Document Summarization

CS-626 Seminar

November 8th, 2013

Kumar Pallav 100050047
Pawan Nagwani 100050049
Pratik Kumar 100100018

What is summarization?

- A summary is a text that is produced from one or more texts and contains a significant portion of the information in the original text is no longer than half of the original text.

Contents

- Motivation
- Genres and types of summaries
- Approaches and paradigms
    - Simple Techniques
    - Graph based approaches
        - Degree Centrality
        - Text Rank / Lex Rank
    - Linguistic Approach
        - Lexical Chains
    - Semantic Approach
        - WordNet Based Approach
- Evaluating summaries
- Conclusions

Motivation

- The advent of WWW has created a large reservoir of data

- A short summary, which conveys the essence of the d ocument, helps in finding relevant information quickly

- Document summarization also provides a way to cluster similar documents and present a summary

Genres

- Indicative vs. informative
  - used for quick categorization vs. content processing.
- Extract vs. abstract
  - lists fragments of text vs. re-phrases content coherently.
- Generic vs. query-oriented
  - provides author’s view vs. reflects user’s interest.
- Background vs. just-the-news
  - assumes reader’s prior knowledge is poor vs. up-to-date.
- Single-document vs. multi-document source
  - based on one text vs. fuses together many texts.

Simple Techniques

1. This representation abstracts the source text into a frequency table
2. Another method also based on linguistic information is t he cue phrase method, which uses meta-linguistic mar kers (for example, “in conclusion”, “the paper describes ”) to select important phrases. The cue phrase method i s based on the assumption that such phrases provide a “rhetorical” context for identifying important sentences.
3. The location method relies on the following intuition — headings, sentences in the beginning and end of the t ext, text formatted in bold, contain important informati on to the summary

Graph Based Approach

- Degree Centrality
- TextRank and LexRank

Graph Based Approach

- Several criteria to assess sentence salience
  - All approach are based on the concept of **“prestige”** in social networks, which has also inspired many ideas in **computer n etworks** and **information retrieval**.
- A cluster of documents can be viewed as a **network of sentences** that are related to each other.
  - They hypothesize that the sentences that are similar to many of the other sentences in a cluster are more central (or salient) to the topic.
- To define similarity, they use the **bag-of-words model** to represent each sentence as an N-dimensional vector, where N is the number of all possible words in the target language.
- A cluster of documents may be represented by a **cosine similarity matrix** where each entry in the matrix is the similarity between the corresponding sentence pair.

Degree Centrality

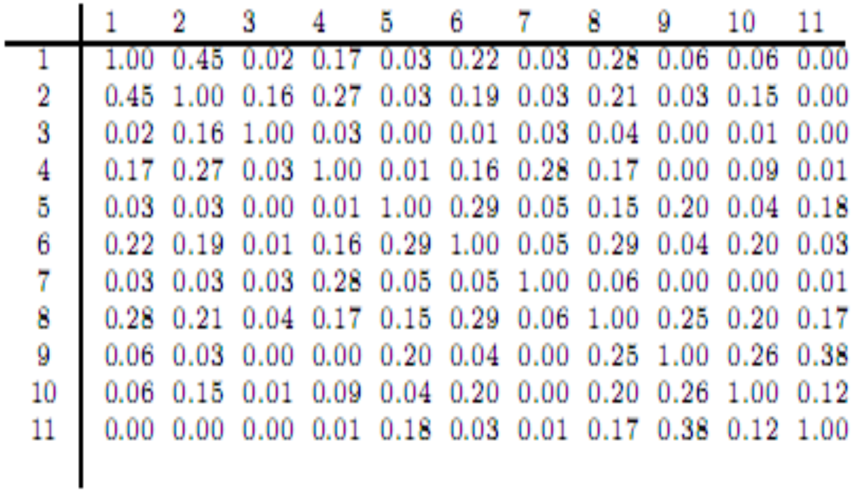
- Represent each sentence by a vector
- Denote each sentence as the node of a graph
- Cosine similarity determines the edges between nodes

