Text Summarization

Slides borrowed from Pawan Goyal, CSE, IITKGP

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.

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

Application: Generating Snippets

Robert O'Neill taking credit for killing Osama bin Laden sparks debate

Hindustan Times - 1 hour ago

Some special operations service members and veterans are unhappy that one of their own has taken credit publicly for killing Osama bin Laden.

It's been special knock as wait has been long: Rayudu

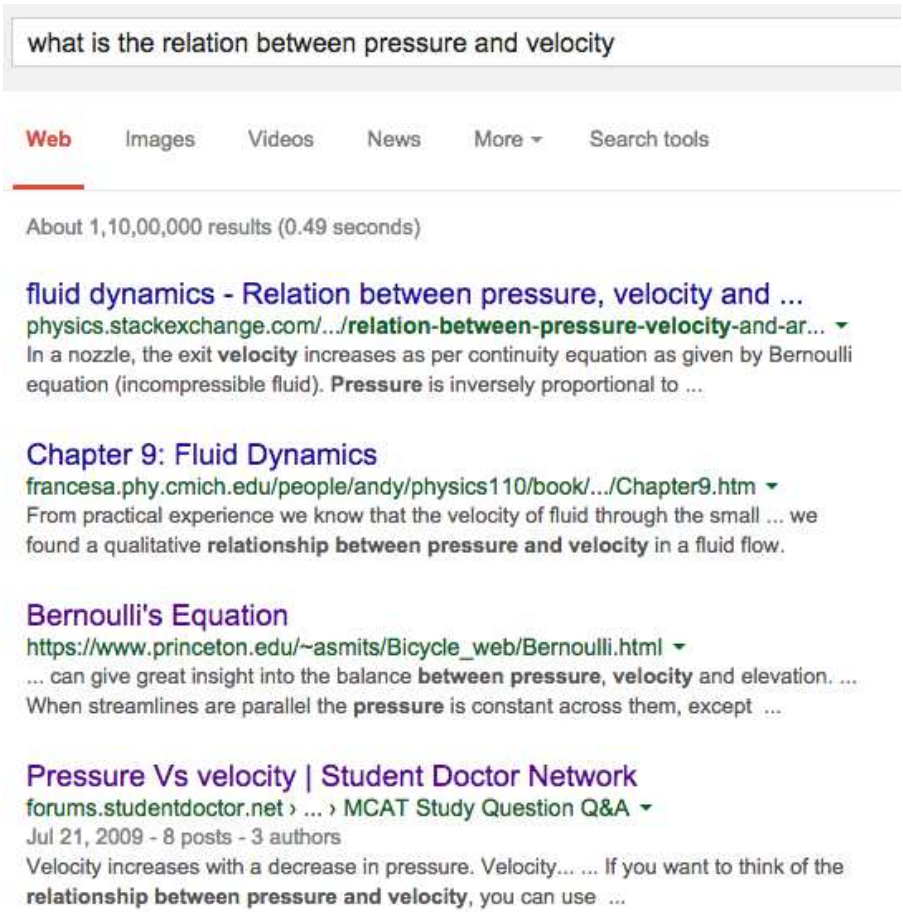
Hindustan Times - 1 hour ago

An elated Ambati Rayudu said Friday that his maiden hundred in international cricket will certainly be a "special one" as it took a long time to come.

Application: Generating Snippets

Robert O'Neill taking credit for killing Osama bin Laden sparks debate
Hindustan Times - 1 hour ago
Some special operations service members and veterans are unhappy that one of their own has taken credit publicly for killing Osama bin Laden.

It's been special knock as wait has been long: Rayudu
Hindustan Times - 1 hour ago
An elated Ambati Rayudu said Friday that his maiden hundred in international cricket will certainly be a "special one" as it took a long time to come.



Automatic Text Summarization

Genres of Summary

- Extract vs. Abstract

  . . .lists fragments of text vs. 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.

Automatic Text Summarization

Genres of Summary

- Extract vs. Abstract

  . . .lists fragments of text vs. 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

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

The most basic algorithm only does the first stage, *content selection*.

