IR-Related Tasks: Segmentation, Summarization, and Information Extraction

Lecture 16

CS 410/510

Information Retrieval on the Internet

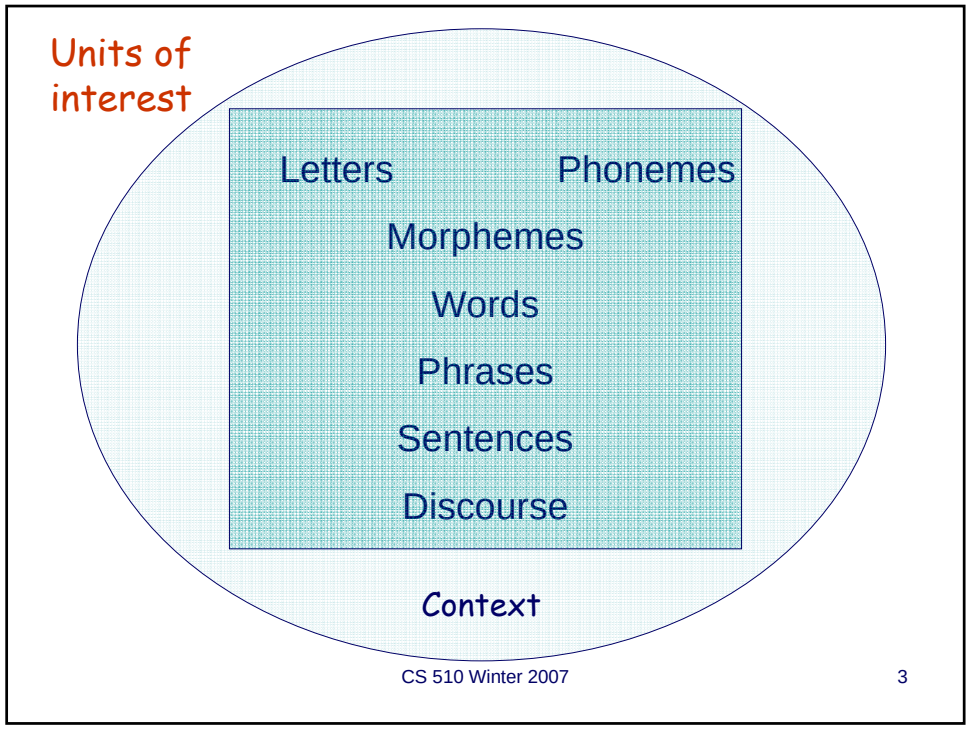
Natural Language Processing

"**Natural language processing (NLP)** ... studies the problems of automated generation and understanding of natural human languages. ... natural language understanding systems convert samples of human language into more formal representations that are easier for computer programs to manipulate."

– http://en.wikipedia.org/wiki/Natural_Language_Processing

Note: much of the material for this lecture is taken from:

Moens, Marie-Francine. *Automatic Indexing and Abstracting of Document Texts*. Kluwer Academic Publishers, Boston, 2000.



Natural Language Processing

- For IR-related tasks, mostly concerned with:
  - Morphological analysis: determining meaningful components of words
    - Stems, affixes, inflection, parts of compound words
  - Syntax: structural relationships between words
    - Parts of speech; phrases
  - Semantics: meaning of the text
  - Discourse issues: resolving relationships across sentences
    - Anaphoric references (e.g. use of pronouns)

Some useful steps (1)

- Tokenization (identify words, sentences)
- Part of speech tagging
  - Dictionary lookup
  - Probabilistic, based on specific word & its context
- Finding phrases
  - Syntactic templates (e.g. adjective followed by noun)
  - Context-free grammar with rules of allowable syntax
    - But multiple parses often possible
  - May normalize to recognize equivalent phrases
  - May be interested only in noun phrases

Some useful steps (2)

- Parsing
  - Full parsing
    - Attempt to get full global parse may degrade local choices
    - May produce ambiguous results – multiple possible parses
    - Works best in well-defined, constrained space
  - Partial parsing
    - Partition into fragments, such as phrases
    - Often tailored to domain or application needs
      - Recognize only segments defined in a grammar; ignore the rest
    - May map text directly into semantic representation structures without trying to identify syntactic elements (*semantic parser*)

