Genre-informed Unsupervised Extractive Automatic Text Summarization

in Danish, English and Spanish
A Thesis Presented to the
FACULTY OF HUMANITIES
UNIVERSITY OF COPENHAGEN
In Partial Fulfillement of the
Requeriments for the Degree
MASTER OF ARTS
(IT & Cognition)

Esther Sebastián Liso
zxv330@alumni.ku.dk
Under the supervision of:
Héctor Martínez Alonso

November 9, 2015

Contents

Abstract

There is a huge amount of texts in the world, and this amount increases incredibly fast. Therefore, it is very important to develop summarizing tools to help us get the most relevant information quickly without losing much time reading less relevant texts.

Automatic text summarization can be used to read information about a topic much faster, easier and with no redundancy. It is also useful for reading just what it is important in a text, instead of having to read the full document.

It can as well be used to give a short presentation about what a text is about, so that people can decide whether it contains what they are looking for or whether they think it is worth it to read. It can likewise be helpful in order to understand a text in a few words, so that the knowledge is easier for the reader to remember.

This thesis explores a method for extractive automatic text summarization that relies on the PageRank unsupervised graph-based algorithm. The PageRank (Brin and Page, 1998) algorithm has traditionally been used to summarize texts by finding relationships between the sentences in a text (e.g. looking for word overlaps or similar parts of the speech between sentences).

PageRank creates a graph with those relationships between sentences and selects the sentences that have the highest rankings (e.g. that are considered the most important by the algorithm). In order to go more in depth and improve automatic summarization, this thesis investigates how the PageRank algorithm can be extended by some linguistical and heuristic specifications, so that the resulting summary is more accurate.

Since PageRank is an unsupervised algorithm, it has been applied to different languages without modifications and it is known to achieve satisfying results. In this thesis, we will test whether PageRank algorithm works equally well in different types of texts and different languages or whether there are differences.

In addition, we perform evaluations and check whether there are differences among different 1) PageRank-based systems, 2) genres and 3) languages.

Finally, we evaluate the performance of the algorithm and its features using the evaluation toolkit ROUGE in all contexts (different languages and different types of texts). We focus on the analysis of the differences in scores and the most important sentences that were selected using different features.

We show the results compared with a baseline created following the study of Mihalcea (2005) "Language Independent Extractive Summarization" and the study of Ceylan et al. (2010) "Quantifying the Limits and Success of Extractive Summarization Systems Across Domains".

1 Introduction

There is a huge amount of texts in the world, and this amount increases incredibly fast. Therefore, it is very important to develop summarizing tools to help us get the most relevant information quickly without losing much time reading less relevant texts.

When humans read a text, they cannot fully agree on what the best summary is. It is even more difficult to measure how good automatic summaries are using computer-based methods of evaluation.

We argue that a fine-tuning in the unsupervised algorithms that are used in the automatic summarization task can approach relatively better system-generated summaries without humans need of annotating or writing summaries. We do this by looking for general patterns in good summaries across different languages and genres, so that a machine "writes" a good summary.

1.1 Motivation

Texts are different and it is necessary for humans to fully understand a text in order to make a good summary. Accordingly, it is difficult for humans to complete the task of summarizing and, consequently, there is no agreement about what the best summary is and what a good summary must or must not contain.

Many studies have focused the task of automatic summarization on the short documents, especially news articles (Mihalcea (2004 and 2005), Gustavsson (2010) and Hong et al. (2014) among others) and e-mail threads (Loza (2014) or Almeida et al. (2014)). There is research on short stories (Kazantseva, 2006) and full books (Ceylan and Mihalcea (2010), Mihalcea and Ceylan (2007) and Luhn (1958)) and scientific articles and legal documents (Ceylan and Mihalcea, 2010). There are scientific reports that summarize web pages (Zhou, 2003), but to our knowledge there are no studies on Wikipedia articles to date.

In addition, although the greatest amount of summarization studies focus on articles in newspapers, the focus of these documents are on news articles, and not on other type of texts that appear in newspapers as well, such as letters to the editor or editorials.

In this thesis we analyze different types of texts: *descriptive*: Wikipedia articles, *narrative*: novels and fairytales, *expository*: news articles, *argumentative*: letters to the editor and *mixed*: scientific papers. Furthermore, we analyze their corresponding system-generated summaries. We compare these summaries with model summaries extracted on-line and compare their results.

For that purpose, we introduce a new corpus specifically gathered for the evaluation of different genre summaries, which is described thoroughly in section @@@@@@@@@. We analyze the differences between five types of texts and the differences in their summaries to discern general characteristics of our data. We generate three baselines using no learning algorithms, namely 1) selecting FIRST[1] (BF), 2) LAST (BL) and 3) RANDOM (BR) sentences of a text.

Table 1@ reports ROUGE-1 recall baseline scores for every genre in Danish, English and Spanish. Numbers in bold highlight the system with the highest score for each genre and language.

Despite the fact that these results cannot be directly compared across natural languages due to the differences mentioned above and differences in the original documents and their summaries, we get an idea of which systems that work best in this data set: generally the deterministic baseline: FIRST.

[1] Our baseline FIRST is usually named LEAD in the automatic summarization literature.

	DANISH			ENGLISH			SPANISH		
Genre	BF	BL	BR	BF	BL	BR	BF	BL	BR
Descriptive	**.22**	.18	**.22**	**.21**	.15	.19	**.29**	.20	.25
Narrative	**.23**	.20	.18	**.26**	.23	.21	**.30**	.26	.25
Expository	**.53**	.24	.31	**.58**	.20	.26	**.29**	.20	.25
Argumentative	**.48**	.32	.35	**.46**	.43	.43	**.47**	**.47**	.46
Mixed	**.30**	.21	.19	**.30**	.24	.22	**.38**	.28	.27

Table 1: Average ROUGE-1 recall scores for 50-word summaries for each genre in Danish, English and Spanish.

We can see differences in the scores when obtained randomly and when using the last sentences of the text. We expect that these divergences reflect structural differences in text types. This knowledge can be used to improve the automatic summarization task. It appears that sentences located toward the end or the centre of the document can be useful for improving the summary of a document and not only those located at the beginning.

1.2 Research Question

We regard the text structure of a genre and the parameters that an unsupervised PageRank-based system (see section @@PageRank) is set to as the two main drivers of the improvement of the automatic text summarization task. We therefore find it vital to investigate how incorporating the variation of different text structures can improve the quality of the automatic summarization.

We seek to accommodate the fact that automatic text summarization is a complex task with many aspects that influence not only its output but also its evaluation results. There are different methods to evaluate the results of the automatic text summarization task, but we are only using ROUGE to evaluate our summaries[2].

We are solely using a PageRank-based system[3]. The main reason is that we can parametrize PageRank to take into account different a priori probabilities of choosing a sentence for the resulting summary. In other words, we incorporate our expectation of the relevance of sentences given this parametrization in texts for different genres. Thus, our main focus is on the different parameters that we feed our PageRank-based system with in order to generate different outputs and find out which output summary works best.

Our preliminary supposition is that an analysis of the first sentences of a text, the last sentences of a text and a randomly-selected sentences of a text will help us detect where in a text the most relevant or important parts of the genre are. Accordingly, this knowledge should guide us on how to modify a PageRank-based system to the extent that the importance of the different parts of a document in any genre are defined from the outset.

The introduction outlined the importance of different types of text structures. The aim of this thesis is to study different text structures and analyze how automatic text summarization can be improved when taking these structures into account. We have formulated the following research question:

> **What is the importance of text structure and genre in an automatically generated summary?** And, including this knowledge, how can we improve automatically generated summaries?

[2] Another method is ParaEval (Zhou et al. 2006), which uses paraphrases to evaluate summaries automatically. This tool closely resembles ROUGE.

[3] In other studies (Mihalcea, 2004 and Mihalcea and Tarau, 2005) PageRank-based results are compared with other methods such as HITS (Kleinberg, 1999) and Positional Power Function (Herings et al., 2001). There are many other methods not taken into consideration in this thesis (such as CLASSY 04 (Convoy et al. 2004), RegSum (Hong and Nenkova, 2014).

1.3 Scope and Delimitation

We adjust our thesis to extractive Single-Document Summarization. Our method is only based on the PageRank algorithm, with which good results have been obtained in other studies. This will be further explained in section @@@@@related work. Against this background, we evaluate our results with ROUGE, since ROUGE is the most commonly used tool to evaluate system-generated summaries.

1.3.1 Languages

This thesis covers more than one language, namely Danish, English and Spanish. We use English as the reference language, due to the fact that the greater part of the research in automatic text summarization is conducted in English.

We apply our method to other languages, too. We have chosen Spanish, because it is easy to find data in this language. Additionally, we have chosen Danish, a language where there is not as much data available as in either English or Spanish (see section @@@data collection in other languages in DATA@@@). In this way, we encompass three languages with different levels of well-studiedness in Natural Language Processing (NLP).

We also choose more than one language because, despite normally existing in all languages, sometimes genres have certain peculiarities that to a certain extent make them different in each language (see section@@@ data argumentative mas largo en danes, etc@@@). We analyze those differences and see whether they are big enough to notice a deviation in some languages, especially between our multilingual parallel data sets.

1.3.2 Text Types

The genres that are analyzed and evaluated in this thesis are the following: 1) *descriptive*, 2) *narrative*, 3) *expository*, 4) *argumentative* and 5) *mixed*. The first four types of text are well described in Van Dijk (1980) and the mixed type of text is a combination of the others (see section types of text@@@@@) that has properties from more than one genre.

1.3.3 General limitations

A major limitation to the automatic summarization task is the quantity of human resources that generating model summaries entails. If we need at least one annotator to write a model summary for every text document in a big data set, it will take too much time to be worth it.

However, thanks to the Internet we can obtain some already human-made summaries and their corresponding full documents. For example, some Wikipedia articles have a section at the very beginning that can work as a model summary of the article. There are specific web pages designed to upload and download summaries from books. Moreover, abstracts from scientific papers can be used as summaries of the full article.

Nevertheless, this simple and fast method of gathering a data set of model summaries also have disadvantages: model summaries vary in length and style since they are voluntarily done without following a standard pattern. As a consequence, the comparison of results between different data sets is not very accurate because of their differences in length and thereby their differences in ROUGE scores.

Furthermore, it is time consuming to generate summaries from a very large data set with long texts because of the time the algorithm needs to process them lengthens. Additionally, once generated, the summaries need to be evaluated. This is also a time-consuming process.

1.4 Contributions

The findings of this study contribute to the automatic summarization task and NLP. The fact that texts can be grouped into different types according to their text structure can give us a clue about where the most important information can reside. By studying genres and localizing where the most and least relevant information is, we can design more sophisticated algorithms to extract the most relevant information when summarizing a text.

