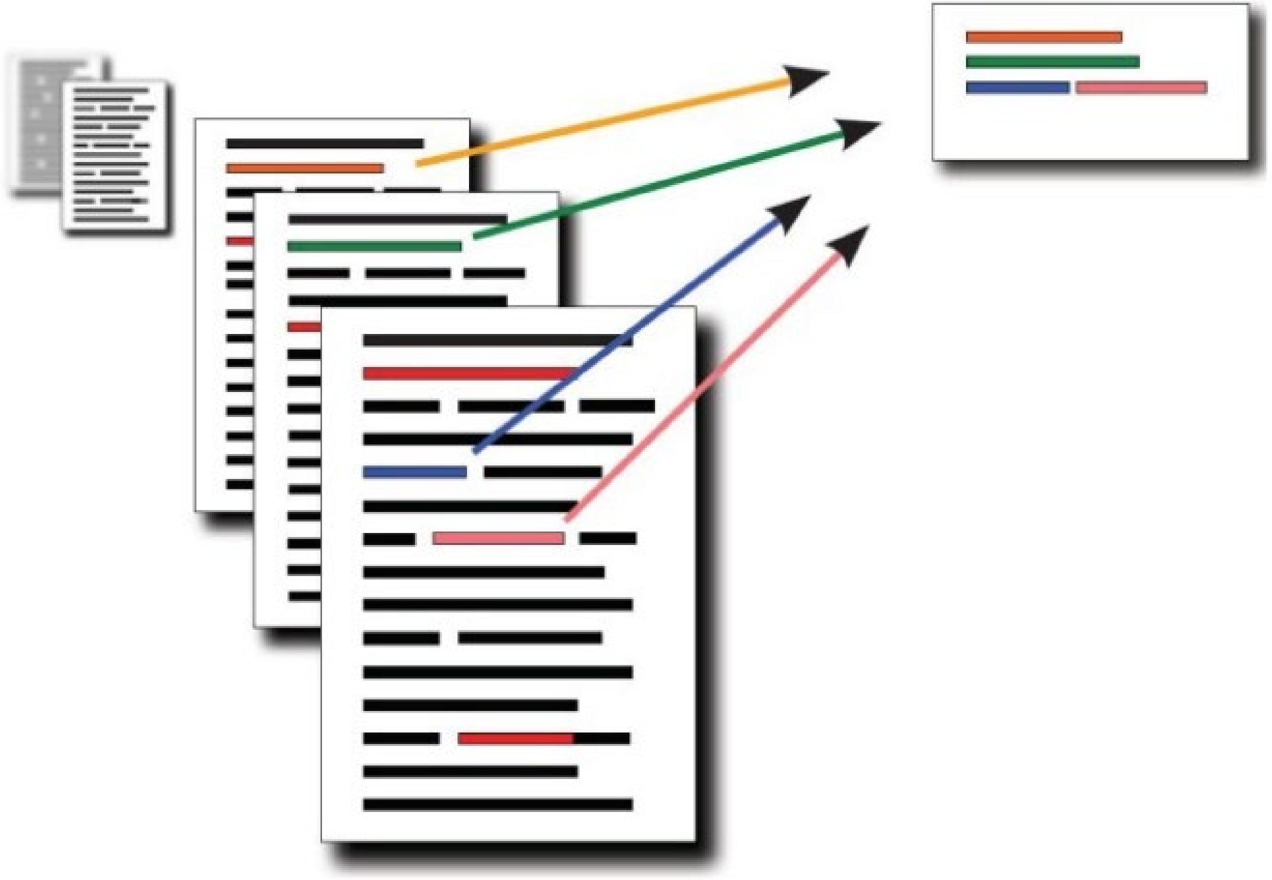
Extended topic - Text Summarization

Summarization

Text summarization is the task of creating a ***shorter, coherent version*** of a longer text while ***preserving its key information and meaning***



Why it is useful?

- Saves time for readers – rapid understanding without reading entire texts
- Helps cope with information overload – Quick consumption of essential information
- Improves accessibility for wider audiences
- Enables translation and multi-lingual communication

Boosts productivity!

