Text Summarization

Text Summarization

What is a summary?

- A summary is a text that is produced from one or more texts, that contains a significant portion of the information in the original text(s), and that is no longer than half of the original text(s). (*Hovy, 2008)*

Text Summarization

What is a summary?

- A summary is a text that is produced from one or more texts, that contains a significant portion of the information in the original text(s), and that is no longer than half of the original text(s). (*Hovy, 2008)*

What is text summarization?

- Text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user or task. (*Mani and MayBury, 2001)*

Automatic Text Summarization

Goal of a Text Summarization System:

- To give an overview of the original document in a shorter period of time.

Summarization Applications:

- *outlines or abstracts of any document, news article etc.*
- *summaries of email threads*
- *action items from a meeting*
- *simplifying text by compressing sentences*

Automatic Text Summarization

Types of Summary

- **Extract vs. Abstract**
  - Extractive: *lists fragments of text*
  - *Abstractive: re-phrases content coherently*
- **Single document vs. Multi-document**
  - *. . .based on one text vs. fuses together many texts.*
- **Generic vs. Query-focused**
  - *. . .provides author's view vs. reflects user's interest.*
- Query-focused summarization can be thought of as a complex question answering system

Summarization: Main stages

- Content Selection
  - Choose sentences to extract from the document

Summarization: Main stages

- Content Selection
  - Choose sentences to extract from the document
- Information Ordering
  - Choose an order to place them in the summary

Summarization: Main stages

- Content Selection
  - Choose sentences to extract from the document
- Information Ordering
  - Choose an order to place them in the summary
- Sentence realization
  - Simplify the sentences

Summarization: Main stages

- Content Selection
  - Choose sentences to extract from the document
- Information Ordering
  - Choose an order to place them in the summary
- Sentence realization
  - Simplify the sentences
- Removing Redundancy
  - Increase diversification by removing redundant sentences

Unsupervised content selection; Luhn (1958)

- Intuition
  - Choose sentences that have salient or informative words
- Two approaches to define salient words
  - *tf-idf: weigh each word w_i in document j by tf-idf*

    weight(w_i) = $tf_{ij} \times idf_i$
  - *Topic signatures: choose a smaller set of salient words, specific to that* domain

    weight(w_i) = 1 if w_i is a specific term (use mutual information)
- Weighing a sentence:

$$weight(s) = \frac{1}{|S|} \sum_{w \in S} weight(w)$$

LexRank: A Graph-based approach

Text Document

Computation is a process following a well defined model ...
A computation can be seen as a purely physical phenomena ...
...

processing

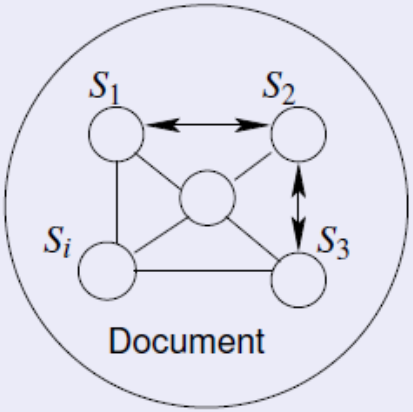
$S_1 \rightarrow \{(computation, 0.1), (process, 0.15), \ldots\}$
$S_2 \rightarrow \{(computation, 0.1), (seen, 0.05), \ldots\}$
$S_3 \rightarrow \ldots$