We claim that a genre-informed gathering of the automatic summarization data sets is beneficial for the study of this task. This concept of genre can catch the different distributions in structure that some texts have and normally share. Therefore, we claim that it is a good approach to group different domains in wider categories: namely genres. This thesis analyzes genre-informed summaries and compares them with other domain-aware summaries.

There are not many data sets that take the genre (and not only the domain) into account. Thus, we have collected the GICorpus, which is another contribution of this thesis. It contains 450 documents and their corresponding model summaries to be used in different tasks, such as automatic summarization, taking the genre of the text into account. This corpus is publicly available for academic use.

Furthermore, we propose a new way of analyzing results in automatic summarization by not only taking into account results obtained from the full document, but also taking into consideration the relative contribution of each section.

We show that by excluding some sections in the text we obtain better results than when analyzing the full document. Moreover, by splitting texts into parts, we can see how the results vary and have a general idea of where the most and least important sentences may lay. This can be useful especially when analyzing long documents.

2 Theoretical Framework

In this section, we define the central concepts in this thesis in the areas of automatic summarization and text structure. We describe our methodology based on the PageRank algorithm and the evaluation system: ROUGE. Finally, after having introduced the necessary terms, we redefine our research question (see section @@@) into a hypothesis.

2.1 What is Automatic Text Summarization?

Automatic Text Summarization is a task by which a computer programme reduces a text into a short version that contains its most important information. It already started with the early approaches from the Natural Language Processing (NLP) community to automatic generation of book abstracts (Luhn, 1958).

Automatic text summarization can be used to read information about a topic much faster, easier and with less redundancy. It is also useful for reading just what it is important in a text, instead of having to read the full document. It can also be used to give a short presentation about what a text is about, so that people can decide whether it contains what they are looking for or whether they think it is worth it to read. It can also be helpful in order to understand a text in a few words, so that the knowledge is easier for the reader to remember.

2.1.1 Single-Document Summarization vs. Multi-Document Summarization

There are different ways to approach the task of summarization, e.g. single-document summarization and multi-document summarization.

1. Single-document summarization is characterized by generating an individual summary for a single document.
2. Multi-document summarization generates one summary for a group of documents (which usually are about the same topic or belong to the same thread).

In this thesis, we deal with single-document summarization.

2.1.2 Extractive summarization vs. abstractive summarization

There are two methods of automatic text summarization: extractive summarization and abstractive summarization.

1. Extractive summarization consists of selecting some sentences, phrases our clauses[4] from the original text in order to create a new summary. Sentences make sense and are grammatically correct. However, there are normally cohesion problems.
2. Abstractive summarization creates a new shorter text based on the information from the original document but it might contain words or syntactical structures that did not appear in it.

In this thesis, we focus on extractive summarization.

2.1.3 Supervised learning vs. Unsupervised learning

There are both supervised and unsupervised methods for automatic summarization.

1. Supervised methods need a large amount of training data from the same domain. They are very domain-dependent, because they do not normally work well in test data sets which are different from their training data set.
2. Unsupervised methods do not need a training data set. In addition, it is very easy to work with unsupervised methods in different domains and use them in any type of text. Nevertheless, some adjustments can be made when running these methods

[4] In some studies, such as Kikuchi et al. 2014, have called them Elementary text Units or EDUs.

in very different types of texts in order to obtain better results. An unsupervised learning method can be used to extract some key features from the data set without any previous labeled data. An unsupervised learning method simply bases its output on text properties and some heuristics that we can determinate a priori, i.e. word overlap, high-frequency content words or sentence position.

In this work, we use the unsupervised algorithm: PageRank (see section@@@@pagerank@@) (Brin and Page, 1999), which is a graph-based method.

2.2 Text structure

All texts have an objective: to communicate a message to a recipient (Jakobson, 1960). This objective is what determines the structure that a text is going to acquire. Although all texts are different and have their own characteristics, it is possible to find technical and linguistic similarities between texts. These similarities allow us to group the texts in general types of text.

There are different kinds of classifications. Nevertheless, it is impossible to create closed categories, since it is not uncommon for texts to have mixed characteristics. Therefore, texts can sometimes be classified in different categories at the same time.

The classification we are using here refers to different types of text: descriptive type of text, narrative type of text, expository type of text, argumentative type of text, defined by van Dijk (1980) under the term superstructure or conventional schema.

These different superstructures are abstract cognitive structures that are socioculturally accepted. These categories have certain rules that texts must follow so that their message is correctly understood. In other words, superstructures are merely the categorical structure that organizes this global content (Van Dijk, 1980).

Some texts, however, do not have a conventional schema because it is necessary that a lot of texts follow the same patterns and that those patterns are accepted by the users of a language. When these patterns are generalized, a new superstructure is formed. In addition, there are some types of texts, like advertisements, modern poems or personal letters that do not normally follow a fixed structure.

Many types of text, as Van Dijk (1980) explains, show some kind of introduction. Despite the fact that this introduction may vary across different superstructures, it always contains general information about what it is going to be dealt with and what is necessary know in advance.

Van Dijk (1980) also found a conclusion in most types of text. A conclusion may have a persuasive function as in arguments, but it can also be a kind of summary of the most important aspects of the text or a closing of the discourse, or a mention of some actions to be taken in the future.

In the middle of these two sections there is the body, which can, according to Van Dijk (1980), be split into some basic meta-categories: problem, solution and evaluation. Nonetheless, some of these sections may be missing in some types of text or have different specific functions. The problem carries some new information. Therefore, after this section, we might find a solution for the given problem. And finally, an evaluation simply adds a moral to the given text (and sometimes may appear in the conclusion).

Example:

Introduction:	”I was in school yesterday.”
Problem:	”I found a wallet on the floor.”
Solution:	”I carried it to the reception.”
Evaluation:	”I think that is what people must do.”
Conclusion:	”They found the owner.”

2.2.1 Text types

Texts are built using different strategies depending on the intention of the addresser (Jakobson, 1980) to transmit the information in a way that the addressee (Jakobson, 1980) understands the message correctly. For example, one possible way to tell about a succession of facts, for example about a journey, would be starting from the first day, continuing by explaining the events of the following days until the last day of the journey.

This structure would be the prototypical structure of a narrative text. However, if we want to convince our addressee about how good our journey was or suggest that he or she should also travel there, we can use another structure where the most relevant information is at the very beginning (more prototypical in expository texts) or at the end (more prototypical in argumentative texts).

Text types, referred to in this thesis as *genres*, are different from literary genres, which are more or less defined by tradition. Examples of literary genres are novels, tales, essays, epic poems or editorials, news and interviews.

The superstructures or *genres* that we analyze and the specific texts we will focus on are the following:

Descriptive:	Wikipedia articles[5]
Narrative:	Novels and fairytales of Hans Christian Andersen
Expository:	On-line news articles
Argumentative:	Letters to the editor
Mixed:	Scientific papers

Descriptive Genre

Descriptions are a type of text which represents features of objects with words. It recreates reality as an intellectual perception of the objects so that the addressee can imagine them. It might be objective or subjective.

The content may follow many different structures such as **composition** (enumeration of the parts or elements that constitute the object), **use** (how the object works or what it is needed so it works), **utility** (what it is used for), etc.

It appears in encyclopedias, historical texts, legal texts, instruction manuals and technical or scientific texts.

Narrative Genre

A narrative text is a telling of one or various real or fictitious facts that have occurred. Normally they are formulated in a chronological order. The most important information is what happened, which is recreated by the addresser with words.

The structure is normally as follows:

1. **Introduction**: conflict and/or happenings before the conflict;
2. **Climax**: development of the conflict;
3. **Denouement**: Resolution of the conflict.

Normally it starts with an introduction, followed by a development and an ending. Novels, tales or stories among others are narrative texts.

Expository Genre

Expositive texts belong to a type of text used to present a topic to the general public or to a specialised addressee. It is characterized by presenting clear and ordered information so that the text is understood. It usually starts with general terms and from then on there is a thematic progression where more elaborated ideas are explained following a deductive or inductive reasoning.

The most frequent pattern consists of:

1. **Introduction**: enunciates and demarcates the topic to talk about, presents information that is necessary to understand the text and some definitions of terms;
2. **Development**: contains the most relevant information, references and ideas. This is usually the largest part;
3. **Conclusion**: it has the form of a summary of the most relevant ideas from the Development part, sometimes missing.

Scientific, technical and humanistic texts often use this type of text. It is also useful to transmit information such as in text books, exams or news.

Argumentative Genre

Argumentation is a type of text that is used to demonstrate a fact or to defend an opinion with proof and reasoning. It is used specially with controversial topics because they admit different opinions. The objective is that the addressee adheres itself to the opinion of the addresser.

Often the structure is organized as follows:

1. **Thesis**: idea pretended to be demonstrated by arguments e.g. "people must stop smoking";
2. **Argumentative body**, reasons that allow the addresser to convince the addressee.

Since the organization of the ideas is fundamental to convince the addressee, there are many different internal structures (for example: an idea following by a conclusion or many the ideas and a final conclusion at the end). An example of arguments for the thesis "People must stop smoking" can be "Tobacco can kill people" or "It is expensive".

The order and complexity of the parts may vary. Sometimes there is a quadratic structure, which means that the thesis appears both at the beginning and at the end of the text.

It is often used in political speech, letters to the editor and editorials in newspapers or advertisements.

Mixed Genre

Most texts actually belong to the mixed genre. Although we have categorized some of them as *pure* descriptive or narrative, it is common to find texts which have parts belonging to different categories. However, we decide that some texts are representative for one specific genre.

Nevertheless, we add a mixed genre category in order to check how these texts behave in our experiments. Some texts, such as scientific papers, are mixed since they usually have differentiated sections. These sections have a different communicative intention, namely that the related work is more descriptive or the conclusion is more argumentative.

Because of this blending, different types of documents within the mixed genre may individually have different structures.

2.3 Algorithm: PageRank

The algorithm we use in this thesis is PageRank (Brin and Page, 1999). PageRank is a graph-based algorithm, which contains information about nodes (a set of vertices) and edges (links which connect pairs of vertices). PageRank can be applied to NLP tasks, such as text summarization, where the sentences of a text become the nodes of the graph and the relations between sentences become the edges of the graph (for a further elaboration of PageRank, see section @@@@2.3.1).

There are differences in how the relationships between sentences in the graph are created: Undirected, directed forward and directed backward. Furthermore, there are differences in which parameters are taken into account in order to create the dependency between sentences. Some examples are: using word overlap, Parts-Of-Speech (POS), synonyms, rhetorical structures, parsing, linking words, anaphorical references and linking words.

2.3.1 PageRank

PageRank was designed as a browsing aid in order to rank by importance every web page of the Internet (Brin and Page, 1998) based on the concept of recommendation.

