Research on Extractive Summarization at UFPE/UFRPE Universities:

Developed Approaches and Perspectives

Rinaldo Lima

Federal Rural University of Pernambuco, Recife/Brazil

rinaldo.jose@ufrpe.br

rjlima01@gmail.com

Dec/2019

UFRPE

Centro de Informática U·F·P·E

Agenda

- Presenting UFPE/UFRPE
- Automatic Text Summarization (ATS)
  - Main Approaches
  - Evaluation Methods, Metrics, and Datasets
- ATS Work at UFRPE/UFPE
- Current Trends in ATS
- Research Collaboration

Presenting:

UFRPE and UFPE

RECIFE (PE/Brazil)

One the Recife's Brigde (downtown)

Old Recife

Olinda

Porto de Galinhas Beach

UFRPE/UFPE

UFRPE

- Headquartered on **Recife Campus** with other 5 campuses in Pernambuco
- More than 1.000 teachers, 900 technicians, 17,000 students
- UFRPE provides courses on Agricultural sciences, Veterinary Medicine, Humanities, Biological, **Computer Science**, Physics, etc.
- The **Computing Department** conducts research on Computational Intelligence and Modelling, Software Engineering, Computational Systems, Text Mining, among others

UFRPE/UFPE

UFPE

- The **Informatics Center (CIn**) at UFPE was established 35 years ago in 1974
- Research at CIn encompasses the following areas: Databases, Computer Architecture, Software Engineering, Computational Intelligence, Distributed Systems, among others
- CIn is ranked among the top quality academic centers in Latin America
- Its faculty members consist of more than 80 PhDs

Automatic Text Summarization

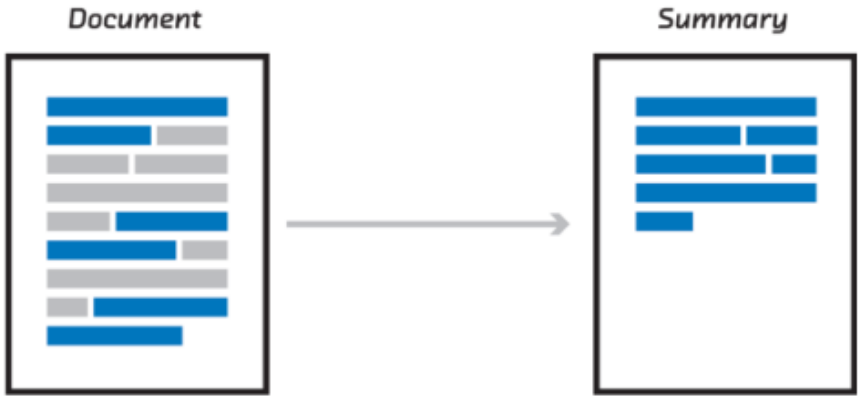
Automatic Text Summarization (ATS/TS)

Definitions

- A summary is a reductive transformation of a source text into a summary text by **extraction** or **generation** [Spärck-Jones and Sakai (2001)]

- A condensed version of a source document having a recognizable genre and a very specific purpose, to give the reader an **exact** and **concise idea** of the contents of the source [Saggion and Lapalme (2002)]

- The **ratio** between the length of the summary and the length of the source document is calculated by the compression rate:



$$\tau = \frac{|\text{Summary}|}{|\text{Source}|}$$

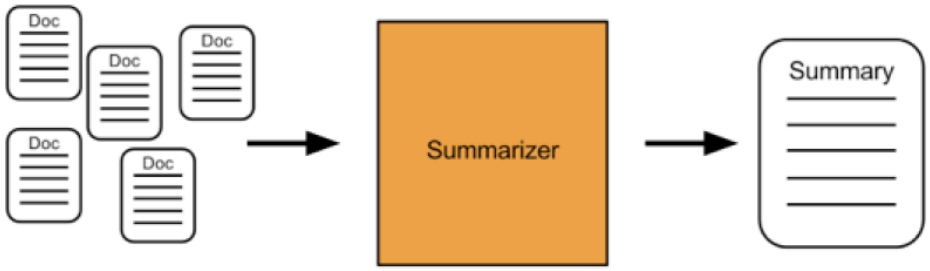
Automatic Text Summarization (ATS/TS)

Motivation to create summaries

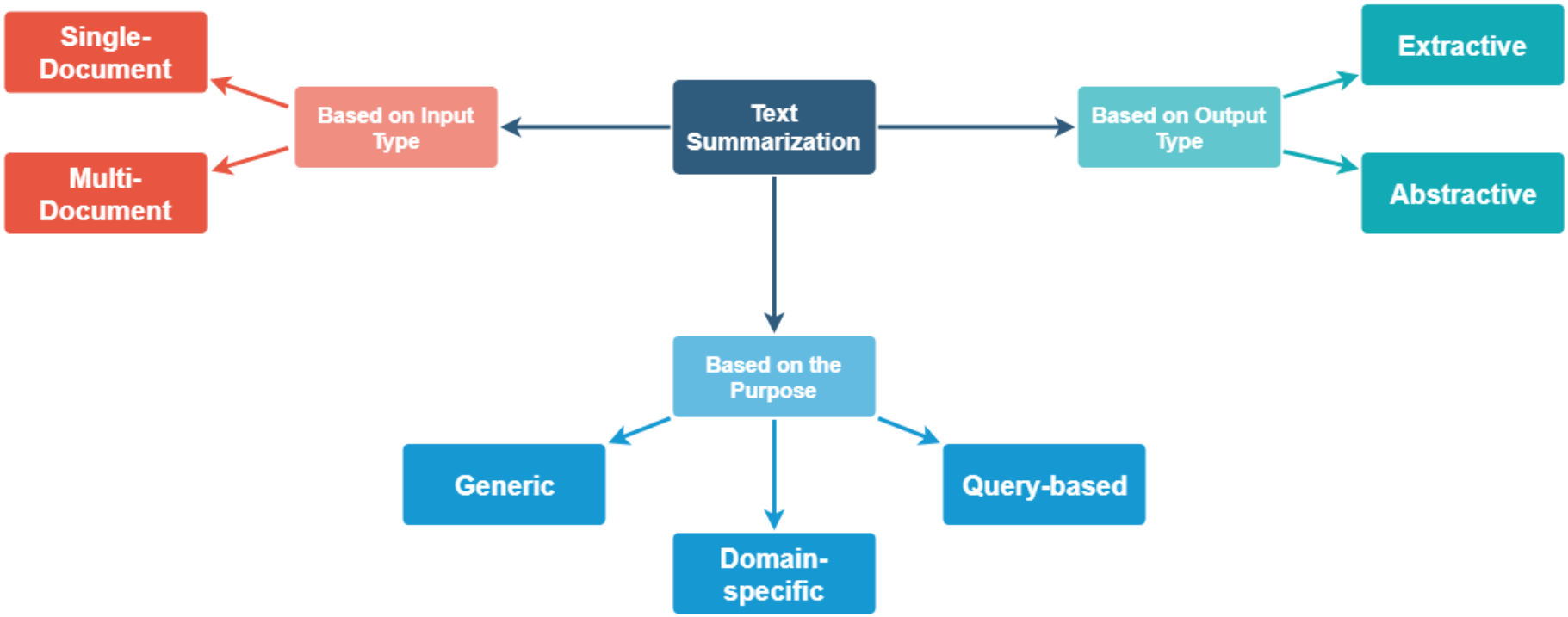
- more and more text data
- less time to process or read it

Main goals in ATS:

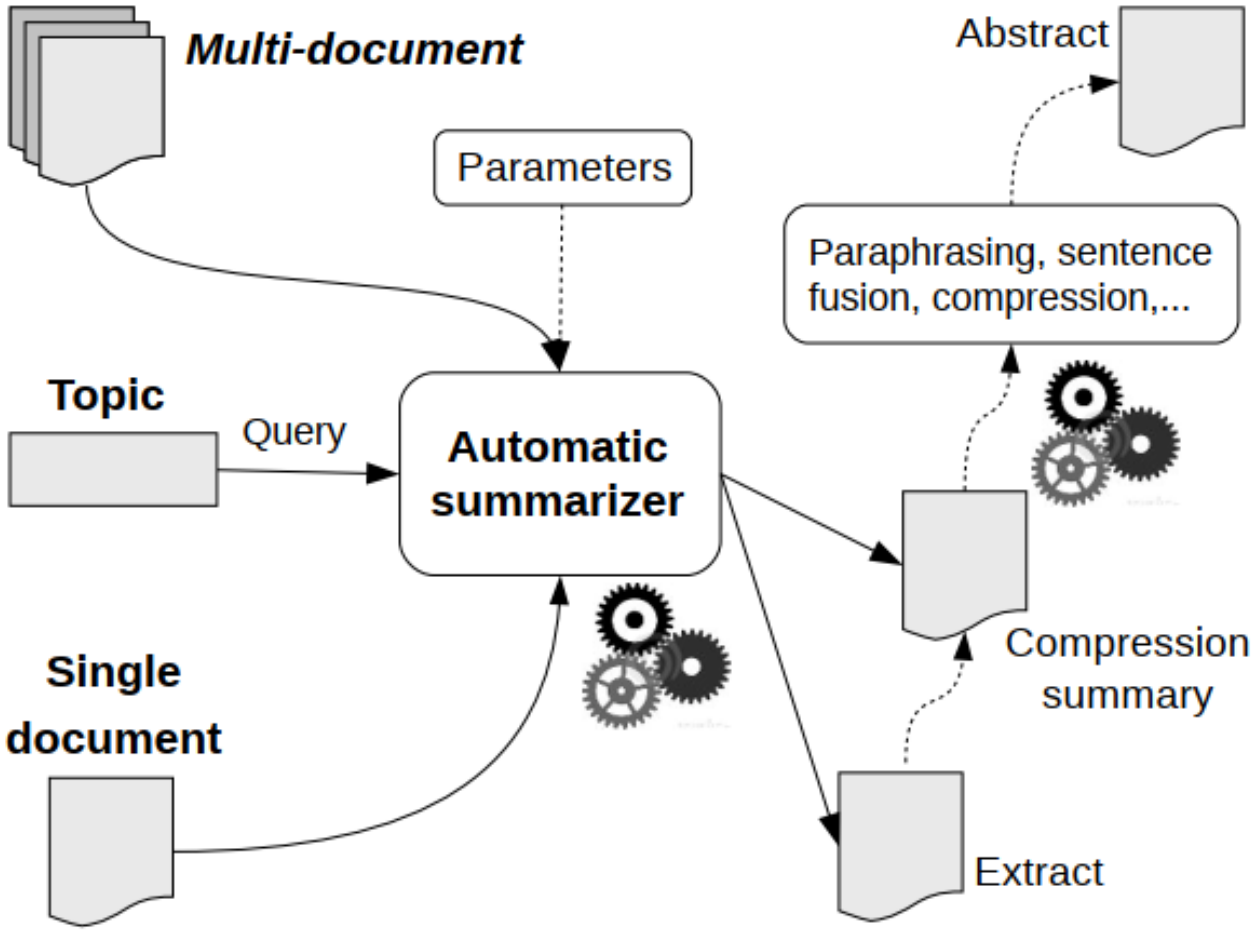
- optimizing topic coverage (**informativeness**)
- optimizing readability
- obtaining cohesive and coherent summaries
- avoiding redundancy



ATS Categorization



Automatic Text Summarization (ATS/TS)



