Automatic Text Summarization

Part I: Orientation

Summarization Everywhere

- Choose a book, turn a “dial” to 2 pages, read the summary
- News headlines
- Abstracts of research papers
- Answers in examinations?

  “unnecessarily long answers will not be corrected”

What is Summarization?

- To take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs.
- Input: one / more source documents
- Output: one summary document

Human Summarization

- Humans are often excellent summarizers
- Summarization – an art?
- Quoting Ashworth:

  "…To take an original article, understand it and pack it neatly into a nutshell without loss of substance or clarity presents a challenge which many have felt worth taking up for the joys of achievement alone. These are the characteristics of an art form…"

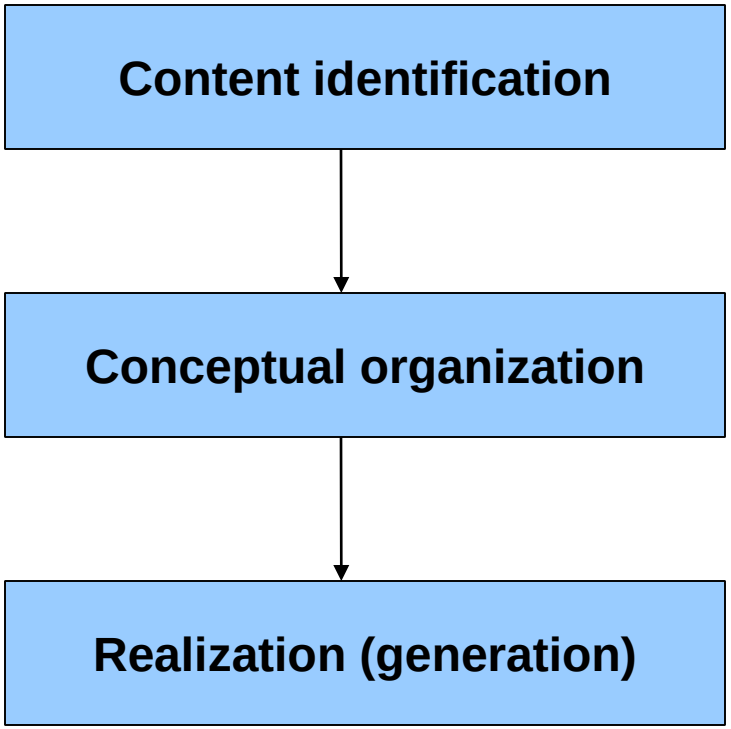
So Why Automatic Summarization?

- Human summarization can be person-specific, context-dependent, varies with human cognition
- Information overload!
- Targeting different audiences and different types of applications
  - Experts / novices
  - Google News, Q-A systems, …

Summary Types and Genres

- Types
  - Form: extract / abstract
  - Dimensions: single / multi-document
  - Context: query-specific / independent
  - Purpose: indicative / informative / critical
- Genres
  - News headlines, minutes, abridgments, movie summaries, chronologies, …

Basic Stages in Summarization



Top-down / Bottom-up Summarization

Top-down

- “I know what I want; give me what I ask for”.
- User needs: only certain types of information
- Particular criteria of interest for focused search
- Templates, term lists

Bottom-up

- “I’m curious to know what’s there in the text”.
- User needs: anything that’s important
- Generic information metrics
- Connectedness of sentences, word frequencies

Summarization Approaches

- Statistical / IR based Approach
  - Operate at lexical level, use word frequencies, similarity measures, etc.
  - Does not support abstraction.
- NLP / IE based Approach
  - Try to “understand” text. Needs rules for text analysis and manipulation.
  - Higher quality, supports abstraction.

