UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

ANA MARIA SCHWENDLER RAMOS

ConceptRank: Extractive summarization based on graph conceptual centrality as salience

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação

Orientador: Prof. Dr. Leandro Krug Wives
Coorientador: Ms. Vinícius Wolozyn

Porto Alegre
2016

“If I have seen further than others,
it is by standing upon the shoulders of giants.“
— Isaac Newton

AGRADECIMENTOS

First of all I would like to thank Professor Leandro Krug Wives for all his work and advice in recent years and Vinicius Woloszyn for all his patience, help and strength in the development of his work. Thanks also to all the professors of the Institute of Informatics of UFRGS. This work would not be possible without all that I have learned over the years in INF. It was a fantastic experience.

Soon after, I would like to thank Liza, Giane, Andressa, Jessica and all my classmates at UFSM for the incredible start of the course that allowed me to get here. Along with this to Professor Giovani Rubert Librelotto for teaching me the first ways of how and why to use code for solving problems.

Later I would like to thank Roberto, Caroline, William Dutra and Otavio Machado, since we've known each other, to know that I had found siblings for life.

Luiza Eitelvein and Eduardo Both, for all this journey in UFRGS, for the support and all the incentive to get here. Without you, I would never have completed this stage, you were and are essential from beginning to end.

Last but not least, I would like to thank my mother, Hedi, for all the teachings since I was a child, and for showing me that I should never give up my goals and that it is best to pursue them until I succeed, regardless of the situation.

CONTENTS

LIST OF FIGURES **6**
LIST OF TABLES **7**
LIST OF ABBREVIATIONS AND ACRONYMS **8**
ABSTRACT **9**
RESUMO **10**
1 INTRODUCTION **11**
1.1 Motivation **11**
1.2 Objective **12**
1.3 Contribuitions **12**
1.4 Text organization **13**
2 BACKGROUND **14**
2.1 ConceptText **15**
2.1.1 Representing documents using concepts 16
2.1.2 Structure and identification of concepts 16
2.1.3 Centrality 17
2.2 Rouge **18**
2.3 CSTNews **20**
2.3.1 RST and CST 20
2.4 Chapter overview **21**
3 RELATED WORK **23**
3.1 CSTSumm **23**
3.2 LexRank **23**
3.2.1 Centrality-based Sentence Salience 24
3.3 Chapter overview **24**
4 EXPERIMENT SETUP **26**
4.1 Chapter overview **27**
5 RESULTS **28**
5.1 Chapter overview **31**
6 CONCLUSION **33**
REFERENCES **34**

LIST OF FIGURES

Figure 5.1 Total average of the experiment....28
Figure 5.2 ROUGE-2 Recall....30
Figure 5.3 ROUGE-3 Recall....31
Figure 5.4 Perfomance of LexRank and ConceptRank....31

LIST OF TABLES

Table 5.1 Average values obtained for LexRank Method 29
Table 5.2 Average values obtained for CSTSumm Method 29
Table 5.3 Average values obtained for ConceptRank Method 29

LIST OF ABBREVIATIONS AND ACRONYMS

ATS	Automatic Text Summarization
SAT	Sumarização Automática de Textos
VSM	Vector Space Model
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
DUC	Document Understanding Conference
RST	Rhetorical Structure Theory
CST	Cross-document Structure Theory
NLP	Natural Language Processing
MDS	Multi-document summarization
UFRGS	Universidade Federal do Rio Grande do Sul

ABSTRACT

Since the access to information is increasing everyday and we can easily acquire knowledge from many resources such as news websites, blogs and social networks, the capacity of processing all this amount of information becomes increasingly difficult. So, a way to deal with this situation is automatically extract the most important sentences, aiming to reduce the amount of text into a shorter version. We can explore this process while preserving the core information content by using a process called Automatic Text Summarization. This work presents a proposal to minimize problems related to the automatic summarization of texts, since some extractive techniques could not totally be prepared to handle with some issues, such as typos, synonyms and other orthographic variations, by evaluating sentences using "concepts" instead of words to represent the content of summaries.

Keywords: Automatic text summarization. summarization based on concepts. extractive summarization. graph conceptual centrality as salience. natural language processing. summary evaluation.

ConceptRank: Sumarização extrativa baseada na centralidade conceitual de um grafo como saliência

RESUMO

Como o acesso à informação está aumentando todos os dias e podemos facilmente adquirir conhecimento de muita fontes, como sites de notícias, blogs e redes sociais, a capacidade de processar essa quantidade de informações torna-se cada vez mais difícil. Sendo assim, uma maneira de resolver esta situação é automaticamente extrair as sentenças mais importantes de um texto, visando reduzir a quantidade de conteúdo em uma versão mais curta. Podemos explorar esse processo, preservando o entendimento da informação, usando um processo chamado Sumarização Automática de Textos. Esta monografia apresenta uma proposta para minimizar os problemas relacionados a sumarização automática de textos, uma vez que algumas técnicas extrativas podem não estar totalmente preparadas para lidar com algumas questões, como erros de digitação, sinônimos e outras variações ortográficas, avaliando as frases usando "conceitos"em vez de palavras para representar o conteúdo dos resumos.

Palavras-chave: sumarização automática de textos. mapeamento por conceitos. sumarização extrativa. centralidade como saliência. processamento de linguagem natural. avaliação de sumários.

1 INTRODUCTION

Automatic Text Summarization (ATS) is a Natural Language Processing (NLP) task that aims to reduce the amount of text into a shorter version while preserving the core information content (NENKOVA; MASKEY; LIU, 2011). This task has become important by the abundance of text available on the Internet and the fact that it is hard for human beings to manually summarize them. The web, in particular, has contributed to the increasing interest on automatic summarization. Nowadays it is easy to access information by Google Trends [1] or Google News[2], examples of specialized news search engines, but the users are overloaded with news information and barely have enough time to digest them in their full form. They deal with a huge amount of information, so the demand for ATS has been massively increasing from the traditional single-document to the more recent multi-document summarization tasks.

