proceedings of CIKM'09*. New York: ACM.

Luhn, H.P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development,* 2 (2), 159-165.

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge, MA: The MIT Press.

Mihalcea, R. & Rau, P. (2004). TextRank: Bringing order into texts. In *Proceedings of EMNLP* (pp. 404-411).

Moens, M.-F. (2006). Using patterns of thematic progression for building a table of content of a text. *Journal of Natural Language Engineering,* 14 (2), 145-172.

Moens, M.-F., Jeuniaux, P., Angheluta, R. & Mitra, R. (2006). Measuring aboutness of an entity in a text. In *Proceedings HLT-NAACL TextGraphs.*

Nenkova, A., Vanderwende, L. & McKeown, K. (2006). A compositional context sensitive multi-document summarizer: Exploring the factors that influence summarization. In *Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 573-580). New York: ACM.

Otterbacher, J., Radev, D. & Kareem, O. (2008). Hierarchical summarization for delivering information to mobile devices. *Information Processing & Management,* 44, 931-947.

Sparck Jones, K. (2007). Automatic summarizing: The state of the art. *Information Processing & Management*, 43, 1449-1481.

Wolf, F. & Gibson, E. (2006). *Coherence in Natural Language. Data Structures and Applications.* Cambridge, MA: The MIT Press.