Examples

- Summary of news articles
- Document summarization (legal, medical, etc.)
- Scientific paper abstract
- Book summaries/ synopsis
- Product reviews summary
- Meeting minutes summary

Types of summarization

Based on:

Input

- Single document summarization (SDS)
- Multiple document summarization (MDS)

Input type:
Single-doc summarization

Document

Cambodian leader Hun Sen on Friday rejected opposition parties ' demands for talks outside the country , accusing them of trying to `` internationalize " the political crisis .
Government and opposition parties have asked King Norodom Sihanouk to host a summit meeting after a series of post-election negotiations between the two opposition groups and Hun Sen 's party to form a new government failed .
Opposition leaders Prince Norodom Ranariddh and Sam Rainsy , citing Hun Sen 's threats to arrest opposition figures after two alleged attempts on his life , said they could not negotiate freely in Cambodia and called for talks at Sihanouk 's residence in Beijing .Hun Sen , however , rejected that .``
I would like to make it clear that all meetings related to Cambodian affairs must be conducted in the Kingdom of Cambodia , " Hun Sen told reporters after a Cabinet meeting on Friday .`` No-one should internationalize Cambodian affairs .
It is detrimental to the sovereignty of Cambodia , " he said .Hun Sen 's Cambodian People 's Party won 64 of the 122 parliamentary seats in July 's elections , short of the two-thirds majority needed to form a government on its own .Ranariddh and Sam Rainsy have charged that Hun Sen 's victory in the elections was achieved through widespread fraud .They have demanded a thorough investigation into their election complaints as a precondition for their cooperation in getting the national assembly moving and a new government formed ....

Summary

Cambodian government rejects opposition's call for talks abroad

Multi-doc summarization

Documents

Fingerprints and photos of two men who boarded the doomed Malaysia Airlines passenger jet are being sent to U.S. authorities so they can be compared against records of known terrorists and criminals. The cause of the plane's disappearance has baffled investigators and they have not said that they believed that terrorism was involved, but they are also not ruling anything out. The investigation into the disappearance of the jetliner with 239 passengers and crew has centered so far around the fact that two passengers used passports stolen in Thailand from an Austrian and an Italian. The plane which left Kuala Lumpur, Malaysia, was headed for Beijing. Three of the passengers, one adult and two children, were American. ……

(CNN) -- A delegation of painters and calligraphers, a group of Buddhists returning from a religious gathering in Kuala Lumpur, a three-generation family, nine senior travelers and five toddlers. Most of the 227 passengers on board missing Malaysia Airlines Flight 370 were Chinese, according to the airline's flight manifest. The 12 missing crew members on the flight that disappeared early Saturday were Malaysian. The airline's list showed the passengers hailed from 14 countries, but later it was learned that two people named on the manifest -- an Austrian and an Italian -- whose passports had been stolen were not aboard the plane. The plane was carrying five children under 5 years old, the airline said. ……

⋮

Vietnamese aircraft spotted what they suspected was one of the doors belonging to the ill-fated Malaysia Airlines Flight MH370 on Sunday, as troubling questions emerged about how two passengers managed to board the Boeing 777 using stolen passports. The discovery comes as officials consider the possibility that the plane disintegrated mid-flight, a senior source told Reuters. The state-run Thanh Nien newspaper cited Lt. Gen. Vo Van Tuan, deputy chief of staff of Vietnam's army, as saying searchers in a low-flying plane had spotted an object suspected of being a door from the missing jet. It was found in waters about 56 miles south of Tho Chu island, in the same area where oil slicks were spotted Saturday. ……

Summary

Flight MH370, carrying 239 people vanished over the South China Sea in less than an hour after taking off from Kuala Lumpur, with two passengers boarded the Boeing 777 using stolen passports. Possible reasons could be an abrupt breakup of the plane or an act of terrorism. The government was determining the "true identities" of the passengers who used the stolen passports. Investigators were trying to determine the path of the plane by analysing civilian and military radar data while ships and aircraft from seven countries scouring the seas around Malaysia and south of Vietnam.

Types of summarization

Based on:

Output

- Extractive summarization
- Abstractive summarization

Output

Extractive summarization

Selects (extracts) from the source text spans that capture the key information

Original text: "The feline, a domesticated member of the family Felidae, often referred to as the domestic cat, is a small carnivorous mammal that is valued by humans for companionship and its ability to hunt vermin and household pests."