Related work in this area relies on Vector Space Model (VSM) (SALTON; WONG; YANG, 1975) to represent the content and relationship of texts. Such representation implies a measure of similarity that often is based upon the cosine similarity. However, the cosine similarity can be influenced and interfered by problems derived from the natural language used inside texts, like orthographic errors, synonyms, homonyms, and other morphological variations, known as vocabulary problems. In this sense, this work present a methodology that use "concepts" instead of words to describe the contents of summaries. Concepts are structures capable of representing document's objects and ideas using a combination of identifiers (LOH; WIVES; OLIVEIRA, 2000a). In this sense, this work aims to show an approach to minimize problems in the area of automation of text summarization since some extractive summarization processes can be influenced by those issues.

1.1 Motivation

The motivation for this work is finding a way to optimize the automatic summarization of texts and the manipulation of the vocabulary contained in documents. Considering that the great part of the languages normally has several verbal conjugations and variations of words, we suggest that those structures can be explored instead of all the

[1] https://www.google.com/trends
[2] https://www.google.com/news

words that a document have, when the extraction of content is executed. As I said in the last section, it is known that using words to evaluate sentences to summarize a text can be interfered by orthographic errors. In this work we expose that this process can be improved by using the concepts presents into the text, since this method is not linked to words, but to the ideas that are associated to the sentences.

1.2 Objective

Based on the motivation of the work, we want to expose a method that suggest another approach to traditional automatic summarization methods. So, the objective of this thesis is to present a comparision between methods of automatic summarization, aiming to expose the results that were obtained with our proposed process that consider "concepts" of sentences instead of the frequency of words.

To reach this goal we developed an analysis where we apply three different automatic summarization algorithms (CSTSumm(3.1), LexRank(3.2) and ConceptRank(2.1)) to a news corpus (CSTNews(2.3)) and evaluate the extracts generated by those methods to a human made extract using the ROUGE(2.2) method of extraction evaluation. More information about the techniques used in the experiment will be explored in the next chapters of this thesis.

1.3 Contribuitions

The expected contribuition of this work is present that: a) a method that uses concepts, instead of words, to automatically summarize texts; b) a comparision between this approach and another tradition automatic summarization methods. With this we can validate a hypothesis that our implementation has advantage over other automatic summarization methods, since when using the concepts of texts is supposed to produce a summary that is closer to a human made summaries. With this we expose an analysis of a method that uses concepts and deals well with lexical diversity in short texts.

1.4 Text organization

This document is structured as it follows. In the next chapter, we introduce some studies that were used as the basis for the development of the work: the method of summarization by concepts, the strategy of evaluation of the summaries and the corpus used to apply our work. The following and third chapter deals with the related works that were used to create a comparison with the work developed, in this case, we selected a summarization method proposed to the corpus, known as CSTSumm, and a traditional summarization method LexRank, a stochastic graph-based method for computing the relative importance of textual units for Natural Language Processing (NLP).

The fourth chapter shows how we developed the project, what technologies were used and how the experiment was made. The fifth chapter shows the results obtained with the experiment, and some interesting observations about the progress of the project. Finally, in the last chapter we present the conclusions obtained with this study, the future work and the final conclusion of this research.

2 BACKGROUND

To better understand Automatic Text Summarization methods, some relevant concepts must be described. In our context, **summarization** is the process of reducing a textual document in order to create a summary that retains the most important points of the original document. A **summary** is the result of the summarization process.

According to (WOLOSZYN, 2015) summaries can be indicative or informative. An **indicative summary** does not claim any role in substituting the source document. Its purpose is to alert the readers, allowing them to decide which part of the document should be read. An **informative summary** can be read in place of the original document. It includes the relevant facts reported in the original text. The purpose of this type of summary is to substitute the original document as far as possible to cover all its information.

Considering the relationship between the summary and the original text, summaries can be extractive or abstractive. An **extractive summary** avoids any efforts on text understanding to generate a summary. It selects a couple of relevant sentences, based upon statistical analysis, from an original document in order to use them in a summary. An **abstractive summary** attempts to understand the central concepts from the text and express those ideas in a totally new text.

Additionally, regarding the number of input documents used to build a summary, the process can be categorized into *single-* or *multi-document* summarization. In **single document** summarization, as the name indicates, sentences are extracted from a single document. However, in **multi-document** summarization, information can be digested from multiple sources into one single document. The significant challenges involving this approach are the repetition of the information, the identification of relevant information from all the documents and the creation of a coherent and non-redundant summary.

Finally, summarizers can be a monolingual, multilingual or even cross-lingual. It refers to the ability of the summarizer to generate summaries in more than one language. In the case of **monolingual**, the output language is the same as the input text. The output language in **multilingual** summarization is the same as the input text, but it can work with more than one language. Finally, a **cross-lingual** summarizer can accept a source text in a particular language and build the summary in another language.

In this work we implement a indicative summary, using extractive summarization with a monolingual multi-document corpus (2.3) written in Brazilian Portuguese. In this chapter we will be introducing the related works that were the basis for this thesis. First

we will be presenting the work that represent the proposal aimed to minimize problems related to the automatic summarization of texts, that use concepts, instead of words to manipulate the sentences of the target texts.

Then we define the ROUGE (LIN, 2004) measure, a method to automatically determine the quality of a summary by comparing it to other ideal summaries created by humans. Lastly, at the end of this chapter, we briefly introduce the set of texts (or *corpus*) that the methods of summarization are apllied.

2.1 ConceptText

According to (WIVES, 2004), one of the major problems of document identification and analysis is the way the features that describe and model documents are chosen. These features are not properly chosen because the characteristics usually used to represent documents are the words they contain. It is clear that we cannot use all the words in a document to represent it, and a selection must be performed. The problem is that, if the choice is made based only on the number of occurrences of the words in the document, it does not accordingly represent the content of that document.

In sequence (WIVES, 2004) suggest that with sentences alone and disconnected, neither the algorithm nor the user can easily understand the contents of the document. Thus, the exploratory analysis of documents, which is the objective of document analysis, is compromised. So, while being able to select the most meaningful sentences, the identification methods do not always obtain correct or easily interpreted results.