Some useful steps (3)

- Named entity recognition
  - People, places, companies, etc.
  - Tools:
    - Lexicons
    - Gazetteers (dictionaries of geographical names)
    - Rules (e.g. capitalization) and context

Text structure

- Organization of discourse can be useful for analysis and manipulation of text
  - Schematic structure
  - Rhetorical structure
  - Thematic structure

Schematic structure

- Ordered text segments
  - Possibly signaled by cue phrases
- Characteristic structure for classes of text
  - Expository text, e.g. scientific articles
    - Purpose, methods, results, discussion, conclusion
  - Narrative text, e.g. news articles
- May be signaled by document structures
  - Titles, subtitles
- Can tailor analysis to document type

Rhetorical structure

- Rhetorical relations between sentences and clauses
- Often indicated by lexical cues, e.g.
  - Succession “I took the test *then* I went home.”
  - Conditionality “I’ll cook *if* you’ll wash dishes.”
  - Motivation “I did it *because* she asked me to.”
  - Circumstance “I bought it *when* I was in Iowa.”
  - Contrast “She is tall *but* her sister is short.”
- Can help identify important sentences, clauses

Thematic structure

- Topical organization
  - Often hierarchical with main topic and subtopics
- Can analyze at document and sentence level
- May see:
  - Sequential progression
  - Progressive detail (e.g. news stories)
  - Repetition
  - Suspension and resumption of a topic (semantic return)

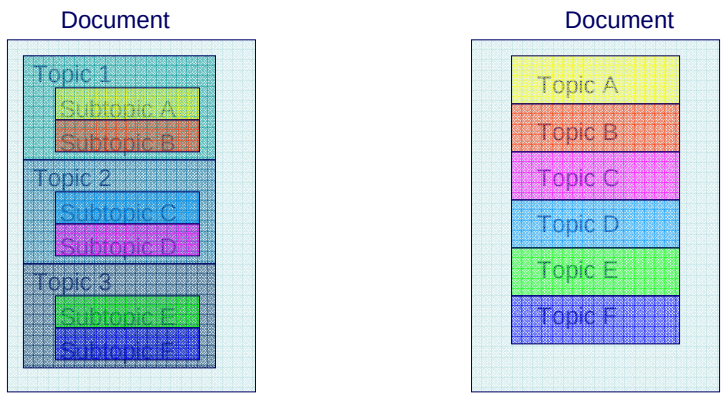
Thematic structure

- Schematic structure orders thematic content
- May coincide with presentation structure (chapters, paragraphs)
- Central theme often indicated in:
  - Sentence at beginning and end of document
  - First or last sentence of paragraph
- Words in text describe topics
  - Word location and frequency
- Linguistic cues can:
  - Indicate topic shifts
  - Indicate thematic roles of phrases, clauses

Text segmentation

- Dividing previously undifferentiated streams of text into components
  - e.g. streams of news stories
- Dividing documents into sections
  - Corresponding to subtopics
  - Use for retrieval or summarization
- Can be linear or hierarchical

Text segmentation



Hierarchical segmentation

Linear segmentation

Text Segmentation

- Define windows of text
  - Could be a sentence, *n* sentences, *m* words, other
- Formulate a "description" for each window
  - Words, phrases, possibly expanded
    - e.g. by using query expansion techniques to find related words and phrases to expand text in original window
- Calculate a similarity between pairs of windows
  - All pairs or pairs within a fixed distance
- Assign boundaries
  - Look for large changes in similarity
- Note: similar to a clustering task except clusters consist of adjacent text

TextTiling: Segmentation example

- Tokenizes text
  - Uses token-sequences of *w* tokens not sentences
- Calculates "lexical score" for gaps between adjacent units: 3 methods
  - Score based on term frequency comparisons between adjacent blocks (windows)
    - Groups *k* token-sequences (of *w* tokens each) into blocks
    - Slides windows by one token-sequence
  - Score based on number of new terms in adjacent token-sequences (vocabulary introduction)
  - Score based on number of active lexical chains
    - Repetitions of morphological variants of same term