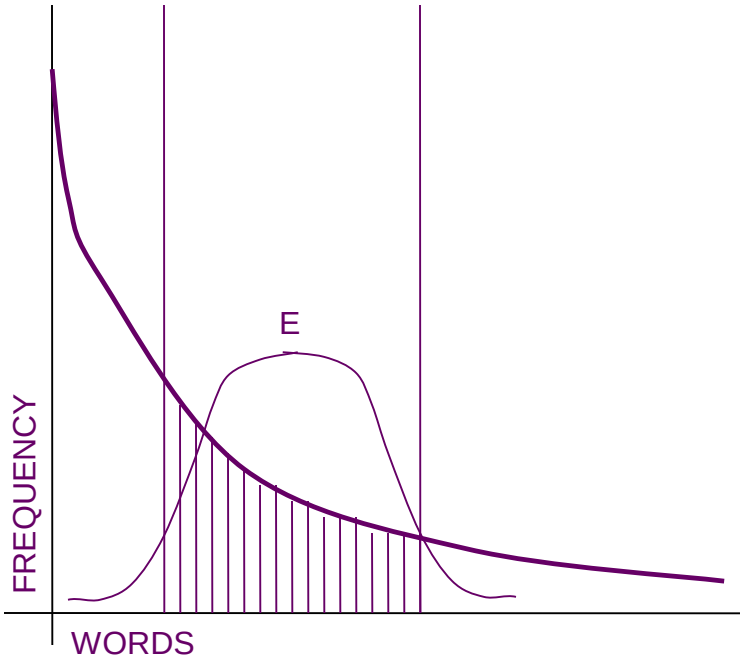
Talk Outline

- Motivation
- Basic concepts in automatic summarization
- Statistical / IR based approaches
- NLP / IE based approaches
- Abstract generation
- Summary evaluation
- Concluding remarks

Part II: Statistical based Approaches

Exploiting Word-frequency Information

- High frequency words are related to the topic of the document
- Of course, this does not include stopwords
- Importance of sentence depends of
  - Number of occurrences of significant words
  - Discriminating power of the words
- Rank sentences and pick the top k



Resolving power of significant words

Using Cue words

- Some words/phrases positively correlated to summary
  - eg. *important*, *to conclude*
- Some words/phrases negatively correlated to summary
  - eg. *for example, exception*

Exploiting Document Structure

- Information from Structure
  - Title words
  - Section, sub-section heading words
- Information from Position
  - Genre dependent
  - First sentence of document, first sentence of paragraph, last sentence of document, etc.

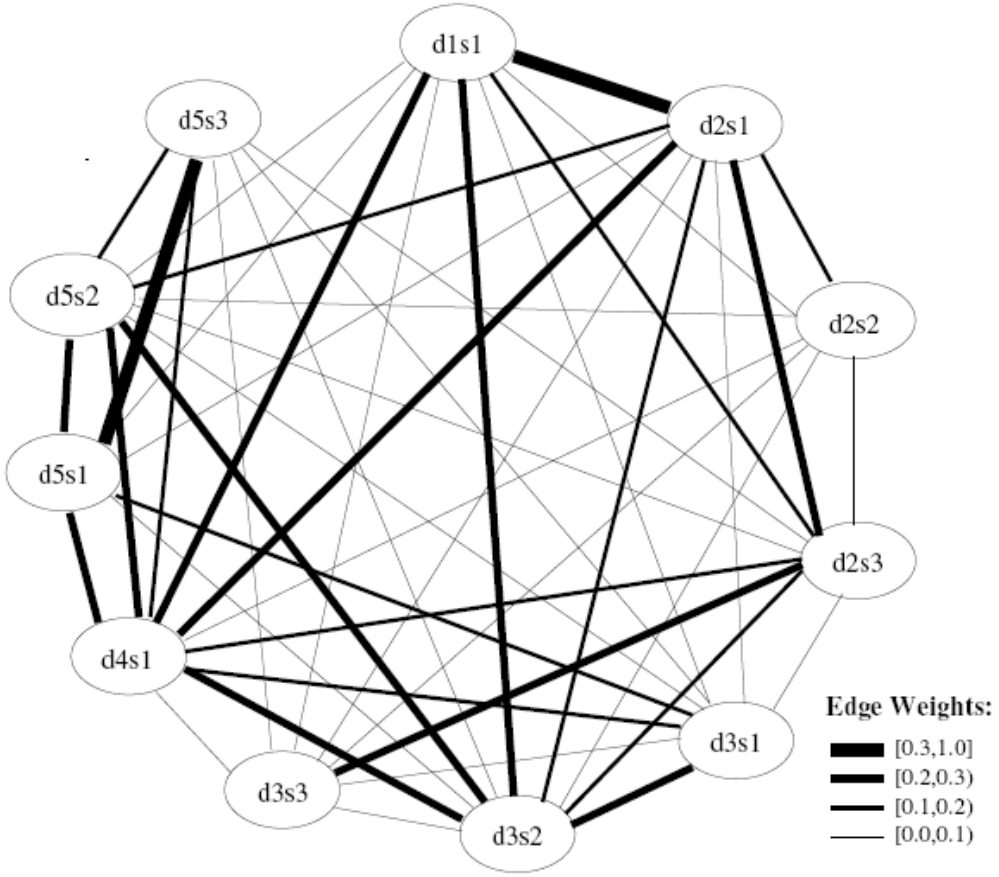
Graph Based Methods

- **Key Idea**: Summarizing sentences are well connected to other sentences
- Connectivity based on similarity with other sentences
- Similarity measure: tf-idf could be used
- Graph G (V,E)
  - V: set of sentences
  - E: similarity between sentences > threshold