It computes the importance of every web page on the Internet by finding both its forward links (or links from one web page that direct to another web page) and backward links (or links from other web page that direct to a web page) and integrates them into a graph.

If we think of a graph, every web page would become a node in the graph and every forward and backward link from that web page would be the edges between nodes. In other words, PageRank takes into consideration which pages that recommend a web page and which pages a web page recommends. Those web pages that become more recommendable because they are recommended by other pages, gather strength and therefore their recommendations are stronger.

Regarding all the relationships between nodes and edges from the entire graph, PageRank produces a relative rank (as a score) of every node, which shows the importance of every node in the graph. Eventually, web pages with a higher rank will be selected as more important.

PageRank has not only been used to find and analyze the most important web pages on the web and its link-structure, but it has also been applied to e.g. citation analysis, social networks, automatic summarization.

PageRank is based on the concept of a random walk model on a graph. At the first iteration there is a uniform distribution of probabilities where all nodes have the same value: 1. After that, PageRank is computed for every node until convergence. Then, we have a stationary distribution of probabilities, which shows the probability of finding the random surfer at a certain node at any time. The values that nodes have at the moment of convergence, is the value that is used to rank the nodes.

Let $PR(V_i)_{k+1}$ be the vertex (or node) being calculated and V_j the node whose outgoing edge (or one of its outgoing edges) is pointing to V_i. Then PageRank's formula is the following:

$$PR(V_i)_{t+1} = \sum_{V_j \in In(V_i)} \frac{PR(V_i)_y}{|OutV_j|}$$

In this formula, we can see how PageRank takes into consideration all incoming and outgoing edges a node has iteratively for every node.

In addition, PageRank adds a damping factor which is a probability, normally set at 0.85, of jumping to any random page of the Internet, so that if there is a small loop of pages, there is a possibility of escaping the loop when calculating the PageRank.

PageRank's formula, where d is the damping factor, is the following:

$$PR(V_i)_{t+1} = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{PR(V_i)_t}{|OutV_j|}$$

2.3.2 Weighted PageRank

Originally, the edges of PageRank were not weighted because it is impossible to find *half* a link from one web page to another. However, PageRank applications are further away from only being used to rank web pages.

In other applications, such as automatic summarization, some connections are stronger than others. In that case, we can add a weight to some edges of the graph in order to create stronger connections.

TextRank (Mihalcea, 2004) (@@see section @@@@related work) is a variation of PageRank for automatic summarization where edges usually have weights. Following Mihalcea (2004 and 2005), we are going to use a weighted PageRank-based system instead of a regular PageRank-based system.

PageRank's weigthed formula, where the weight of every node is taken into account, is the following:

$$PR^{W}(V_i)_{t+1} = (1-d) + d * \sum_{V_j \in In(V_i)} w_{ji} \frac{PR(V_i)_t}{\sum_{V_k \in Out(V_j)} w_{kj}}$$

In this formula, the weights of the incoming nodes are taken into account. Additionally, if the incoming nodes have other connections with other nodes, its weights are also taken into account. In this way, the strength of nodes sending a lot of edges is weaker than nodes that only send one outgoing edge. @@@@@@revisar esto

2.3.3 Personalized PageRank

A Personalized PageRank varies from a regular PageRank by the fact that it beforehand gives more importance to some nodes than to others. This difference is materialized in a non-uniform distribution of the nodes at a starting point, which leads us to a modified final distribution of the graph.

Regular and weighted PageRank algorithms assume that $PR^{W}(V_i)_{t=0} = \frac{1}{|V|}$, where *t=0* is the state of the nodes before PageRank starts computing its values. However, this value can be adjusted so that there is no longer a uniform distribution of values. Personalized PageRank assigns more importance to some nodes before starting iterating. After the first iteration, PageRank is computed iteratively like a regular or weighted PageRank for each node. The implementation of Personalized PageRank that we use[6] implements a *personalization vector* to set the importance of certain sentences at the iteration 0.

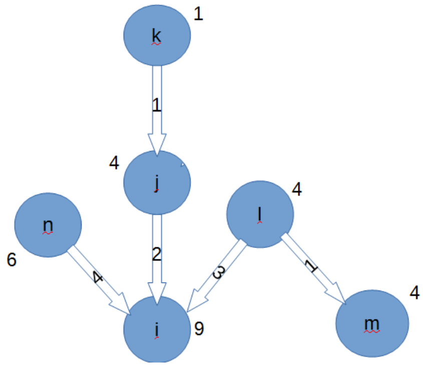

Figure 1: Example of a directed graph showing some nodes with their values outside each node before iterating (*t=0*) and their respective weights, inside the white arrows.

In figure @@@@ we show an example of a directed graph with relations between nodes and their weights and their values before starting iterating. If we apply the weighted PageRank formula, the result for node i would be:

$$PR^{W}(V_i)_{t+1} = (1-0.85) + 0.85 * \left((4+2)4\left(\frac{6}{4}\right) + 2\left(\frac{4}{2}\right) + 3\left(\frac{4}{3+1}\right) \right)$$

$$PR^{W}(V_i)_{t+1} = 71.55$$

[6] documentation can be found at https://networkx.github.io/.

2.4 Evaluation method: ROUGE

ROUGE (Lin, 2004) is an evaluation method for evaluating automatic text summaries. It determines the quality of system-generated summaries by comparing them to one or more model summaries written by humans.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) was created to facilitate the evaluation of summarization tasks and judge how good a system-generated summary is. ROUGE was inspired by BLEU (Papineni et al., 2002), used to evaluate machine translations.

ROUGE compares system-generated summaries by using word count of different sets of literal n-grams (i.e. unigrams, bigrams, etc.) and their overlapping with human-generated summaries. ROUGE has been tested in 17 different parametrization (Lin, 2004), which use different lengths for word n-grams and ”skip-grams”.

ROUGE does not difference between low-informative n-grams such as two stop words one following the other (i.e. ”of the”) and highly informative n-grams (such as proper nouns). ROUGE does not take into account overlap between units such as synonyms (i.e. fair-just), related words (i.e. related-relation), hypernyms (i.e. ”flower” instead of ”rose”) or anaphorical references (i.e. ”The president”-”He”)[7].

ROUGE has shown a good correlation between system-generated summaries and human judgements by comparing computer-generated scores given by ROUGE with human assigned scores[8]. In short, the ROUGE package is effectively used in automatic evaluation of single-document summaries. Because of this, we use this evaluation method to estimate how good our system-generated summaries are.

Different ROUGE variants are:

1. **ROUGE-N**

   ROUGE-N is a method to compare n-grams between system-generated summaries and model summaries (Lin, 2004). It has been tested with n = 1 to 9, where the number of n-grams must co-occur in the same order in the system-generated summary and the model summary (or summaries). Both Lin (2004) and (Lin and Hovy, 2003) have shown that automatic evaluation using n = 1, i.e. ROUGE-1, is well correlated with human evaluations based on various statistical metrics, especially in very short summaries (Lin, 2004).

   Example:

Model:	[A, B, C, D, E]
System-1:	[F, G, H, I, J, K, B, C, D, E]
System-2:	[L, M, F, N, O, P, Q, R, S, T, E]

   In this example, ROUGE-1 finds four words matching between Model and System-1 and one word matching System-2 with Model. ROUGE-2, ROUGE-3 and ROUGE-4 do not match any word between Model and System-2, but it finds three, two and one, respectively, between Model and System-1.

2. **ROUGE-L**

   ROUGE-L stands for Longest Common Subsequence (LCS). ROUGE-L assumes that the higher similarity between a system-generated summary and a model summary,

[7] However, stemmed and stopped versions of the summaries can be created and evaluated by ROUGE.

[8] The evaluation data used was gathered from the Document Understanding Conferences (DUC) in 2001, 2002 and 2003 and it contained single-document summaries of about 100 words, single-document very short summaries of about 10 words, multi-document summaries of about 10 words, 50 words, 100 words, 200 words and 400 words.

the more similar (and good) the system-generated summary is. In other words, the ROUGE-L perfect score corresponds to two sentences that have exactly the same words in the same order.

Since words must appear in the same word order (although they might have other words interspersed), ROUGE-L takes into account sentence level structure and gives lower scores to sentences with the same words in different positions of the sentence. The idea is that the summaries to evaluate might not contain the same meaning as in the model summaries.

The advantage of using ROUGE-L instead of ROUGE-N is that ROUGE-L does not need a specific n-gram length since it automatically includes the longest in-sequence common n-grams. However, as a disadvantage ROUGE-L can only reflect the longest subsequence, so if there is more than one subsequence, only the first subsequence is considered (Lin, 2004).

Example:

Model:	[A, B, C, D]
System-1:	[A, E, C, D]
System-2:	[C, D, E, A]
System-3:	[C, D, A, B]

In this example, System-1 and System-2 have the same ROUGE-2 score. However, System-1 has a ROUGE-L score of 3/4 = 0.75, while System-2 has a ROUGE-L score of 2/4 = 0.5, since in System-1 there are three units that appear in the same order as in Model, but in System-2 there are only two units in the same order as in Model. In System-3 we can see that there are two chains of words that appear in the same order as in the model, but only the first subsequence is taken into account to calculate the ROUGE-L score of the sentence.

3. **ROUGE-W**

   ROUGE-W works as ROUGE-L but the difference is that it weights consecutive in-sequence matches differently. ROUGE-W remembers the length of consecutive word matches encountered. The purpose is to differentiate an LCS, which appears in the same word order as in the model summaries, from an LCS that appear interspersed.

   Example:

Model:	[A, B, C, D, E, F, G]
System-1:	[A, B, C, D, H, I, J]
System-2:	[A, H, B, K, C, I, D]

In this example, both System-1 and System-2 have the same ROUGE-L score. However, System-1 has a higher score than System-2 using ROUGE-W since the units that match between System-1 and the Model appear consecutively in both Model and System-1 but not in System-2.

4. **ROUGE-S**

   ROUGE-S means Skip-Bigram Co-Occurrence Statistics. It indicates that any pair of words that appears both in the system-generated summary and in the model summary in the same sentence order is considered, even when there are words interspersed between them.

Although this system does not require consecutive words that match, it is still sensitive to word order (Lin, 2004). Nevertheless, applying Skip-Bigrams in a loose manner is risky since matches such as ”the the” or ”a of” may be considered to be good matches. In order to avoid this, skip distance between two words can be limited. In Lin (2004), ROUGE-S was tested with a maximum skip distance of 1, 4 and 9.