Trying to minimize these problems, (WIVES, 2004) proposes the use of more adequate structures to represent the contents of the documents rather than the words contained in them, using a format that can model and represent the objects and ideas present in the documents, facilitating their comprehension.

In his thesis, he proposes that a way to approach the content of a document to the user vocabulary, is using a controlled vocabulary, standardized, that uses a set of words known by the user. However, this controlled vocabulary also has words that, despite being in a context known by the user, are disconnected, not correctly representing the contents of the document.

Something like a controlled vocabulary is to use "concepts" to represent documents, which are fragments of knowledge that human beings use to represent ideas, opinions, and thoughts. They can be expressed through the language with the use of specific

terms (LOH; WIVES; OLIVEIRA, 2000b; SOWA, 2000), that is, words that when found, indicate the presence of the concept. The use of concepts places the representation of document content at a higher level of abstraction, helping the user to better understand the results of any method of document discovery and analysis.

2.1.1 Representing documents using concepts

The idea of representing documents through concepts can be understood as a process of vocabulary standardization, in which we map their words to a higher level representation scheme. To do this, methods of manipulating the vocabulary contained in documents are used. In order to standardize the terms used in the documents, a research in the literature of the area was proposed by (WIVES, 2004), and it was identified that the existing methods could be divided into two main groups: those that perform **statistical analysis** and those that do **natural language processing** (NLP).

Analyzing those methods, the ones from the NLP group are more complex, so the methods of statistical analysis were the most used for a time. Nevertheless, NLP techniques were not discarded, but improved and minimized in terms of effort and complexity. For correct understanding, analysis and processing of texts using NLP, some support knowledge is needed, also called background knowledge, usually expressed through production rules, grammatical rules, morphological dictionaries or even using ontologies.

2.1.2 Structure and identification of concepts

To work with concepts, the first task to do is to identify or extract concepts from documents (in our case, from sentences). To perform such identification, (LOH; WIVES; OLIVEIRA, 2000a) suggest applying an automatic categorization task. The categorization is guided by a set of rules that describe how a concept should be identified. These rules include cue terms that once found in a document may indicate the presence of concept. The terms may include synonyms, lexical variations and derivations, and semantic related works. Each term has an associated weight that describes the relative importance of this term to indicate the presence of the corresponding concept in the document. Weights range from 0 (totally irrelevant) to 1 (totally relevant). This weight can be manually assigned (by an expert) or by a learning process.

Using a fuzzy reasoning about the cue (terms) in a document, it is possible to calculate the likelihood of a concept being present in that document. Once the concepts of documents are identified, we must identify the similarity among the concepts of each document (or sentences). To perform this, we have chosen the following equation, which calculates the grade of similarity (gs) among two vectors (V_1,V_2) representing the concepts present on two sentences being compared. This equation was already used in previous work with promising results (PRADO et al., 2005; WIVES; OLIVEIRA; LOH, 2008; WIVES; LOH; OLIVEIRA, 2009).

$$gs(V_1, V_2) = \frac{\sum_{h=1}^{k} gi(a, b)}{n} \tag{2.1}$$

Where k is the number of concepts that sentences V_1 and V_2 have in common; n is the total number of concepts on both sentences, and gi is the grade of equality among the weighs of the h^{th} element (a in V_1 and b in V_2), which is calculated by the following equation:

$$gi(a, b) = \frac{1}{2}\Big[(a \to b) \wedge (b \to a) + (\overline{a} \to \overline{b}) \wedge (\overline{b} \to \overline{a})\Big] \tag{2.2}$$

where $\overline{x} = 1 - x$, $a \to b$ = max $\{c \in [0, 1] | a * c <= b\}$, and $\wedge$ = min.

2.1.3 Centrality

According to (ERKAN; RADEV, 2004) the centrality of a sentence is often defined in terms of the centrality of the words that it contains. A commom way of assessing word centrality is by looking at the centroid of the cluster in a vector space. The centroid of a cluster is a pseudo-document that consists of words that have tf x idf scores above a predefined threshold. In this case tf is the frequency of a word in the cluster, and idf values that are tipically computed over a much larger and similar dataset.

Erkan's work (ERKAN; RADEV, 2004) uses cosine scores to count the degree of centrality of sentences, but different scores dramatically influence the interpretation of centrality. Too low values may mistakenly take weak similarities into consideration while too high ones may lose many of the similarity relations. After computing the centrality degree, it uses a variation of the PageRank method, where each edge is a vote to determine the overall centrality value of each node. They call this new measure the "lexical PageRank", or LexRank.

In our work we use the concept approach, present above, to calculate the sentence distance and the PageRank method to extract the graph salience, it permits us to select the most salient sentences of the document, as it is show in our experiment (4).

2.2 Rouge

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It includes measures to automatically determine the quality of a summary by comparing it to other (ideal) summaries created by humans. The measures count the number of overlapping units such as n-gram, word sequences and word pairs between the computer-generated summary to be evaluated and the ideal summaries created by humans. Traditionally evaluation of summarization involves human judgments of different quality metrics, for example, coherence, conciseness, grammaticality, readability and content (MANI, 2001). Knowing about that we can conclude that the ROUGE name aim to present a method of finding the main idea of a summary, comparing its content to summaries created by humans. Gisting, in this case, is related to the central meaning or theme of a document.

In the package article (LIN, 2004) suggests that even simple manual evaluation of summaries on a large scale over a few linguistic quality questions and content coverage as in the Document Understanding Conference (DUC) (Over and Yen, 2003) would require over 3,000 hours of human efforts. This is very expensive and difficult to conduct in a frequent basis. Therefore, how to evaluate summaries automatically has drawn a lot of attention in the summarization research community in recent years.There are many evaluation methods that measure similarity between summaries. However, they did not show how the results of these automatic evaluation methods correlate to human judgments.

ROUGE has several automatic evaluation methods that measure the similarity between summaries. In this work, we are exploring two specific methods that will be explained above: ROUGE-S and ROUGE-SU. This choice was made because those methods allow us to use arbitrary gaps between sentences.