$$\mathrm{idf}_i = \log\left(\frac{N}{n_i}\right)$$

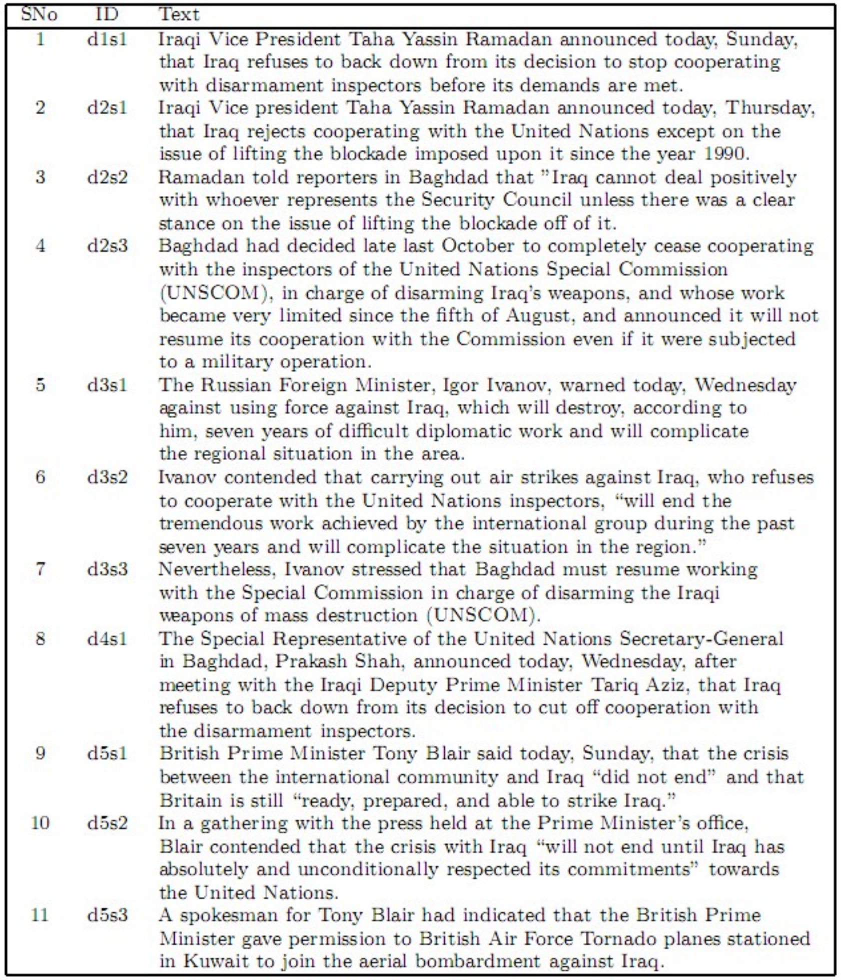
$$\text{idf-modified-cosine}(x, y) = \frac{\sum_{w \in x,y} \mathrm{tf}_{w,x}\mathrm{tf}_{w,y}(\mathrm{idf}_w)^2}{\sqrt{\sum_{x_i \in x}(\mathrm{tf}_{x_i,x}\mathrm{idf}_{x_i})^2} \times \sqrt{\sum_{y_i \in y}(\mathrm{tf}_{y_i,y}\mathrm{idf}_{y_i})^2}}$$

Degree Centrality

- Sentence ID dXsY indicates the Y th sentence in the Xth document.



	1	2	3	4	5	6	7	8	9	10	11
1	1.00	0.45	0.02	0.17	0.03	0.22	0.03	0.28	0.06	0.06	0.00
2	0.45	1.00	0.16	0.27	0.03	0.19	0.03	0.21	0.03	0.15	0.00
3	0.02	0.16	1.00	0.03	0.00	0.01	0.03	0.04	0.00	0.01	0.00
4	0.17	0.27	0.03	1.00	0.01	0.16	0.28	0.17	0.00	0.09	0.01
5	0.03	0.03	0.00	0.01	1.00	0.29	0.05	0.15	0.20	0.04	0.18
6	0.22	0.19	0.01	0.16	0.29	1.00	0.05	0.29	0.04	0.20	0.03
7	0.03	0.03	0.03	0.28	0.05	0.05	1.00	0.06	0.00	0.00	0.01
8	0.28	0.21	0.04	0.17	0.15	0.29	0.06	1.00	0.25	0.20	0.17
9	0.06	0.03	0.00	0.00	0.20	0.04	0.00	0.25	1.00	0.26	0.38
10	0.06	0.15	0.01	0.09	0.04	0.20	0.00	0.20	0.26	1.00	0.12
11	0.00	0.00	0.00	0.01	0.18	0.03	0.01	0.17	0.38	0.12	1.00



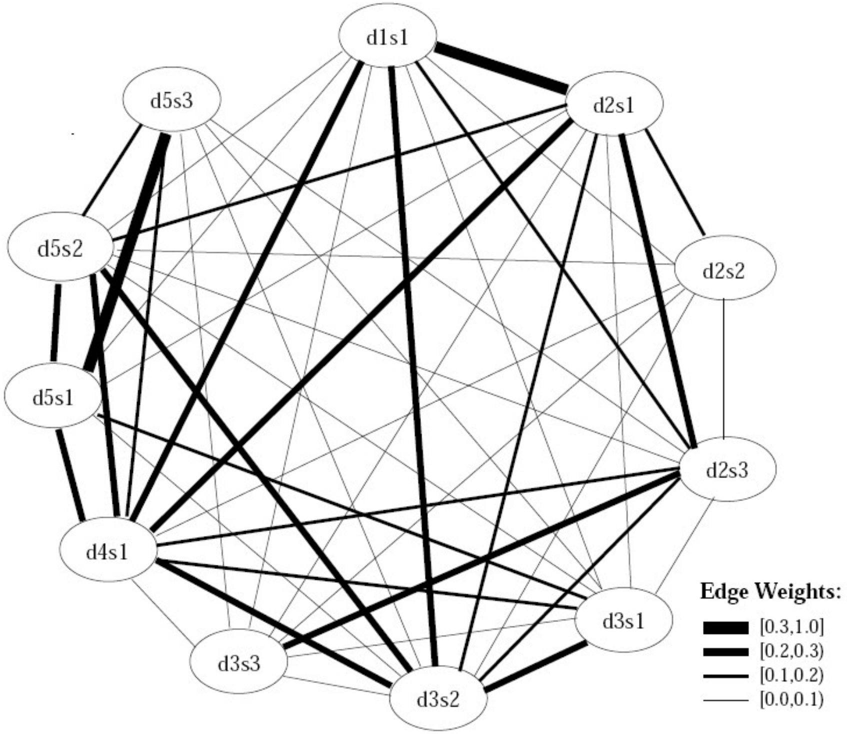
SNo	ID	Text
1	d1s1	Iraqi Vice President Taha Yassin Ramadan announced today, Sunday, that Iraq refuses to back down from its decision to stop cooperating with disarmament inspectors before its demands are met.
2	d2s1	Iraqi Vice president Taha Yassin Ramadan announced today, Thursday, that Iraq rejects cooperating with the United Nations except on the issue of lifting the blockade imposed upon it since the year 1990.
3	d2s2	Ramadan told reporters in Baghdad that ”Iraq cannot deal positively with whoever represents the Security Council unless there was a clear stance on the issue of lifting the blockade off of it.
4	d2s3	Baghdad had decided late last October to completely cease cooperating with the inspectors of the United Nations Special Commission (UNSCOM), in charge of disarming Iraq's weapons, and whose work became very limited since the fifth of August, and announced it will not resume its cooperation with the Commission even if it were subjected to a military operation.
5	d3s1	The Russian Foreign Minister, Igor Ivanov, warned today, Wednesday against using force against Iraq, which will destroy, according to him, seven years of difficult diplomatic work and will complicate the regional situation in the area.
6	d3s2	Ivanov contended that carrying out air strikes against Iraq, who refuses to cooperate with the United Nations inspectors, "will end the tremendous work achieved by the international group during the past seven years and will complicate the situation in the region."
7	d3s3	Nevertheless, Ivanov stressed that Baghdad must resume working with the Special Commission in charge of disarming the Iraqi weapons of mass destruction (UNSCOM).
8	d4s1	The Special Representative of the United Nations Secretary-General in Baghdad, Prakash Shah, announced today, Wednesday, after meeting with the Iraqi Deputy Prime Minister Tariq Aziz, that Iraq refuses to back down from its decision to cut off cooperation with the disarmament inspectors.
9	d5s1	British Prime Minister Tony Blair said today, Sunday, that the crisis between the international community and Iraq "did not end" and that Britain is still "ready, prepared, and able to strike Iraq."
10	d5s2	In a gathering with the press held at the Prime Minister's office, Blair contended that the crisis with Iraq "will not end until Iraq has absolutely and unconditionally respected its commitments" towards the United Nations.
11	d5s3	A spokesman for Tony Blair had indicated that the British Prime Minister gave permission to British Air Force Tornado planes stationed in Kuwait to join the aerial bombardment against Iraq.