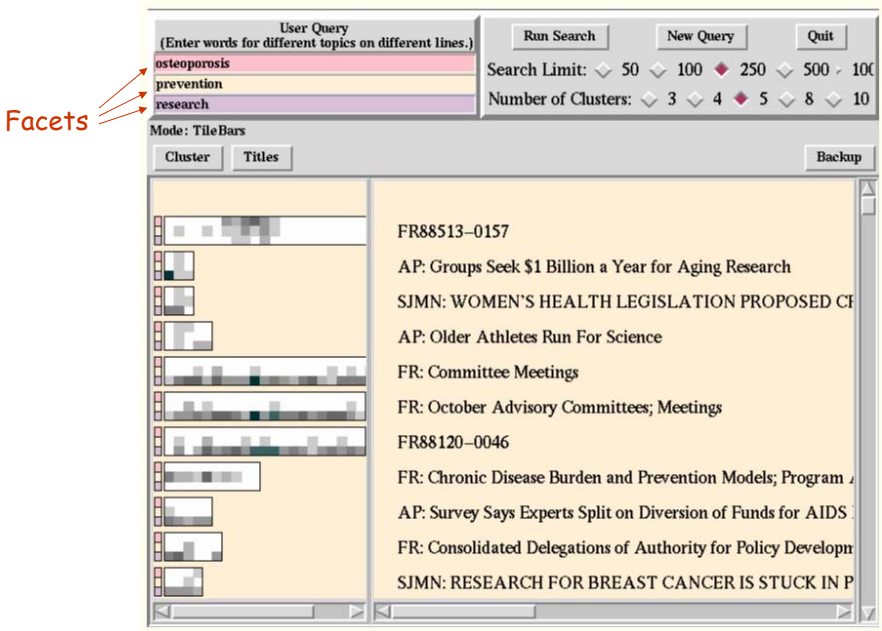
TextTiling: Segmentation example

- Assigns boundaries at “valleys” of lexical score
  - Depth score measures distance between peak and valley (with some smoothing)
- Results with block algorithm reasonably good
- Applied technique to IR by segmenting documents, querying against segments
  - Used a TREC collection
  - Found summing scores of segments better than using whole-document scores for ranking documents

TileBars: Using segmentation

- Displays information about frequency of query terms in document segments
  - Also document length (# segments)
- Input is faceted query
  - Implicit OR among terms for a given facet
  - Implicit AND between facets
- Rows of boxes represent segments
  - One box per segment
  - One row per query facet (set of terms)
- Gray scale represents term frequency
  - Darker boxes → more term occurrences
- User scans patterns and selects documents to view

TileBars example Screenshot (cropped) from:
http://www.ischool.berkeley.edu/~hearst/images/tb-example.html



Automatic summarization

- Two main paradigms
  - Extracts
    - Keywords
    - Sentences
  - Abstracts
    - Coherent generated representation
- Two main approaches
  - Statistical analysis (word frequency)
    - Shallow; little domain or linguistic knowledge needed
  - Knowledge-based
    - Uses natural language processing techniques
    - May use domain knowledge

Automatic summarization

- Important tradeoffs:
  - Brevity versus readability
  - Condensation versus possible loss of meaning/accuracy
- Can apply to single document or group of documents
- Summarization of multiple documents may involve finding repetition, possibly agreement and disagreement

Automatic summarization

- Steps in automatic summarization
  - Text analysis
    - Word frequency analysis
    - Parsing; structured knowledge representation
  - Transformation (condensation)
    - Selection
    - Generalization of content
      - May use thesaurus, ontology

girls with dolls
boys with trains
} children with toys

  - Synthesis (generation of summary text)

Statistical summarization: example

- Salton, et al. used vector-space similarity[1]
  - Discover abstract structure of text
  - Generate text summaries
- Idea: most texts composed of multiple functional (or semantic) units
- Paragraphs within a unit more highly linked than paragraphs in different units
- Used similarity to link paragraphs
  - Vocabulary overlap

[1]Salton et al. Automatic text structuring and summarization. Information Processing & Management 33(2), pp193-207, 1997.

