Text Summarization: News and Beyond

Kathleen McKeown
Department of Computer Science
Columbia University
Today

- HW3 assigned
- Summarization (switch in order of topics)
- WEKA tutorial (for HW3)
- Midterms back
What is Summarization?

- Data as input (database, software trace, expert system), text summary as output

- Text as input (one or more articles), paragraph summary as output

- Multimedia in input or output

- Summaries must convey maximal information in minimal space
Types of Summaries

- Informative vs. Indicative
  - Replacing a document vs. describing the contents of a document
- Extractive vs. Generative (abstractive)
  - Choosing bits of the source vs. generating something new
- Single document vs. Multi Document
- Generic vs. user-focused
Types of Summaries

- **Informative** vs. **Indicative**
  - Replacing a document vs. describing the contents of a document
- **Extractive** vs. **Generative**
  - Choosing bits of the source vs. generating something new
- **Single document** vs. **Multi Document**
- **Generic** vs user-focused
Questions (from Sparck Jones)

- Should we take the reader into account and how?

- “Similarly, the notion of a basic summary, i.e., one reflective of the source, makes hidden fact assumptions, for example that the subject knowledge of the output’s readers will be on a par with that of the readers for whom the source was intended. (p. 5)”

- Is the state of the art sufficiently mature to allow summarization from intermediate representations and still allow robust processing of domain independent material?
Foundations of Summarization – Luhn; Edmundson

- Text as input
- Single document
- Content selection
- Methods
  - Sentence selection
  - Criteria
Sentence extraction

- Sparck Jones:
  - `what you see is what you get`, some of what is on view in the source text is transferred to constitute the summary
Luhn 58

- Summarization as sentence extraction
  - Example

- Term frequency determines sentence importance
  - TF*IDF
  - Stop word filtering
  - Similar words count as one
  - Cluster of frequent words indicates a good sentence
TF*IDF

- Intuition: Important terms are those that are frequent in this document but not frequent across all documents
Term Weights

- Local weights
  - Generally, some function of the frequency of terms in documents is used
- Global weights
  - The standard technique is known as inverse document frequency

\[ \text{idf}_i = \log \left( \frac{N}{n_i} \right) \]

N = number of documents; ni = number of documents with term i
**TFxIDF Weighting**

- To get the weight for a term in a document, multiply the term’s frequency derived weight by its inverse document frequency.

  \[ TF \times IDF \]
Edmunson 69

Sentence extraction using 4 weighted features:

- Cue words (“In this paper..”, “The worst thing was ..”)
- Title and heading words
- Sentence location
- Frequent key words
Sentence extraction variants

- Lexical Chains
  - Barzilay and Elhadad
  - Silber and McCoy

- Discourse coherence
  - Baldwin

- Topic signatures
  - Lin and Hovy
Lexical Chains

- “Dr.Kenny has invented an anesthetic machine. This device controls the rate at which an anesthetic is pumped into the blood.“
- “Dr.Kenny has invented an anesthetic machine. The doctor spent two years on this research.“
- Algorithm: Measure strength of a chain by its length and its homogeneity
  - Select the first sentence from each strong chain until length limit reached
- Semantics needed?
Discourse Coherence

- Saudi Arabia on Tuesday decided to sign...
- *The official Saudi Press Agency reported that King Fahd made the decision during a cabinet meeting in Riyadh, the Saudi capital.*
- The meeting was called in response to ... the Saudi foreign minister, that the Kingdom...
- An account of the Cabinet discussions and decisions at the meeting...
- The agency...
- It
**Topic Signature Words**

- Uses the log ratio test to find words that are highly descriptive of the input
- the log-likelihood ratio test provides a way of setting a threshold to divide all words in the input into either descriptive or not
  - the probability of a word in the input is the same as in the background
  - the word has a different, higher probability, in the input than in the background
- Binomial distribution used to compute the ratio of the two likelihoods
- The sentences containing the highest proportion of topic signatures are extracted.
Summarization as a Noisy Channel Model

- Summary/text pairs
- Machine learning model
- Identify which features help most
To summarize is to reduce in complexity, and hence in length while retaining some of the essential qualities of the original. This paper focusses on document extracts, a particular kind of computed document summary. Document extracts consisting of roughly 20% of the original can be as informative as the full text of a document, which suggests that even shorter extracts may be useful indicative summaries. The trends in our results are in agreement with those of Edmundson who used a subjectively weighted combination of features as opposed to training the feature weights with a corpus. We have developed a trainable summarization program that is grounded in a sound statistical framework.
Statistical Classification Framework

- A training set of documents with hand-selected abstracts
  - Engineering Information Co provides technical article abstracts
  - 188 document/summary pairs
  - 21 journal articles
- Bayesian classifier estimates probability of a given sentence appearing in abstract
  - Direct matches (79%)
  - Direct Joins (3%)
  - Incomplete matches (4%)
  - Incomplete joins (5%)
- New extracts generated by ranking document sentences according to this probability
Features

- Sentence length cutoff
- Fixed phrase feature (26 indicator phrases)
- Paragraph feature
  - First 10 paragraphs and last 5
  - Is sentence paragraph-initial, paragraph-final, paragraph medial
- Thematic word feature
  - Most frequent content words in document
- Upper case Word Feature
  - Proper names are important
Evaluation

- Precision and recall
- Strict match has 83% upper bound
  - Trained summarizer: 35% correct
- Limit to the fraction of matchable sentences
  - Trained summarizer: 42% correct
- Best feature combination
  - Paragraph, fixed phrase, sentence length
  - Thematic and Uppercase Word give slight decrease in performance
Questions (from Sparck Jones)

- Should we take the reader into account and how?

- “Similarly, the notion of a basic summary, i.e., one reflective of the source, makes hidden fact assumptions, for example that the subject knowledge of the output’s readers will be on a par with that of the readers for whom the source was intended. (p. 5)”

- Is the state of the art sufficiently mature to allow summarization from intermediate representations and still allow robust processing of domain independent material?