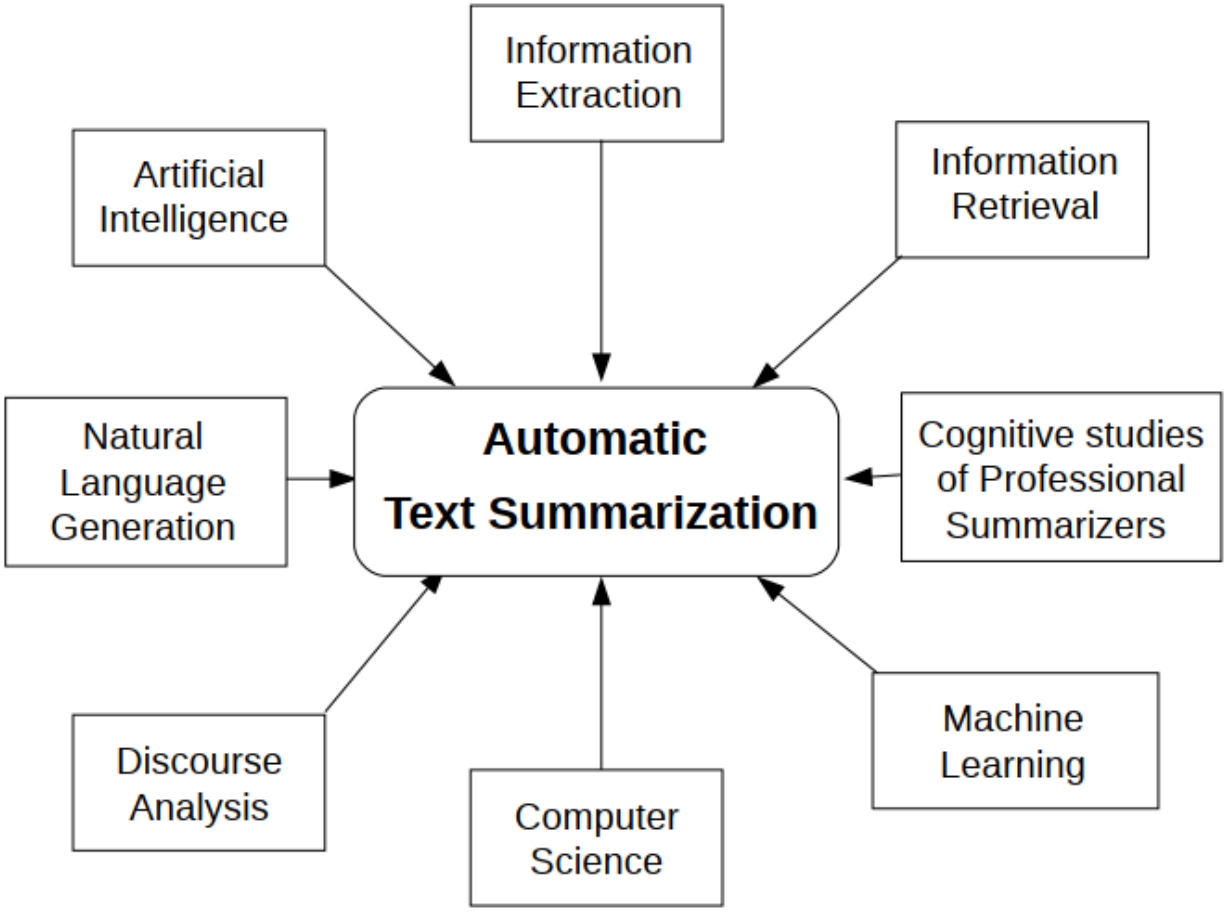
Simplified General Summarization Process

Automatic Text Summarization (ATS/TS)

TS Applications

- increasing the performance of traditional IR and IE systems (coupled with Question-Answering system)
- opinion summarization
- news summarization
- blog, tweet, web page, email thread summarization
- report summarization for business men, politicians, researchers, etc.;
- meeting summarization;
- biographical extracts
- automatic extraction and generation of titles and papers abstracts
- domain-specific summarization (domains of medicine, chemistry, law, etc.)
- …

Automatic Text Summarization (ATS/TS)



Fields of research which have influenced the development of ATS

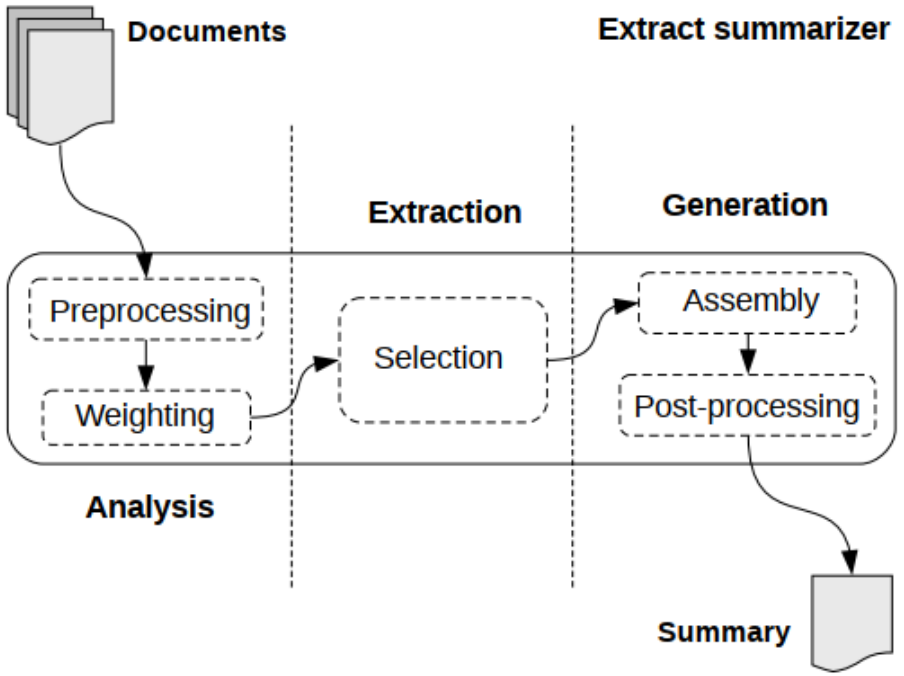
Text Summarization

Main Approaches

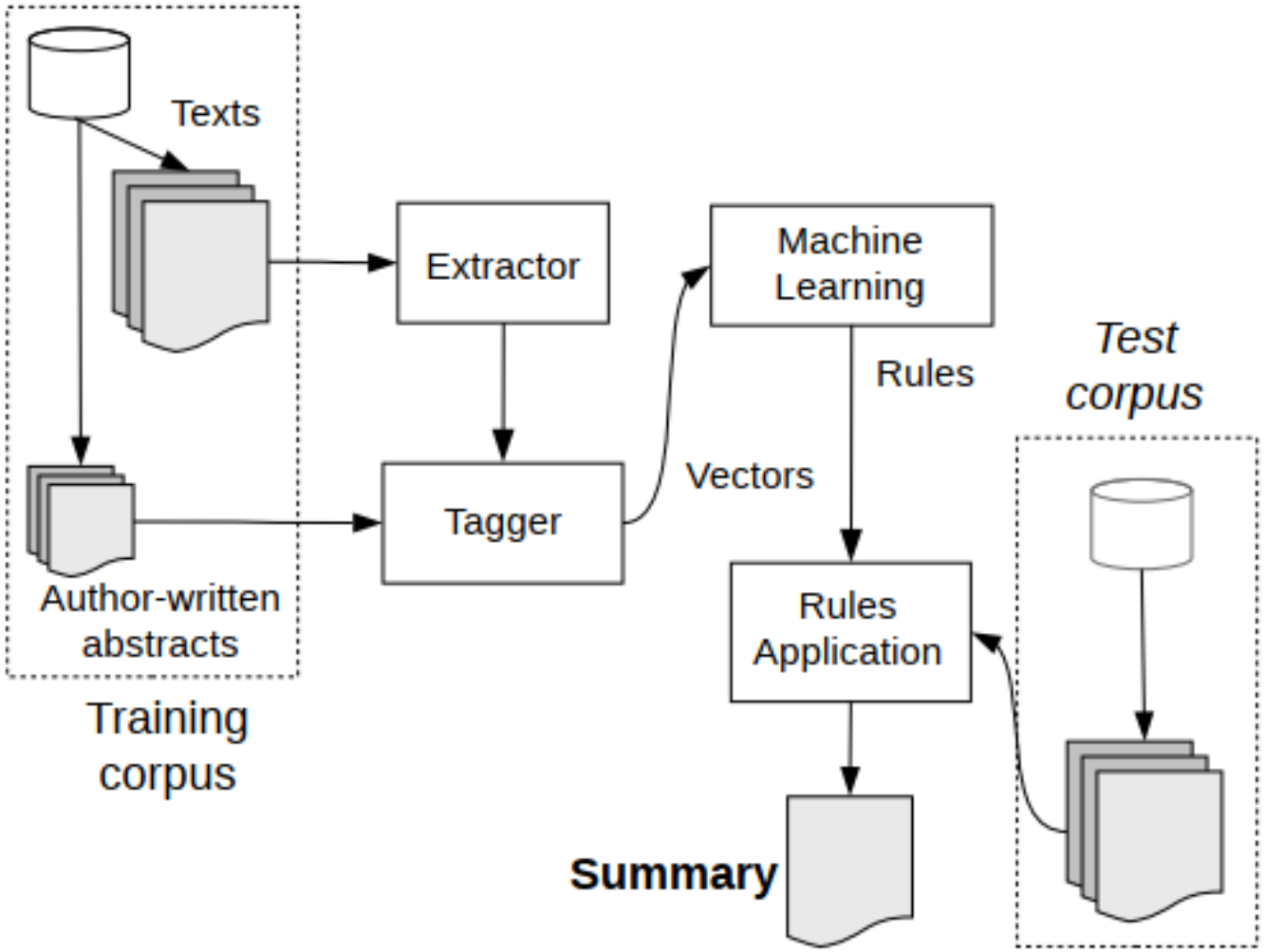
Main Approaches to ATS

Mono and Multidocument

- **Supervised Content Selection**
- **Unsupervised Content Selection (topic-based)**
- **Graph-based (LexRank)**
- **Lexical Chains**
- **Deep NLP-based**



Supervised Content Selection



16

Supervised Content Selection

- Given:
  - a labeled training set of good summaries for each document
- Align:
  - the sentences in the document with sentences in the summary
- Extract features
  - position (first sentence?)
  - length of sentence
  - word informativeness, cue phrases
- Train
  - a binary classifier (put sentence in summary? **yes** or **no**)

- Problems:
  - hard to get labeled training data
  - alignment difficult
  - performance not better than unsupervised algorithms
- So in practice:
  - **Unsupervised content selection is more common**

Unsupervised content selection

- Intuition dating back to Luhn (1958):
  - Choose sentences that have **salient** or **informative** words
- Two approaches to defining salient words
  1. **tf-idf**: weigh each word w_i in document j by tf-idf

$$weight(w_i) = tf_{ij} \times idf_i$$

  2. **topic signature:** choose a smaller set of salient words
     - mutual information
     - log-likelihood ratio (LLR) Dunning (1993), Lin and Hovy (2000)

$$weight(w_i) = \begin{cases} 1 & \text{if } -2\log \lambda(w_i) > 10 \\ 0 & \text{otherwise} \end{cases}$$

18

Graph-based (Ranking Methods)

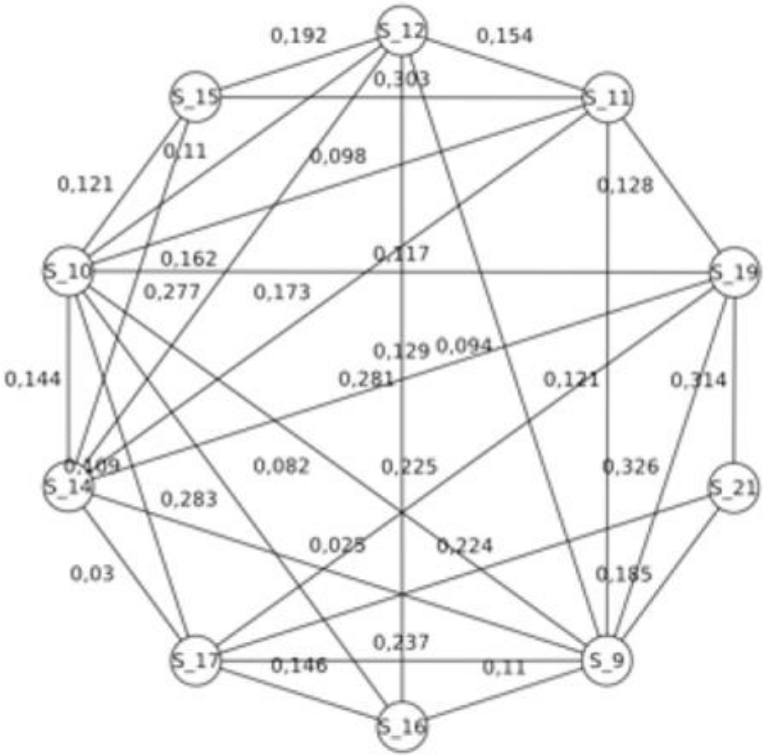
Graph-based Approach (Ranking idea)