Degree Centrality

- In a cluster of related documents, many of the sentences are expected to be somewhat similar to each other since they are all about the same topic.
- Since they are interested in significant similarities, they can **eliminate some low values** in this matrix by defining a threshold so that the cluster can be viewed as an (undirected) graph
  - each sentence of the cluster is a node, and significantly similar sentences are connected to each other
- They define **degree centrality of a sentence** as the **degree of the corresponding node** in the similarity graph.

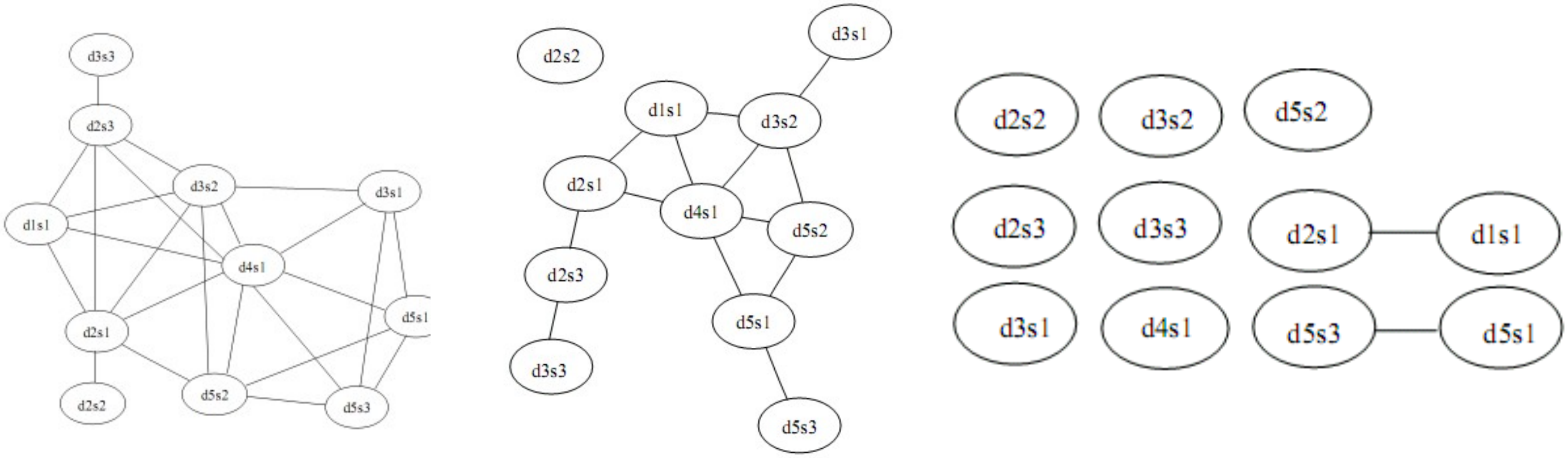
Degree Centrality

- Since we are interested in significant similarities, we can eliminate some low values in this matrix by defining a threshold.