Basically, ROUGE was created to enable a direct comparison between an automatically generated summary and its human references. ROUGE calculates a score among 0 and 1 based on sets of words (e.g., the n-grams that may vary from 1 to 4) in common between human summaries and automatically generated summaries, producing precision, recall/coverage, and f-measure values. Precision indicates the proportion of reference n-grams in the automatic summary; recall indicates the proportion of reference summary;

f-measure is a unique measure of performance, combining precision and recall. These are detailed bellow.

- ROUGE-S: Skip-Bigram Co-Occurrence Statistics

  Skip-bigram is any pair of words in their sentence order, allowing for arbitrary gaps. Skip-bigram co-occurrence statistics measure the overlap of skip-bigrams between a candidate translation and a set of reference translations. For example the sentence *"police killed the gunman"* has six skip-bigram of size two: *"police killed"*, *"police the"*, *"police gunman"*, *"killed the"*, *"killed gunman"* and *"the gunman"*.

  Skip-bigram counts all in-order matching word pairs. To reduce spurious matches such as "the the" or "of in" we can limit the maximum skip distance, between two in-order words that is allowed to form skip-bigram.

- ROUGE-SU: Extension of ROUGE-S

  Conforming to (LIN, 2004), one potential problem for ROUGE-S is that it does not give any credit to a candidate sentence if the sentence does not have any word pair co-occurring with its references. The extended version is called ROUGE-SU. For example if we consider the sentence used before: *"police killed the gunman"* and its inverse *"gunman the killed police"* there is no skip bigram match between them. To accomodate this, they extend ROUGE-S with the addition of unigram as counting unit. With that we are able to compare a larger variety of skip-bigrams.

- Precision, recall and f-score

  Since we are evaluating many methods of automatic summarization, we need to start to calculate the quality of the clustering techniques we employed. This may be assessed by external measures that indicate how close the automatically produced summaries are in relation to the human extracts. For this evaluation, its used precision (2.3), coverage (recall) (2.4) and f-measure (f-score) (2.5). Precision indicates the proportion of correct segments there is inside each cluster; coverage shows the proportion of correct segments there are in each cluster in relation to was predicted in the reference clusters; f-measure is a unique performance measure, combining precision and coverage values.

$$Precision = \frac{|human\ document - automatic\ document|}{|automatic\ document|} \tag{2.3}$$

$$Recall = \frac{|human\ document - automatic\ document\ |}{|human\ document|} \tag{2.4}$$

$$F - Score = \frac{|precision * recall|}{|precision + recall|} * 2 \tag{2.5}$$

2.3 CSTNews

In our experiments we used the CSTNews corpus (CARDOSO et al., 2011) and evaluate the summarization between human extracts and CSTSumm, Concept Rank and LexRank methods. The CSTNews corpus is composed of 50 groups of news articles written in Brazilian Portuguese collected from several sections (Politics, Sports, World, Daily News, Money and Science) of mainstream online news agencies (Folha de São Paulo, Estadão, O Globo, Jornal do Brasil and Gazeta do Povo).

In conformity to (CARDOSO et al., 2011) the texts are annotated in different ways for discourse organization, following both the Rhetorical Structure Theory (RST) and Cross-document Structure Theory (CST) - both are detailed in Section 2.3.1. The corpus is a result delivered within the context of the SUCINTO Project[1], which aims at investigating summarization strategies and developing tools and resources for that purpose. Below, we detail the discourse annotation of the corpus, which aims at supporting the investigation of deep strategies on single and multi-document summarization for Brazilian Portuguese texts. Besides the subjective of RST and CST the annotation experience showed that is possible to obtain some level of systematization of the task, which allows reaching acceptable levels of agreement.

2.3.1 RST and CST

As reported by (CARDOSO et al., 2011) the Rhetorical Structure Theory (RST) was proposed by Mann and Thompson (1987) as a theory of text organization based upon its underlying propositions and their functions. More specifically, the theory prescribes a way to retrieve and generate the relationships among propositions under the assumption that the writer rhetorically organizes a text based upon his/her intentions towards

[1] http://conteudo.icmc.usp.br/pessoas/taspardo/sucinto/

the reader. Propositions express basic meaningful units, which are usually expressed by clauses or sentences at the surface of a text.

RST also defines what is called nuclearity for each relation. The propositions in a relation are classified as nuclei (i.e., more important propositions) or satellites (i.e., complementary information), and this classification reflects the author's intention. Relations with one nucleus and one satellite are said to be mononuclear relations. Relations that only have nuclei (where all propositions are equally important) are said to be multinuclear relations. Sequence, Contrast, List, Joint and Same-Unit are multinuclear relations, the others are mononuclear relations.

Inspired by RST and related work, the Cross-document Structure Theory (CST) was proposed by Radev (2000) as a way of relating text passages from different texts on the same topic. Therefore, differently from RST, CST was devised mainly for dealing with multi-document organization, and my be used to solve several problems such as summarization and question-answering ones. It may provide the means for a more intelligent information processing, particularly if we consider that it allows for dealing with redundancy and other different multi-document phenomena conveyed by a group of texts.

Although CST seems simpler than RST, it involves very difficult issues concerning its set of relations. Besides the possibility of relating every segment pair, ambiguity often takes place, which may be due to different text interpretations or to sub-specified information in the texts (for example, when the publication dates of news are not specified in several newspapers, it is difficult to determine the appropriate order to reproduce some events).

2.4 Chapter overview

In this chapter we present an overview of the background used to develop the work, exploring the previous works that inspired the construction of our work. As it follows we define some notions about automatic text summarization, that must be clarified before finally exposing our experiment, defining that this work implement a indicative summary, using extractive summarization with a monolingual corpus written in Brazilian Portuguese. Then, we expose the strategy that is used to perform the automatic summarization in this thesis, using concepts 2.1, ideas of the text, instead of words to extract the sentences.