- Sentence as vertices
- Similarity as edges

Iterative ranking

- LexRank, TextRank (inspired by PageRank)

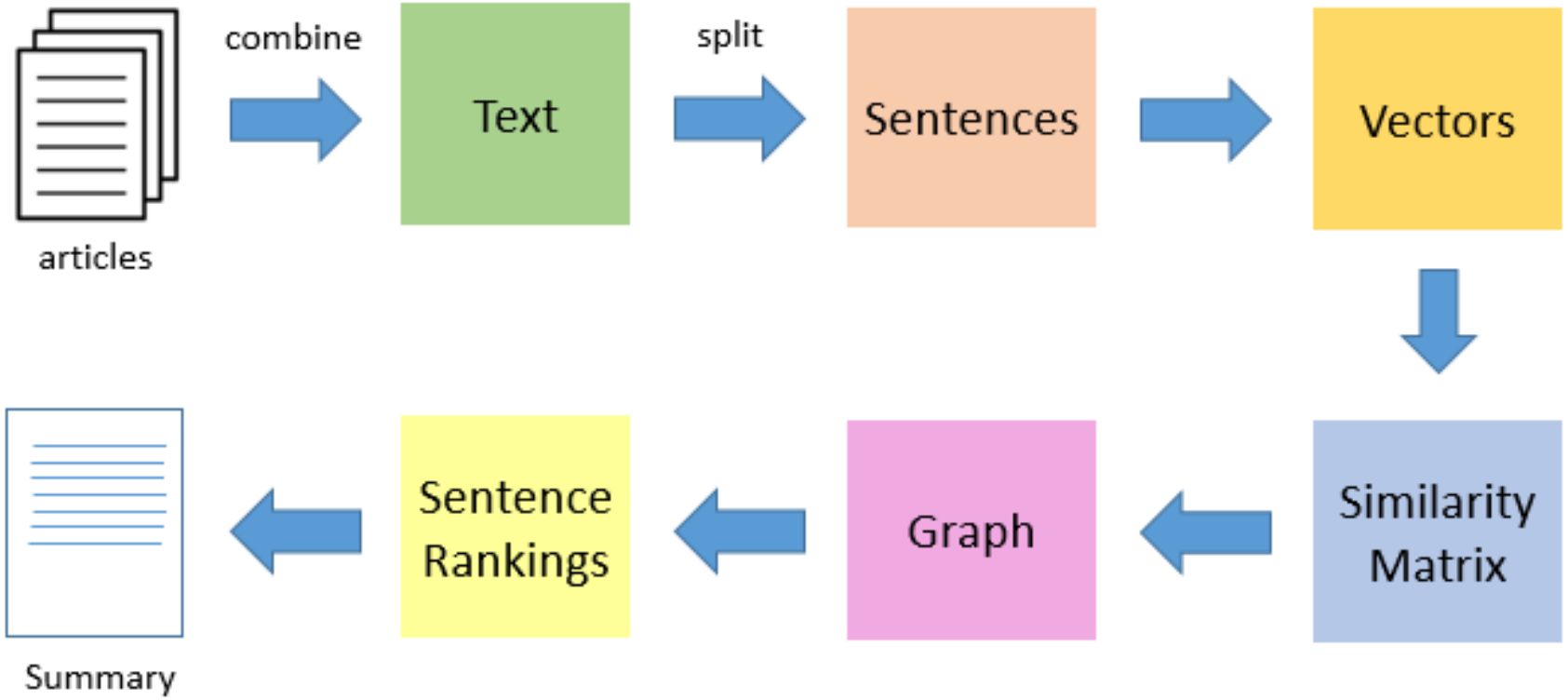
The main idea is that sentence "recommends" another one in the graph



$$\omega(s_i) = \frac{d}{N} + (1-d) \sum_{s_j \in In(s_i)} \frac{\text{sim}(s_i, s_j)}{\sum_{s_k \in \text{Out}(s_j)} \text{sim}(s_k, s_j)} \omega(s_j)$$

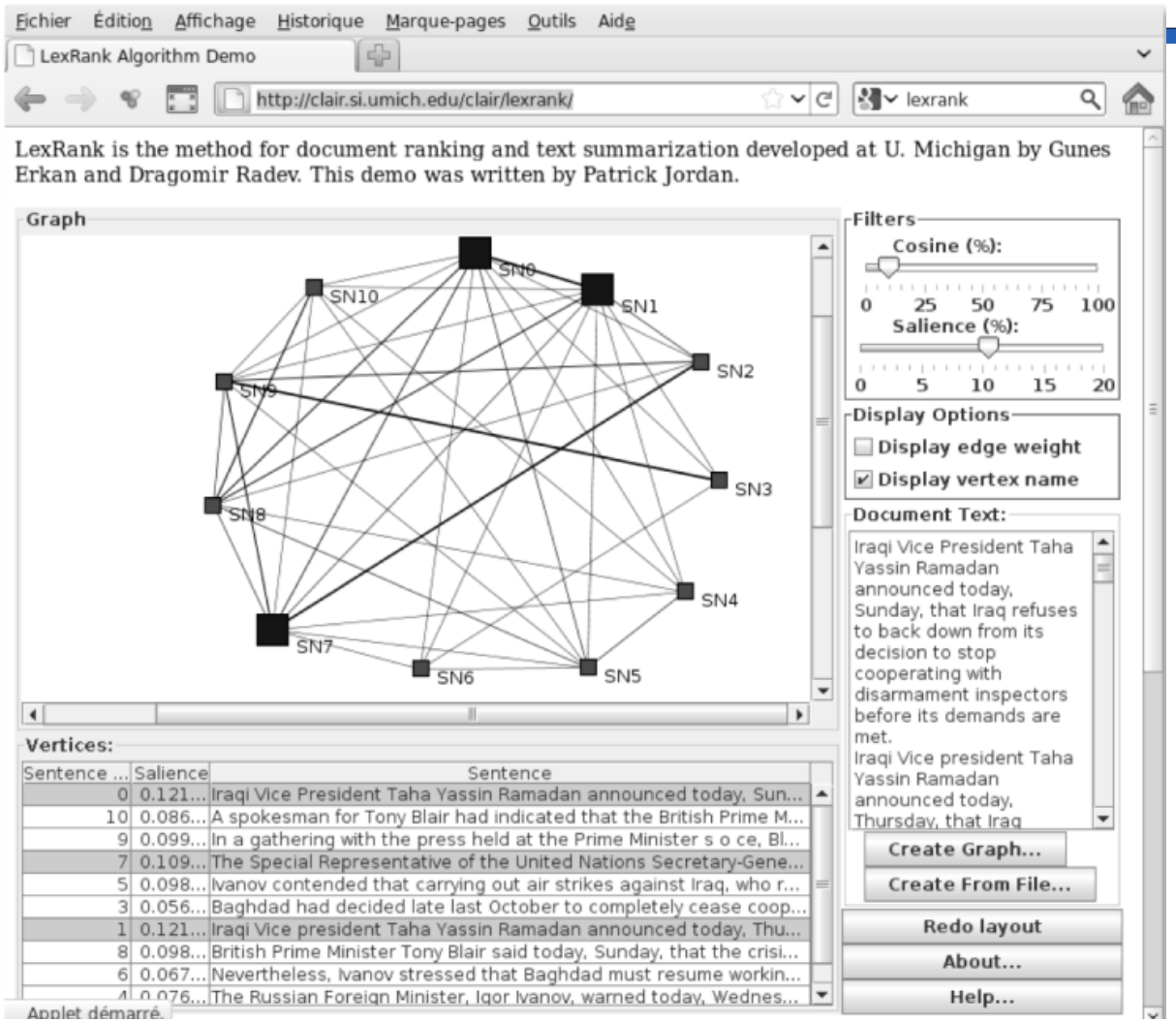
N = number of vertices
d = damping factor

Graph-based ATS



Components in a Graph-based ATS system

Graph-based (LexRank Demo)



- https://yale-lily.github.io/demos/demos/lexrank/lexrankmead.html
- https://github.com/rrajasek95/lexrank-demo (Python code)

Lexical Chains

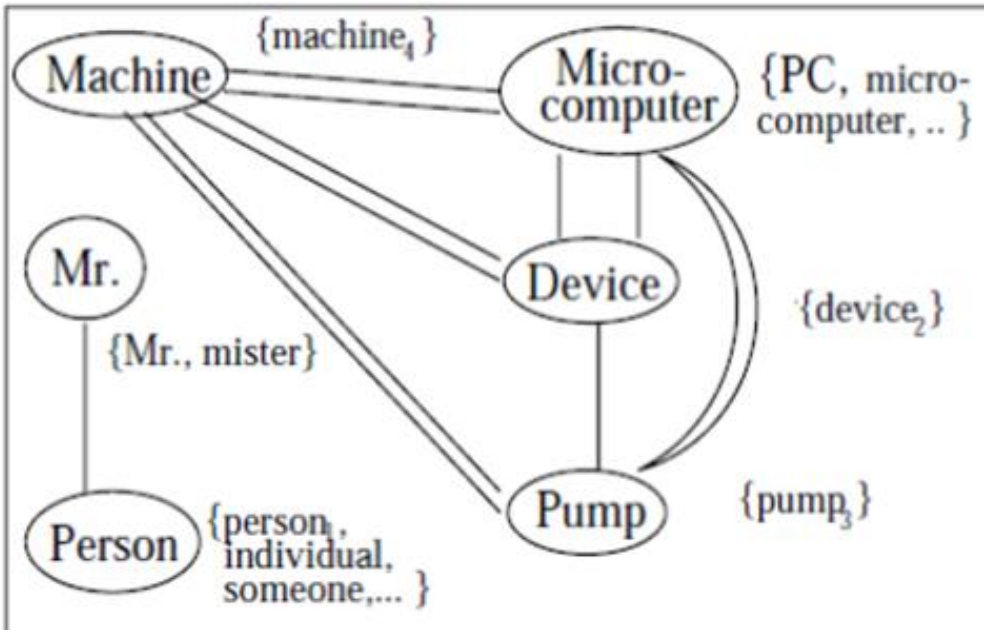
In Lexical Chain summarisation, first the chains are formed

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

Lexical Chains (Example)

*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.*



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)

A sentence is extracted from strong chains

Rethorical Analysis for ATS

Rhetorical Structure Theory (RST) makes the assumption that a text is divided into a hierarchical structure of **Elementary Discourse Units** (EDUs)

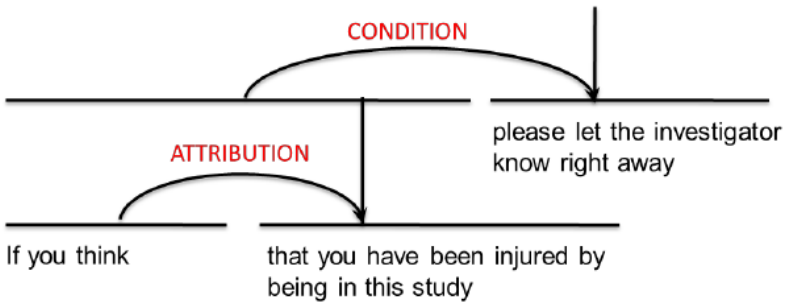
The EDUs are either **satellites** or **nuclei**

- A satellite needs a nucleus to be intelligible, but the opposite is not true

An algorithm is applied which **weights** and orders each EDU for the tree structure of the discourse

- The higher the element in the structure, the more significant its weight

In RST, there is a limited number of universal and fixed **relations**, including elaboration, condition, illustration, opposition, concession, justification, etc.



