Paper presentation of

TextRank: Bringing Order into Texts

by Rada Mihalcea and Paul Tarau

Course: CS 657 INFORMATION RETRIEVAL

Group Members:

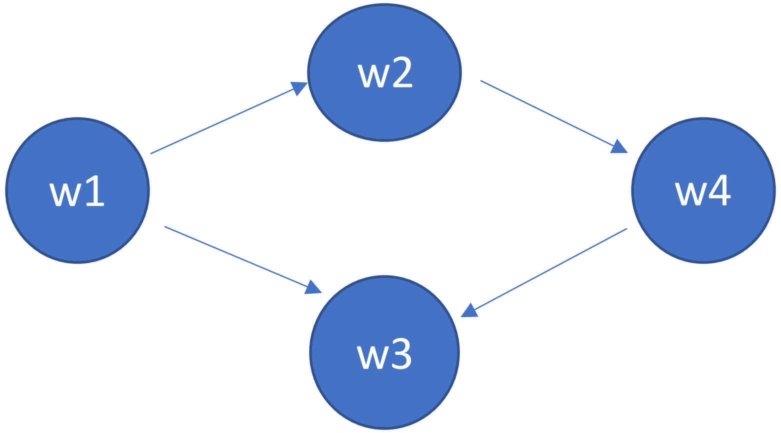
Shashank Bhusan 21111074,Sunil Dhaka 17817735,Punyabrat Kalwar 21211403,Vikas Lodhi 21111068, Spoorthi S.L. 160704

Introduction

- We present TextRank – a graph-based ranking model for text processing, and show how this model can be successfully used in natural language applications. We investigate and evaluate the application of TextRank to two language processing tasks consisting of unsupervised keyword and sentence extraction.
- Graph-based ranking algorithms like HITS and PageRank have triggered a paradigm-shift in the field of Web search technology, by providing a Web page ranking mechanism that relies on the collective knowledge of Web architects. Applying a similar line of thinking to lexical or semantic graphs extracted from natural language documents, results in a graph-based ranking model that can be applied to a variety of natural language processing applications, where knowledge drawn from an entire text is used in making local ranking/selection decisions.

The TextRank Model

- Graph-based ranking algorithms are essentially a way of deciding the importance of a vertex within a graph, based on global information recursively drawn from the entire graph. The basic idea implemented by a graph-based ranking model is that of "voting" or "recommendation". Links = votes.



Vertices= Pages

Edges=Links

In(w3)=2

Out(w4)=1

The higher the number of links for a vertex, the higher the importance of the vertex. Moreover, the importance of the vertex casting the vote determines how important the vote itself is, and this information is also taken into account by the ranking model.

Score of Vertex Vi:

$$S(V_i) = (1-d) + d * \sum_{j \in In(v_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

- S(Vi) - the weight of webpage i
- d - damping factor, in case of no outgoing links
- In(Vi) - inbound links of i, which is a set
- Out(Vj) - outgoing links of j, which is a set
- |Out(Vj)| - the number of outbound links

This can also be extended to weighted graphs as:

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

Creating Graph from Text Data

- To be able to use, graph based algorithms like pagerank, first we build a graph(G) that represents the text. Graph(G) = (V,E).
- Steps to create graph for natural language texts:

1. text units needs to be identified and put as vertices in the graph.
    - text unit depends on the task at hand. some examples: words, keyphrases, entire sentences etc.

2.Now to draw connections between such text units(vertices) we need to identify relations (edges). This also depends on the task that we are undertaking.