Example:

Model:	[A, B, C, D]
System-1:	[$\underline{\text{A}}$, E, $\underline{\text{C, D}}$]
System-2:	[$\underline{\text{C, D}}$, E, A]
System-3:	[$\underline{\text{C, D}}$, $\underline{\text{A, B}}$]

In this example, System-1 has three Skip-Bigram matches with Model (”A-C”, ”A-D”, ”C-D”); System-2 has one Skip-Bigram match with Model (”C-D”); and System-3 has two Skip-Bigram matches with Model (”C-D” and ”A-B”). Therefore, the ROUGE-S score for System-1 is the best of the three and System-2 receives the lowest score.

- **ROUGE-SU**

  ROUGE-SU is an extended version of ROUGE-S that takes into account the addition of unigrams as counting unit (Lin, 2004). ROUGE-S does not recognize units if there is not a bigram that co-occurs in the same sentence order both in the system-generated summaries and in the model summary.

  However, ROUGE-SU makes a distinction between a summary where the same words appear in a different order than in the model summary and a summary that does not contain any of the words from the model summary.

  Considering a System-4: [D, C, B, A] (the exact same words as in Model), ROUGE-S would overlook the similarity between Model and System-4. In any case, System-4 is more similar to Model than a sentence that is not related at all (i.e. System-5: [E, F, G, H]).

2.4.1 Determining optimal ROUGE variants

We have seen that different ROUGE parameters give different scores for the same sentences. Many studies have used different parameters. In order to compare our results with the results of other studies, we apply some of the most popular parameters, namely ROUGE-1, ROUGE-2 and ROUGE-SU4.

ROUGE-1 has been used by Kai Hong et al. (2014), Liang Zhou et al. (2006), Milhalcea (2004, 2005), Milhalcea and Tarau (2005), Kikuchi (2014) and Ceylan et al. (2010) among others. According to Lin (2004), ROUGE-1 functions well for very short summaries, so it will be a good parameter to use especially for measuring argumentative and expository texts, that are also very short. ROUGE-1 has as well been characterised as having the highest correlation with human judgements (Milhalcea, 2005).

ROUGE-2 has been applied in many studies such as Almeida et al. (2014@@@@), Ceylan et al. (2010), as well as Liang Zhou et al. (2006) and was recommended by Hong et al. (2014). Lin (2004) argues that ROUGE-2 is the best parametrization for 100-word summaries by Lin (2004). Therefore, we will use ROUGE-2 specifically for evaluating descriptive and narrative as well as scientific papers@@@@@.

ROUGE-SU4 is another parametrization used extensively. Examples are Almeida et al. (2014) and Ceylan et al. (2010). ROUGE-SU4 is furthermore recommended by Zhou et al. (2006). It is not as popular as ROUGE-1 and ROUGE-2, but it was described as a good parametrization by Lin (2004) in order to analyze how good a short summary is. ROUGE-SU4 has likewise been good for evaluating summaries of 100 words.

In this study we solely use recall scores and not precision scores or the F-score. We do this due to the fact that it was only recall scores that were presented in most studies and the Document Understanding Conferences (DUC) within the ROUGE literature.

2.4.2 Measure of accuracy: F-score

ROUGE calculates the precision and recall measurements of relevance as well as the F-score, which is the harmonic mean of precision and recall. However, most research conducted in the area of automatic summarization only takes the recall scores as their measure of accuracy. We compare ROUGE-generated recall scores in every genre and language.

In the automatic summarization context, the recall score is the fraction of relevant words of a text that are retrieved. In other words, the higher the recall score, the more relevant summaries we have generated. For example, if we have a 25-words system-generated summary (compared with a 25-words model summary) where five words are relevant, our recall score is 5/25 or 20%. If we have a 50-words system-generated summary (compared with two 50-words model summaries) and we successfully retrieve 30 words, our recall score is 30/50 or 60%.

We must notice, as Mihalcea (2004) pointed out, that our maximum recall scores are always going to be below 100% when doing extractive summarization if summaries are generated not only taking into account the sentences from the original document.

2.5 Hypothesis

After having analyzed the theoretical framework of automatic text summarization and text structure and different genres, we can formulate our research question as the following hypothesis:

> Incorporating the variation of text structure of a texts genre can improve the quality of its automatic summarization.

More specifically, we claim that we can incorporate document-structure factors in a graph-based method. Building the graph, which represents a document, and adding the variation in relevance of the different sections for different genres, we can locate and fetch the most important information of a text and leave the least important information.

Figure 1 shows the idealized relative relevance curves for different parts of a text in each genre.

Analyzing text types can help us understand the structure of a text and retrieve the relevant information since the location of information in a text is governed by its genre. Because of that, it is important to be able to locate and extract that information.

There are methods to incorporate genre-informed document structures in an unsupervised summarization system. We apply a Personalized PageRank-based method in order to accomplish this task.

Ideally, each genre will lend itself better to a certain graph representation, where the graph incorporates both sentence-to-sentence relations and sentence-to-document relevance.

Specifically, important information normally tend to be placed at the beginning of a text, but relatively relevant information is placed in different parts of a text according to the text genre. We analyze the relative relevance of different parts of texts more in detail in section 5.4@@@@@@.

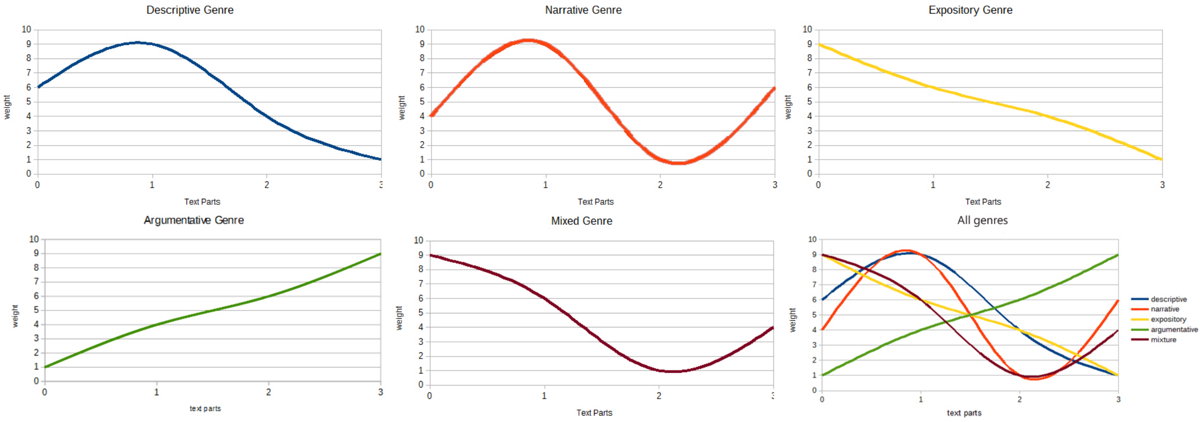

Figure 2: Graphs showing the relative relevance of different text parts. 0 stands for the first sentence of the document (or the title) and 1, 2 and 3 are the first (introduction), second (middle) and third parts (end) of the document.

2.6 State-of-the-Art

This section explains the related work that has been conducted on Automatic Summarization, one of the major tasks in the field of NLP.

2.7 Automatic Summarization

A great amount of research has been done in the field of automatic summarization. In the Document Understanding Conferences (DUC)[9], we can find different approaches and many different algorithms for this task.

Some graph-based methods applied to generate automatic summaries are TextRank (Mihalcea, 2004) and LexRank (Erkan, 2004). Nowadays, there are even online tools[10] that generate automatic summaries. Besides, Microsoft Word also has a summarizer for documents.

2.7.1 TextRank

TextRank is a widely applied unsupervised automatic summarization method (Mihalcea, 2004). TextRank is a graph-based ranking algorithm designed for summarizing texts. Following the same conditions that PageRank uses in order to select the most important web pages (i.e. the random walk model), TextRank selects the most important sentences in a text.

It is assumed that sentences that are recommended by other sentences are more likely to be more important for a text. So, the more connections and the more important connections a sentence has, the higher rank it gets and the more important it is to the text.

TextRank creates a graph, in which the structure of the text is encoded. Thus, TextRank computes the importance of every sentence (node) by finding relationships between sentences (edges) and integrates them into a graph.

As PageRank, when the system converges in a stationary distribution of probabilities, we obtain values of probabilities for each node. We rank those values in order to create a ranking of sentences. In a final instance, sentences with higher ranks will be selected for the final summary.

[9] http://duc.nist.gov/pubs.html

[10] such as AutoSummarizer in http://autosummarizer.com/index.php

TextRank is based on the global information of the structure of a text and not on any labeled training data. Because it is an unsupervised method, TextRank can be used in any type of text regardless of topic, genre, language or intrinsic characteristics. The output is simply produced taking into consideration the internal connections (sentences that recommend other sentences) in every text.

Since sentences (and relations between sentences) provide more complex information than links in web pages [11], the TextRank algorithm usually weights some relations between sentences in order to give more strength to some of the sentences.

There are methods to weight particular edges of the graph depending on the characteristics of the sentences. The results of TextRank algorithms have shown to be different from unweighted graphs, since some sentences obtain a higher rank than if they were not weighted.

There are many ways of creating relations between sentences: some concepts such as word overlap, similar concepts or synonyms can be taken into account.

Finally, while PageRank often utilizes directed graphs because it aims to capture the temporal dimension of click-through behaviour, TextRank normally uses undirected graphs. The main reason why undirected graphs are preferred in automatic text summarization have to do with the fact that a priori we do not know whether relationships between sentences are anaphorical or cataphorical or it is not possible to detect it.

Previous work using TextRank

Mihalcea (2004) already showed that TextRank is a competitive state-of-the-art summarization method by evaluating a corpus of 567 news articles and their corresponding 100-word summaries. Mihalcea (2004) claimed that a weigthed TextRank was a well-functioning unsupervised summarization system because it can identify important sentences, which humans also found important, because of the concept of recommendation.

Additionally, Mihalcea (2004) proposed that TextRank could be used both in short text and long documents due to the fact that it ranks over all sentences of a text. Therefore, it is only necessary to pick the desired number of words in order to create a summary.

In 2005, Mihalcea investigated the differences between different types of graphs in TextRank: 1) undirected and 2) directed forward and 3) directed backward. Their results show, (Figure 2@@), that backward tends to perform better than the other systems because of the data set used: news articles.

Mihalcea (2005) concluded that because of the characteristics of this type of texts, the most important information was found at the beginning of the documents, and, therefore, a backward direction of the graph was favoured. Furthermore, after analyzing both an English and a Portuguese corpus, Mihalcea suggested that TextRank could be used in any text, independently of the language.

Mihalcea and Ceylan (2007) analyzed the domain of books more in depth. Although they refer to their texts as "books", we deduct that they are focusing on long novels[12]. Their new corpus was compounded by long documents and it was thus very different from the previously used news data set.