Degree of Centrality

- Rank sentences by their degree
- Pick top k as summarizing sentences
- Sensitive to distortion by 'rogue' sentences

Sentence Clusters based on Similarity



LexRank

- Inspired by PageRank
- Value connections from highly connected neighbours
- Random Markov Walk over the graph
- $LR(u) = \sum LR(v) / deg(v)$

  where v is a neighbour of u

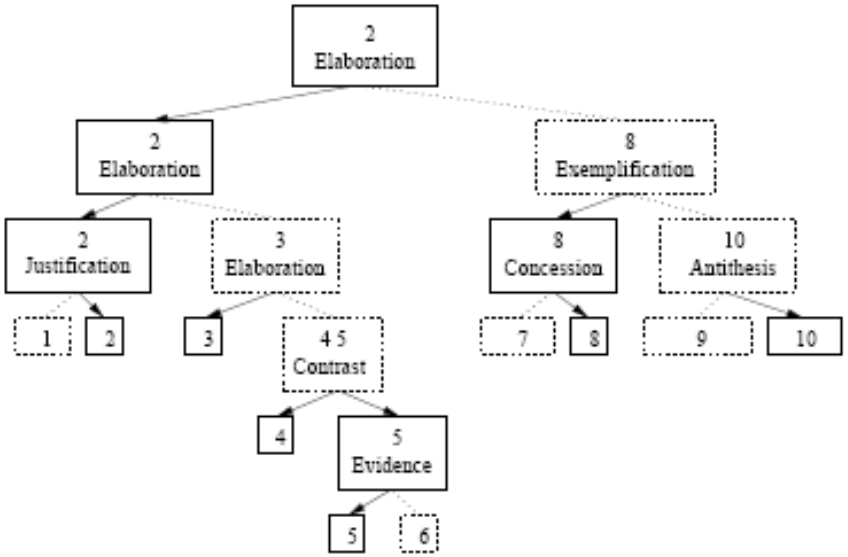
Part III: NLP based Approaches

Rhetoric based Summarization

- Rhetoric Relation
  - Between two non overlapping spans of text
  - Nucleus : core idea
  - Satellite : arguments to favor core idea
- Rhetoric relation is a relation between Nucleus and Satellite.

  E.g. Justification, elaboration, contrast, evidence, etc.

Rhetoric based Summarization (2)



Rhetoric Structure Tree

[With its distant orbit — 50 percent farther from the sun than Earth — and slim atmospheric blanket,[1]] [Mars experiences frigid weather conditions.[2]] [Surface temperatures typically average about −60 degrees Celsius (−76 degrees Fahrenheit) at the equator and can dip to −123 degrees C near the poles.[3]] [Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion,[4]] [but any liquid water formed in this way would evaporate almost instantly[5]] [because of the low atmospheric pressure.[6]]

[Although the atmosphere holds a small amount of water, and water-ice clouds sometimes develop,[7]] [most Martian weather involves blowing dust or carbon dioxide.[8]] [Each winter, for example, a blizzard of frozen carbon dioxide rages over one pole, and a few meters of this dry-ice snow accumulate as previously frozen carbon dioxide evaporates from the opposite polar cap.[9]] [Yet even on the summer pole, where the sun remains in the sky all day long, temperatures never warm enough to melt frozen water.[10]]

Rhetoric based Summarization (3)

- Summarization Method
  - Generate Rhetoric Structure Tree
    - Because of rhetoric ambiguity there are multiple trees
  - Pick best tree using
    - Clustering-based metric
    - Shape-based metric etc.
  - Pick up top K nodes nearest to the root, where K is no. of sentences expected in summary

Wordnet based Summarization

- Pick up a subgraph of wordnet
  - Mark each word in wordnet
  - Traverse hyperymy direction up to suitable level and mark intermediate nodes
  - Mark synsets

Wordnet based Summarization (2)

- Ranking Synsets
  - R : Vector of nodes in subsidized wordnet graph
  - A : Square matrix of size |R|x|R|
    - A[i][j] = 1/predecessors(j) if j is descendant of i
      = 0 otherwise
  - Repeat R_new = R_old * A /| R_old*A | until R_new becomes small enough