Extractive summary: "The feline, often referred to as the domestic cat, is valued by humans for companionship and its ability to hunt vermin and household pests."

Abstractive summarization

Generates new text that encapsulates the key information from the source text

Words in the summary may not be present in the source text (s)!

Original text: "The feline, a domesticated member of the family Felidae, often referred to as the domestic cat, is a small carnivorous mammal that is valued by humans for companionship and its ability to hunt vermin and household pests.“

Abstractive summary: "Cats are popular pets known for their companionship and pest control abilities."

Types of summarization

Based on:

Approach

- Supervised
- Unsupervised

Approach

Supervised summarization

- Uses labelled training data (pairs of full texts and their summaries)
- Learns to generate summaries based on examples
- Uses Neural Network based models (E.g. BART, GPT)
- Requires large amounts of labelled data, computationally expensive, can hallucinate

Unsupervised summarization

- Does not require labeled training data
- Relies on statistical or linguistic features of the text itself
- Uses graph based methods (E.g. TextRank), LSA, Clustering, LDA, etc.
- Uses existing sentences, may lack coherence, struggles to capture abstract concepts

Types of summarization

Based on:

Focus

- Generic
- Query-focused

Focus

Generic summarization

- Summarize the content of the doc(s)

Query-focused summarization

- Summarize a doc w.r.t. a user's query
- Complex question-answering (answer a question by summarizing a doc that has the information to construct the answer)

Query-focused summarization - Example

Query-focused summarization

- Creates answers to complex questions, summarizing multiple documents
- Instead of giving a snippet for each document, creates a cohesive answer that combines information from multiple documents

Q: Does this seem like a job for extractive or abstractive summarization?

Abstractive summarization

- More complex than extractive summarization
- Challenges in factual accuracy, coherence, computational efficiency, and handling domain-specific vocabulary
- Typically done using neural network architectures (which is our focus in the next lecture)
- There is increasing focus on *multi-modal* summarization (E.g. summarizing text + images)

Multimodal summarization

Q: What is an example application for *multimodal summarization*?

Multimodal Video Summarization via Time-Aware Transformers

Xindi Shang*
shangxin@comp.nus.edu.sg
National University of Singapore
Singapore

Zehuan Yuan†
yuanzehuan@bytedance.com
ByteDance Inc.
China

Anran Wang
anran.wang@bytedance.com
ByteDance Inc.
Singapore

Changhu Wang
wangchanghu@bytedance.com
ByteDance Inc.
China

ABSTRACT

With the growing number of videos in video sharing platforms, how to facilitate the searching and browsing of the user-generated video has attracted intense attention by multimedia community. To help people efficiently search and browse relevant videos, summaries of videos become important. The prior works in multimodal video summarization mainly explore visual and ASR tokens as two separate sources and struggle to fuse the multimodal information for generating the summaries. However, the time information inside videos is commonly ignored. In this paper, we find that it is important to leverage the timestamps to accurately incorporate multimodal signals for the task. We propose a Time-Aware Multimodal Transformer (TAMT) with a novel short-term order-sensitive attention mechanism. The attention mechanism can attend the inputs differently based on time difference to explore the time information inherent inside video more thoroughly. As such, TAMT can fuse the different modalities better for summarizing the videos. Experiments show that our proposed approach is effective and achieves the state-of-the-art performances on both YouCookII and open-domain How2 datasets.

CCS CONCEPTS

• **Computing methodologies → Video summarization.**

KEYWORDS

video description, time-aware, multimodal summarization

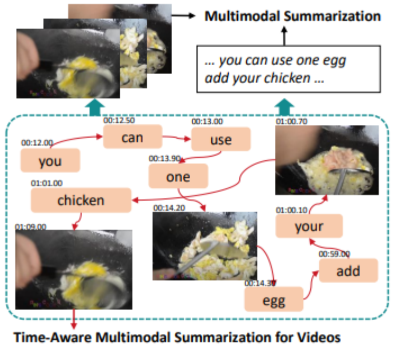


Figure 1: Illustration of multimodal video summarization. As can be seen, time information offers important clues to better fuse the vision and speech modalities for the task.