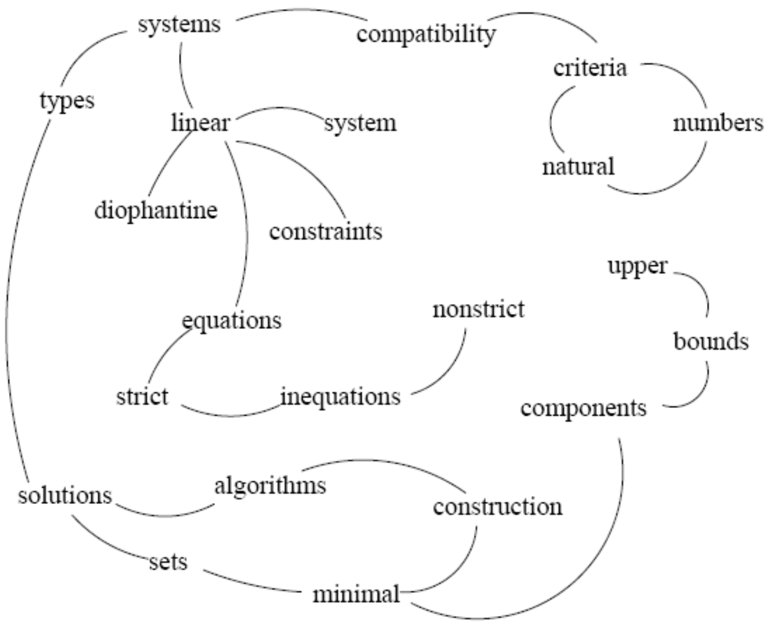
- For sentences we could use content overlap similarity. content overlap of two sentences can be determined simply as the number of common tokens between sentences.
- We can also use cosine similarity between vector representations of sentences, that we get from word embedding.
- When we use keyphrases as vertices in graph, we can use syntactic similarity b/w them to identify relation.

3.We iteratively apply the graph based ranking algorithms on the constructed graph, until convergence.

4.Based on the final values of the importance score, sort vertices and use it for selection decision.

Keyword Extraction

- The graph employed to extract keywords has:
  - Nodes – unigrams
  - Edges – formulation like Levenstein distance that captures syntactic similarity between the two words



Keyword Extraction

- Syntactic filter is used to get the nodes
- Uses a co-occurrence relation in order to add edges between two nodes
- Apply ranking algorithm
- Uses pagerank until it converges and generates scores for each node
- Sample top k notes and apply post processing step
- During post-processing, all lexical units selected as potential keywords by the textrank algorithm are marked in the text, and sequences of adjacent keywords are collapsed into a multi-word keyword.

Keyword Extraction Evaluation

- About 500 abstracts from the Inspec database are used for the experiment.
- The maximum recall that can be achieved on this collection is less than 100%, since indexers were not limited to keyword.
- The system consists of the TextRank approach, with a co-occurrence window size set to two, three, five, or ten words.
- TextRank is completely unsupervised and it relies exclusively on information drawn from the text itself, making it easily portable to other text collections, domains, and languages.

Sentence Extraction

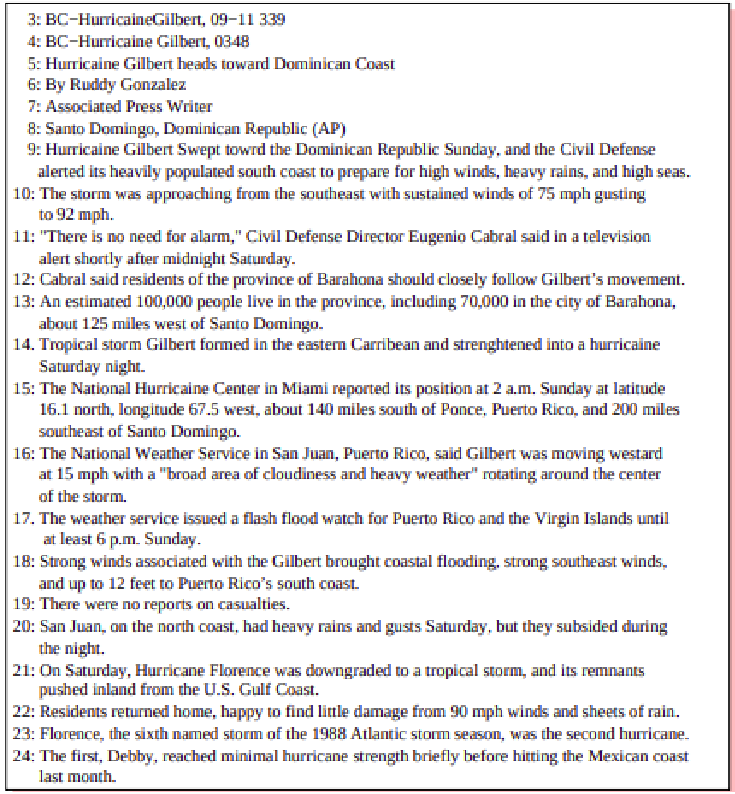
- Similar to keyword extraction
  - Nodes – sentences in document
  - Edges - similarity between sentences
- Similarity
  - Let a sentence be represented by word