Degree Centrality

- Similarity graphs that correspond to thresholds 0.1, 0.2, and 0.3, respectively, for the cluster



0.1 0.2 0.3

- The choice of **cosine threshold** dramatically influences the interpretation of centrality.
- Too **low thresholds** may mistakenly take **weak similarities** into consider ation while too **high thresholds** may **lose many of the similarity relatio ns** in a cluster

Degree Centrality

- Compute the degree of each sentence
- Pick the nodes (sentences) with high degrees
- Remove its neighbours with high similarity
- Repeat the above process

Issue with Degree Centrality

- When computing degree centrality, we treat **each edge as a vote** to determine the overall centrality value of each node.
- In many types of social networks, **not all of the relationships are considered equally important**.
  - The **prestige** of a person does not only depend on **how many friends** he has, but also depends on **who his friends** are.
- Taking the centrality of the voting nodes into account in weighting each vote.
  - A straightforward way of formulating this idea is to consider every node having a centrality value and distributing this centrality to its neighbors.

TextRank and LexRank

- This approach models the **document as a graph** and uses an algorithm similar to **Google's PageRank** alg orithm to find top-ranked sentences.
- The key intuition is the notion of **centrality or prestige** in social networks i.e. a **sentence should be highly r anked if it is recommended by many other highly r anked sentences.**

TextRank and LexRank : Intuition

- *“If Sachin Tendulkar says Malinga is a good batsman, he should be regarded highly. But then if Sachin is a g entleman, who talks highly of everyone, Malinga might not really be as good.”*

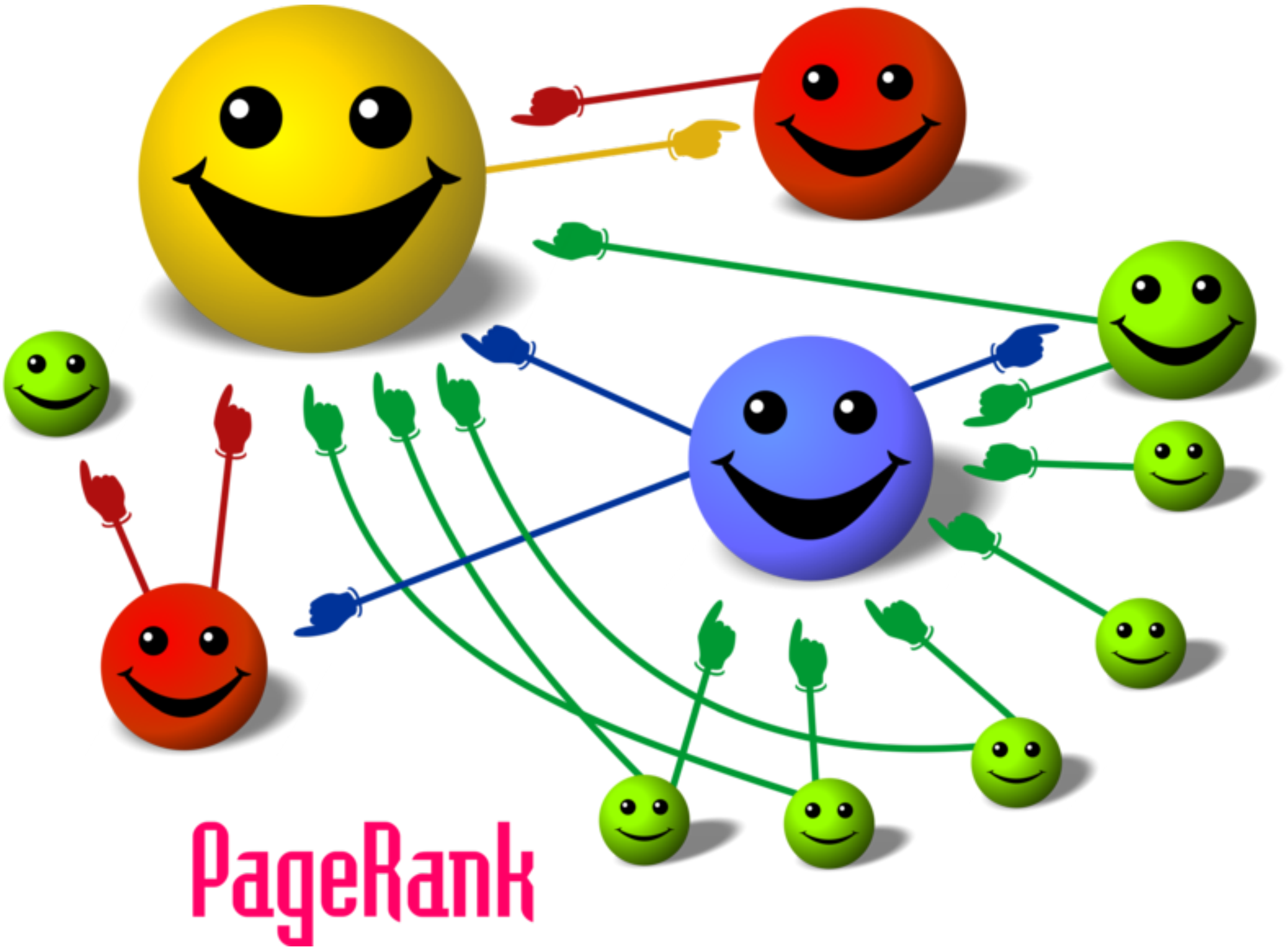
- Formula

$$PageRank\ of\ site = \sum \frac{PageRank\ of\ inbound\ link}{Number\ of\ links\ on\ that\ page}$$