Machine-readable format

Document Representation

Underlying Hypothesis

Sentences that convey the theme of the document are more similar to each other

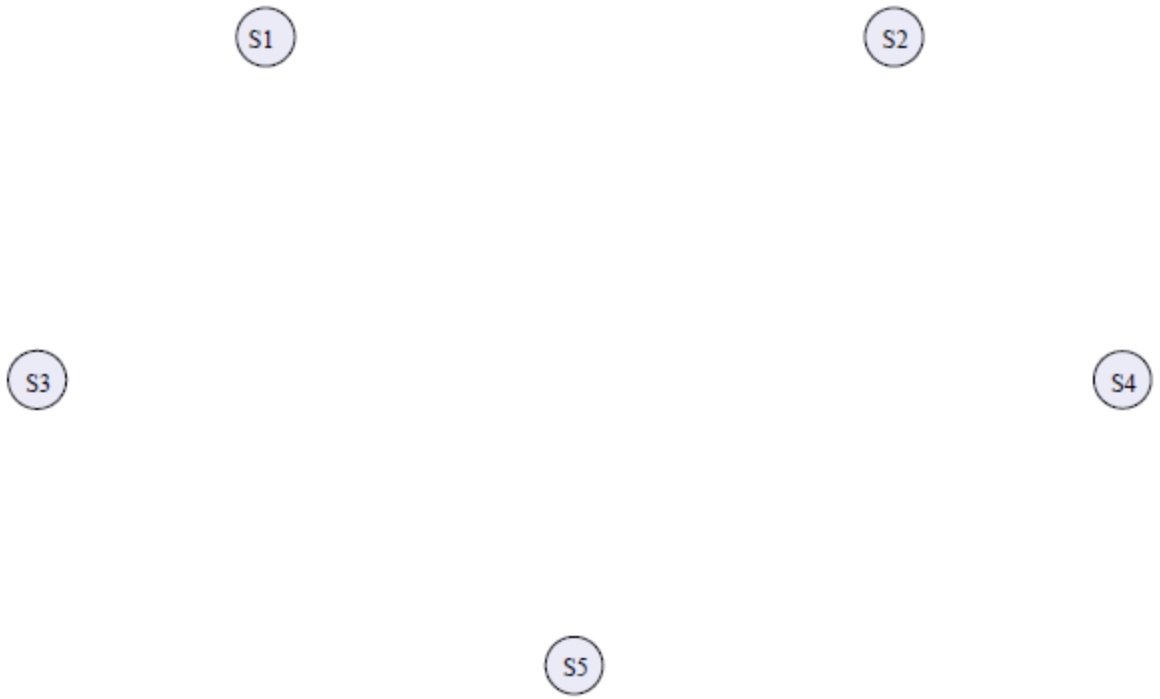


Finding the most salient sentences

Sentence Centrality Measure

Finding the most salient sentences:

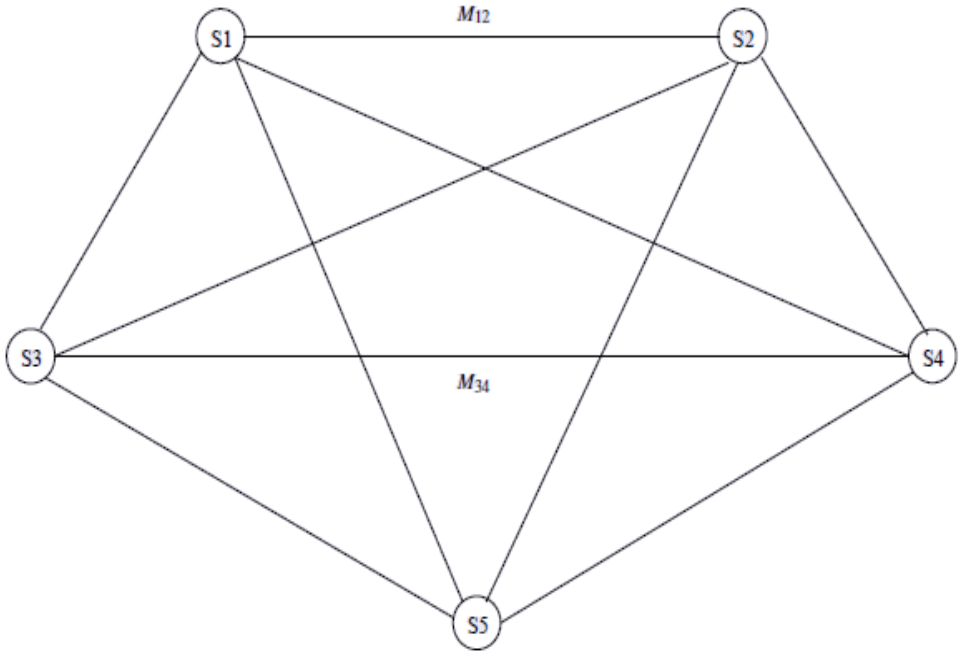
- A document graph is constructed with sentences as the vertices



Sentence Centrality Measure

Finding the most salient sentences:

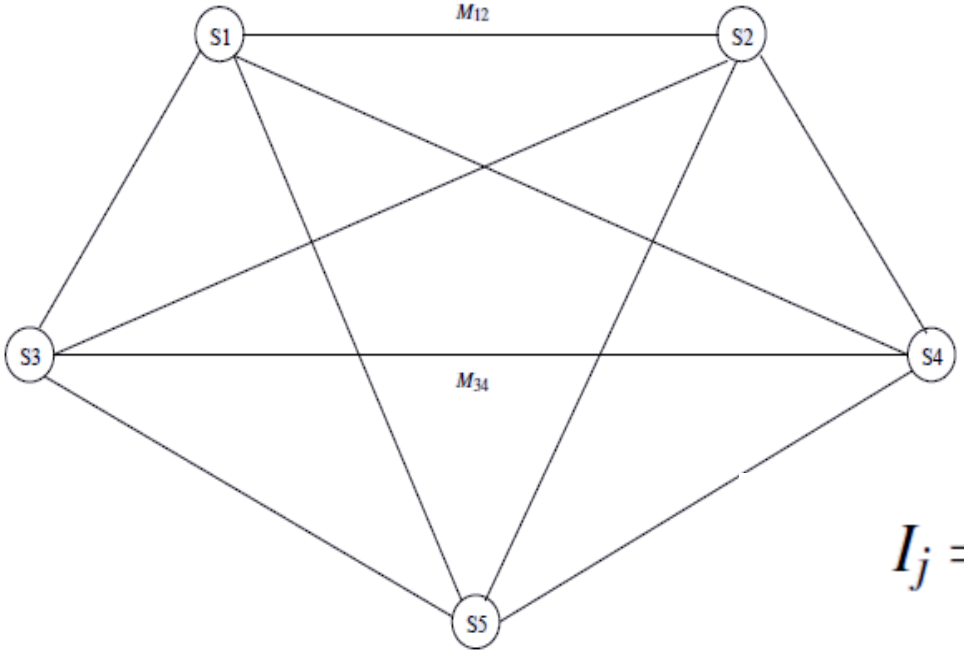
- A sentence similarity function is used to calculate the edge weights



$$\tilde{M} = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.4 & 0.1 \\ 0.5 & 0.0 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \\ 0.4 & 0.0 & 0.4 & 0.0 & 0.2 \\ 0.3 & 0.0 & 0.0 & 0.7 & 0.0 \end{bmatrix}$$

Sentence Centrality Measure

- **Finding the most salient sentences:**
  - PageRank based algorithm is used to compute the sentence centrality vector I.



$$\tilde{M} = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.4 & 0.1 \\ 0.5 & 0.0 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \\ 0.4 & 0.0 & 0.4 & 0.0 & 0.2 \\ 0.3 & 0.0 & 0.0 & 0.7 & 0.0 \end{bmatrix}$$

$$I_j = \mu \cdot \sum_{\forall k \neq j} I_k \cdot \tilde{M}_{k,j} + \frac{1-\mu}{|S|}$$

$$I = \begin{bmatrix} 0.22 & 0.18 & 0.2 & 0.3 & 0.1 \end{bmatrix}$$

Node which has highest score has highest information score

Removing Redundant Sentences

Maximal Marginal Relevance:

- An iterative method for content selection from a selected list of important sentences
- Iteratively choose the best sentence to insert in the summary that is minimally redundant with the summary so far (Summary):

$$\mathbf{Inf(s)_{MMR} = max_{s \in D}(Inf\ (s) - \lambda * sim(s, Summary))}$$

where Inf (s) denotes the informativeness score of a sentence

Sentence Ordering

Chronological ordering: the simplest method

- List the sentences in the order, they appear in the document

Coherence

- Choose orderings that make neighboring sentences similar (by cosine)
- Choose orderings in which neighboring sentences discuss the same entity