Text Summarization

Summary Evaluation

Summary Evaluation

- Challenging task because human judges differ in scores assigned to same summaries
- There no sense of having an “ideal summary”
  - It depends on the domain and user's intentions
- DUC (2001-2007)/ TAC(2008-2011) evaluation campaigns proposed many types of tasks and evaluation metrics

Summary Evaluation

Evaluation Methods

Categorized acording to the level of human effort in the evaluation:

- Manual Evaluation
- Semi-automatic methods
  - ROUGE
  - Pyramid
- Automatic methods
  - FRESA

Summary Evaluation

ROUGE (Recall Oriented Understudy for Gisting Evaluation)

[Lin and Hovy,2003]

Intrinsic metric for automatically evaluating summaries

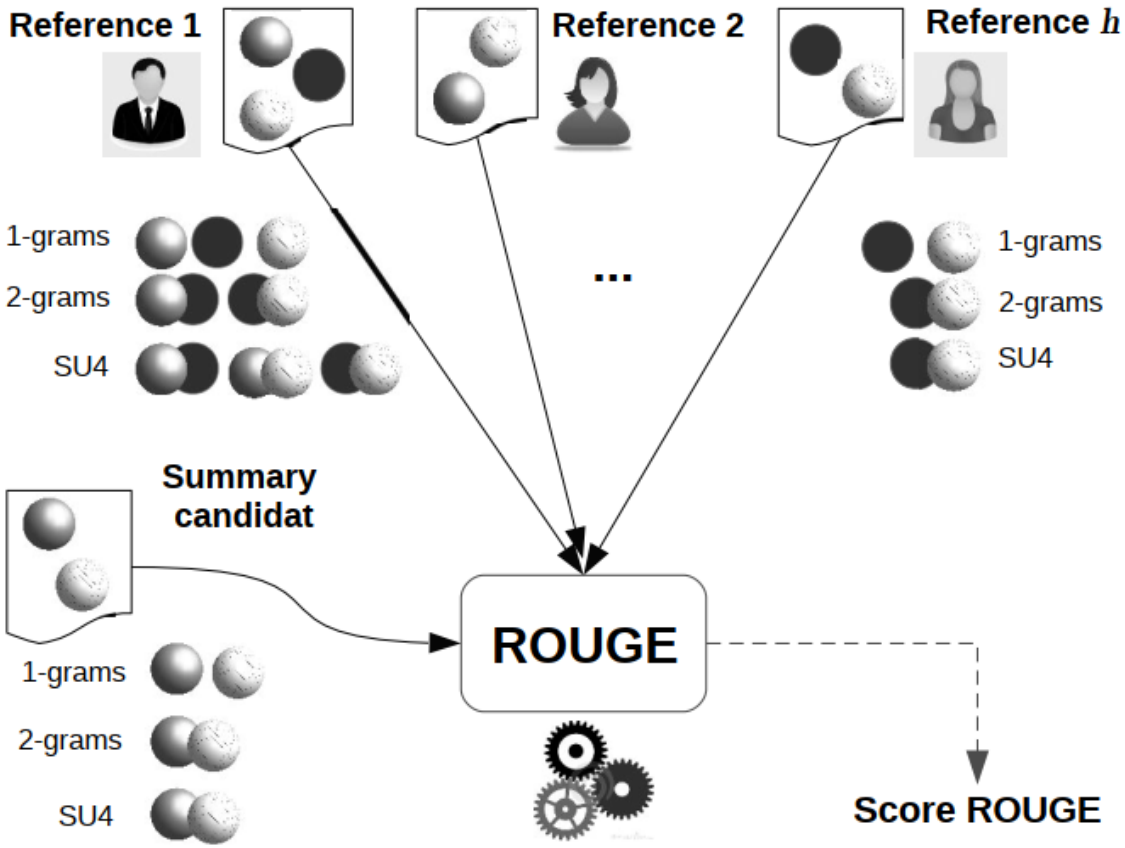
- Based on BLEU (a metric used for machine translation)
- Not as good as human evaluation
- But much more convenient

Given a document D, and an automatic summary X:

1. Have N humans produce a set of reference summaries of D
2. Run system, giving automatic summary X
3. What percentage of the bigrams from the reference summaries appear in X?

Summary Evaluation

ROUGE
(Basic Idea)



$$\text{ROUGE-}n = \frac{\sum_{n\text{-grams}} \in \{\text{Sum}_{\text{can}} \cap \text{Sum}_{\text{ref}}\}}{\sum_{n\text{-grams}} \in \text{Sum}_{\text{ref}}}$$

RxNLP

remedies for taming text

Home Web APIs Open Source Tools

ROUGE 2.0 – A Java Package For Automatic Summary Evaluation

ROUGE 2.0 is a Java Package for Evaluation of Summarization Tasks building on the Perl Implementation of ROUGE.

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It consists of a set of metrics for evaluating automatic summarization of texts as well as machine translation. It works by comparing an automatically produced summary or translation against a set of reference summaries (typically human-produced) or translations.

ROUGE 2.0 is a lightweight open-source tool that allows for easy evaluation of summaries or translation by limiting the amount of formatting needed in terms of reference summaries as well as system summaries. In addition, it also allows for evaluation of unicode texts known to be an issue with other implementations of ROUGE. One can also add new evaluation metrics to the existing code base or improve on existing ones.

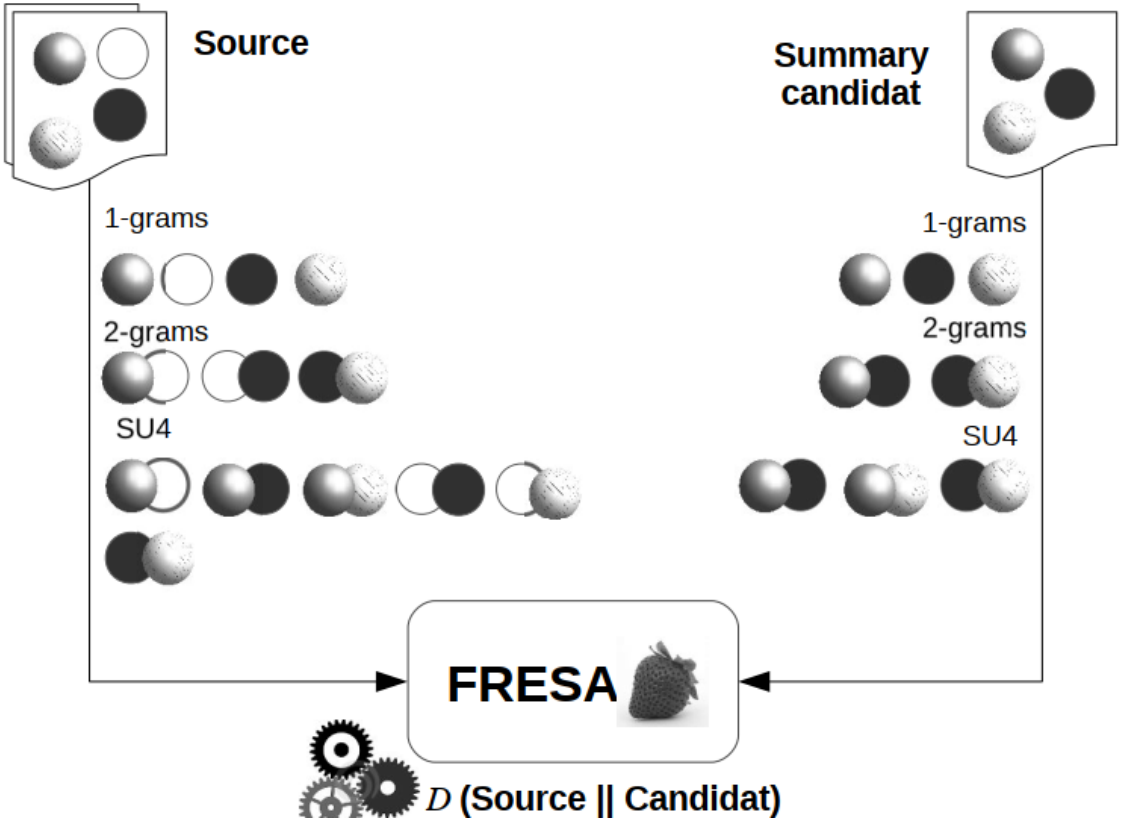
www.rxnlp.com/rouge-2-0

Summary Evaluation

FRESA – Framework for evaluating summaries automatically [TORRES-MORENO ET AL., 2010)

It is an automatic method based on information theory, which evaluates summaries **without** using human references

Standard preprocessing of documents (filtering and lemmatization) before calculating the **probability distributions** between the summary and the original document



https://github.com/fabrer/automatic-summarization/tree/master/EVALUATION

Text Summarization Research

at UFRPE/UFPE

Functional Summarization – UFPE Team Funded by HP (2012-2016)

Rafael Lins
Coordinator

Eduardo Silva
Project Manager

Bruno Ávila
Technical Leader

George Cavalcanti
Research Fellow

Fred Freitas
Research Fellow

Gabriel Silva
Researcher

Hilário Oliveira
Researcher

Luciano cabral
Researcher

Rafael Mello
Researcher

Rinaldo Lima
Researcher

Ensemble Approach to ATS

Automatic Text Summarization based on Concepts via Integer Linear Programming (PhD Thesis 2018)

Research Goal

Proposing and evaluating a concept-based approach using ILP and regression for the tasks of mono/multidocument summarization of news articles

The architecture of the proposed solution is composed by two stages:


Hilário Oliveira

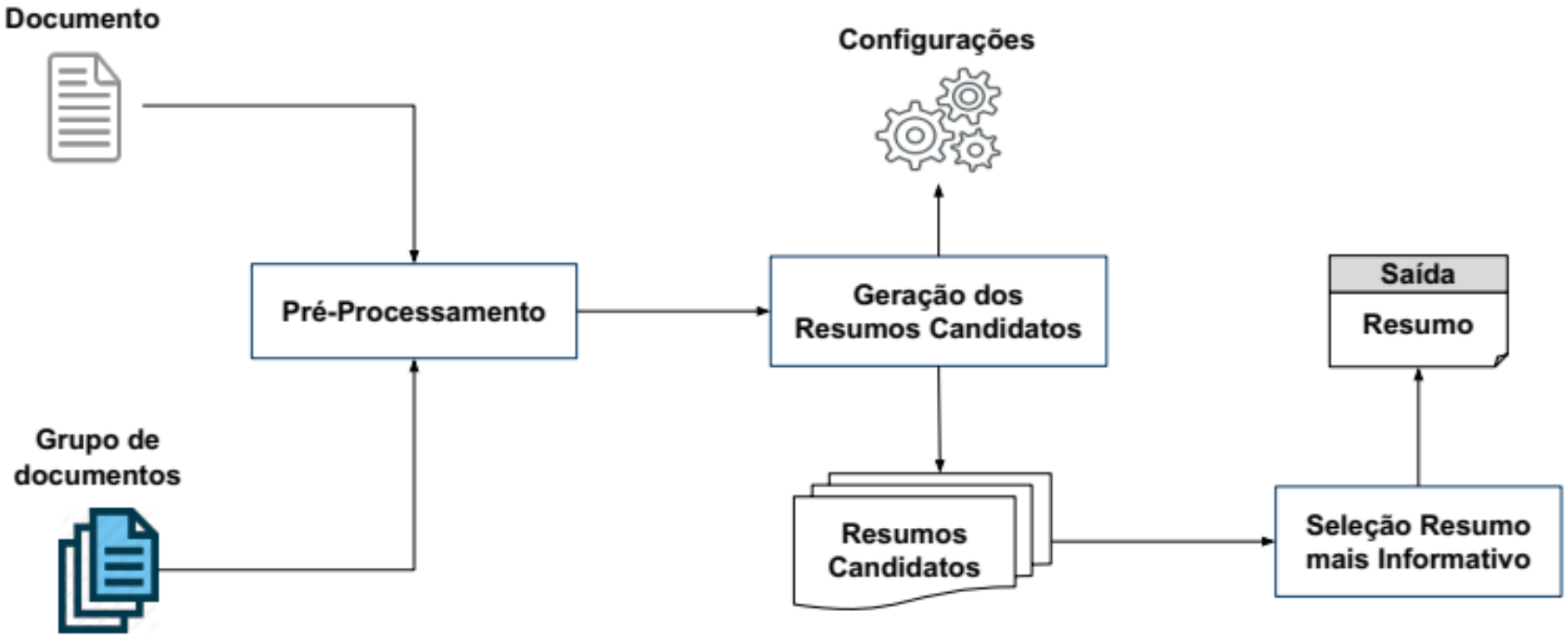
1. The generation of many **candidate summaries**.
   - Adopting a concept-based approach using ILP
2. The estimation and selection of the **most informative summary**.
   - Applying the regression algorithm

Research Questions:

1. Is it possible to increase the **informativeness** of the generated summaries by combining **ILP and regression**?
2. Can this approach also take into account both the **redundancy** and the **cohesion** of the generated sumaries?