OR

$$PR(u) = (1 - d) + d \times \sum \frac{PR(v)}{N(v)}$$

TextRank and LexRank : Example



TextRank and LexRank

$$p(u) = \sum_{v \in adj[u]} \frac{p(v)}{deg(v)}$$

$$\mathbf{B}(i,j) = \frac{\mathbf{A}(i,j)}{\sum_k \mathbf{A}(i,k)}$$

$$\mathbf{p} = \mathbf{B}^{\mathrm{T}}\mathbf{p}$$

$$\mathbf{p}^{\mathrm{T}}\mathbf{B} = \mathbf{p}^{\mathrm{T}}$$

p(u) - LexRank of Node u

A - Adjacency Matrix

TextRank and LexRank

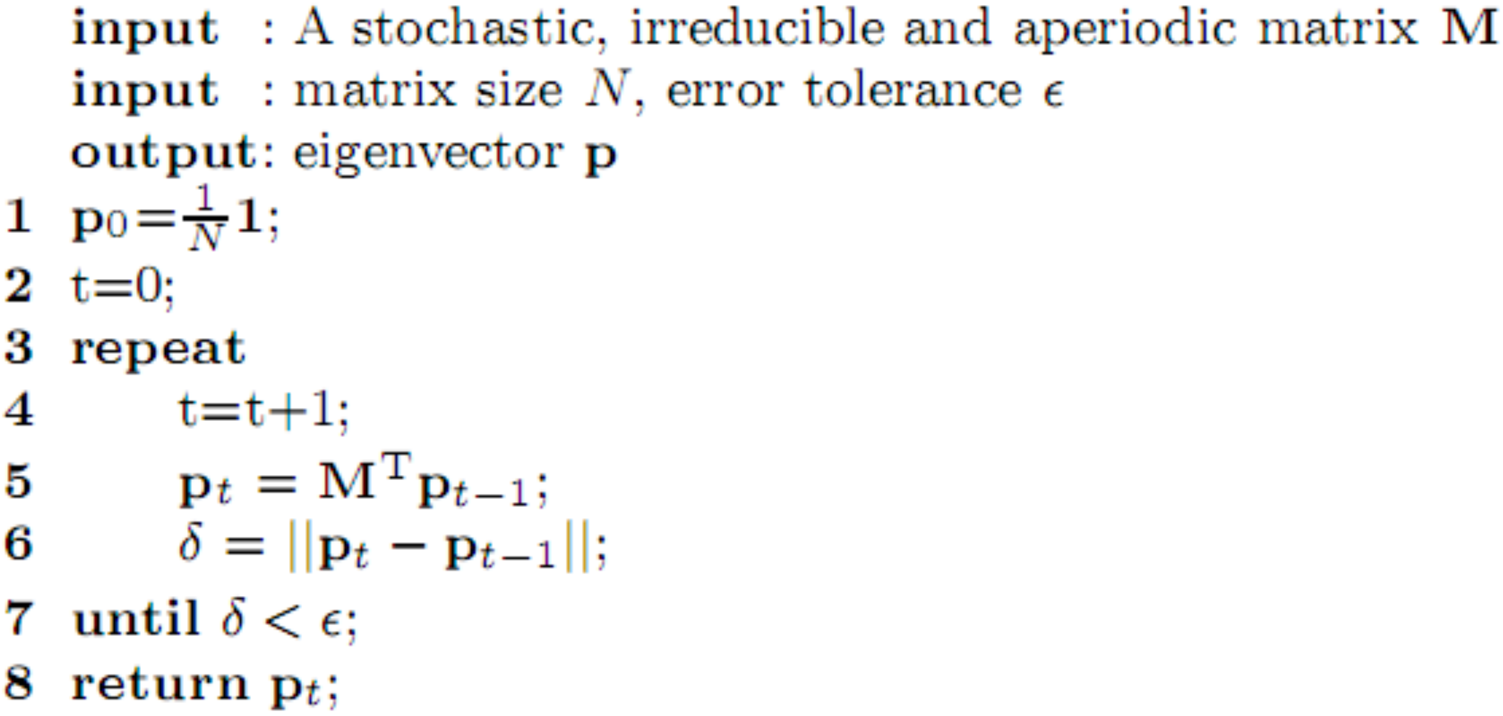
- Damping
  - Because we want B to be Stochastic Matrix