After that, we describe the method of evaluation of the quality of summaries that

we generated, ROUGE 2.2, that are methods of measuring to automatically determine the quality of summaries by comparing them to human made summaries. Finally we demonstrate the corpus that is used to generate automatic summaries, CSTNews 2.3, that is corpus of fifth cluster of news, that are annotated in different ways for discourse organization, and follows the Rhetorical and Cross-document Structure Theory.

3 RELATED WORK

The aim of this work is to propose and evaluate the ConceptRank method to other traditional methods of automatic summarization. In what follows, we introduce related work and concepts that are the basis of this study. We briefly review some summarization methods that were our base for comparison. The approaches that will be described are CSTSumm (RIBALDO; CARDOSO; PARDO, 2016), a multi-document summarization (MDS) method, and LexRank (ERKAN; RADEV, 2004), a stochastic graph-based method for computing relative importance of textual units for Natural Language Processing (NLP).

3.1 CSTSumm

The first algorithm considered by our work is the nominated CSTSumm, a technique that explore a strategy for generating multi-document summaries that represent texts as graphs. As reported by (RIBALDO; CARDOSO; PARDO, 2016), the method investigated to perform multi-document automatic extractive summarization is the Segmented Bushy Path. The algorithm is organized by few steps. Firstly, the algorithm preprocess the source texts and computes the lexical similarity among their sentences to build a graph, where the vertices are the sentences and links have numeric values that indicate how lexically close the sentences are (using cosine mesure).

Then, the algorithm divides each text into subtopics, using TextTiling, that is a technique for subdiving texts multi-paragraph units that represents passages, or subtopics (HEARST, 1997). Once the texts are segmented, (RIBALDO; CARDOSO; PARDO, 2016) states that the next step is to identify and cluster commom subtopics within and across the documents, with those clusters, the segmented bushy path is used to select the relevant information for the summary, perfoming the content extraction.

3.2 LexRank

The other method we selected to compare to our approach is the LexRank (ERKAN; RADEV, 2004), a stochastic graph-based method for computing relative importance of textual units for Natural Language Processing. As reported by (ERKAN; RADEV, 2004),

LexRank introduce a way for computing sentence importance based on the concept of eigenvector centrality in a graph representation of sentences. In (MURGANTE et al., 2014) eigenvector centrality is defined by a measure of the influence of a node in a network, in this case, it calculates the importance of a node in a graph. It associate scores to all graph nodes based on that connections to high-scoring nodes add more to other nodes than equal connections to low-scoring nodes. In this model, the adjacency matrix graph representation of the sentences is represented by a connective matrix based on intra-sentence cosine similarity.

3.2.1 Centrality-based Sentence Salience

There are several ways of computing sentence centrality using the cosine similarity matrix and the corresponding graph representation. Conforming to (ERKAN; RADEV, 2004), the hypothesis of LexRank is that the sentences that can better describe a text (in other words, are **similar** to the other sentences in the text) are more central (or *salient*) in the graph. In order to define similarity, it is used 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. The similarity of sentences is calculated by the cosine between two corresponding vectors:

$$idf - modified - cosine(x, y) = \frac{\sum_{w \in x,y} tf_{w,x} tf_{w,y} (idf_w)^2}{\sqrt{\sum_{x_i \in x} (tf_{x_i,x} idf_{x_i})^2} \times \sqrt{\sum_{y_i \in y} (tf_{y_i,y} idf_{y_i})^2}} \quad (3.1)$$

where $tf_{w,s}$ is the number of occurrences of the word w in the sentence s, and the idf_i is the *inverse document frequency* a mesure used to asses the importance of the words in a sentence, that is defined by the formula (JONES, 1972) :

$$idf_i = \log(\frac{N}{n_i}) \quad (3.2)$$

3.3 Chapter overview

This chapter deals with the related works that are used to level this thesis work. Knowing about this, we describe the sophisticated method of automatic extractive summa-

rization proposed by (RIBALDO; CARDOSO; PARDO, 2016) nominated as CSTSumm, which tries to represent in a summary the main subtopics from the source texts. The we explore the traditional method LexRank, that uses the cosine similarity between sentences.

4 EXPERIMENT SETUP

After discussing the methods to represent textual content, assess their similarity and evaluate results, the next step is to design an experiment and define the corpus that will be used as "gold-standard". The aim of this chapter is to provide an overview of the strategy that we used to implement the setup of the experiment.

In this procedure, we use the CSTNews(2.3) as corpus to apply the methods for automatic texts summarization. This corpus is composed by fifth groups of news articles written in Brazilian Portuguese from several sections of news such as politics, sports, world, daily news, money and science. Each news is collected from several mainstream online agencies (Folha de São Paulo, Estadão, O Globo, Jornal do Brasil e Gazeta do Povo). For example, an article about a hurricane in Japan is collected from many online news agencies to create a variety of sentences, about the same subject, to be explored by our summaries.

To generate the automatic summaries we used the three method explored in the lasts chapters of this thesis: the algorithm suggested by CSTSumm(3.1), a traditional method of summarization: LexRank(3.2), and the approach that we are proposing that uses "concepts" instead of words to generate a summary, nominated at ConceptRank((2.1)).

When the three summaries for each news are ready, we need to find a way to evaluate the quality of the summaries, aiming to compare their performance individually, seeking for a way to show that the method using concepts excels in some structure of evaluation. According to (UMAM et al., 2015), a good summary should contain the main topics of the original document (**coverage**) while keeping the redundancy to a minimum (**high diversity**) and have smooth connection among sentences (**high coherence**). Several methods have been proposed to evaluate summaries' quality, understanding these methods is essential to correctly evaluate systems.

The corpus that was selected to be consider in this experiment also provided a human generated extract, so we decided that the evaluation of the summaries quality could be made by comparing our outcome to the gold-standard summary that is human made. In order to compare the quality of the automatic and the human generated extract, we used the evaluation method ROUGE (2.2), this measure count the number of overlapping units such as n-gram, word sequences and word pairs between the computer-generated summary to be evaluated and the ideal summaries created by humans.