Method 1: Luhn

Unsupervised content selection; Luhn (1958)

Intuition

Choose sentences that have salient or informative words

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

$$weight(w_i) = 1 \text{ if } w_i \text{ is a specific term (use mutual information)}$$

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

$$weight(w_i) = 1 \text{ if } w_i \text{ is a specific term (use mutual information)}$$

Weighing a sentence

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

Method 2: LexRank

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

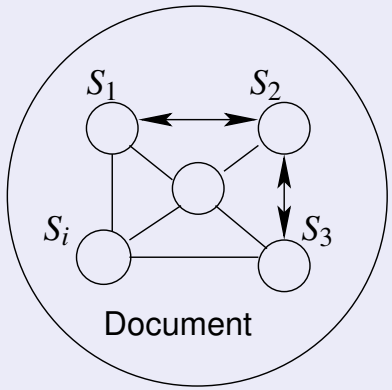
$S_1 \rightarrow \{(computation, 0.1), (process, 0.15), \ldots\}$
$S_2 \rightarrow \{(computation, 0.1), (seen, 0.05), \ldots\}$
$S_3 \rightarrow \ldots$

Machine-readable format

Document Representation

Underlying Hypothesis

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

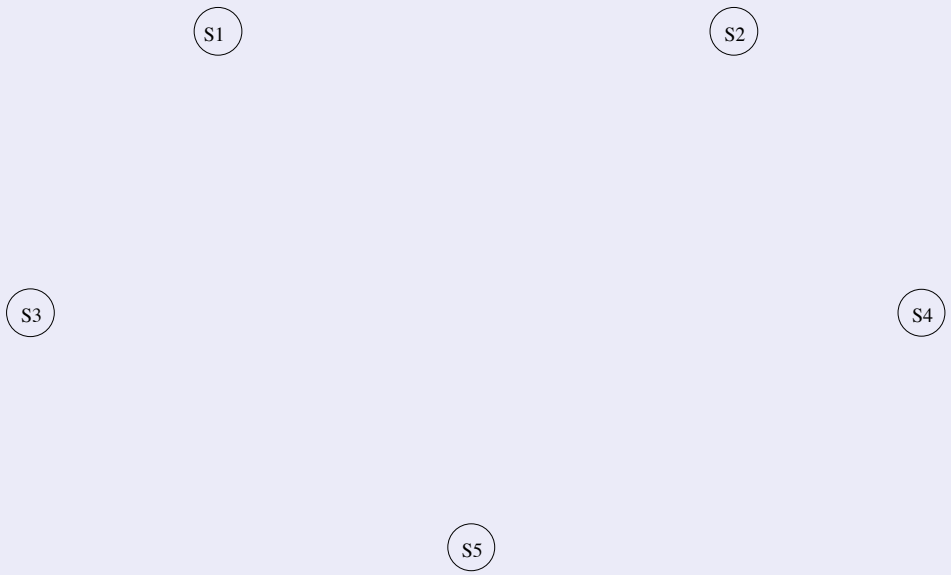


Finding the most salient sentences

Sentence Centrality Measure

Finding the most salient sentences

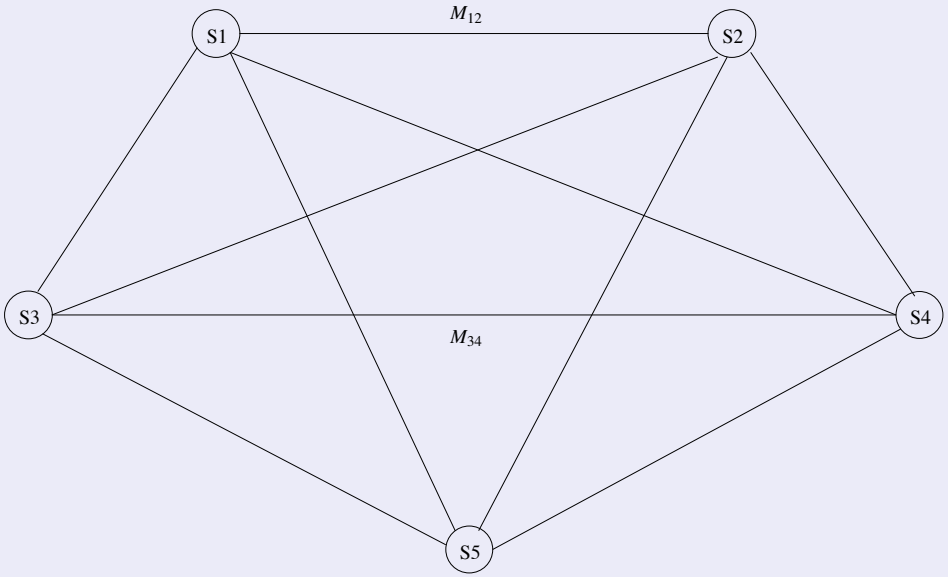
A document graph is constructed with sentences as the vertices