$$p(u) = \frac{d}{N} + (1-d) \sum_{v \in adj[u]} \frac{p(v)}{deg(v)}$$

$$\mathbf{p} = [d\mathbf{U} + (1-d)\mathbf{B}]^{\mathrm{T}}\mathbf{p}$$

TextRank and LexRank : Algorithm

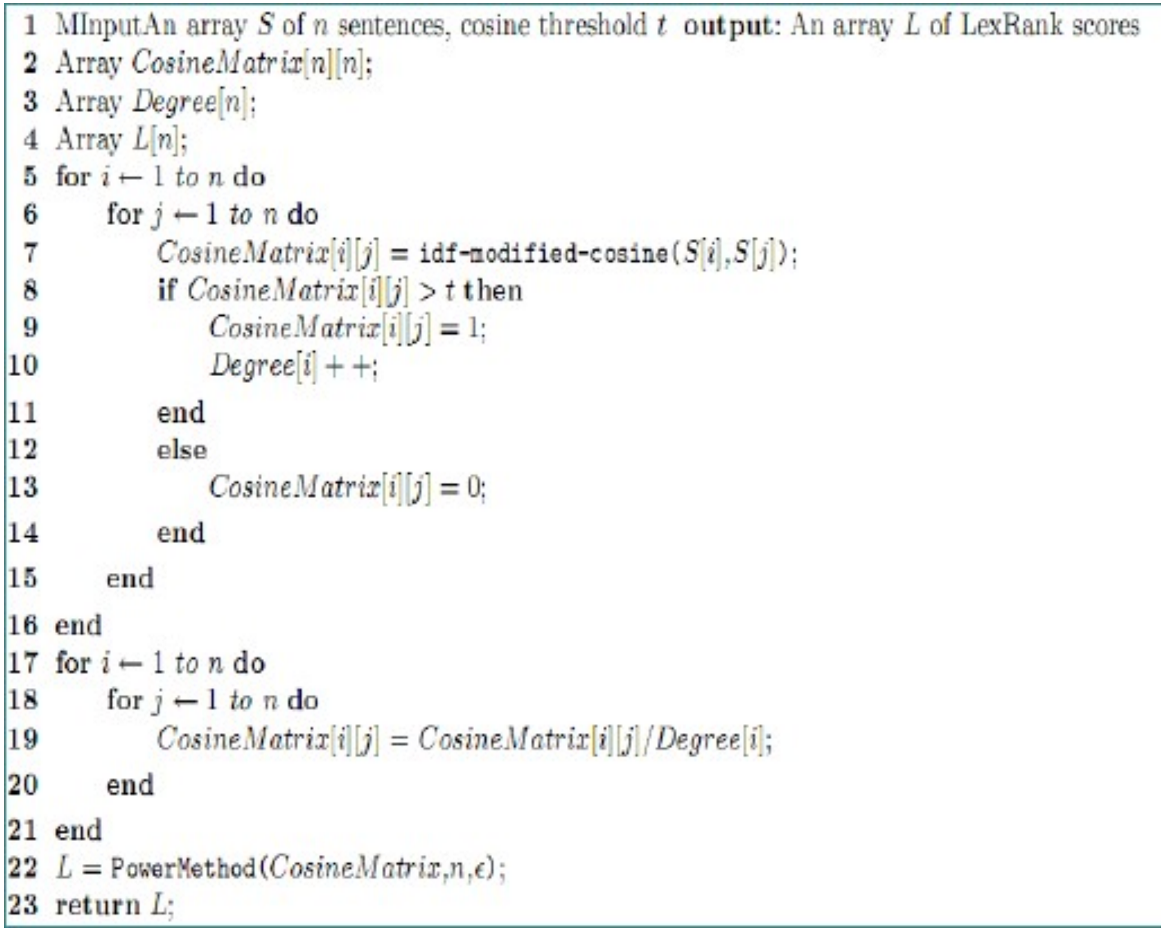
- Power Method



```
input  : A stochastic, irreducible and aperiodic matrix $\mathbf{M}$
input  : matrix size $N$, error tolerance $\epsilon$
output : eigenvector $\mathbf{p}$
$\mathbf{p}_0 = \frac{1}{N}\mathbf{1}$;
t=0;
repeat
    t=t+1;
    $\mathbf{p}_t = \mathbf{M}^{\mathrm{T}}\mathbf{p}_{t-1}$;
    $\delta = ||\mathbf{p}_t - \mathbf{p}_{t-1}||$;
until $\delta < \epsilon$;
return $\mathbf{p}_t$;
```

TextRank and LexRank : Algorithm

- Main Algorithm



```
MInputAn array $S$ of $n$ sentences, cosine threshold $t$ output: An array $L$ of LexRank scores
Array $CosineMatrix[n][n]$;
Array $Degree[n]$;
Array $L[n]$;
for $i \leftarrow 1$ to $n$ do
    for $j \leftarrow 1$ to $n$ do
        $CosineMatrix[i][j]$ = idf-modified-cosine($S[i]$,$S[j]$);
        if $CosineMatrix[i][j] > t$ then
            $CosineMatrix[i][j] = 1$;
            $Degree[i] + +$;
        end
        else
            $CosineMatrix[i][j] = 0$;
        end
    end
end
for $i \leftarrow 1$ to $n$ do
    for $j \leftarrow 1$ to $n$ do
        $CosineMatrix[i][j] = CosineMatrix[i][j]/Degree[i]$;
    end
end
$L$ = PowerMethod($CosineMatrix$,$n$,$\epsilon$);
return $L$;
```

Why TextRank works?

- Through the graphs, **TextRank** identifies connections between various **entities**, and implements the **concept of recommendation**.
- A text unit **recommends** other related text units, and th e **strength of the recommendation is recursively co mputed** based on the importance of the units making t he recommendation.
- The sentences that are highly recommended by other sentences in the text are likely to be more informative

Linguistic Methods

- ➔ **Lexical Chains**

  Collecting ideas from the text as correlated chain of words

- ➔ **Rhetorical Analysis**

  Finding Rhetorical relations - between two non-overlapping text snippets, Nucleus - Core Idea, Writers Purpose, Satellite - referred in context to nu cleus for Justifying, Evidencing, Contradicting etc.

Lexical Chain Summarisation

- In Lexical Chain summarisation, first the chains are formed
  a. a set of candidate words are selected
     The words selected are nouns and noun-compounds
  b. for each candidate word, find an appropriate chain relying on a relatedness criterion among members of the chains
  c. if it is found, insert the word in the chain and update it accordingly

- So for a document, a set of chains is formed, each with a different central theme

Example Chain Formation

*Mr. Kenny is the **person** that invented an anesthetic **machine** which uses **micro-computers** to control the rate at which an anesthetic is pumped into the blood. Such **machines** are nothing new. But his device uses two **micro-computers** to achieve much closer monitoring of the **pump** feeding the anesthetic into the patient.*

The words in bold are under consideration.