4.1 Chapter overview

In this chapter we expose the strategy used to accomplish our experiment. In short, we present that we use a corpus (CSTNews(2.3)) to perform the application of automatic summarization methods, in this case CSTSumm(3.1), LexRank(3.2) and ConceptRank(2.1). With this, it is possible to draw some conclusions about our approach by evaluating the performance of our method when compared to other automatic summarization methods.

At the end is stated that to compare the performance of the methods individually, is used the method ROUGE, that count the number of overlapping units existing in a text, being a way to compare the computer-generated summary and the ideal summaries created by humans.

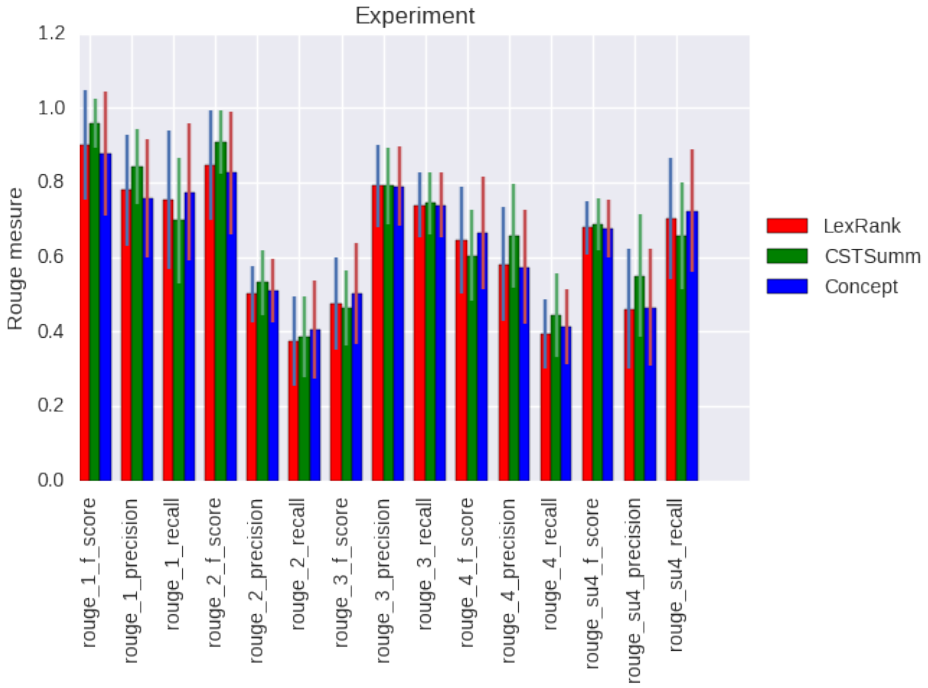
5 RESULTS

As already stated in the experiment setup chapter(4), after creating the automatic summaries for all fifth groups of news, we evaluate the outcome with the gold-standard extract made by human using the ROUGE method. The size of the human summaries that the corpus provide corresponds to 30% of the size of the longest article in each group of news, resulting in a compression rate of 70%. To match out summaries to the human made extract, we decided to use the same compression rate, in order to the comparision using ROUGE fair.

ROUGE calculates a score among 0 and 1 based on set of words (e.g. the n-grams that may vary from 1 to 4) in common between humam summaries and automatically generated summaries, producing the **precision** that indicates the proportion of reference n-grams in the automatic summary, the **recall** that indicate the proportion of reference n-grams in the automatic summary in relation to the reference summary and the **f-measure**, measure of performance, combining precision and recall.

After performing the experiment, our average results with standard deviation are the ones presented in Figure 5.1:

Figure 5.1: Total average of the experiment



For a better comprehension about the results, the three methods will be explored using tables.

Below, the average results that we obtained using the LexRank Method:

Table 5.1: Average values obtained for LexRank Method

LexRank	**Quality Evaluation Methods**		
	Precision	**Recall**	**F-Score**
ROUGE-1	0.9013	0.7545	0.7907
ROUGE-2	0.7802	0.6451	0.6785
ROUGE-3	0.5801	0.4757	0.5006
ROUGE-4	0.4599	0.3746	0.3946
ROUGE-SU4	0.8467	0.7035	0.7389

Then, we present the average results that we obtained using the corpus summarization method, named CSTSumm Method:

Table 5.2: Average values obtained for CSTSumm Method

CSTSumm	**Quality Evaluation Methods**		
	Precision	**Recall**	**F-Score**
ROUGE-1	0.9590	0.6986	0.7914
ROUGE-2	0.8425	0.6042	0.6888
ROUGE-3	0.6562	0.4646	0.5324
ROUGE-4	0.5491	0.3862	0.4438
ROUGE-SU4	0.9080	0.6557	0.7453

At last, the average results that we obtained using the proposed method using concept instead number of words, ConceptRank Method:

Table 5.3: Average values obtained for ConceptRank Method

ConceptRank	**Quality Evaluation Methods**		
	Precision	**Recall**	**F-Score**
ROUGE-1	0.8782	0.7743	0.7898
ROUGE-2	0.7578	0.6639	0.6779
ROUGE-3	0.5736	0.5015	0.5113
ROUGE-4	0.4642	0.4054	0.4128
ROUGE-SU4	0.8262	0.7245	0.7400

Studying those results we obtained by the application of the automatic summarization methods for each news group, we can observe that the ROUGE method evaluation cover what we are proposing since the beginning of this thesis. All the automatic summarization methods shows very similar results from all evaluations, but looking forward, the average results we can see that observing the precision evaluation, that evaluates the number of references in the summary itself, methods that use word frequency (LexRank, CSTSumm) shows better results since that using the more similar words, the summary generated has more references when compared to itself.

Noticing the recall evaluation, that compares the automatic summaries to the gold-standard human made summary, we see that the proposed method using concepts instead of words have a little advantage, what makes sense, since the use of concepts explore the ideas of the texts, what makes the extraction of sentences more closer to the human selection.

