Recent Trends in Text Summarization

Ali Asad Lotia
lotia@umich.edu

University of Michigan
LING 541
Fall 2005 Term

December 8 2005
What's the problem?

• Availability of information becoming increasingly greater
• Consumers of information overwhelmed by the amount of available information
• Some way needed to reduce the burden on information consumers
• Enter text summarization
What is a summary?

- Text that contains information from one or more source texts, but shorter than original
  - At most 50% (Radev et al. 2002) of the length of the original.
- Core ideas of original text are retained, dispensing with peripheral information, yielding shorter version
- Possible since information content is “bursty” (Radev et. al)
Types of summaries and basic approaches to summarization

- Summary types:
  - Indicative
  - Informative
  - Topic-oriented
  - Generic

- Approaches to summarization:
  - Extraction
  - Abstraction

- Single and multiple document summaries
Why use computers to do it?

- Experts needed for each subject area
- Speed - Summarization takes time
- Scalability - Increase in information volume requires more manpower
- Consistency in selection criteria
- On demand summarization
- Computers better at producing machine parsable formats
Early approaches

● Luhn 1958
  – Generated abstracts of news articles sentence ranking
  – Used occurrence of significant words as major criterion for sentence rank calculation

● Baxendale 1958, refined by Edmundson
  – Topic sentences
  – Phrase selection

● Summarization very domain specific at that time
Discourse models

- Addressing different knowledge domains
  - Analysis of discourse structure employed to make summarization less domain specific. (Hobbs, 1979) (Ono et al. 1994)
  - Discourse models used to locate argumentative content in text (Ono)
  - Inter-paragraph and Intra-paragraph analysis
  - 74% of most important sentences represented in extracted summaries using Ono et al. model.
Machine learning

• Application of machine learning Kupiec et al. (1995)
  – Used human generated abstracts in training corpus
  – Used journals from several fields to widen knowledge domain
  – System analyzes features in original text that result in their selection for the summary.
  – Human generated summaries in training data compared with machine generated summaries and system tuned to give high match percentage
Lexical Chains

- Sets of words and their various senses present in text
- Brunn et al. (2001) and Alemany and Fort (2003) recent implementations
- Select candidate words/phrases from parsed output
- Used WordNet to obtain different senses of selected words
- Sentences containing segments with most chain members selected for summary and segments
Summarization for small displays

- Proliferation of net enabled mobile devices poses display challenges
- Buykkokten et al. (2001) summarize web pages to display in a custom browser running on a Palm
- Pages divided into Semantic Textual Units (STU) and different sized summaries created for each STU
- Similarly for forms, selecting only required fields for initial summary
Multi-document summarization

- Challenging because styles of documents may be different even when dealing with similar topics
- Similarity in theme/focus hard to determine
- Radev et al. (2000) developed MEAD which uses “cluster centroids”
- Cluster centroids are words central to all articles in a cluster of documents
- MEAD different from previous approaches to multi-document efforts
Multi-document summarization (2)

- MEAD used to build NewsInEssence
  http://www.newsinessence.com
- Columbia system developed by McKeown et al. (2001) employs multiple strategies and selects the most appropriate one for the type of task
- *Router* used to select between several pre-defined categories of relatedness
Headline generation

- Strictly extractive
- *Headlines* distinct from other written text
- Zajic and Dorr (2002) use HMM to select appropriate story words from body text
- Model created using corpus of newspaper stories
  - Employs unigram model of “story words”
  - Bigram model of headlines
Evaluation of summarization

- Challenging task because human judges differ in scores assigned to same summaries.
- Addressed by Radev et al. (2003)
- Sentence co-selection
  - Precision and recall of co-selected sentences
- Content based Similarity
  - More fine grained than sentence level
- Relevance correlation
  - Decrease in document retrieval when using summary index instead of article index
Limitations of Summarization

- Abstraction not as easy to do since it requires semantic understanding of text
- Machine generated summaries still not as well composed as human generated ones
- Shorter documents harder to summarize because of scarcity of material to extract from
- Better evaluation methods need to be devised, particularly for multi-document systems