Topical ordering

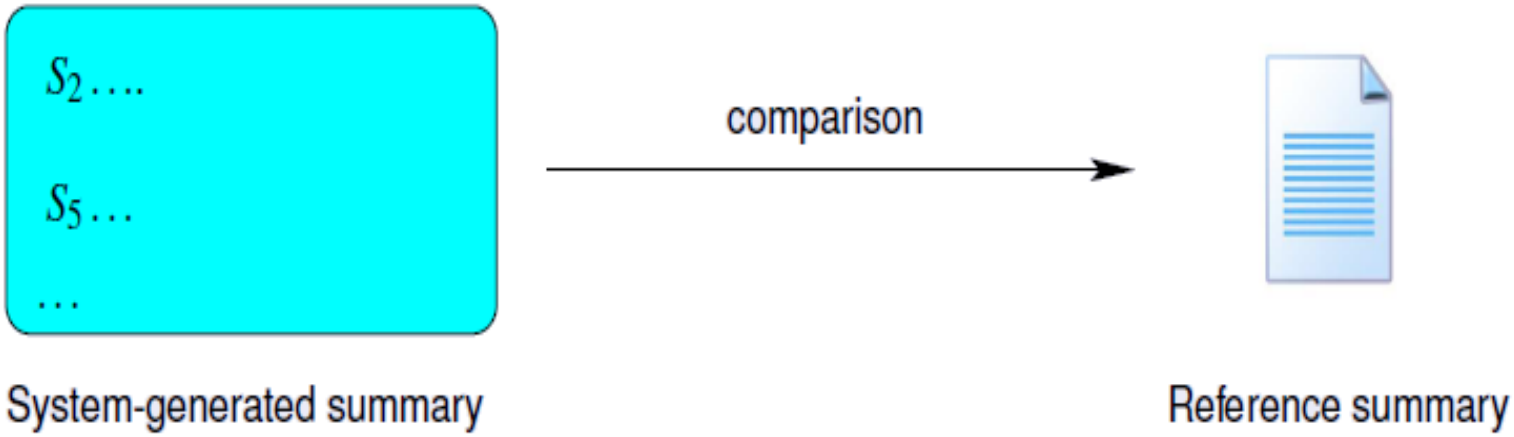
- Learn the ordering of topics in the source documents

Simplifying Sentences

- Parse sentences, use rules to decide which modifiers to prune
  - **Initial adverbials:** For example, on the other hand, as a matter of fact, at this point, ...
  - **PPs without named entities:** The commercial fishing restrictions in Washington will not be lifted unless the salmon population increases [PP to a sustainable number]
  - **Attribution clauses:** Rebels agreed to talks with government officials, international observers said Tuesday

Summarization: Evaluation

System Evaluation



System Evaluation

Evaluation Criteria:

ROUGE : Recall Oriented Understudy for Gisting Evaluation

- Not as good as human evaluation but much more convenient
- Toolkit available for download.

ROUGE for evaluation

- Given a document D, and an automatic summary X:
  - Have N humans produce a set of reference summaries of D (N >=1)
  - Run system, giving automatic summary X
  - What percentage of the n-grams from the reference summaries appear in X?

$$ROUGE-2 = \frac{\sum_{S \in \{RefSums\}} \sum_{bi-gram \in S} Count_{match}(bi-gram)}{\sum_{S \in \{RefSums\}} \sum_{bi-gram \in S} Count(bi-gram)}$$

ROUGE Example

Reference Summaries :

- **Human 1:** water spinach is a green leafy vegetable grown in the tropics.
- **Human 2:** water spinach is a semi-aquatic tropical plant grown as a vegetable.
- **Human 3:** water spinach is a commonly eaten leaf vegetable of Asia

System Summary:

- water spinach is a leaf vegetable commonly eaten in tropical areas of Asia.

ROUGE-2 : $\frac{3+3+6}{10+10+9} = 12/29 = 0.413$

Micro-average and Macro-average Methods

- Sum up individual TPs, FPs and FNs of the system for different sets and then apply them to get statistics.
- For different sets of data:
- Set1 : TP1, FP1, FN1
- Set2: TP2, FP2, FN2
- …

Average Precision using Micro Average Method:

Micro-average precision: $\frac{TP1+TP2}{TP1+TP2+FP1+FP2}$

Micro-average recall : $\frac{TP1+TP2}{TP1+TP2+FN1+FN2}$

Macro Average Method

- For different sets of data:
- Set1 : P1, R1
- Set2: P2, R2
- …

Macro Average Precision : $\frac{P1 + P2}{2}$

Macro Average recall : $\frac{R1 + R2}{2}$

Macro-average F-score is harmonic mean of the two