Ensemble Approach to ATS

ILP & Regression-based Approach (ensemble)



Ensemble Approach to ATS

Bigrams are used to represent the notion of concepts

- It has showed better results in the preliminary experiments
- Bigrams formed only by stopwords are removed (Boudin et al., 2015)

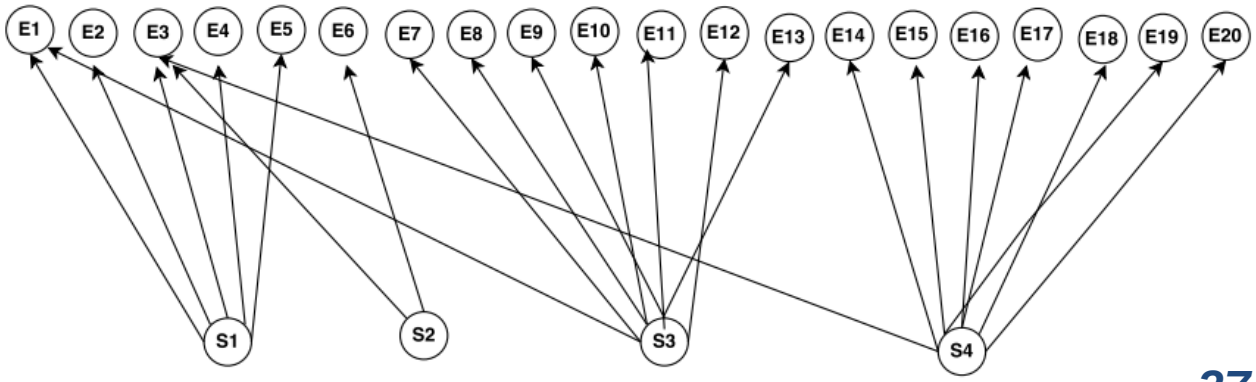
Entity Graph

- We adopted the **Entity Graph model** (Guinaudeau, Strube,2013) to estimate an approximate **local cohesion score** for the sentences in texts

The weight of a **concept** is estimated based on its **position** and **sentence coverage**

- Explore the weights of your neighbors (**Concept Graph**).
- Edges denote **adjacency relations** between two concepts
- New weight distribution strategy (contribution)

Concept Graph
of 4 sentences and
its entities



Ensemble Approach to ATS

Cohesion constraints are included in the ILP model to minimize two typical problems in Extractive Summarization

- **Coreference Resolution**
- **Explicit discourse dependencies** between a given pair of sentences.

ILP-based Sentence Selection Optimization Problem

$$MAX \quad \sum_{c_i \in C} w_i c_i + \sum_{s_j \in S} co_j s_j \tag{4a}$$

$$s.t. \quad \sum_{s_j \in S} l_j s_j \leq L \tag{4b}$$

$$s_j Occ_{ij} \leq c_i \quad \forall i, j \tag{4c}$$

$$\sum_{s_j \in S} s_j Occ_{ij} \geq c_i \quad \forall i, j \tag{4d}$$

$$Ds_j \leq \sum_{s_d \in S_D} s_d \quad \forall j, d \tag{4e}$$

$$\sum_{s_r \in S_r} s_r \leq 0 \tag{4f}$$

$$c_i, s_j, s_d, s_r, Occ_{ij} \in \{0, 1\} \quad \forall i, j, d, r \tag{4g}$$

Ensemble Approach to ATS

Comparative Results on 2 datasets :

Multidocument Scenario

Sistemas	DUC 2003		DUC 2004	
	R-1	R-2	R-1	R-2
PLI	40,47† (6,00)	10,95† (4,12)	39,05 (4,42)	10,04 (3,25)
PLI + Regressão	**41,39** (6,12)	**11,22** (4,51)	**40,12** (4,19)	**10,49** (3,35)
Greedy-KL	40,35† (5,76)	9,20 (3,99)	38,27 (4,73)	8,96 (3,09)
ICSISumm	40,07† (4,88)	10,95† (4,00)	38,42 (4,14)	9,80 (3,17)
LLRSum	36,94 (6,05)	8,87 (3,21)	35,90 (5,01)	8,06 (3,12)
ProbSum	37,60 (7,11)	9,28 (4,05)	35,37 (4,41)	8,18 (3,00)
Sume	39,36† (5,57)	9,81† (3,95)	37,29 (4,24)	8,83 (2,71)
Sist. 12/Classy 04	38,44 (5,25)	9,11 (3,95)	37,69 (4,08)	8,98 (3,08)
Oráculo	*46,34* (5,22)	*14,01* (4,84)	*43,32* (3,73)	*11,77* (3,07)

Ensemble Approach to ATS

Comparative Results on 3 datasets :

Multidocument Scenario

Sistemas	CNN	
	R-1	R-2
PLI	57,54 (20,09)	41,08† (25,38)
PLI + Regressão	**58,58** (19,87)	**41,36** (25,43)
Baseline	45,99 (21,77)	33,49 (25,00)
Classifier4J	46,63 (20,32)	32,15 (23,13)
HP-UFPE FS	50,71 (20,34)	34,58 (24,38)
Oráculo	*74,13* (16,96)	*65,71* (19,00)

Sistemas	DUC 2001		DUC 2002	
	R-1	R-2	R-1	R-2
PLI	45,32 (9,74)	20,25 (11,52)	48,85 (8,45)	23,30 (9,76)
PLI + Regressão	**46,37** (9,76)	**21,10** (11,71)	**49,78** (8,40)	**23,92** (9,74)
Baseline	43,75 (10,47)	19,57 (11,64)	46,94 (9,20)	22,14 (10,01)
Classifier4J	44,44 (9,85)	19,86 (11,34)	47,09 (8,93)	22,12 (9,87)
HP-UFPE FS	35,91 (11,78)	11,78 (9,78)	45,70 (9,31)	20,59 (9,88)
Sistema T/28	44,53 (9,23)	20,27† (10,75)	48,07 (8,90)	22,88 (9,96)
Oráculo	*53,33* (8,81)	*28,02* (12,36)	*56,16* (7,87)	*30,39* (10,50)

Towards Coherent Single-Document Summarization

Motivation

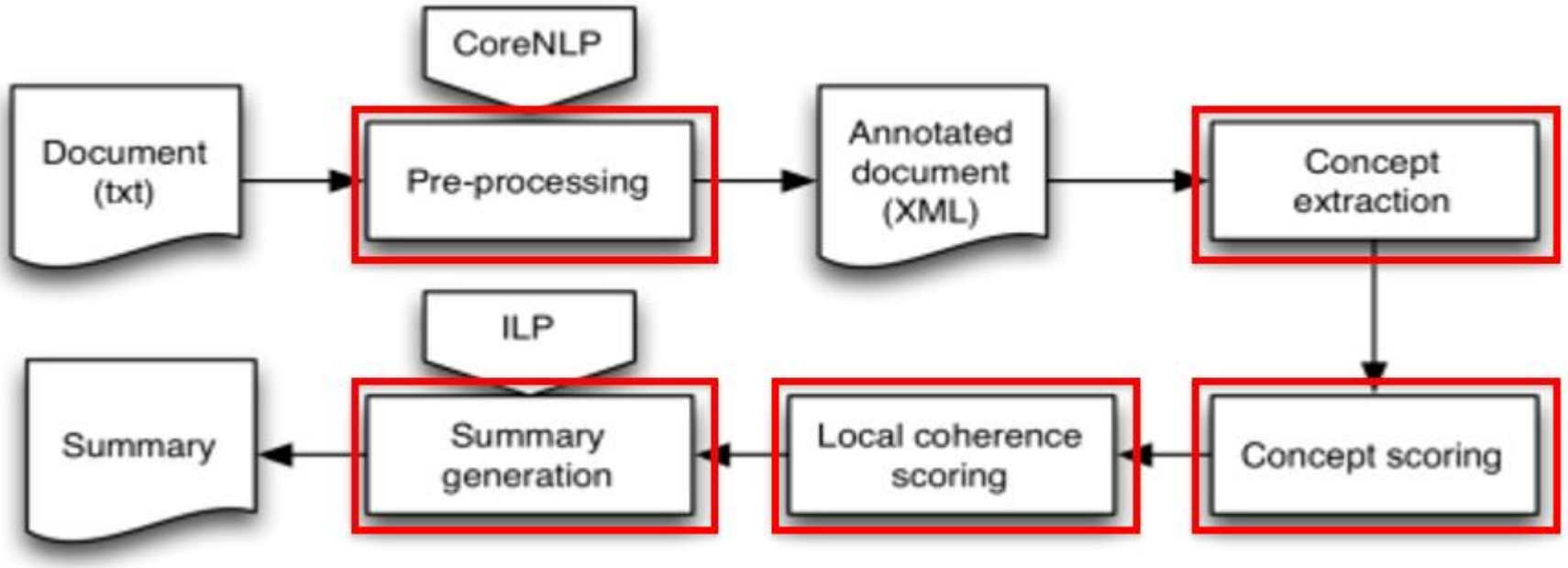
- The **extractive approach** is the most studied approach in the literature.
- However, most current extractive summarization systems usually produce summaries with 2 major issues:
  - **loose sentences** lacking relationship among them,
  - **dangling coreferences** that breaks the natural discourse flow.

➢ Adding **coherence** in extractive summarization is one of the main open problems in ATS.