Sentence Centrality Measure

Finding the most salient sentences

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

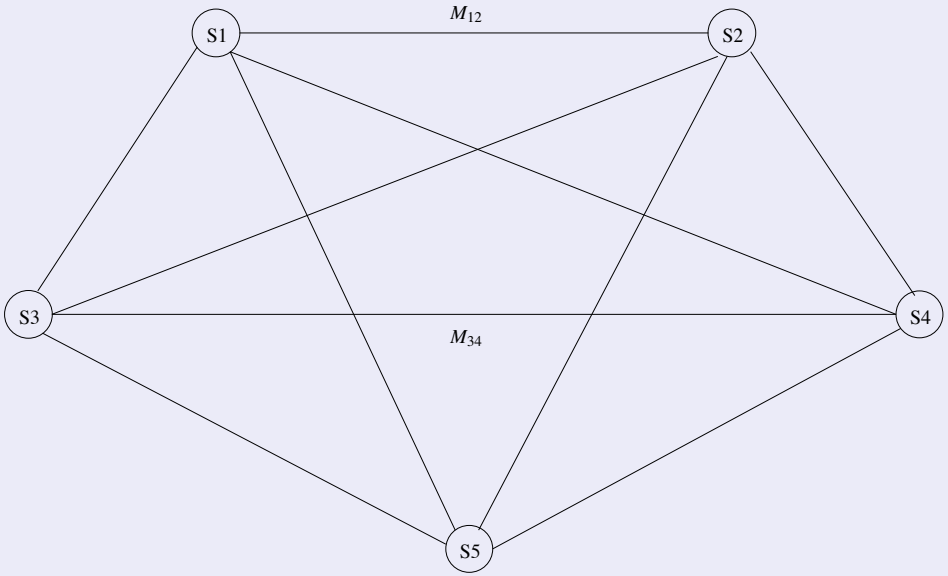


$$\tilde{M} = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.4 & 0.1 \\ 0.5 & 0.0 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \\ 0.4 & 0.0 & 0.4 & 0.0 & 0.2 \\ 0.3 & 0.0 & 0.0 & 0.7 & 0.0 \end{bmatrix}$$

Sentence Centrality Measure

Finding the most salient sentences

PageRank based algorithm is used to compute the sentence centrality vector I.



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

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

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

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 (Sum)

$$Inf(s)_{MMR} = max_{s \in D}(Inf(s) - \lambda \cdot sim(s, Sum))$$

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

Optimization based approach for summarization

Global Inference

- Let us define document D with t_n textual units

$$D = t_1, t_2, \ldots, t_{n-1}, t_n$$

- Let $Rel(i)$ be the relevance of t_i to be in the summary
- Let $Red(i,j)$ be the redundancy between t_i and t_j
- Let $l(i)$ be the length of t_i

Inference Problem

- The inference problem is to select a subset S of textual units from D such that summary score of S, i.e., $s(S)$, is maximized.
- $S = \arg\max_{S \subseteq D} \left[\sum_{t_i \in S} Rel(i) - \sum_{t_i, t_j \in S, i<j} Red(i,j) \right]$ such that $\sum_{t_i \in S} l(i) \leq K$, where k denotes the maximum length of the summary

A Greedy Solution

1. Sort D so that $Rel(i) > Rel(i+1) \forall i$
2. $S = \{t_1\}$
3. while $\sum_{t_i \in S} l(i) < K$
4. $t_j = \arg\max_{t_j \in D-S} s(S \cup \{t_j\})$
5. $S = S \cup \{t_j\}$
6. return S

Integer Linear Programming (ILP)

- Greedy algorithm is an approximate solution
- Use exact solution algorithms with ILP
- ILP is a constrained optimization problem
- Many solvers on the web
- Define the constraints based on relevance and redundancy for summarization