1 INTRODUCTION

Videos tend to comprise multiple modalities these days. In fact, the

DOC2PPT: Automatic Presentation Slides Generation from Scientific Documents

Tsu-Jui Fu[1], William Yang Wang[1], Daniel McDuff[2], Yale Song[2]
[1] UC Santa Barbara [2] Microsoft Research
{tsu-juifu,william}@cs.ucsb.edu {damcduff, yalesong}@microsoft.com

Abstract

Creating presentation materials requires complex multimodal reasoning skills to summarize key concepts and arrange them in a logical and visually pleasing manner. Can machines learn to emulate this laborious process? We present a novel task and approach for document-to-slide generation. Solving this involves document summarization, image and text retrieval, and slide structure to arrange key elements in a form suitable for presentation. We propose a hierarchical sequence-to-sequence approach to tackle our task in an end-to-end manner. Our approach exploits the inherent structures within documents and slides and incorporates paraphrasing and layout prediction modules to generate slides. To help accelerate research in this domain, we release a dataset of about 6K paired documents and slide decks used in our experiments. We show that our approach outperforms strong baselines and produces slides with rich content and aligned imagery.

Introduction

Creating presentations is often a work of art. It requires skills to abstract complex concepts and conveys them in a concise and visually pleasing manner. Consider the steps involved in creating presentation slides based on a white paper or manuscript: One needs to 1) establish a storyline that will connect with the audience, 2) identify essential sections and components that support the main message, 3) delineate the structure of that content, e.g., the ordering/length of the sections, 4) summarize the content in a concise form, e.g.,

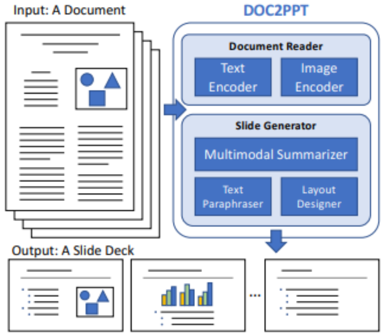


Figure 1: We introduce DOC2PPT, a novel task of generating a slide deck from a document. This requires solving several challenges in the vision-and-language domain, e.g., visual-semantic embedding and multimodal summarization. In addition, slides exhibit unique properties such as concise text (bullet points) and stylized layout.

and *visual-centric* (e.g., figures are first-class citizens, the visual layout plays an important role, etc.).

Ideas for creative assignments for AT2!

CNN/ Dailymail data set

https://huggingface.co/datasets/cnn_dailymail

A commonly used data set for summarization

Dataset Summary

The CNN / DailyMail Dataset is an English-language dataset containing just over 300k unique news articles as written by journalists at CNN and the Daily Mail. The current version supports both extractive and abstractive summarization, though the original version was created for machine reading and comprehension and abstractive question answering.

Dataset Split	Number of Instances in Split
Train	287,113
Validation	13,368
Test	11,490

Summarization articles on the CNN/ Dailymail data set

Text Summarization with Pretrained Encoders

Yang Liu and **Mirella Lapata**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
yang.liu2@ed.ac.uk, mlap@inf.ed.ac.uk

SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization

Yixin Liu
Carnegie Mellon University
yixinl2@cs.cmu.edu

Pengfei Liu *
Carnegie Mellon University
pliu3@cs.cmu.edu

Get To The Point: Summarization with Pointer-Generator Networks

Abigail See
Stanford University
abisee@stanford.edu

Peter J. Liu
Google Brain
peterjliu@google.com

Christopher D. Manning
Stanford University
manning@stanford.edu

Find the state-of-the-art (SOTA) papers for any NLP task:

- https://scholar.google.com/
- https://nlpprogress.com/
- paperswithcode.com/sota
- connectedpapers.com

Evaluation metric: ROUGE-N

Recall-Oriented Understudy for Gisting Evaluation (**ROGUE**) is a commonly used *intrinsic metric* for automatically evaluating summaries

ROGUE-N measures the *overlap of n-grams* between the generated summary and one or more reference summaries

- Scores range from 0 to 1, where 1 indicates a perfect match
- Not as good as human evaluation (of course)

Toy example:
Reference: "The cat sat on the mat"
Generated: "The cat sat on the rug"

ROUGE-1 = 5/6 = 0.833 (5 matching unigrams out of 6 in the reference: the, cat, sat, on, the)

How is ROGUE-N calculated?