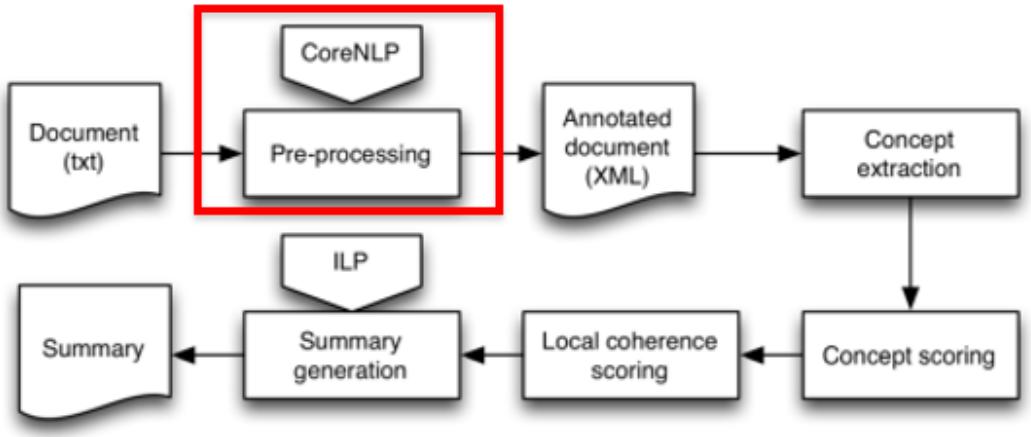
Towards coherent single-document summarization: an integer linear programming-based approach
Rodrigo Garcia, Rinaldo Lima, Bernard Espinasse, Hilario Oliveira
Proceedings of the 33rd Annual ACM SAC, 2018

Towards Coherent Single-Document Summarization

General Architecture of the proposed solution:



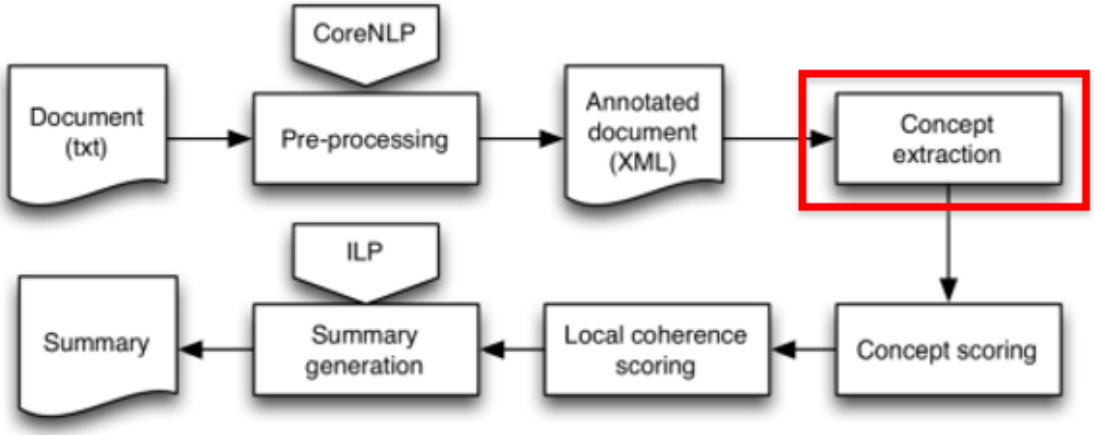
Preprocessing Task: annotation



Stanford CoreNLP Toolkit is used for annotating the input documents:

- tokenization, sentence splitting, POS tagging, lemmatization, dependency parsing, and Coreference Resolution (CR)
- We remove all stopwords
- CR is employed to find anaphoric expressions referring to the same concept in the text: All personal pronouns, except "it", are replaced by the referring entities.

Concept Extraction Task



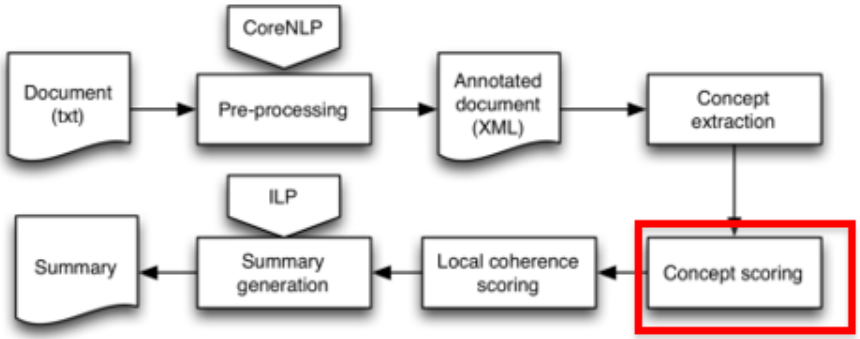
Unigram/Bigram nouns are extracted as concepts.

Ex. Consider the sentence:

"Paul wrote a book on Artificial Intelligence"

- The following concepts will be extracted: *"Paul", "book", "Artificial Intelligence".*
- This reflects the intuition that such terms are important in ATS because they likely describe real world entities.

Concept Scoring Task



To weigh the importance of concepts, we use Term Frequency (TF), **Normalized Term Frequency** (NTF), **and Term Frequency/Inverse Sentence Frequency (TF-ISF) :**

- Given **TF** the number of occurrences of a term in a document **d**

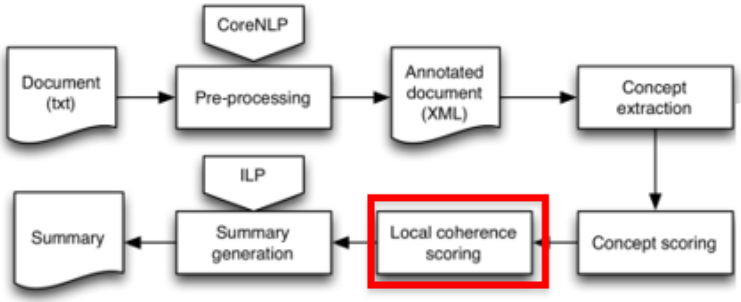
$$NTF(w,d) = \frac{TF(w)}{n}$$

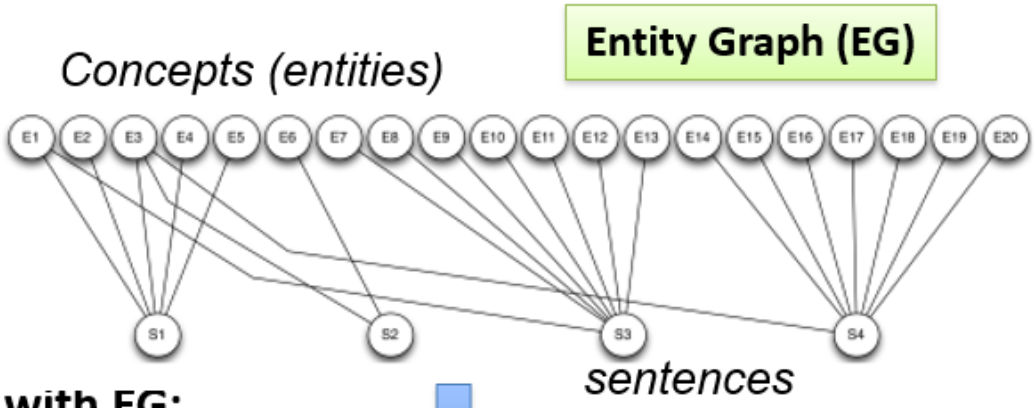
- where **n** is the number of sentences in document **d**

$$TF - IFS(w) = TF(w) \times \log(\frac{n}{TS(w)})$$

- **TS(w)** is the total number of sentences in which **w** occurs, and **n** defined as above.

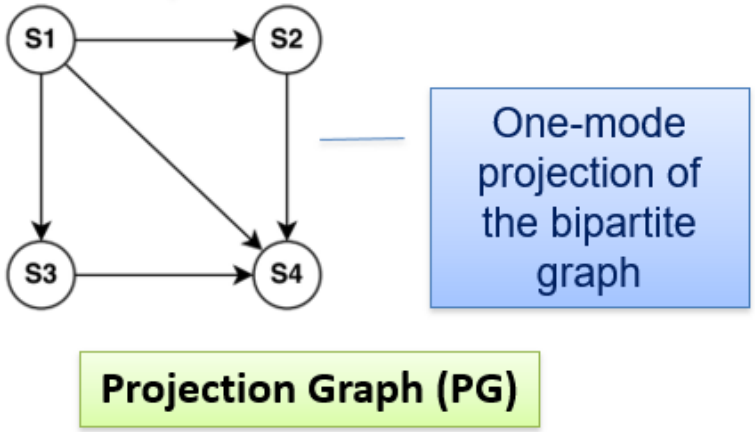
Local coherence scoring task: Entity Graph



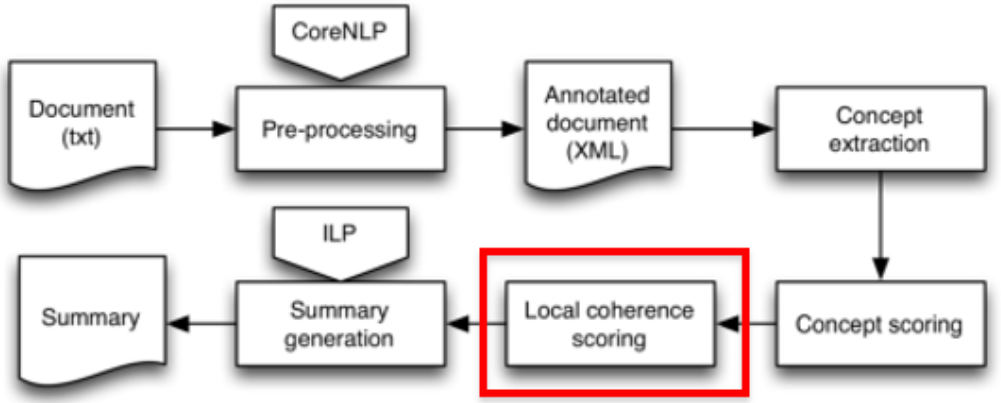


1. Capture the distribution of discourse concepts across sentences with EG:

- The ***EG*** denotes sentences and concepts as 2 separated clusters of nodes
- Concepts with syntactic role of subject, object and other are assigned weights 3, 2, 1, respectively.
- The ***EG*** is transformed (by one-mode projection) into a new graph (***Projection Graph***) containing only nodes sharing common concepts.



Local coherence scoring task: Entity Graph



2. Coherence Scoring on the Projection Graph (PG)

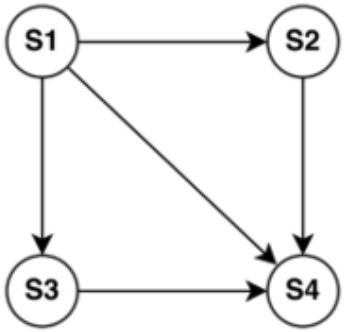
- The coherence score for a given sentence Si is given the centrality measure:

$$coherence(s_i) = Outdegree(s_i, PG)$$

- The higher the centrality value for a sentence S, the more connected S is with other sentences.

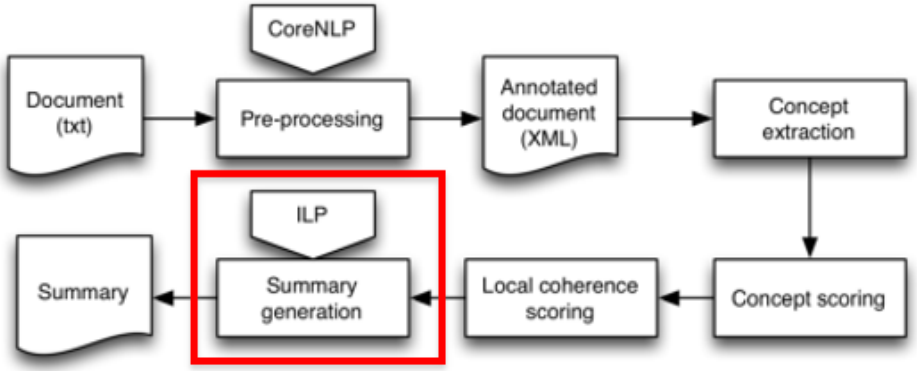
Entity Graph (EG)

One-mode projection of the bipartite graph



Projection Graph (PG)

Summary generation task: ILP optimization



Integer Linear Programming (ILP) optimization model for both informativeness and coherence :

Selects the maximun number of allowed concepts per summary

$$Maximize : \sum_{c_i \in C} w_i \cdot c_i + \sum_{s_j \in S} Rank(s_j) \cdot s_j$$

The Rank() function denotes the coherence score based on the shared entities among the sentences captured by the Entity Graph Model

Objective function

constraints

$$\sum_{s_j \in S} l_j \cdot s_j \leq L$$

$$s_j Occ_{ij} \leq c_i$$

$$\sum_{s_j \in S} s_j Occ_{ij} \geq c_i$$

$$c_j, s_j, Occ_{ij} \in \{0,1\} \forall_{i,j}$$

- Constraints on the size of the summaries
- selecting a sentence means that it has also to select all the concepts it contains