    $$S_i = w_1^i, w_2^i, ..., w_{N_i}^i$$

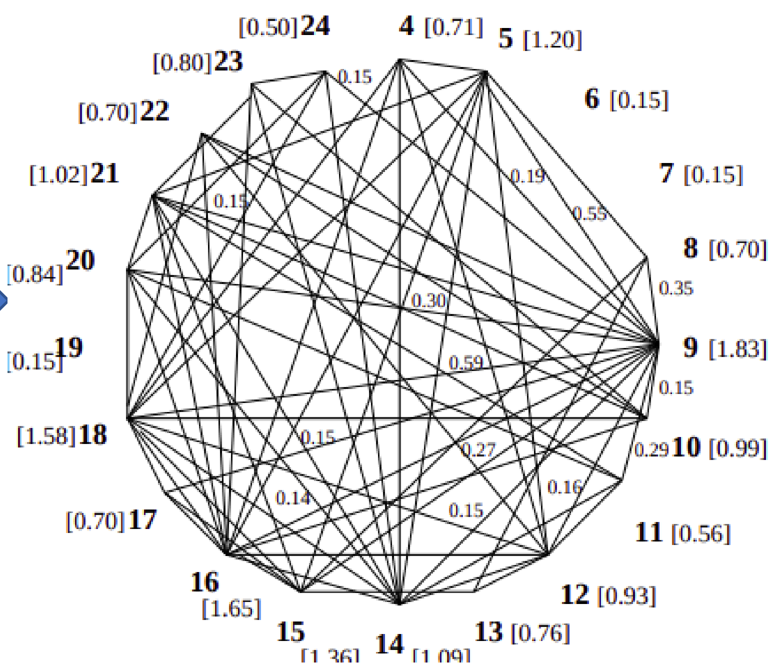
  - Define $$Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)}$$
  - Any other similarity can be used such as cosine similarity and longest common subsequence can be used
- This forms a weighted undirected graph

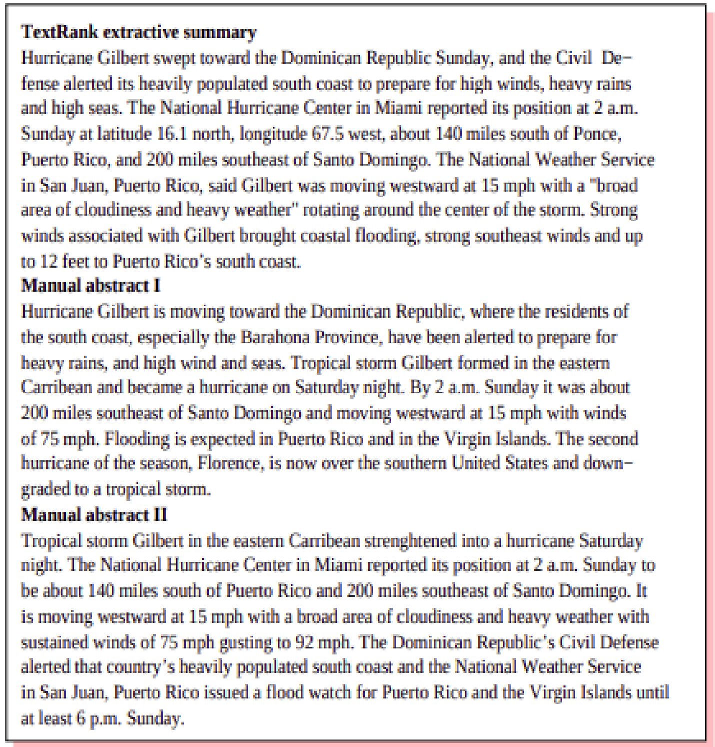
Sentence Extraction

- Page ranking algorithm is run on the graph,
- Sentences are sorted in reversed order of their score,
- Top ranked sentences are selected for inclusion