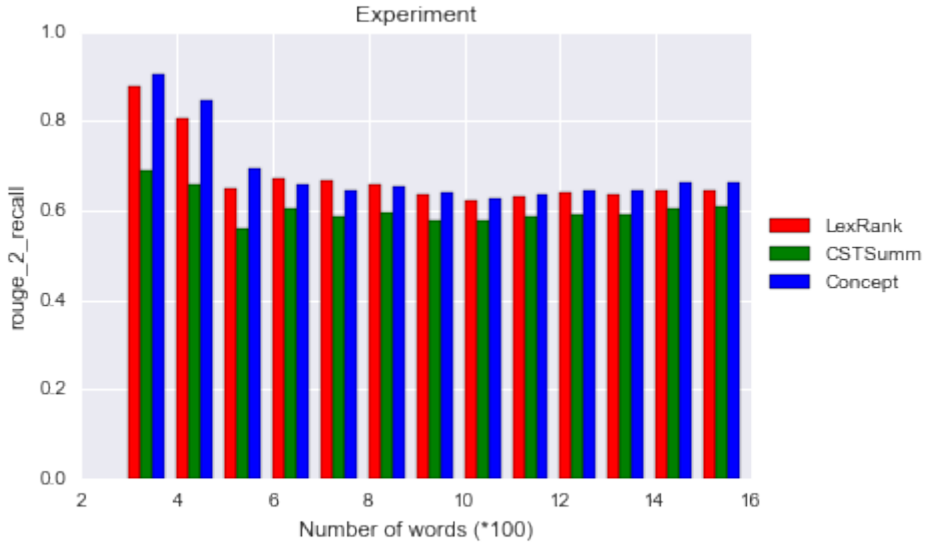
The f-score mesure show a performance evaluation, selecting the method with better average results, combining precision and recall, as we can see, CSTSumm has the best results of all three methods, because its precision is very high compared to the other approaches, but the aim of this work is focused in show the best outcome compared to the human summary, which we can conclude looking at the graphic shape.

Extending the evaluation results, we can conclude some more observations for our method.

- The recall measure of our approach, using concepts, is always higher, especially for texts with less than 500 words.

  We can notice that by looking for the average graphic, but we can verify that at the Figure 5.2:

Figure 5.2: ROUGE-2 Recall

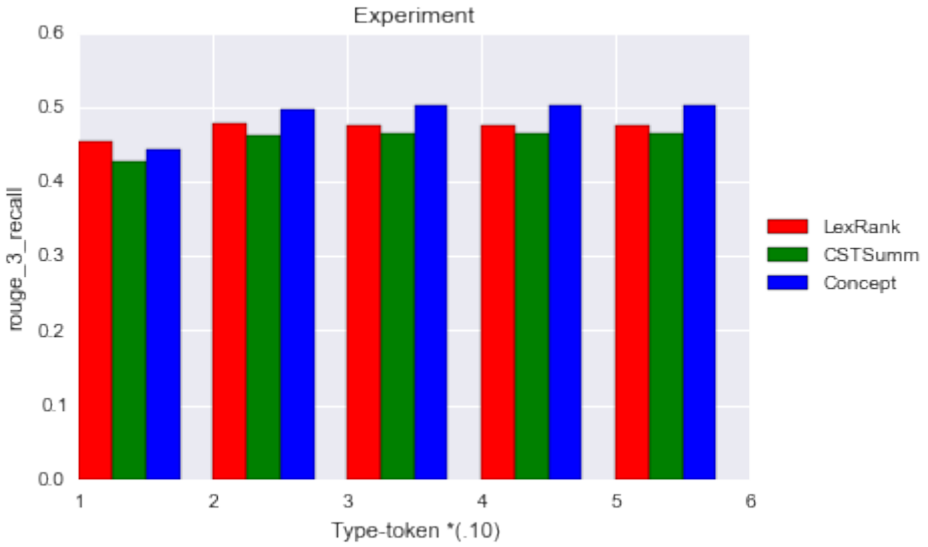


- The concept approach get better results when the relation type-token is higher than 3, as it follows in Figure 5.3:

  The type-token distinction separates types (representing abstract descriptive concepts) from tokens (representing objects that instantiate concepts).[1]

[1] http://plato.stanford.edu/entries/types-tokens/

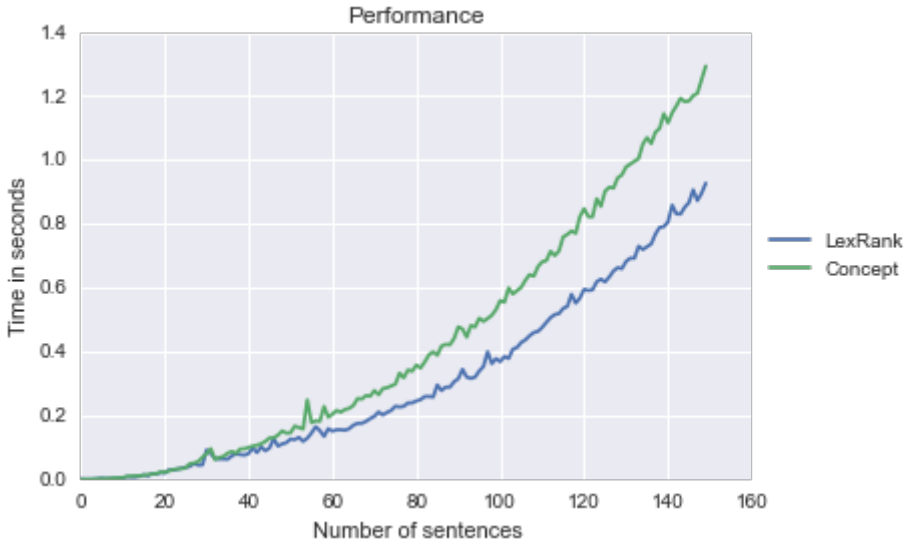
Figure 5.3: ROUGE-3 Recall



- The performance of ConceptRank and LexRank is similar where our approach is better.

  Although the concept method presents a higher runtime when compared to LexRank, this time may not be a problem for short texts, where the results of our approach better.