DUC Datasets and ROUGE measure

The DUC 2001 and 2002 datasets:

- used for evaluating generic single-document summarizers
- composed of news articles written in English
- contain abstractive summaries with approximately 100 words each one

Dataset	N. of Docs	N. of Sent	N. of Words
DUC 2001	309	11.026	269.990
DUC 2002	576	14.370	348.012

Evaluation measure

- ROUGE-N provided by the ROUGE package
- consists of an **n-gram recall** between the summary generated by an ATS system, and a set of golden standard summaries produced by humans

Comparative Assessment: Informativeness

Selected systems in the comparison:

- the best summarizes participating in the DUC 2001 (**System T**) and DUC 2002 (**System 28**) conferences
- **AutoSummarizer** (2016)
- **Classifier4J** (Lothian, 2003)
- **HP-UFPE FS** (highest R-1 performance reported in (Batista et al., 2015)
- **TextRank** (Mihalcea & Tarau, 2004)
- **Parvens's summarizer** (Parveen & Strube, 2015)

The proposed summarizer obtained very competitive results

DUC 2001

Summarizer	R-1	R-2
AutoSummarizer	41.92	16.63
Classifier4J	44.44	19.86
HP-UFPE FS	35.91	11.78
System T	44.53	**20.27**
TextRank	40.66	15.09
Proposed summarizer	**45.00**	17.91

DUC 2002

Summarizer	R-1	R-2
AutoSummarizer	43.79	19.17
Classifier4J	47.09	22.12
HP-UFPE FS	45.70	20.55
System 28	48.07	22.88
TextRank	43.93	18.66
Parveen's summarizer	**48.50**	**23.00**
Proposed summarizer	47.36	20.96

Comparative Assessment: Coherence (1)

- **ROUGE** measures cannot take into account text properties such as **readability** and **mainly coherence** (Guinaudeau & Strube, 2013).
- We performed a **preliminary human comparison** of the **coherence level** of some summaries

Summary 3: Doc. WSJ910107-0139/DUC 2001 - 102 sentences

Classifier4J

S1: Under a microscope, parts of his brain, riddled with little holes, looked like a sponge.

S2: Thus began another chapter in one of medicine's most bizarre mysteries, a tale of sick sheep and mad cows, cannibals and Pennsylvanians, ancient life forms and a cat named Max.

S3: The plot revolves around a family of brain diseases, probably variations of a single disorder, called spongiform encephalopathy.

S4: Spongiform research already has raised questions about a cornerstone of biology and spawned a Nobel Prize.

S5: The uncanny nature of the disorder sometimes grips scientists with a kind of obsessive fascination, notes NIH researcher D. Carleton Gajdusek.

AutoSummarizer

S1: Yet sheep have never been strongly implicated in cases of the human form of the disease, Creutzfeldt-Jakob disease, or CJD. Indeed, CJD is a nefarious trickster; the first reported case of the disease, which entered medical texts in the 1920s, really wasnt CJD after all, scientists now believe.

S2: Based on the clues, Dr. Gajdusek, back at NIH, led studies in the 1960s showing kuru, CJD and scrapie to be essentially the same infectious disease, studies that won the 1976 Nobel Prize for medicine.

S3: Brain tissue of infected animals could transmit the disease when injected into different animals brains, yet microscopes revealed no signs of infectious microbes.

S4: Still, Dr. Brown and other scientists arent much worried about mad cows because their animal studies show spongiform disease is very hard to transmit orally.

Comparative Assessment: Coherence (2)

- Our finding: the longer the input documents, the more likely is to obtain a less cohesive summary
- The summary generated by the proposed system has superior coherence among the sentences

Summary 3: Doc. WSJ910107-0139/DUC 2001 - 102 sentences

Proposed System

S1: The plot revolves around a family of brain diseases, probably variations of a single disorder, called spongiform encephalopathy.
S2: Scientists agree, however, that the disease in animals probably poses little danger to people.
S3: But the infectious agent continued to baffle scientists.
S4: It seemed like "biological spontaneous combustion," says NIH researcher Paul Brown.
S5: Now more than 10 British cats have died of it, suggesting brains from infected sheep or cows got into cat food.

Text Summarization Research

Proposed Datasets for Extractive TS

DocEng 2019
23.-26.9.2019 Berlin, Germany
ACM SYMPOSIUM ON DOCUMENT ENGINEERING

The CNN-Corpus
A Large Textual Corpus for Single-Document Extractive Summarization

Rafael Dueire Lins, Hilario Oliveira, Luciano Cabral,
Jamilson Batista, Bruno Tenorio,
Rafael Ferreira, Rinaldo Lima,
Gabriel Pereira e Silva , Steven J. Simske

VIRTUS IMPAVIDA
UFRPE
Colorado State University

Enterovirus D68 spreads to Canada

By **Jacque Wilson**, CNN
updated 10:10 AM EDT, Wed September 17, 2014

Virus sends hundreds to hospital

Part of complete coverage on
Cold & Flu Season

STORY HIGHLIGHTS

- Canada confirms three cases of Enterovirus D68 in British Columbia
- A fourth suspected case is still under investigation
- Enterovirus D68 worsens breathing problems for children who have asthma
- CDC has confirmed more than 100 cases in 12 U.S. states since mid-August

(CNN) -- Canadian health officials have confirmed three cases of Enterovirus D68 in British Columbia. A fourth suspected case from a patient with severe respiratory illness is still under investigation.

Two of the confirmed cases are children between 5 and 9, said Dr. Danuta Skowronski, lead epidemiologist on emerging respiratory viruses at the British Columbia Centre for Disease Control. The third is a teen between 15 and 19.

Enteroviruses are common, especially in the summer and fall months. The U.S. Centers for Disease Control and Prevention estimates that 10 million to 15 million infections occur in the United States each year. These viruses usually appear like the common cold; symptoms include sneezing, a runny nose and a cough.

Most people recover without any treatment. But Enterovirus D68


A respiratory virus called Enterovirus D68 has sent hundreds of children to the hospital. CNN's Elizabeth Cohen explains.

What is Enterovirus D68?
updated 3:33 PM EDT, Tue September 9, 2014


This type of enterovirus is uncommon but not new. We've seen less than 100 cases in the United States since it was identified.

What parents should know
updated 4:50 PM EDT, Tue September 9, 2014


What are the symptoms of Enterovirus D68? When

The CNN - Corpus

Highlights:

1. Canada confirms three cases of Enterovirus D68 in British Columbia.
2. A fourth suspected case is still under investigation.
3. Enterovirus D68 worsens breathing problems for children who have asthma.
4. CDC has confirmed more than 100 cases in 12 U.S. states since mid-August.

Gold Standard:

1, 2, 13, 26

Text:

1. Canadian health officials have confirmed three cases of Enterovirus D68 in British Columbia.
2. A fourth suspected case from a patient with severe respiratory illness is still under investigation.
3. Two of the confirmed cases are children between 5 and 9, said Dr. Danuta Skowronski, lead epidemiologist on emerging respiratory viruses at the British Columbia Centre for Disease Control.
4. The third is a teen between 15 and 19.
5. Enteroviruses are common, especially in the summer and fall months.
6. The U.S. Centers for Disease Control and Prevention estimates that 10 million to 15 million infections occur in the United States each year.
7. These viruses usually appear like the common cold; symptoms include sneezing, a runny nose and a cough.
8. Most people recover without any treatment.
9. ...

www.cnn.com

Overview statistics of the CNN-corpus

Categories	Articles	Avg.Sentence/ Summary	Avg.Words/ Story High.	Avg.Sentences/ Text	Avg.Words/ Sentences	Avg. Sentences/ Gold. Summary	Avg.Words/ Gold. Sentence
Business	161	3.3	14.1	30.8	21.6	3.4	25.5
Health	290	3.3	11.7	47.0	18.6	3.4	23.0
Justice	224	3.7	11.9	35.6	20.2	3.5	25.6
Living	98	3.6	12.9	53.3	19.3	3.7	26.9
Opinion	192	3.8	13.5	43.8	20.7	3.9	26.1
Politics	195	3.5	12.2	37.8	21.7	3.5	26.7
Showbiz	241	3.5	11.6	28.8	19.0	3.5	23.2
Sport	148	3.7	11.6	31.3	20.9	3.6	27.0
Technology	132	3.4	12.2	39.1	19.0	3.4	25.4
Travel	171	3.3	12.6	55.4	17.7	3.5	24.7
US	160	3.6	11.9	39.7	18.8	3.6	23.6
World	988	3.7	12.2	35.6	20.6	3.7	25.2
Total/Average	3,000	3.6	12.3	38.4	19.9	3.6	25.0

- Strict quality policies were enforced: every summary agreed on by, at least, two experts.
- Avoiding human subjectivity and mistakes.

The CNN-Corpus was used in the:

DocEng'19 Competition on Extractive Text Summarization

Rafael Dueire Lins, Rafael Ferreira , Steven J. Simske

The same development methodology was used for:

a Large Corpus for Extractive Text Summarization in the Spanish Language

Text Summarization Research

Trends in TS

Limitations of Extractive Summarization

- Abstraction not as easy to do since it requires **semantic understanding** of text
- No coherence/cohesion treatment in most of the current ATS systems
- Redundancy problem
- Better (semi)automatic evaluation methods need to be devised particularly for multi-document systems

Main Task	**Linguistic Quality**	**Global Score**
Human Abstracts	**8.915**	**8.830**
Human Extracts	7.477	6.341
Best system (ICSI)	*5.932*	*5.159*
Update Task	**Linguistic Quality**	**Global Score**
Human Abstracts	**8.807**	**8.506**
Human Extracts	7.250	6.114
Best system (ICSI)	*5.866*	*5.023*

Human performance vs system performance TAC'09

Neural-based Sentence Compression

- The next logical step is to eliminate redundant or less informative content from the extracted sentences
- A possible way to achieve this is by **sentence compression**

Current approaches

- Sentence Compression by **Deletion**
- Sentence Compression as a **Sequence to Sequence** Problem

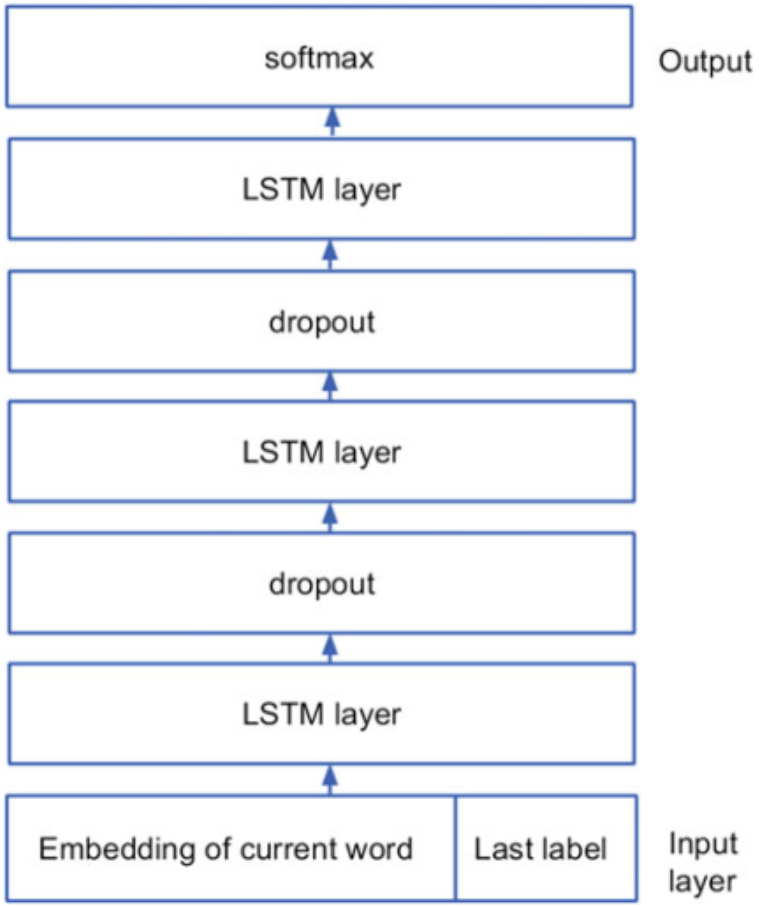
Neural-based Sentence Compression

Neural-based Sentence Compression

The **LSTM-based** sentence compression model

It uses a **parallel corpus of 2 million** sentence-compression instances (pairs) from a news corpus