Statistical summarization: example

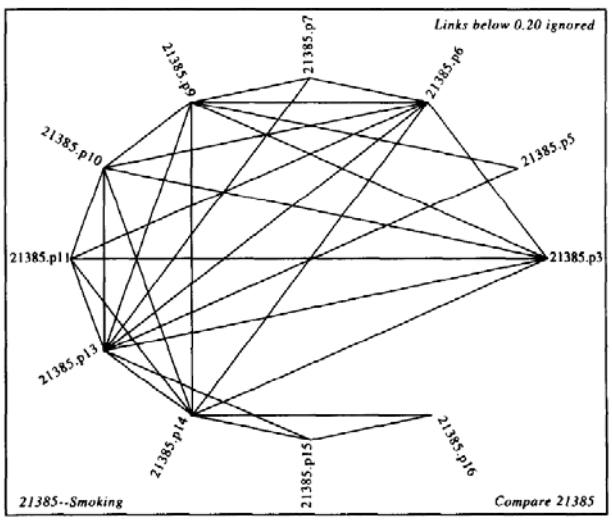


Fig. 2. Paragraph relationship map (smoking—16 paragraphs, 29 links above 0.20).

Paragraph relationship map

Figure from: Salton et al. Automatic text structuring and summarization. Information Processing & Management 33(2), pp193-207, 1997.

Statistical summarization: example

- Paragraph relationship map
  - Use threshold to control link number
  - Eliminate long-distance links
  - Produce small number highly-connected segments
- Number of links may indicate importance
- Extract important paragraphs from map
  - *n* bushiest nodes (*best performance)
  - Depth-first path
    - Start at first or most bushy node
    - Select next most similar node following current node
  - Bushy nodes from each segment
  - May need to add transition between segments

Knowledge-based summarization: example

- SUMMA summarization system[1]
- Analysis
  - Part of speech tagging, partial parsing
  - Main topic of sentences
  - Term distributions
  - Hierarchical topic tree
    - Desired summary length determines level of detail

[1]Moens et al. Generic technologies for single- and multiple-document summarization. Information Processing & Management 41, pp. 569-586, 2005

Knowledge-based summarization: example

- Transformation
  - Sentence condensation
    - Shortest phrase that spans most topic terms in a clause **[ ]**
    - Delete subclauses
    - Eliminate auxiliary verbs and determiners

"A Massachusetts **[**scientist ~~who has helped dispel the once widely held belief that schizophrenia came from bad mothering~~ ~~has~~ won ~~an~~ Albert Lasker award**]**" → *scientist won Albert Lasker award*

  - Eliminate redundant content

Knowledge-based summarization: example

- Generation of summary
  - Single documents:
    - Select first sentence of each topical segment
  - Multiple documents:
    - Cluster sentences of single document summaries (using term vectors)
    - Select most representative sentence in cluster for the summary

How do you know if a summary is good?

- Evaluation types:
  - Intrinsic: informativeness, coverage, correctness
  - Extrinsic: effectiveness for intended purpose
- Example: Salton paper (extracts)
  - Manually-extracted summaries (50 documents)
    - Two manual extracts per document
  - Reported amount of overlap between manual and automatic extracts
  - Average overlap of two ***manual*** extracts only 46%

How do you know if a summary is good?

- Example: Document Understanding Conference (DUC)
  - Manual comparison of system results to manually-created abstracts
  - Quality of summary
    - Grammaticality
    - Cohesion (sentences fit with surrounding sentences)
    - Organization/coherence (content expressed and arranged effectively)
  - Coverage
    - Compared content to human summary
  - Human judges found evaluation difficult

Information Extraction

- Extract pieces of information related to a particular scenario
- Example[1]: Scenario related to changes in company management

Text excerpt	Output template
A.C. Nielsen Co. said George Garrick, 40 years old, president of Information Resources Inc.'s London-based European Info Services operation, will become president of Nielsen Research, a unit of Dun & Bradstreet Corp. He succeeds John Costello, who resigned in March.	<Succession_event> = PersonIn: George Garrick PersonOut: John Costello Post: President Org: Nielsen Research

[1]Example is abbreviated from Turmo et al. Adaptive Information Extraction. ACM Computing Surveys 38(2), July 2006.

Information Extraction

- Task often represented as a template to be filled
- Structured summarization?
- Typically do only partial parsing of text
- Extraction phase is domain-specific
  - Often use pattern matching to extract elements
- May use machine learning techniques to learn extraction rules