Figure 5.4: Perfomance of LexRank and ConceptRank



5.1 Chapter overview

The aim of this chapter is to show the results we obtained comparing automatic summarization methods with ROUGE evaluation approach, that show a score from 0 to

1 of overlapping units such as n-grams. Analyzing the average outcome collected from the evaluation of each group of news, we validated some propositions made from the beginning of the thesis. We use three quality scores: precision, recall and f-score to present that: a) methods with frequency word extraction have a high precision score, since this score measure the number of references with itself; b) our proposed method using concepts has an advantage when considering the recall measure, since this score aims to compare our extraction to human summary, what gives the impression that we wanted to prove since the beginning of the work.

Exploring more evaluation results we concluded that the concept method recall score is always higher, especially when the corpus of texts has less than 500 words. In addition to this our approach get better results when the relation type-token (a relation that distinct types from tokens) is higher than 3. Finally, we present a performance graph comparing the concept approach to the LexRank method.

6 CONCLUSION

This works has the objective to expose a method that use concepts, instead of words, to perform automatic text summarization. With this approach we analyze a process of automatic extraction of sentences that aim to prove that using concepts we can get a summary that is more related to human made summaries. By that, we explore the Automatic Text Summarization (ATS), a task that became important since the abundance of texts available on the internet is growing. To validate our idea, we perform the application of some methods of automatic text summarization to a corpus that consists in a group of fifth news articles. By doing that we conclude that the method using concepts is not attached to words and has advantage when the lexical diversity is high and when compared to human made summaries.

In addition to this our method results for texts with number of words up to 500 terms are higher than the other approaches explored in this thesis. Observing all those conclusions we can finally declare that our method is an automatic summarization option when we have short texts with high lexical diversity.

For future work we propose an experiment with a longer texts set of texts, in order to investigate our method perfomance to longer texts, exploring which type of text better works with our approach.

REFERENCES

CARDOSO, P. C. et al. Cstnews-a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In: **Proceedings of the 3rd RST Brazilian Meeting**. [S.l.: s.n.], 2011. p. 88–105.

ERKAN, G.; RADEV, D. R. Lexrank: graph-based lexical centrality as salience in text summarization. **Journal of Artificial Intelligence Research**, p. 457–479, 2004.

HEARST, M. A. Texttiling: Segmenting text into multi-paragraph subtopic passages. **Computational linguistics**, MIT Press, v. 23, n. 1, p. 33–64, 1997.

JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. **Journal of documentation**, MCB UP Ltd, v. 28, n. 1, p. 11–21, 1972.

LIN, C.-Y. Rouge: A package for automatic evaluation of summaries. In: BARCELONA, SPAIN. **Text summarization branches out: Proceedings of the ACL-04 workshop**. [S.l.], 2004. v. 8.

LOH, S.; WIVES, L. K.; OLIVEIRA, J. P. M. de. Concept-based knowledge discovery in texts extracted from the web. **SIGKDD Explor. Newsl.**, ACM, New York, NY, USA, v. 2, n. 1, p. 29–39, jun. 2000. ISSN 1931-0145. Available from Internet: <http://doi.acm.org/10.1145/360402.360414>.

LOH, S.; WIVES, L. K.; OLIVEIRA, J. P. M. de. Concept-based knowledge discovery in texts extracted from the web. **ACM SIGKDD Explorations Newsletter**, ACM, v. 2, n. 1, p. 29–39, 2000.

MANI, I. **Automatic summarization**. [S.l.]: John Benjamins Publishing, 2001.

MURGANTE, B. et al. **Computational Science and Its Applications-ICCSA 2014: 14th International Conference, Guimarães, Portugal, June 30-July 3, 204, Proceedings**. [S.l.]: Springer, 2014.

NENKOVA, A.; MASKEY, S.; LIU, Y. Automatic summarization. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings...** [S.l.]: Association for Computational Linguistics, 2011. p. 3.

PRADO, H. A. D. et al. Text mining in the context of business intelligence. In: KHOSROW-POUR, M. (Ed.). **Encyclopedia of Information Science and Technology, First Edition**. [S.l.]: Hershey: IGI Global, 2005. p. 2793–798.

RIBALDO, R.; CARDOSO, P. C. F.; PARDO, T. A. S. Exploring the subtopic-based relationship map strategy for multi-document summarization. **Revista de Informática Teórica e Aplicada**, v. 23, n. 1, p. 183–211, 2016.

SALTON, G.; WONG, A.; YANG, C. S. A vector space model for automatic indexing. **Commun. ACM**, ACM, New York, NY, USA, v. 18, n. 11, p. 613–620, nov. 1975. ISSN 0001-0782. Available from Internet: <http://doi.acm.org/10.1145/361219.361220>.

SOWA, J. F. Ontology, metadata, and semiotics. In: SPRINGER. **International Conference on Conceptual Structures**. [S.l.], 2000. p. 55–81.

UMAM, K. et al. Coverage, diversity, and coherence optimization for multi-document summarization. **Jurnal Ilmu Komputer dan Informasi**, v. 8, n. 1, p. 1–16, 2015.

WIVES, L. K. **Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos**. Thesis (PhD) — UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL, 2004.

WIVES, L. K.; LOH, S.; OLIVEIRA, J. PALAZZO M. de. A comparative study of clustering versus classification over reuters collection. In: 8TH INTERNATIONAL WORKSHOP ON PATTERN RECOGNITION IN INFORMATION SYSTEMS. **Proceedings...** [S.l.], 2009. p. 231–236.

WIVES, L. K.; OLIVEIRA, J. P. M. de; LOH, S. Conceptual clustering of textual documents and some insights for knowledge discovery. In: PRADO, H. d.; FERNEDA, E. (Ed.). **Text Mining: Techniques and Applications**. [S.l.]: Information Science Reference Hershey, PA, 2008. p. 223–243.

WOLOSZYN, V. **Tell me why: uma arquitetura para fornecer explicações ricas sobre revisões**. Dissertation (Master) — Universidade Federal do Rio Grande do Sul, 2015.