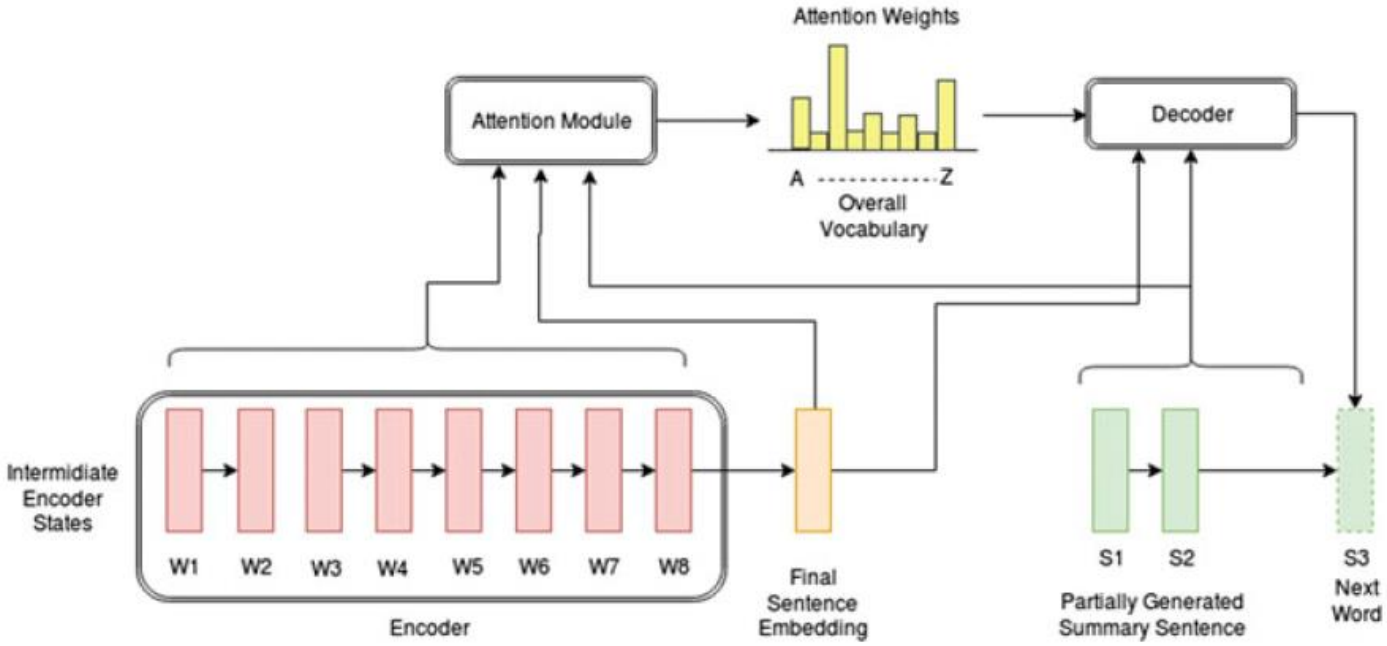
One-hot encoding or word embedding of the input dictionnary



Filippova et al., 2015. Sentence compression by deletion with LSTM. In: Proc. of the 2015 Conference on EMNLP. ACL, Lisbon, Portugal (2015)

Neural-based Sentence Compression

Sequence to Sequence Model



- Sentence **encoder** is responsible to sequentially read the word representations and generate intermediate sentence representations for each state
- The **decoder** module generates output sentence one word at a time
- At a given step in the decoding process, the **attention module** assigns weights to each of the input steps (optional)

Text Summarization Research

Research Collaborations

Text Ming Group at UFRPE

Research areas:

Information Extraction

Sentiment Analysis

Relational Learning for Text Mining

Deep Learning for Text Mining

Automatic Summarization

Text Simplification

Text Classification and Text Clustering

NLP applied to Augmentative and Alternative Communication (AAC) Systems

Text Analytics

tmgufrpe.github.io

Future Research Projects Proposal

- Doctoral and Master thesis co-supervision
- Funded bilateral projects options:
  - COFFECUB
  - CNPq/CAPES
  - FACEPE

References

Most cited papers of our Team

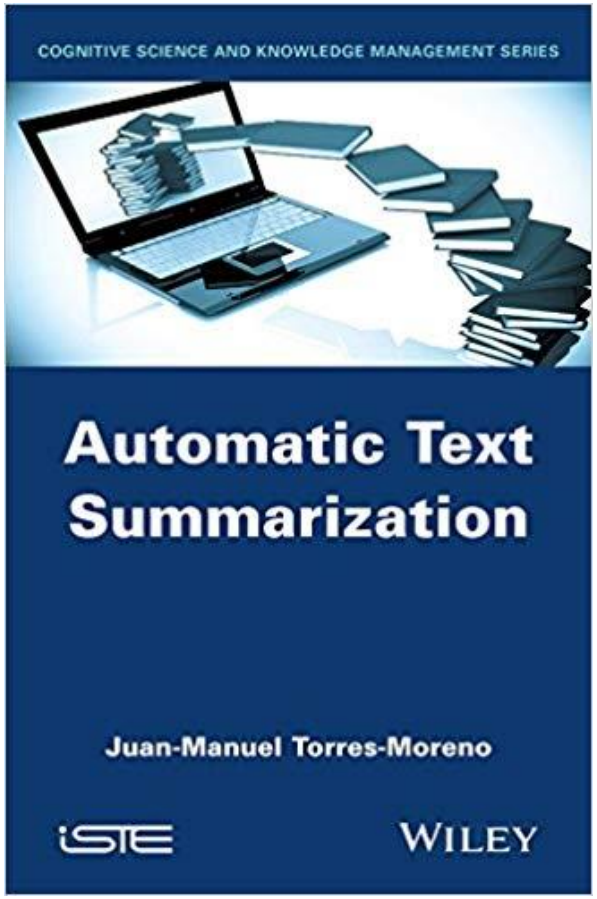
PAPER TITLE	CITED BY	YEAR
Assessing sentence scoring techniques for extractive text summarization R Ferreira, L de Souza Cabral, RD Lins, GP e Silva, F Freitas, ... Expert systems with applications 40 (14), 5755-5764	212	2013
A multi-document summarization system based on statistics and linguistic treatment R Ferreira, L de Souza Cabral, F Freitas, RD Lins, G de França Silva, ... Expert Systems with Applications 41 (13), 5780-5787	85	2014
A context based text summarization system R Ferreira, F Freitas, L de Souza Cabral, RD Lins, R Lima, G França, ... 2014 11th IAPR International Workshop on Document Analysis Systems, 66-70	50	2014
A four dimension graph model for automatic text summarization R Ferreira, F Freitas, L de Souza Cabral, RD Lins, R Lima, G França, ... 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI …	35	2013
Assessing sentence similarity through lexical, syntactic and semantic analysis R Ferreira, RD Lins, SJ Simske, F Freitas, M Riss Computer Speech & Language 39, 1-28	34	2016
Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization H Oliveira, R Ferreira, R Lima, RD Lins, F Freitas, M Riss, SJ Simske Expert Systems with Applications 65, 68-86	29	2016
A new sentence similarity assessment measure based on a three-layer sentence representation R Ferreira, RD Lins, F Freitas, SJ Simske, M Riss Proceedings of the 2014 ACM symposium on Document engineering, 25-34	23	2014
A quantitative and qualitative assessment of automatic text summarization systems J Batista, R Ferreira, H Tomaz, R Ferreira, R Dueire Lins, S Simske, ... Proceedings of the 2015 ACM Symposium on Document Engineering, 65-68	14	2015

References

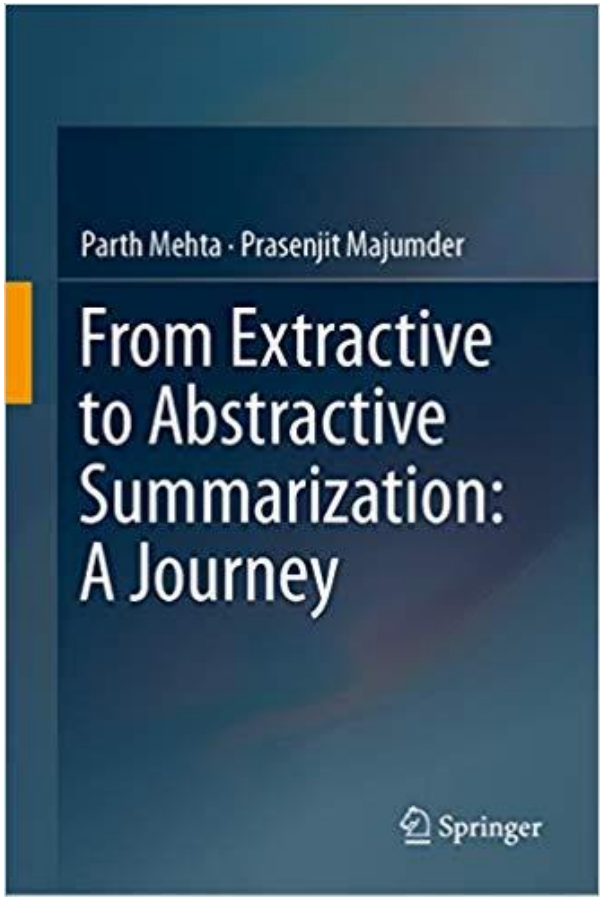
Most cited papers of our Team

Automatic text document summarization based on machine learning G Silva, R Ferreira, RD Lins, L Cabral, H Oliveira, SJ Simske, M Riss Proceedings of the 2015 ACM Symposium on Document Engineering, 191-194	13	2015
Automatic Summarization of News Articles in Mobile Devices L Cabral, R Lima, R Lins, M Neto, R Ferreira, S Simske, M Riss 2015 Fourteenth Mexican International Conference on Artificial Intelligence …	6	2015
Appling Link Target Identification and Content Extraction to improve Web News Summarization R Ferreira, R Ferreira, RD Lins, H Oliveira, M Riss, SJ Simske Proceedings of the 2016 ACM Symposium on Document Engineering, 197-200	4	2016
DocEng'19 Competition on Extractive Text Summarization RD Lins, RF Mello, S Simske Proceedings of the ACM Symposium on Document Engineering 2019, 4	2	2019
The CNN-Corpus: A Large Textual Corpus for Single-Document Extractive Summarization RD Lins, H Oliveira, L Cabral, J Batista, B Tenorio, R Ferreira, R Lima, ... Proceedings of the ACM Symposium on Document Engineering 2019, 16	2	2019
Automatic Document Classification using Summarization Strategies R Ferreira, RD Lins, L Cabral, F Freitas, SJ Simske, M Riss Proceedings of the 2015 ACM Symposium on Document Engineering, 69-72	2	2017

Recent Books on ATS



2014



2019

Links to Survey Papers on ATS

https://www.cs.bgu.ac.il/~elhadad/nlp16/nenkova-mckeown.pdf

https://arxiv.org/pdf/1707.02268.pdf

https://www.cs.cmu.edu/~nasmith/LS2/das-martins.07.pdf

https://dergipark.org.tr/en/download/article-file/392456

https://link.springer.com/article/10.1007/s10462-016-9475-9

https://wanxiaojun.github.io/summ_survey_draft.pdf

http://jad.shahroodut.ac.ir/article_1189_28715967fcd8b7bfb463ab90aca5a9f7.pdf

Thank you for your attention

Any questions?