They found that the information that can be extracted by the sentence position was not helpful when handling long documents, such as books, because of the great number of topic shifts that can take place in a long text. Specifically, their ROUGE-1 recall score when they

[11] A priori all relationships are equally important.

[12] Understanding the word with the modern meaning, as we can find in the Cambridge Dictionary: "a long printed story about imaginary characters and events".

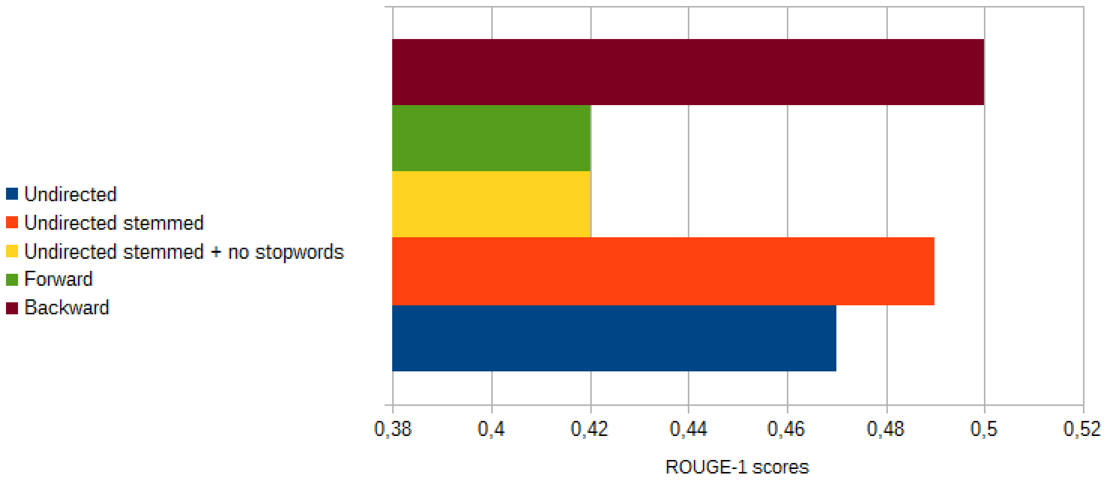

Figure 3: Mihalcea (2004 and 2005) ROUGE-1 scores when applying different directed and undirected graphs in a news data set.

take into account sentence position is 0.32. When not considering sentence position as a feature the recall score is 0.33.

Mihalcea and Ceylan (2007) realized that when checking individual results, the quality of the summaries of longer texts were improving much more than shorter texts: +0.07 for the longest texts against 0.01 for the shortest texts compared to the baseline. Additionally, they found cases where their texts did not surpass their baseline.

They also claim that using more than one model summary does not influence ROUGE scores.

Another important study for our thesis was conducted by Ceylan et al. (2010). They looked at the possible search space of extractive system-generated summaries across four different domains.

Analyzing all the possible combinations to create an extractive system-generated summary, they found that TextRank was always very close to the optimal possible results in their search space, except for the literary domain, where TextRank did not perform as well as in the other domains.

On the basis of these results, Ceylan et al. (2010) conclude that extractive summarization cannot to a large extent improve the task of automatic text summarization. Therefore, they suggested that other techniques should be taken into consideration.

Ceylan et al. (2010) discuss the evaluation conducted with ROUGE as well. They pointed out that ROUGE does not take into account cohesion or coherence in system-generated summaries, so other summary quality metrics should be considered in the future in order to obtain a better evaluation.

Finally, they realized that compression ratios in model summaries influenced the results in automatic summarization. As a result, they suggested that further research on different compression ratios should be carried out.

Figure 3@@ shows that there are differences in the ROUGE-1 scores of Ceylan et al. (2010) when analyzing different domains. Nonetheless, we cannot compare these differences directly since the data sets used were very different.

However, we can see that the scores of system-generated summaries of legal text are much higher than the scores of the other domains.

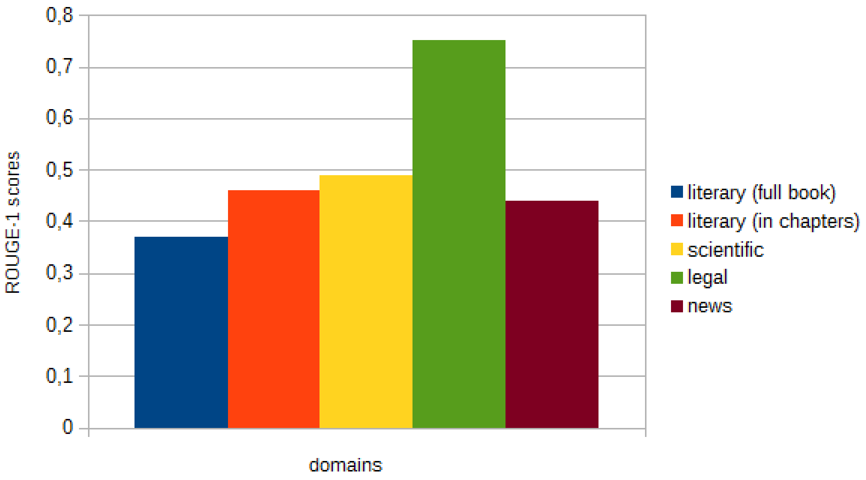

Figure 4: Ceylan et al. (2010) ROUGE-1 scores in different domains

If we compare the news data set from figure 2@@, compounded by 567 news articles, with the news data set from Figure 3, compounded by 50 news articles, we can see that the results vary. So, even when analyzing texts in the same domain, we cannot compare results directly.

2.8 Data sets for summarization

The DUC (2002) data set[13] has been used in many studies such as Mihalcea (2004 and 2005) for single and multiple automatic summarization in English. In particular, news articles and their corresponding 100-word model summaries are used.

Mihalcea (2005) added a Portuguese Corpus TeMrio (Pardo and Rino, 2003), which consists of 100 news articles in Brazilian Portuguese and their human-generated model summaries. These model summaries do not contain 100 words, but they were made respecting the same compression ratio for all documents.

Mihalcea and Ceylan (2007) generated a book data set containing 50 books from English literature. They gathered their model summaries on-line from web pages such as Grade Saver [14] and Cliff's notes [15].

They picked texts that had a full on-line version of the book and two human-generated summaries available. Their documents and their summaries gathered on-line varied a lot in length and compression ratio.

The Ceylan et al. (2010) data set might be the corpus that resembles mostly our corpus (GICorpus), since it is divided into different domains: namely newswire, literary, scientific and legal. Although these domains do not directly correspond with the genres used in our thesis, we follow the approach of Ceylan et al. (2010).

The data set of Ceylan et al. (2010) contains 50 documents per domain[16] and one model summary per domain, except for the newswire domain, where they collected two model summaries.

Ceylan et al. (2010) mention that model summaries are very different in writing style and length of the sentences. In their study, they also realized that it was impossible to handle

[13] retrieved from http://www-nlpir.nist.gov/projects/duc/data.html.

[14] http://www.gradesaver.com/

[15] http://www.clifssnotes.com/

[16] 50 news articles from the DUC 2002 data set, five chapters of 10 novels that are literature classics, 50 scientific papers and 50 law documents.

the huge graphs that were generated from long text documents. Therefore, they decided to split the texts into sections and compare those sections individually.

2.9 Summary

We have gone through the theoretical framework of Automatic Summarization and the related work that has been conducted in this field. After having read all this literature, we have concluded that we will focus on single-document extractive unsupervised graph-based summarization, using the weighted PageRank algorithm as most of the studies we have presented did. We will use ROUGE as our evaluation method since it is the most used metric to evaluate summaries.

We consider the background in text structure and different text types (or genres) and the idealized curves of our hypothesis as a starting point to go in depth into a Personalized PageRank-based system in both directed and undirected graphs. We take especially into account the sentence position and the relative relevance of sentences in our texts.

Our data set consists of 50 documents per genre like Ceylan et al. (2010) in five different genres: descriptive, narrative, expository, argumentative and mixed. As Mihalcea and Ceylan (2007) and Ceylan et al. (2010) did, we collect documents and model summaries from already existing on-line ressources. Additionally, 25, 50 and 100-word summaries are created so that we can compare our results with the studies mentioned before.

3 Data: the GICorpus

In this section, we describe our newly gathered data set: the GICorpus, created for the purpose of analyzing and comparing system-generated summaries belonging to different genres. This section is based on the work presented for the LREC 2016 conference.

Three baseline systems (BF, BL and BR) and PageRank-based systems (PF, PB and PU) were created. Three different summaries for each document to check the relative relevance of each section of these documents were also created (the title of the texts, the remaining part of the text after removing the title and the two last thirds of the document). Finally, Personalized PageRank-based system (PPF, PPB and PPU) summaries were produced with a personalization vector formed by heuristic values that we deduced from the previously created summaries. We can see this number of system-generated summaries in Table 17 @@ (see section @@).

In Table 2@@, we can see the distribution of the documents in the GICorpus.

Type of document	Number of documents
full documents	450
model summaries	450
baseline summaries	450 x 3
PageRank summaries	450 x 3
Relative Relevance summaries	450 x 3
Personalized PageRank summaries	450 x 3

Table 2: Table showing the number of documents and summaries generated for each document in the GICorpus.

In total, the GICorpus contains 900 documents and 5,400 system-generated summaries. Both the GICorpus and the system-generated summaries are publicly available for academic use and can be used especially for genre-informed research purposes.

3.1 Description of the documents

The data used in this thesis is the Genre-Informed single-document summarization Corpus (GIC) for written Danish, English and Spanish.

The GICorpus contains 450 documents and their corresponding 450 model summaries. These are divided into five *genres*. It contains 30 documents in all three languages for each genre. We chose text documents that maximized comparability regarding topics and length of the documents and model summaries in all three languages.

Even though NLP works are often domain-aware, we advocate for genre-aware summarization. *Genre* is understood in this corpus as the language and structure that is utilized in a specific type of text, e.g. news articles or scientific papers, which are likely to have a certain structure, regardless of language or geographical or historical context. These genres correspond approximately to different types of text or ”superstructure” (Van Dijk, 1980).

The genres that forms the GICorpus of this thesis are the following: *descriptive*, *narrative*, *expository*, *argumentative* and *mixed*.

In Table 1@@@, we can see how the average length varies from genre to genre and from language to language. We can see that argumentative texts in Danish are much longer than argumentative texts in English and Spanish. It is also visible that Spanish texts in the mixed genre are much shorter than mixed texts in Danish or English.

3.1.1 Descriptive genre: Wikipedia articles

This genre is constituted by five articles for six different topics: animals, capitals of Europe, diseases, elements of the periodic table, everyday objects and members of royal families in

Genre	Danish	English	Spanish
Descriptive	2,433	5,522	6,315
Narrative	3,108	4,741	4,328
Expository	631	715	828
Argumentative	337	115	141
Mixed	7,415	7,660	3,120

Table 3: Average number of words in each genre and language.