Given a document D, and an automated summary X

- Get N humans to produce a set of reference summaries of D
- Run model to produce X

4 ROUGE

The state-of-the-art automatic summarization evaluation method is ROUGE (Recall Oriented Understudy for Gisting Evaluation, (Hovy and Lin 2002)), an n-gram based comparison that was motivated by the machine translation evaluation metric, Bleu (Papineni et. al. 2001). This system uses a variety of n-gram matching approaches, some of which allow gaps within the matches as well as more sophistcated analyses. Surprisingly, simple unigram and bigram matching works extremely well. For example, at DUC 05, ROUGE-2 (bigram match) had a Spearman correlation of 0.95 and a Pearson correlation of 0.97 when compared with human evaluation of the summaries for responsiveness (Dang 2005). ROUGE-n for matching $n-$grams of a summary X against h model human summaries is given by:

$$R_n(X) = \frac{\sum_{j=1}^{h} \sum_{i \in N_n} \min(X_n(i), M_n(i,j))}{\sum_{j=1}^{h} \sum_{i \in N_n} M_n(i,j)},$$

where $X_n(i)$ is the count of the number of times the n-gram i occurred in the summary and $M_n(i,j)$ is the number of times the n-gram i occurred in the j-th model (human) summary. (Note that for brevity of notation, we assume that lemmatization (stemming) is done *apriori* on the terms.)

https://dl.acm.org/doi/pdf/10.5555/1273073.1273093

ROUGE-N

- Query: “What is water spinach?”
- Model’s output: “Water spinach is a leaf vegetable commonly eaten in tropical areas of Asia.”
- Human Summaries (gold truth):

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.

$$\text{ROUGE-2} = \frac{3+3+6}{10+9+9} = 12/28 = .43$$

Time to put it into action!

tinyurl.com/ANLPColab3Part1

Let’s run an extractive summarization step-by-step (Using a graph-based ranking algorithm, unsupervised)

1. Text Preprocessing
    1. Tokenize the text into sentences.
    2. Remove stopwords and perform any necessary text cleaning.
    3. Optionally, apply stemming or lemmatization to words.
2. Create a Graph:
    1. Each sentence becomes a node in the graph.
    2. Edges are created between sentences based on their similarity.
3. Calculate Sentence Similarity:
    1. For each pair of sentences, compute their similarity.
    2. Common methods include: a) Cosine similarity of word vectors b) Overlap of words (Jaccard similarity) c) Embedding-based similarity (e.g., using word2vec or BERT embeddings)
4. Build the Similarity Matrix:
    1. Create an NxN matrix where N is the number of sentences.
    2. Each cell [i,j] contains the similarity score between sentences i and j.

5. Apply the TextRank Algorithm:
   - Initialize each node (sentence) with a default score (usually 1).
   - Iteratively update the score of each node using the TextRank formula: $WS(V_i) = (1-d) + d * \Sigma(W_{ji} / \Sigma(W_{jk}) * WS(V_j))$ Where:
     - $WS(V_i)$ is the score of the current node
     - d is a damping factor (typically 0.85)
     - W_{ji} is the weight of the edge between nodes j and i
     - $\Sigma(W_{jk})$ is the sum of weights of edges from node j to all its neighbors
     - $WS(V_j)$ is the score of node j
6. Iterate Until Convergence:
   - Repeat step 5 until the scores stabilize (change less than a small threshold).
   - This usually takes 20-30 iterations.
7. Rank Sentences:
   - Sort the sentences based on their final TextRank scores.
8. Select Top Sentences:
   - Choose the top N sentences with the highest scores.
   - N is determined by the desired summary length.
9. Reorder Selected Sentences:
   - Arrange the selected sentences in their original order from the text.
10. Generate Summary:
    - Combine the selected sentences to form the final summary.

Toy example demo

Original text:

"TextRank is a graph-based algorithm. It is used for text summarization. The algorithm is inspired by PageRank. Graph nodes represent sentences. Edges represent similarity between sentences.”

S1: "TextRank is a graph-based algorithm."

S2: "It is used for text summarization."

S3: "The algorithm is inspired by PageRank."