Sentence Level ILP Formulation

Optimization Function

maximize $\sum_i \alpha_i Rel(i) - \sum_{i<j} \alpha_{ij} Red(i,j)$

Sentence Level ILP Formulation

Optimization Function

maximize $\sum_i \alpha_i Rel(i) - \sum_{i<j} \alpha_{ij} Red(i,j)$

Constraints

such that $\forall i,j$:

- $\alpha_i, \alpha_{ij} \in \{0,1\}$
- $\sum_i \alpha_i l(i) \leq K$
- $\alpha_{ij} - \alpha_i \leq 0$
- $\alpha_{ij} - \alpha_j \leq 0$
- $\alpha_i + \alpha_j - \alpha_{ij} \leq 1$

Sentence Level ILP Formulation

Optimization Function

maximize $\sum_i \alpha_i Rel(i) - \sum_{i<j} \alpha_{ij} Red(i,j)$

Constraints

such that $\forall i,j$:

- $\alpha_i, \alpha_{ij} \in \{0,1\}$
- $\sum_i \alpha_i l(i) \leq K$
- $\alpha_{ij} - \alpha_i \leq 0$
- $\alpha_{ij} - \alpha_j \leq 0$
- $\alpha_i + \alpha_j - \alpha_{ij} \leq 1$

Is generic enough

Depending on your task, you can define your own optimization function and constrains.

The next steps: Sentence Ordering

The next steps: Sentence Ordering

Chronological ordering: the simplest method

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

The next steps: 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

The next steps: 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

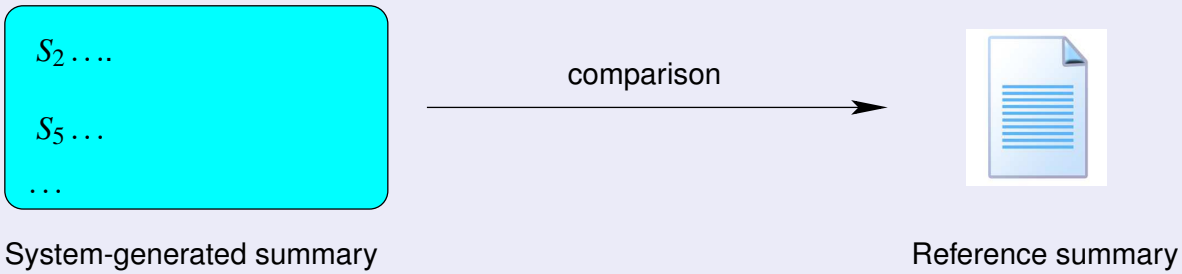
Topical ordering

Learn the ordering of topics in the source documents

Evaluation of summarization algorithms

System Evaluation

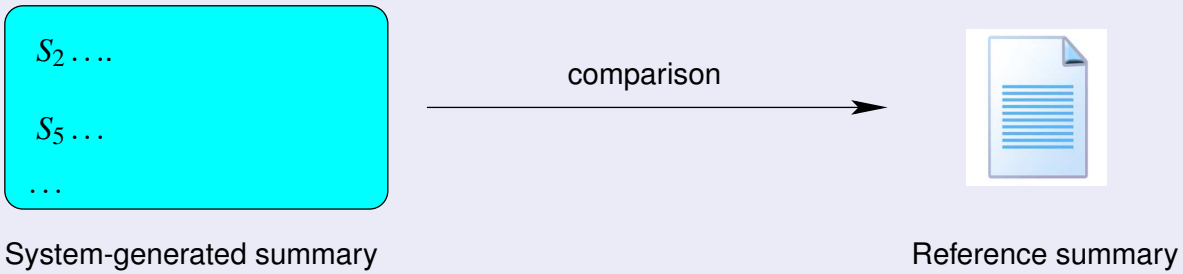
Evaluation Criteria



System Evaluation

System Evaluation

Evaluation Criteria



System Evaluation

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 ($N \geq 1$)
- 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 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

$$\frac{3+3+6}{10+10+9} = 12/29 = 0.413$$

Further discussions

Further Discussions

- Multi-document summarization

Further Discussions

- Multi-document summarization
- Query-specific summarization

Further Discussions

- Multi-document summarization
- Query-specific summarization
- Abstractive summarization