European countries.

For each of these topics, we have selected five articles. These articles are correlative articles in Danish, English and Spanish. We have chosen articles that maximize the size of their shortest variant (namely the Danish versions).

3.1.2 Narrative genre: Novels and tales

Our narrative genre is constituted by a collection of five chapters[17] of four classic novels and ten of Hans Christian Andersentype fairytales[18].

The novels chosen fall in two categories: 1) two novels originally published in English and their corresponding translations into Spanish; and, 2) two novels originally published in Spanish and their matching translations into English.

3.1.3 Expository genre: News articles

This genre is represented by a collection of 30 news articles in Danish[19], English[20] and Spanish[21] published from March through July 2015.

The topics are many and various and include politics, migration, terrorism, food and the weather.

3.1.4 Argumentative genre: Letters to the editor

In this thesis, this genre is constituted by 30 letters to the editor in Danish[19], English[22] and Spanish[21] in July 2015.

The topics discussed in these letters are very different. The main topics of this genre are politics, the economy and culture & education.

3.1.5 Mixed genre: Scientific papers

Scientific papers correspond to a mixture of text types. @@@@@hector dice que expand@@@@ Different parts of these text have different functions (i.e. to describe, to argument).

This genre contains 30 different articles in each language. The main topics are in the social sciences (politics and sociology) and medicine.

[17] Chapters are chosen in order to correspond to different parts of the book (the beginning, three chapters spread out in the middle chapters of each novel and the end of the book.

[18] It was not possible to find public human-generated summaries of the novels in Danish. In order to reach 30 articles in this genre in Danish, we collected 20 additional Hans Christian Andersen's fairytales.

[19] from jyllands-posten.dk

[20] from metro.co.uk

[21] from elpais.es

[22] from theglobeandmail.com

3.2 Description of the Model summaries

We also provide a collection of human-generated model summaries. There is one model summary per document[23] either gathered from on-line sources or written by a human annotator when on-line resources were not available.

Genre	Danish	English	Spanish	on-line
Descriptive	207	425	377	100%
Narrative	379	255	198	100%
Expository	66	94	58	22%
Argumentative	41	52	66	0%
Mixed	189	218	117	100%

Table 4: Average number of words in model summaries for each genre and language and the percentage of the data that was fetched on-line.

In Table 2, we can see that the length of the model summaries is very different in some genres and languages. Individual human-generated summaries are inevitably very different in word length, writing style and content. This is especially the case in the narrative genre. We can also see the percentage of on-line fetched human-generated summaries.

In Table 3, we show the average compression ratio for each genre as expressed in Ceylan et al. (2010). We can see that there is discrepancy between genres and among languages inside every genre.

Genre	Danish	English	Spanish
Descriptive	92%	92%	94%
Narrative	94%	92%	94%
Expository	91%	91%	89%
Argumentative	80%	64%	63%
Mixed	98%	98%	94%

Table 5: Average Compression Ratio Original Document-Model Summary for each genre.

As a result of this way of collecting human-generated summaries, model summaries are very different in terms of length, compression ratio, style, etc., especially in the descriptive type of text, narrative type of text and the mixture of text types. However, human summary generation is very time consuming. Therefore, finding these resources on-line, albeit less accurate when evaluating, should help to ease this resource-intensive task.

In order to solve the evaluation problem, we have also generated 24 extra model summaries by randomizing the sentences in every model summary. This is done in order not only to consider the first sentences of the summaries in the evaluation, but more variations of the same summaries.

This problem has been addresed in other studies such as Ceylan et al. (2010) and Mihalcea et al. (2007). They solved the problem by keeping one model summary as the main reference and using an extra model summary as a way to decide on the length of the system-generated summaries (Mihalcea et al., 2007). Their system-generated summaries have the same length and are evaluated against each other (possibly with a different length). [24].

[23] Mihalcea and Ceylan (2007) show that the use of more than one model summary seems not to influence the ROUGE scores.

[24] An alternative solution proposed by Mihalcea (2007) was to determine the length of the system-generated summaries using a predefined compression rate (e.g, 10%) for all the summaries. Nonetheless, there would still be variations across the system-generated summaries and model summaries regarding length. In a final instance, it would be difficult to interpret variations across the ROUGE scores.

3.2.1 Descriptive genre: Wikipedia articles

We apply the introduction section of Wikipedia articles (which appear in the beginning of every Wikipedia article) as model summaries for the entire original Wikipedia article.

3.2.2 Narrative genre: Novels and tales

Model summaries of novels and tales were gathered from different on-line websites specialized in book summaries[25]. When the required summaries were not available, we searched for summaries in other websites from individuals, reviews or plot sections in Wikipedia articles.

3.2.3 Expository genre: News articles

It was not always possible to find model summaries of news articles in all three languages on-line. However, sometimes we can find a short summarizing paragraph before the body of news articles. When long enough, those paragraphs were used as a model summary.

When that information was not available, an annotator wrote a model summary in Danish, English and Spanish. The annotators were instructed to write about 40-50 words (in at least two sentences) summaries.

3.2.4 Argumentative genre: Letters to the editor

Model summaries for letters to the editor were not available, but it was the same annotators that wrote the Danish, English and Spanish model summaries, respectively. The annotators were instructed to write about 40-50 words long summaries (in at least two sentences) and not to copy verbatim from the original documents.

3.2.5 Mixed genre: Scientific papers

The abstracts of the academic articles were used as model summaries in this genre. Because of author preferences when writing abstracts, the model summaries of this genre are again very different in length and content, e.g. some abstracts seem more like an introduction to the study while others focus on the results and conclusions of their papers.

[25] such as http://www.sparknotes.com/ (in English), http://www.monografias.com/ (in Spanish) and http://www.litteratursiden.dk/ (in Danish).

4 Experimental Setup

In this section, we explain how our baselines and weighted PageRank-based system are obtained. We also explain how we complement our basic baseline and PageRank-based method with other different data handling methods in order to find the relative relevance of certain parts of our texts: namely the title and the remaining parts of the document; and the relative importance of the middle and lasts parts. Finally, we explain how we perform our evaluation with ROUGE (see section @@@@).

The nomenclature for our systems used in the all the tables of this thesis is the following:

BF:	**B**aseline, **F**irst *n* sentences of a document
BL:	**B**aseline, **L**ast *n* sentences of a document
BR:	**B**aseline, **R**andom *n* sentences of a document
PF:	weighted **P**ageRank directed **F**orward
PB:	weighted **P**ageRank directed **B**ackward
PU:	weighted **P**ageRank **U**ndirected
PPF:	**P**ersonalized **P**ageRank directed **F**orward
PPB:	**P**ersonalized **P**ageRank directed **B**ackward
PPU:	**P**ersonalized **P**ageRank **U**ndirected

4.1 Baselines

Table 6@@ shows three different baselines that require no learning algorithm. This table was initially mentioned in section @@Motivation.

	Danish			English			Spanish		
Genre	BF	BL	BR	BF	BL	BR	BF	BL	BR
Descriptive	**.22**	.18	**.22**	**.21**	.15	.19	**.29**	.20	.25
Narrative	**.23**	.20	.18	**.26**	.23	.21	**.30**	.26	.25
Expository	**.53**	.24	.31	**.58**	.20	.26	**.29**	.20	.25
Argumentative	**.48**	.32	.35	**.46**	.43	.43	**.47**	**.47**	.46
Mixed	**.30**	.21	.19	**.30**	.24	.22	**.38**	.28	.27

Table 6: Average ROUGE-1 50-word summary recall scores for every genre in Danish, English and Spanish.

The baselines BF and BR have been commonly used in automatic summarization. We add another baseline (BL) to check the differences across genres focusing on their final sentences.

We can see the results for these baselines in Table 1@@ for Danish, English and Spanish, respectively. Out of these three baselines, BF is the system that generally works best. We can also see that BR performs better than BL most of the time and that the results from all BF, BL and BR are not so far away from each other.

4.2 Weighted PageRank

We apply a weighted PageRank-based system to examine differences across genres. We generate connections between sentences by checking the N-gram overlap between two sentences. If they share a word, then we create a link.

Following Ceylan et al. (2010) or Mihalcea (2004), we assign specific weights when the connection of two sentences is stronger. In this case, for every word two sentences share, we add 1 to the weight.

We remove stopwords[26] and punctuation. All words were lower-cased and, initially, there is no stemming. Therefore, only words that match exactly are used to create connections between sentences.

We test whether directionality in graphs affects the results. We examine 1) an undirected (PU) graph and two directed graphs: 2) forward (PF) and 3) backward (PB). The results for each genre and language appear in Table 5@@.

We can see that the scores for the PB system are usually higher than the scores of PF and PU systems. In addition, we can see that PF tends to be the worst of these three systems, except for the argumentative genre in Spanish, where it is the best system.

	Danish			English			Spanish		
Genre	PF	PB	PU	PF	PB	PU	PF	PB	PU
Descriptive	.20	**.23**	.22	.18	**.23**	.20	.25	**.29**	.25
Narrative	.19	**.20**	**.20**	.22	**.23**	**.23**	.27	**.29**	.27
Expository	.26	**.48**	.32	.22	**.59**	.35	.27	**.45**	.33
Argumentative	.34	**.44**	.37	.43	**.44**	.42	**.50**	.46	.48
Mixed	.22	**.26**	.22	.25	**.27**	.22	.30	**.34**	.26

Table 7: Average ROUGE-1 50-word summary recall scores for every genre in Danish, English and Spanish.

4.3 Estimation of the relative relevance of the sections in a text

In this thesis, we focus our attention on the different parts of a document in every genre and how relevant these parts are for the generated summary. In order to find out where the most important information is located in a text, we do not only check which baseline and weighted PageRank-based systems perform best. We also analyze the relative relevance of different parts of texts in different genres: namely the title, the introduction, the development and the end.

We have seen in our baseline and weighted PageRank-based systems that there were differences in the results of the different systems and genres. One example is in the argumentative genre, where both BL and BR, and PF and PU systems generally obtained higher results than in the other systems. Additionally, in many genres such as in the descriptive and narrative genres (and to a certain extent in the mixed genre), the results of the different systems were not so far away from each other. They were more differentiated in the expository genre, where the best working system performed much better than the other genres.

By analyzing how important the title of a document is and the importance of the different parts of the document, we can use this information in order to create a better summary.

The previous Baseline and weighted PageRank-based systems in English are shown in Table 8@@@. Additionally, the relative relevance of the different sections in a text has been added in each genre in order to compare the results in the English language.