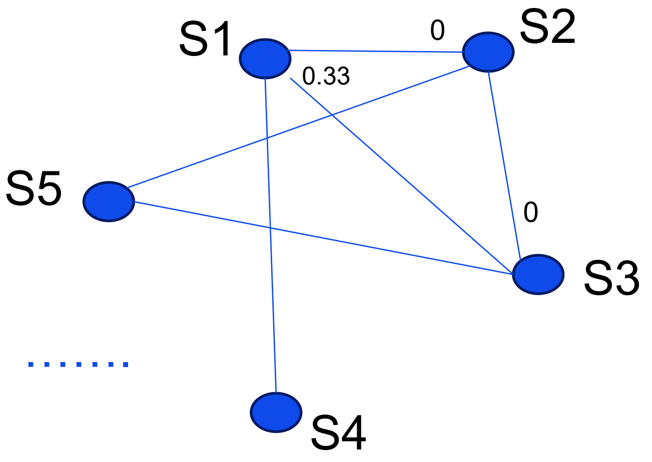
S4: "Graph nodes represent sentences."

S5: "Edges represent similarity between sentences.“

	textrank	graph	algorithm	based
S1	1	1	1	1
S2	0	0	0	0
S3	0	0	1	0
…	…..			

Cosine similarity matrix

	S1	S2	S3	…
S1	1	0	0.33	…
S2	0	1	0	
S3	0.33	0	1	
…	…..			



$$Similarity(S_i, S_j) = \frac{|W_k|W_k \in S_i \& W_k \in S_j|}{log(|S_i|) + log(|S_j|)}$$

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.

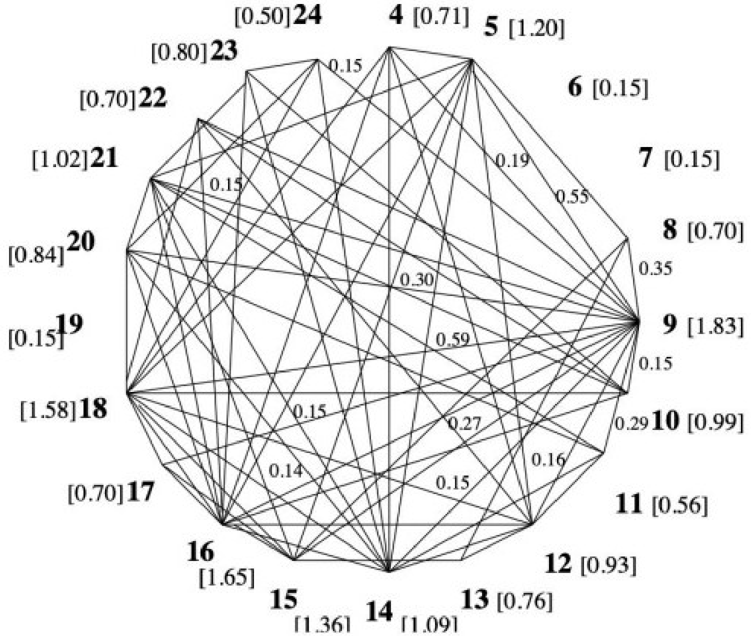


Figure 1: Sample graph build for sentence extraction from a newspaper article.

Rada Mihalcea, ACL 2004

Sample final ranking

S4: 0.4989 - "Graph nodes represent sentences."

S1: 0.7468 - "TextRank is a graph-based algorithm."

S5: 0.6340 - "Edges represent similarity between sentences."

S3: 0.8410 - "The algorithm is inspired by PageRank."

S2: 0.3345 - "It is used for text summarization."

Iterating scores

Initial scores: [1, 1, 1, 1, 1]
Iteration 1: S1: 0.15 + 0.85 * (0.33/0.33 + 0.41/1.04) = 0.7785 S2: 0.15 + 0.85 * (0.22/0.85) = 0.37 S3: 0.15 + 0.85 * (0.33/0.74) = 0.5285 S4: 0.15 + 0.85 * (0.41/0.74 + 0.63/0.85) = 0.8975 S5: 0.15 + 0.85 * (0.22/1.22 + 0.63/1.04) = 0.6985
.....

If we're selecting the top 2 sentences, we would choose:
S3: "The algorithm is inspired by PageRank."
S1: "TextRank is a graph-based algorithm."

Re-order based on original position

Extractive summary:
"TextRank is a graph-based algorithm. The algorithm is inspired by PageRank."