CSC 9010: Text Mining Applications
Document Summarization

Dr. Paula Matuszek
Paula.Matuszek@villanova.edu
Paula.Matuszek@gmail.com
(610) 647-9789
Document Summarization

• Document Summarization

• Provide meaningful summary for each document

• Examples:
  • Search tool returns “context”
  • Monthly progress reports from multiple projects
  • Summaries of news articles on the human genome

• Often part of a document retrieval system, to enable user judge documents better

• Surprisingly hard to make sophisticated

• Surprisingly easy to make effective
Document Summaries: What

• Goal of summarization:
  • Indicative: what’s it about?
  • Informative: substitute for reading it

• Number of documents: single vs multi-document

• Desired summary:
  • paragraph
  • keywords
  • template
  • focused
  • update
Document Summarization -- How

- Two general approaches:
  - Abstractive: Capture in abstract representation, generate summary
    - Useful in well-defined domains with clearcut information needs.
    - Still a research topic in the more general case
  - Extractive: Extract representative sentences/clauses.
    - Useful in arbitrarily complex and unstructured domains; broadly applicable, and gets "general feel".
Abstractive: Capture and Generate

- Documents can have arbitrary format
- Knowledge needed is well-defined.
- Often information need is for summarizations across multiple documents
- Example:
  - Summarizing restaurant reviews. Take newspaper articles and produce price range, kind of food, atmosphere, quality, service.
Example Reviews

Minella's Diner

Directions | Write a review

Address: 320 W Lancaster Ave, Wayne, PA 19087
Phone: (610) 687-1575
Prices: $$$
Hours: Thursday Open 24 hours - See all
Menu: menupages.com

Reviews
3.7 ★★★★☆ 71 Google reviews

Minella's Main Line Diner:
Restaurant Reviews

320 Lancaster Ave, Wayne, PA 19087
(610) 687-1575
Website | Menu | Update restaurant details

Ranked #9 of 94 restaurants in Wayne
80 Reviews

Cuisines: American, Diner
Dining options: Breakfast/Brunch, Late Night
Most recent review
"Great diner / mainline staple"
- Mar 11, 2013

2 visitor photos

80 reviews from our community

62% Recommend

Excellent | Very good | Average | Poor | Terrible
20 | 30 | 21 | 4 | 5

Rating summary

Food | Service | Value | Atmosphere

copyright Paula Matuszek 2013
Capture and Generate: Methods

- State of the art:
  - Create "template" or "frame"
    - Represent the knowledge you want to capture
  - Extract Information to fill in frame
    - Standard information extraction problem
    - Typically relatively large frames with relatively few relations; mostly facts.
  - Generate based on template
    - Relatively simple "fill-in-the-blank"
    - More complex based on parse tree.
- Still basically research: parse entire document into parse tree tied to rich semantic net; apply rules to trim tree; generate continuous narrative.
Capture and Generate: Advantages and Disadvantages

• Advantages:
  • Produces very focused summaries.
  • Can readily incorporate multiple documents.
  • Not dependent on authors

• Disadvantages
  • Assumes information need is clearly defined.
  • Information extraction component development time is significant.
  • Document parsing slow; probably not real-time.

• Comment:
  • Makes no attempt to capture author's intent
Document format can be arbitrary
Document content can also be arbitrary; information need not clearcut
Summarization consists of text extracted directly from document.

Examples:
- Context returned by Google for each hit
- Google News summaries.
Find Representative Sentences: Method

- Typically, choose representative individual terms, then broaden to capture sentence containing terms. The more terms contained, the more important the sentence.

- If in response to a search or other information request, the search terms are representative.

- If no prior query, TF*IDF and other BOW approaches. May use pairs or n-ary groups of words.
Additional Features

- Other features can include
  - position
  - nearness to specific phrases such as “in summary” : cues

- Centrality: How similar is the sentence to the overall document? (Any of our similarity measures used for clustering)
Other Factors

- Once we have sentences, there are new aspects to consider:
  - Sentence Ordering
    - Can make a big difference in how good a summary is perceived to be
  - Sentence Revision
    - named entity substitution
    - combining information from multiple sentences: fusing
    - eliminating information from some sentences
Find Representative Sentences: Advantages and Disadvantages

- **Advantages**
  - Can be applied anywhere.
  - Relatively fast (compared to full parse)
  - Provides a good general idea or feel for content.
  - Can do multiple-document summaries.