Wordnet based Summarization (3)

- Sentence Selection
  - Matrix R : sentences Vs nodes of subsidized wordnet
  - R[i][j] = R[j] if node j of graph is reached from words of sentences of sentence i

    = 0 othewise

Wordnet based Summarization (4)

- PCA
  - Take eigen value decomposition of matrix R
  - Order eigen vectors on decreasing value of its corresponding eigen values
  - Project sentences on eigen vectors
  - Pick up top N_sentences sentences for particular eigen vector based on their projection on that eigen vector
    - Where $N\_sentences = \lambda(i) / \sum \lambda(j) * N$

Part IV: Abstraction Summarization

Extraction Summarization: Pros and Cons

- Lack of fluency and coherence
- Anaphora: presence of pronouns and undefined references
- Multi-doc summarization: possible contradiction between sources

Abstraction Summarization

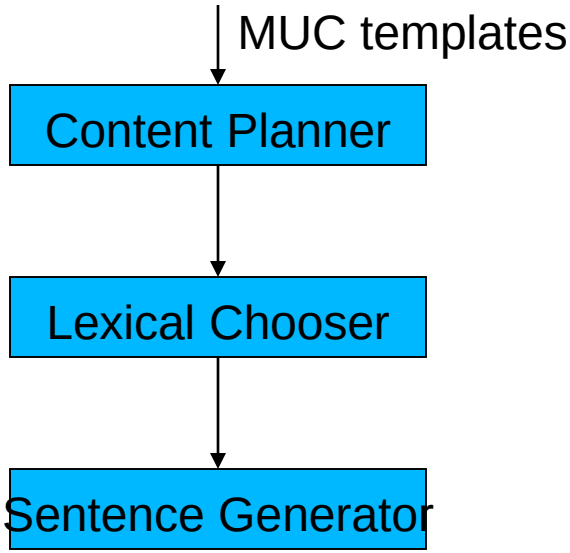
- Motivation
- Steps
  - Topic identification
  - Topic interpretation
  - Summary generation

Topic Interpretation

- Concept generalization
  - John bought some apples, pears and orange
  - John bought some fruits
- Topic Signature
  - TS=[head,(w1,s1),(w2,s2),...]
  - [restaurant-visit,(eat,s(eat)), (table,s(table)), (pay,s(pay)),..]

Summary Generation

- Conceptual processing
  - Content / Paragraph planner
- Linguistic processing
  - Lexical chooser
  - Sentence generator



SUMMONS Architecture

Part V: Summary Evaluation

Criteria for Summary Evaluation

- Fluency / coherence
- Informativeness
- Compression ratio

Evaluation methods

- Intrinsic – with summary itself
  - Reference summary
  - Summarization input
    - Semantic
    - Surface
- Extrinsic – with other task which uses the summary

Concluding Remarks

- Large amount of research in the field
- More maturity in extraction summarization
- Evaluation is difficult

References

- *The Automatic Creation of Literature Abstracts,* HP Luhn, IBM Journal of Research and Development, 1958
- *New Methods in Automatic Extracting*, HP Edmundson, Journal of the ACM, 1969
- *LexRank: Graph-based Lexical Centrality as Salience in Text Summarization,* G Erkan and D R Radev, Journal of Artificial Intelligence Research, 2004
- *Generic Text Summarization using WordNet*, Kedar Bellare, Anish Das Sarma, Atish Das Sarma, Navneet Loiwal, Vaibhav Mehta, Ganesh Ramakrishnan, Pushpak Bhattacharyya, LREC 2004, Barcelona, 2004
- *Generating Natural Language Summaries from Multiple On-Line Sources*, Dragomir R. Radev, Kathleen R. McKeown, Journal of Computational Linguistics, 1998
- *Summarization Evaluation: An Overview*, Inderjeet Mani, NAACL, 2001
- *Automated Text summarization and the SUMMARIST SYSTEM*, Eduard Hovey. and Chin-Yew Lin, 1998

References (2)

- *Identifying Topics by Position*, Lin, C-Y. and E.H. Hovy, In Proceedings of the Applied Natural Language Processing Conference (ANLP-97), 283\u2013290. Washington, 1997
- *Improving summarization through rhetorical parsing tuning,* Daniel Marcu, 1998.
- *Text Summarization Portal*: http://www.summarization.com/
- *Rhetoric Structure Theory*: http://www.sfu.ca/rst/