4.3.1 Relevance of the title

We examine, on the one hand, the relevance of the title of the document and, on the other hand, the relevance of the text without the title of the document.

We compare the results of the title of the document (just one sentence) with the results of the last sentence of the text and a randomly chosen sentence of the text. If the results of the first sentence (or title) of the text are better than a random sentence or the last sentence of the text, we consider the title more relevant than other sentences in the document.

[26] We remove them by using the Stopwords Corpus (Porter et al. @@@@) from the NLTK package.

The results of the document without the title are compared with the results of the full document. If the results of the document without the title are better than with the title, we claim that the title is not as relevant as the other sentences of the document. If the results without the title are worse than the results of the full document, we say that the title is a very relevant part of the document.

The results are shown in Table 8@@@@ for the English language. We only observe the English language to have an idea of which system works best and extract heuristic values afterwards because it is the language where most research has been conducted. Results highlighted in bold letters show the system that works best for the full document.

If the results from the relative contribution of specific text parts (title, introduction, development or end) surpass the score for the full document, then this is also highlighted in bold and red. Scores in magenta mean that the relative parts of the text surpass the score of the system used in the full document but not the highest score achieved from the best system.

4.3.2 Relative relevance of each section

In addition, we test the relative relevance of the sentences in the three different sections. When analyzing our baseline and weighted PageRank-based systems, we have seen that the introduction is the most important part of the documents in all genres and languages.

Genre	length	BF	BL	BR	PF	PB	PU
Descriptive	full doc	.22	.18	.22	.20	**.23**	.22
	1 sent	.03	.11	.09	.15	.15	**.24**
	no title	**.27**	-	-	**.24**	**.28**	**.25**
	2/3	-	-	-	**.24**	**.25**	**.25**
Narrative	full doc	**.26**	.23	.21	.22	.23	.23
	1 sent	.08	.10	.08	.14	.13	.19
	no title	.23	-	-	**.22**	.23	.22
	2/3	-	-	-	**.22**	.22	**.23**
Expository	full doc	**.58**	.20	.26	.22	.47	.21
	1 sent	.21	.08	.11	.10	.20	.20
	no title	.49	-	-	**.22**	**.50**	**.33**
	2/3	-	-	-	**.23**	.25	**.25**
Argumentative	full doc	**.46**	.31	.43	.43	.44	.42
	1 sent	.06	.26	.26	.25	.25	.37
	no title	**.57**	-	-	**.55**	**.57**	**.54**
	2/3	-	-	-	**.54**	**.54**	**.55**
Mixed	full doc	**.30**	.24	.22	.19	.27	.22
	1 sent	.11	.15	.11	.17	.14	**.22**
	no title	.25	-	-	**.25**	.25	**.22**
	2/3	-	-	-	**.25**	.22	.21

Table 8: Average ROUGE-1 50-word summary recall scores for every genre in English. The ROUGE-scores for the full document are shown, only taking into consideration one sentence (the first sentence, the last sentence and a random sentence), the first 50 words without the title and, additionally, the ROUGE-scores of PageRank-based system for the last two thirds of the document.

After analyzing our baseline and weighted PageRank results, we test the relative relevance of the middle part of the documents and the last part of the documents. To be able to do that, we test the last two thirds of the documents. In other words, we remove the first part and the title of the documents (roughly the introduction) in order to compare the scores of the middle and last sections. We compare the scores of the middle and last sections with the scores of the weighted PageRank-based system in full documents and see how the results vary.

We expect the results of the two last third parts of the document to be worse than the results of the full document. However, if the results are higher, it means that the middle and last parts of the document were more important than the introduction. Nevertheless, weighted PageRank-based systems were not giving the most appropriate scores to the sentences in the middle and last parts of the document.

Moreover, if the last two thirds of the document are analyzed and the weighted PageRank-based system PB is performing better than PF, then the middle part is more important than the last part. Conversely, if the PageRank-based system PF is performing better than PB, then the last part is more important than the middle part.

After generating the corresponding summaries and having evaluated them, we check whether the results are better or worse than the results of the full document. Table 8 @@ shows these results for every genre in English. Further analysis of these results is discussed in section @@@@ 4.4.1.

4.4 Personalized PageRank

We have seen in Table 8@@@ that different sections in documents have different relative relevance. Additionally, this relative relevance depends on the genre the document belongs to. Therefore, we build a Personalized PageRank-based system in order to give a priori more relevance to some sentences than to others regarding their sentence position in the text: whether it is the title, the first section, the middle section or the last section.

We split as explained each document into four different parts: the title (the first sentence) and 2) three equally divided parts (roughly introduction, development and end)[27].

In the same way, we create a personalization vector in our Personalized PageRank-based system, which contains four values: one for each of the previously split sections. These vector values or "preferences" as Aktas et al. (@@@@) call them would be decided based on heuristic values extracted from our idealized curves (see @@ section hypothesis).

We give a value to the title, a value to the first k sentences, a value for the second k sentences and a value for the last k sentences. This splits each document of *1+3k* sentences in 4 sections. The values for the weight factor are between 0 and 10.

In Table 9@@@@ we can see a normalized weight factor of these numbers. We have decided to give the same sum for all personalization vectors (20), so that it is easier to see the relative relevance of each section in every genre and how it is distributed according to the most and least relevant parts of a text. The individual values selected are 1, 4, 6, and 9, i.e a possible personalization vector can be [1, 4, 6, 9] or [4, 6, 1, 9].

Following the values showed in Table 9 @@, we can see that the vector [1, 4, 6, 9] would be equivalent to the vector [.05, .20, .30, .45].

weight factor	0	**1**	2	3	**4**	5	**6**	7	8	**9**	10
normalized weight factor	0	**.05**	.10	.15	**.20**	.25	**.30**	.35	.40	**.45**	.50

Table 9: Correspondence between weight factor (between 0 and 10) and its normalized weight factor (between 0 and 1).

An example of the way we split the documents is the following: if we assign the personalization vector [1, 4, 6, 9] to a document of seven sentences, the first sentence would be the title and would have the initial value of 1, the next two sentences would be the introduction and would have the initial value of 4. From then on the next two sentences would be the

[27] When the remainder of the division was not zero, the last part would have one or two extra sentences, depending on the remainder.

development and would have the initial value of 6 and the last two sentences would be the end and have the initial value of 9. This correspondence between sentences and initial values is shown in Table 10 @@@.

	Title	Introduction		Development		End	
sentence number	0	1	2	3	4	5	6
initial value	1	4	4	6	6	9	9

Table 10: Initial values in a Personalized PageRank-based system for every sentence in a 7-sentence document with the personalization vector [1, 4, 6, 9]. The title is given the sentence number 0.

If a document belonging to the same genre as the document in the previous example had 11 sentences, then the title would be the first sentence with the initial value of 1, the introduction would be the next three sentences with the initial value of 4, the development would be the next three sentences with the initial value of 6 and, finally, the end would be the last four sentences with the initial value of 9. Table 11 @@@ shows the correspondence between initial values and sentences.

	Title	Introduction			Development			End			
sentence number	0	1	2	3	4	5	6	7	8	9	10
initial value	1	4	4	4	6	6	6	9	9	9	9

Table 11: Initial values in a Personalized PageRank-based system for every sentence in a 11-sentence document, with the personalization vector [1, 4, 6, 9]. The title is given the sentence number 0.

4.4.1 Determining optimal weighting function for each genre

The N combinations of the parameter space yield k effectively different personalization vectors i.e. probability distributions of node selection.

This approach is very parameter sensitive in the sense that this set of parameters have many possible combinations. We apply heuristic values that approximate the shape of the idealized curve (@@@@as we discussed in our motivation) of relevance per item in order to solve this issue.

In an unsupervised set-up we can apply the following method to estimate parameters: a brute-force approach (an exhaustive exploration of the parameter space) would have an estimate of 81.6 hours for the argumentative genre, which contains the shortest documents[28] in a university server.

However, we aim to estimate which parameters that would be ideal for one genre/language combination by analyzing our baseline and weighted PageRank-based system results we mentioned in section 1.1 (N.B. it is a small data set).

The results vary across languages. However, we take the English language results as our starting point to personalize our system due to the fact that this is the language where scientific research is often conducted. Moreover, the English language is standard practice in NLP.

The results also depend on the number of words that we evaluate with ROUGE. Nevertheless, the results tend to increase or decrease more or less consistently. We pay special attention to the 50-word summary scores. As the next step, we analyze whether the result scores increase or decrease, and whether the results change consistently in all three languages.

[28] There are 4,896 different distributions to check.

Descriptive genre

We propose a vector [6, 9, 4, 1] where the introduction is more important than the middle and last parts since PB is the system that works best. This is shown in Table 6@@.

The middle part is important as well. When analyzing the last two third parts of the document, we obtain better results in both PB and PU than in PF. Additionally, we can see that BR has as good results as BF in the full document.

Despite not finding the title so relevant, we do not consider it appropriate to give a very low value to the title. In this data set, the title consists of only one word, namely the object to be described, which should be relevant. We assume that the results of the title are poor due to the fact that PageRank–based systems tend to give more importance to long sentences.

Narrative genre

The vector to be created in this genre is [4, 9, 1, 6]. This reflects a situation in which the introduction is the most relevant part of the text. Table 6@ shows that BF is the system that works best. In addition, both PB and PU obtained better results than PF.

The middle part is the less important part of the text. This is clear when analyzing the last two thirds of the document. PB works worse in this contexts than with the full document, showing that these sentences are not so relevant.

The last part of the text is slightly more important than the middle part of the text because the last sentence was found more important than the title or a random sentence.

Expository genre

The suggested vector for this genre is [9, 6, 4, 1] where the beginning of these documents start being the most important part and its relevance declines gradually afterwards.

Table 6 shows that BF has by far the best results, even when comparing it with PB. Moreover, when removing the title, the results were worse, which means that it is an important part of the document.

We can see that PF has worse results than PB or PU and that the last line is even less important than a random sentence. Therefore, the last part of this genre has the lowest value in the personalization vector.

Argumentative genre

In Table 6@@ it is shown that removing the introduction and especially, when removing the title, we obtain much better results than when considering the full document. Therefore, we suggest the vector [1, 4, 6, 9], where the title has very little relevance, the introduction is less relevant and the last two thirds of the document are the most relevant.

We can see that scores for PF, PB and PU are more or less equal in this genre. Moreover, it is a bit contradictory that when taking into account the full document, BF has higher results than BL and BR and that BL is the system that performs the worst.

Mixed genre

We propose a vector [9, 6, 1, 4] where the introduction is an important part and the last part is more relevant than the middle part. Table 6@@ shows that we give relevance to the title because the results get worse when removing it. We give some importance to the introduction since BF and PB are the systems that work best in the full document.