3: BC−HurricaineGilbert, 09−11 339
4: BC−Hurricaine Gilbert, 0348
5: Hurricaine Gilbert heads toward Dominican Coast
6: By Ruddy Gonzalez
7: Associated Press Writer
8: Santo Domingo, Dominican Republic (AP)
9: Hurricaine Gilbert Swept towrd the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains, and high seas.
10: The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph.
11: "There is no need for alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly after midnight Saturday.
12: Cabral said residents of the province of Barahona should closely follow Gilbert's movement.
13: An estimated 100,000 people live in the province, including 70,000 in the city of Barahona, about 125 miles west of Santo Domingo.
14. Tropical storm Gilbert formed in the eastern Carribean and strenghtened into a hurricaine Saturday night.
15: The National Hurricaine Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo.
16: The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westard at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm.
17. The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday.
18: Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds, and up to 12 feet to Puerto Rico's south coast.
19: There were no reports on casualties.
20: San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsided during the night.
21: On Saturday, Hurricane Florence was downgraded to a tropical storm, and its remnants pushed inland from the U.S. Gulf Coast.
22: Residents returned home, happy to find little damage from 90 mph winds and sheets of rain.
23: Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane.
24: The first, Debby, reached minimal hurricane strength briefly before hitting the Mexican coast last month.





TextRank extractive summary

Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas. The National Hurricane Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo. The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm. Strong winds associated with Gilbert brought coastal flooding, strong southeast winds and up to 12 feet to Puerto Rico's south coast.

Manual abstract I

Hurricane Gilbert is moving toward the Dominican Republic, where the residents of the south coast, especially the Barahona Province, have been alerted to prepare for heavy rains, and high wind and seas. Tropical storm Gilbert formed in the eastern Carribean and became a hurricane on Saturday night. By 2 a.m. Sunday it was about 200 miles southeast of Santo Domingo and moving westward at 15 mph with winds of 75 mph. Flooding is expected in Puerto Rico and in the Virgin Islands. The second hurricane of the season, Florence, is now over the southern United States and downgraded to a tropical storm.

Manual abstract II

Tropical storm Gilbert in the eastern Carribean strenghtened into a hurricane Saturday night. The National Hurricane Center in Miami reported its position at 2 a.m. Sunday to be about 140 miles south of Puerto Rico and 200 miles southeast of Santo Domingo. It is moving westward at 15 mph with a broad area of cloudiness and heavy weather with sustained winds of 75 mph gusting to 92 mph. The Dominican Republic's Civil Defense alerted that country's heavily populated south coast and the National Weather Service in San Juan, Puerto Rico issued a flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday.

Sentence Extraction
Evaluation

- Text-Rank sentence extraction algorithm on a single-document summarization task.
- Document Understanding Evaluations 2002 (DUC, 2002)-consist of using 567 news articles
- For each article, Text Rank generates an summary and compared with summary written by expert
- ROUGE evaluation toolkit, which is a method based on Ngram
- TextRank, top 5 (out of 15) DUC 2002

Conclusions

- This paper, introduced TextRank – a graphbased ranking model for text processing, and show how it can be successfully used for natural language applications.
- In particular,it proposed and evaluated two innovative unsupervised approaches for keyword and sentence extraction, and showed that the accuracy achieved by TextRank in these applications is competitive with that of previously proposed state-of-the-art algorithms.
- The important aspect of TextRank is that it does not require deep linguistic knowledge, nor domain or language specific annotated corpora, which makes it highly portable to other domains, genres, or language