- **Disadvantages**
  - Often choppy or hard to read
  - Does poorly when document doesn't contain good summary sentences.
  - Can miss major information
Multi-Document Summarization

• When we have multiple documents on a single subject, we might want:
  • one summary with all common information
  • similarities and differences among documents
  • support and opposition to specific ideas/concepts
Data-Driven Summarization

- Extend “representative sentences” approach.
- May also require some significant NLP
Frequency as indicator of importance

- The topic of a document will be repeated many times

- In multi-document summarization, important content is repeated in different sources
Greedy frequency method

- Compute word probability from input
- Compute sentence weight as function of word probability
- Pick best sentence
Sentence clustering for theme identification

1. PAL was devastated by a pilots' strike in June and by the region's currency crisis.

2. In June, PAL was embroiled in a crippling three-week pilots' strike.

3. Tan wants to retain the 200 pilots because they stood by him when the majority of PAL's pilots staged a devastating strike in June.

Based on slides from Ani Nenkova, www.cis.upenn.edu/~nenkova/Courses/cis430/Lecture01.ppt
• Cluster sentences from the input into similar themes

• Choose one sentence to represent a theme

• Consider bigger themes as more important

Based on slides from Ani Nenkova, www.cis.upenn.edu/~nenkova/Courses/cis430/Lecture01.ppt
Using machine learning

- Ask people to select sentences
- Use these as training examples for machine learning
  - Each sentence is represented as a number of features
  - Based on the features distinguish sentences that are appropriate for a summary and sentences that are not
- Run on new inputs

Based on slides from Ani Nenkova, www.cis.upenn.edu/~nenkova/Courses/cis430/Lecture01.ppt
Machine Learning Difficulties

- Getting the training data is expensive
- Human raters are not very consistent
  - Any of several sentences may be equally good
  - Judgment depends in part on focus and intent of rater
- Can be very domain-specific
  - TF*IDF scores vary widely across content domains
  - Features like position vary widely across format domains: chats vs newspaper vs scientific articles
Automatic summary edits

- Some expressions might not be appropriate in the new context
  - References:
    - he
    - Putin
    - Russian Prime Minister Vladimir Putin
  - Discourse connectives
    - However, moreover, subsequently
- Requires more sophisticated NLP techniques

Based on slides from Ani Nenkova, www.cis.upenn.edu/~nenkova/Courses/cis430/Lecture01.ppt
Before

Pinochet was placed under arrest in London Friday by British police acting on a warrant issued by a Spanish judge. Pinochet has immunity from prosecution in Chile as a senator-for-life under a new constitution that his government crafted. Pinochet was detained in the London clinic while recovering from back surgery.

Based on slides from Ani Nenkova, www.cis.upenn.edu/~nenkova/Courses/cis430/Lecture01.ppt
Gen. Augusto Pinochet, the former Chilean dictator, was placed under arrest in London Friday by British police acting on a warrant issued by a Spanish judge. Pinochet has immunity from prosecution in Chile as a senator-for-life under a new constitution that his government crafted. Pinochet was detained in the London clinic while recovering from back surgery.
Turkey has been trying to form a new government since a coalition government led by Yilmaz collapsed last month over allegations that he rigged the sale of a bank. Ecevit refused even to consult with the leader of the Virtue Party during his efforts to form a government. Ecevit must now try to build a government. Demirel consulted Turkey's party leaders immediately after Ecevit gave up.
Turkey has been trying to form a new government since a coalition government led by Prime Minister Mesut Yilmaz collapsed last month over allegations that he rigged the sale of a bank. Premier-designate Bulent Ecevit refused even to consult with the leader of the Virtue Party during his efforts to form a government. Ecevit must now try to build a government. President Suleyman Demirel consulted Turkey's party leaders immediately after Ecevit gave up.
Getting There in Some Domains

- http://newsblaster.cs.columbia.edu/
Some systems to try

- This system has demos showing several forms of summarization.
  - http://summarizer.intellexer.com

- These systems will take a search term and produce a summary:
  - http://ultimate-research-assistant.com/
  - http://iResearch-reporter.com (only one search without registering.)