Example Chain Formation

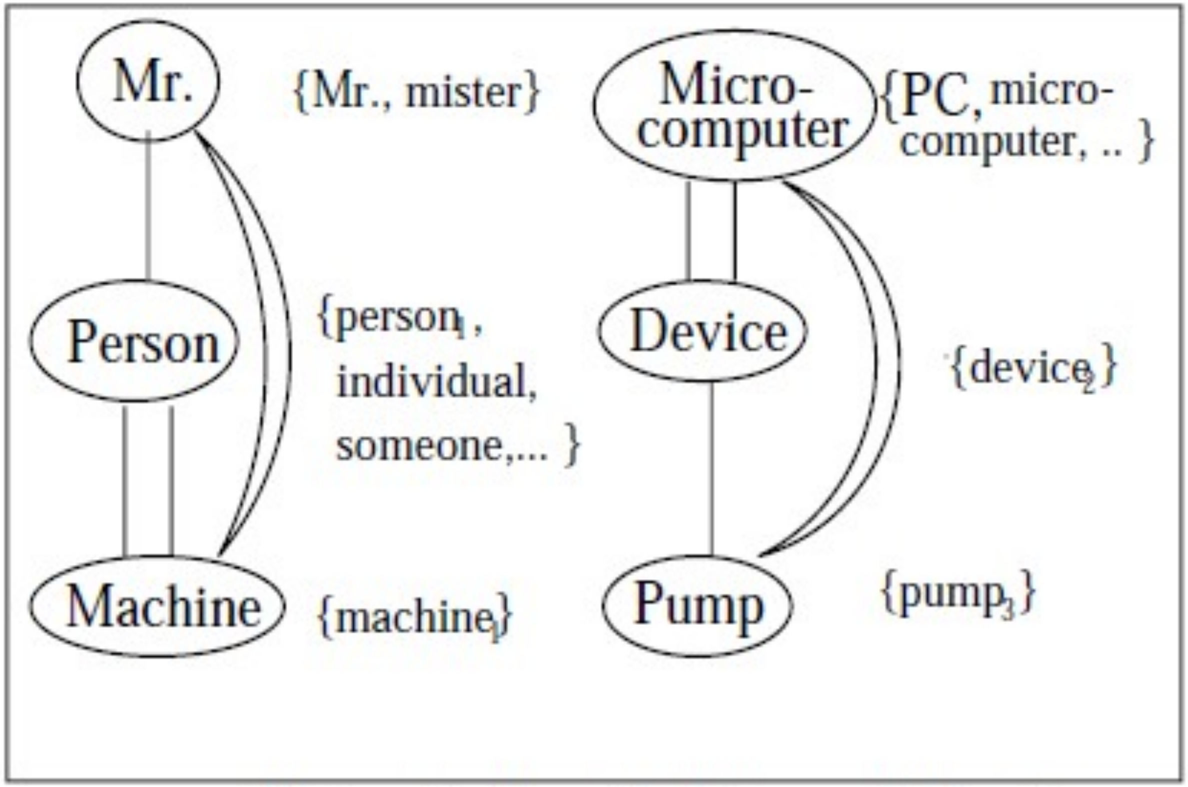


Figure 6: Step 3, Interpretation 1

machine$_1$ is *"an efficient person"*

Example Chain Formation

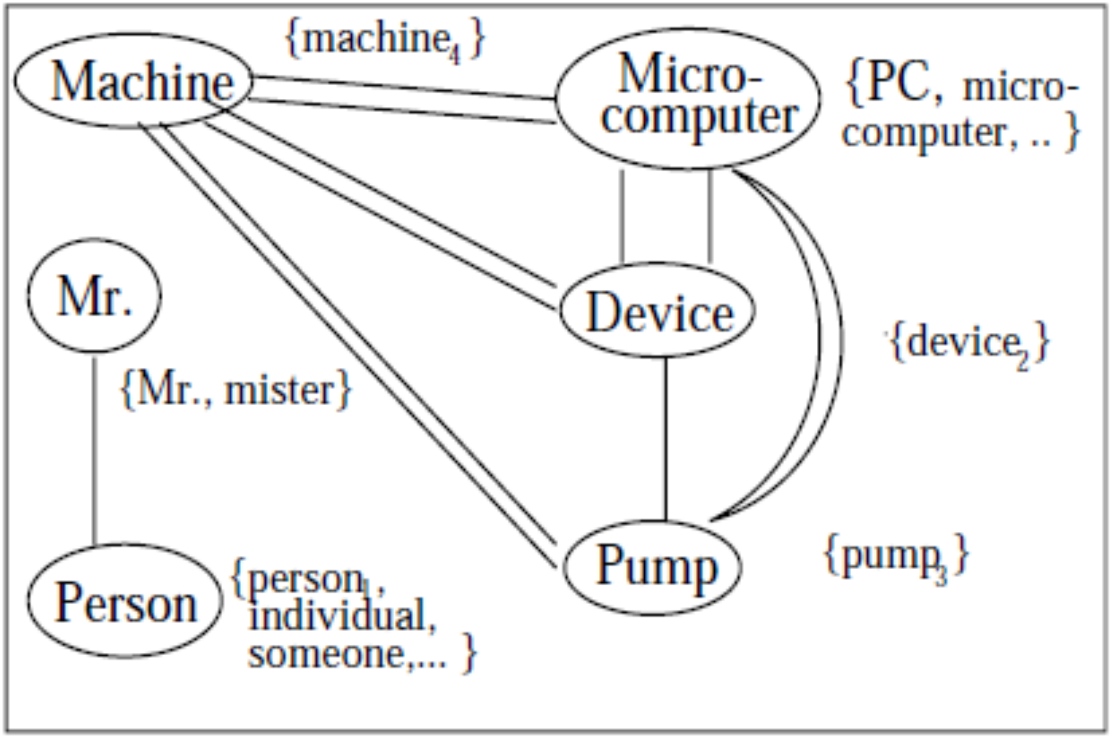


Figure 7: Step 3, Interpretation 2

machine$_4$ is similar to the general sense

Scoring of Lexical Chains

- Of the possible chains, those with **strong** scores are chosen
- Some good parameters for scoring are found to be:
  - **Length**

    The number of occurrences of members of the chain
  - **Homogeneity index**

    1 - (the number of distinct occurrences) / length
- Scoring functions

  $Score(Chain) = Length * HomogeneityIndex$
- Strength criterion

  $Score(Chain) > Average(Scores) + 2 * StandardDeviation(Scores)$

Extracting significant sentences

- A sentence is extracted for each of the strong chain
- Various heuristics can be employed for this purpose:
    a. Choosing the sentence that has the first occurrence of a word from the chain
    b. Choosing the sentence that has the first occurrence of the key word of the chain
    c. Choosing the sentence that has the first occurrence in the text unit, which has high density of chain members
- Though, the third heuristic is more complex, in most of the test cases, the second heuri stics produced the best result.

WordNet based Summarization

- This approach to text summarization selects sentences based on their semantic content and its relative importance to the semantics of the whole text.
- Also this method avoids selecting too many sentences with same semantic content and thus reduces redundancy in summary.
- Language generation is not involved, only representative sentences are chosen.

WordNet based Summarization

Major steps involved are :

- ➢ Preprocessing of text
- ➢ Constructing sub-graph from WordNet
- ➢ Synset Ranking
- ➢ Sentence Selection
- ➢ Principal Component Analysis

Preprocessing

- Break text into sentences

  This approach involves selection of sentences.

- Apply POS tagging

  Helps determine correct sense of the word.

- Identify collocations in the text

  Collocations are treated as one entity.

- Remove the stop words

  Words like It, The etc. do not add much to semantics.

- Sequence of the steps is important

Constructing sub-graph from Wordnet

- Mark all the words and collocations in the WordNet graph which are present in the text
- Traverse the generalization edges up to a fixed depth, and mark the synsets you visit
- Construct a graph, containing only the mark ed synsets as nodes and generalization edg es as edges.

Synset Ranking

- Rank synsets based on their relevance to text

- Construct a Rank vector, corresponding to each node of the graph

  initialized to 1/√n,   where n=no of nodes in graph

- Create an authority matrix,

  A(i,j) = 1/(num_of_predecessors(j)), if j is a child of i

  = 0 otherwise

Synset Ranking

- Update the R vector iteratively as,

$$R_{new} = R_{old}*A / |R_{old}*A|$$

- Stop when $|R_{new}|$ changes less than threshold

- Higher value implies better rank and higher relevance

Sentence Selection

- Construct a matrix, M with m rows and n columns where m is number of sentences and n is number of nodes in subgraph.

- For each sentence S_i
  - Traverse graph G, starting with words present in S_i and following generalization edges find set of reac hable synsets Sy_i
  - For each $sy_j \in Sy_i$ $M[S_i][sy_j] = R[sy_j]$

Principal Component Analysis

- Apply PCA on matrix M and get set of principal components or eigen vectors
- Eigen value of each eigen vector is measure of relevance of eigen vector to the meaning
- Sort eigen vectors according to Eigen values and for each eigen vector, find its projection on each sentence

Principal Component Analysis

- Select top k sentences for each eigen vector
- k is proportional to the eigen values of the eigen vectors.

WordNet based Summarization

- This method gives good results when compared to manually generated su mmaries.

Manual summaries used	Average Cosine Similarity
DUC-200	0.7402
DUC-400	0.7875

Evaluation

- Recall based score to compare system generated summary with one or more human generated summaries.
- Can be with respect to n-gram matching.
- Unigram matching is found to be the best indicator for evaluation.
- ROUGE-1 is computed as division of count of unigrams in reference that appear in sy stem and count of unigrams in reference s ummary.

Conclusion

Most of the current research is based on extractive multi-document summarization.

Current summarization systems are widely used to summarize NEWS and other online articles.

Conclusion

Rule-based vs. ML/ Statistical

➢ Most of the early techniques were rule-based whereas the current one apply statistical approaches.

Keyword vs. Graph-based

➢ Keyword based techniques rank sentences based on the occurrence of relevant keywords.

➢ Graph based techniques rank sentences based on content overlap.

Refrences

- Regina Barzilay and Michael Elhadad, 1997, *“Using Lexical Chains for Text Summarization”*, Proceedings of ACL’97/EACL’97 workshop on Intelligent Scalable Text Summarization
- Gunes Erkan and Dragomir R. Radev, 2004, “LexRank: Graph-based Lexical Centrality as Salience in Text Summarization”, Journal of Artificial Intelligence Research
- Kedar Bellare, Anish Das Sharma, Atish Das Sharma, Navneet Loiwal and Pushpak Bhattacharyya, 2004, Generic Text Summarization Using Wordnet, Language Resources Engineering Conference (LREC 2004), Barcelona, May, 2004.
- www.wikipedia.com