Moreover, we decide to give more importance to the last part rather than to the middle part because the final sentence of the document has a higher score than a random sentence. Furthermore, when regarding only the last two thirds of the document, we can see that PF performs better than PB.

4.4.2 Realization of the hypothesis

We create some vectors from the heuristic values based on the idealized curves of the hypothesis and the results of our baselines, PageRank-based method and the relative relevance of each section.

In Figure 4@@, we can see the idealized curves introduced in our hypothesis overlap with the values of our personalized vector. @@@@expo y argu en la suma@@@

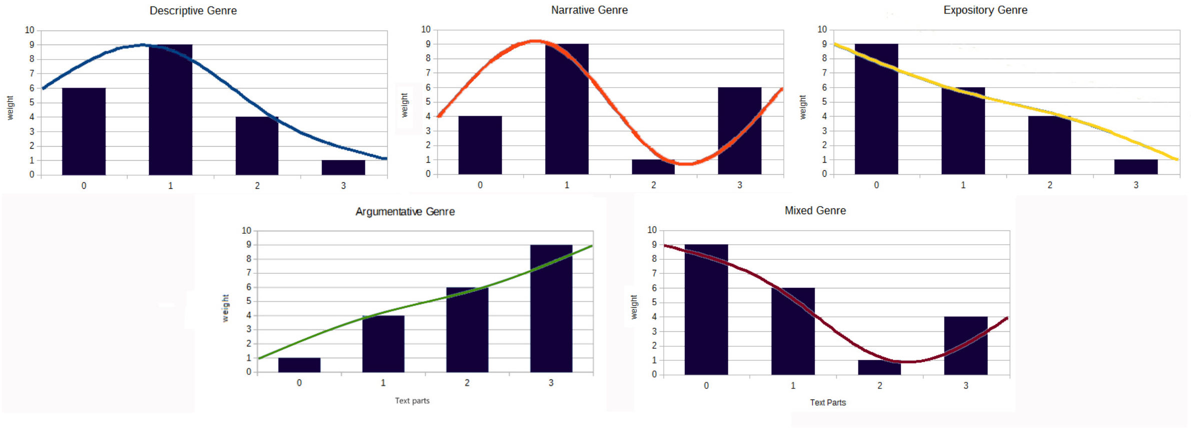

Figure 5: Graphs showing the relative relevance of different text parts.

4.5 Evaluation (ROUGE)

We use the ROUGE metrics (Lin, 2004)[29] to evaluate the system-generated summaries. Tables@@@@ 4, 5 and 6 reports ROUGE-1 recall scores for every genre in Danish, English and Spanish.

Although these results cannot be directly compared to inter-language due to the differences mentioned above and differences in the original documents and their model summaries, we get an idea of which systems work best in this data set: generally BF.

[29] We used version 1.5.5. with the following parameters: -n 2 -2 4 -U -a -x -c 95 -r 1000 -f A -p 0.5 -t 0 -l 400 -a.

5 Results

In this section, we present our results in different genres and languages. We analyze them individually and make a comparison, where relevant.

Firstly, we compare our baselines and Weighted PageRank-based method. Secondly, we compare the results of our Personalized PageRank-based system with the previous results and test our hypothesis. Thirdly, we discuss our results and which factors that come into play. Finally, we analyze the factors that can influence our results.

		Danish			English			Spanish	
	PPF	PPB	PPU	PPF	PPB	PPU	PPF	PPB	PPU
Gen	25 50 100	25 50 100	25 50 100	25 50 100	25 50 100	25 50 100	25 50 100	25 50 100	25 50 100
Des	.13 .19 .26	.16 **.23** .28	.14 .18 .25	.15 .20 .26	**.18 .23 .29**	.17 .22 .28	.19 .24 .31	.23 .28 **.35**	.20 .25 .33
Nar	.15 .19 .28	.14 .20 .28	.13 .20 .26	16 .22 .29	.19 .23 .29	.16 .23 .29	.21 .27 .33	.23 .28 .34	.20 .26 .34
Exp	.18 .27 .46	.39 .50 .63	.24 .34 .52	.15 .24 .37	.46 **.59 .67**	.27 .38 .54	.22 .26 .37	.38 .46 .62	.27 .36 .49
Arg	.24 .33 .49	.30 .44 .56	.27 .28 .54	.35 **.56 .73**	.38 **.58 .74**	.37 **.57 .73**	.33 .49 .71	.31 .44 .70	.30 .45 .70
Mix	.15 .22 .31	.21 .27 .35	.16 .22 .31	.19 .25 .32	.23 .27 **.34**	.18 .22 .27	.22 .30 .37	.29 .34 .39	.22 .27 .35

Table 12: ROUGE-1 50-word summaries scores from our Personalized PageRank-based systems Undirected, directed Forward and directed Backward for every genre in English.

5.1 Overview

We analyze our five genres one by one and see in which genres our Personalized PageRank-based systems perform better than any of the previous systems.

In Table 8 @@, we can see the results of the three Personalized PageRank-based systems for every genre in English. In blue, we can see the systems that surpass our baseline and weighted PageRank-based system. The letters in green show when the results obtained where the equal.

We can see that only some systems perform better than the baseline or weighted PageRank-based system. In fact, all systems that outperform previous results are in English. When applying the same personalization vectors to Danish or Spanish the results are normally worse (or the same). This is probably due to the fact that we obtained information about the English language in order to create our personalization vectors. When applying this information to other languages even in the same genre, the values are not adequate.

In Figure 5@@, we can see the improvement achieved from our Personalized PageRank-based systems in comparison with our best performing baseline and weighted PageRank-based systems in Danish, English and Spanish.

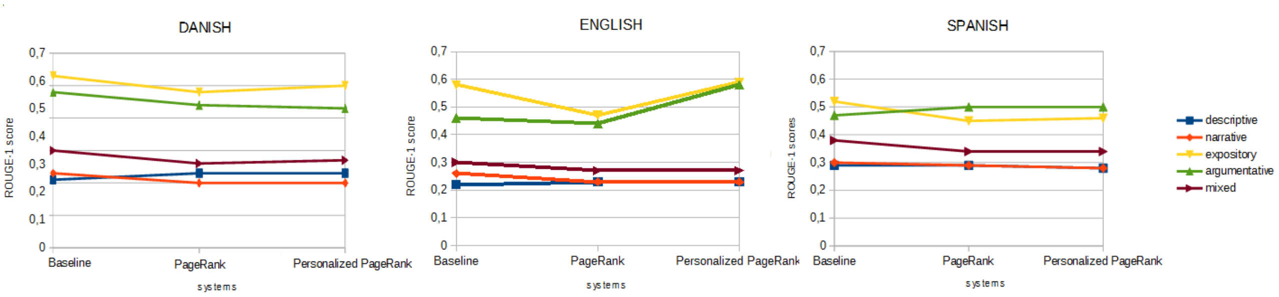

Figure 6: Graphs showing the improvement in Danish, English and Spanish from the best baseline ROUGE-1 50-word summary score and the best performing weighted PageRank-based system to the best Personalized PageRank-based system.

1. Descriptive Genre

   In Table 8, we can see that our Personalized PageRank-based system does not perform that well. PPB works just as well as our PageRank-based system (PB) for 25, 50 and 100-word summaries in English. However, despite not performing much better, these systems improved their baseline results. Table 5@@ shows these results.

   In Danish and Spanish the results are very similar. In some system-generated summaries, we obtain the same score than in the baseline (BF) or weighted PageRank-based system (PB) or less, although not very far away from each other.

2. Narrative Genre

   Table 8@@ shows that our Personalized PageRank-based system performs even worse than our baseline (BF). Our weighted PageRank-based method, though, already obtained worse results than our baseline.

   However, we can see in Figure 5 that our best Personalized PageRank-based system (PPB) works just as well as our weighted PageRank-baseds system (PB). Therefore, we can see that our personalization vector does not influence our ROUGE scores.

   In Danish and Spanish, despite not being far away from each other, the results from our Personalized PageRank-based system are always worse than the baseline (BL).

3. Expository Genre

   We can see in Table 8@@ that our Personalized PageRank-based system (PPB) improves a bit our results from our best working baseline system (BF).

   As shown in Figure 5@, our weighted PageRank-based system performed much worse than our baseline system. Taking into consideration our personalization vector, we can not only improve its results but reaching a higher score than our baseline in 50 and 100-word summaries.

   Although our Personalized PageRank-based system (PPB) outperforms the weighted PageRank-based system (PB) in Danish and Spanish, it does not work as well as the baseline (PF).

4. Argumentative Genre

   In this genre we can find the clearest improvements after applying all our Personalized PageRank-based systems. This makes PPB the best working system (Table 8@). It outperforms our Baseline best scores (BF).

   In Figure 5@, we can see that our weighted PageRank-based system does not work as well as the baseline, but our personalization vector improved the system a lot.

   The Personalized PageRank-based systems in Danish work worse than the weighted PageRank-based system (PB) and, consequently, the Baseline BF. In Spanish, the Personalized PageRank-based system (PPF) performs better than the baselines, but it does not have better results than the weighted PageRank-based system (PF).

5. Mixed Genre

   Finally, in this genre there were no improvements with respect to the baseline scores (BF) that were the highest. We can see that when considering a 100-word summary, our PPB system just reaches the same score as the BF system and surpasses the weighted PageRank-based system (PB).

For the 50-word summary (Figure 5@@), the results from our weighted PageRank-based system are worse than our Baseline and, additionally, our Personalized PageRank-based system were even worse.

The pattern is the same in both the Danish and the Spanish data sets.

5.2 Error Analysis/Analysis of Results

After having shown our results, we go more into details with regards to our ROUGE recall scores. We start analyzing our baselines, and them comparing its results to our weighted PageRank. We also discuss how our weighted PageRank works.

Then, we analyze our Personalized PageRank system and we compare the best working system with it. We discuss differences in all genres and languages. Finally, we test whether our hypothesis was validated. Furthermore, we explain how different factors have affected our results and we compare our results with other studies.

5.2.1 Performance of Baseline and PageRank-based systems

In Figure 6@@, we can see some differences in our baseline and weighted PageRank-based systems results. However, in most cases the more words we take into account in our summaries, the better results we obtain.

The exception is the argumentative genre, where the results are worse in the 100-word summaries than in the 50-word summaries. We can see that our 50-word summaries have a higher ROUGE-1 score than our 25-word summaries in all cases.

The reason for these differences is the total number of words that our model summaries in this genre contain. When our system generates more words than the number of words that the model summary contains, we risk that the performance worsens. This is shown in our results.

Our model summaries in the argumentative genre in English had an average of 52 words and 94 words in the expository genre. If we analyze more than that number of words, it implies a loss in the ROUGE-1 scores. That is the reason why in some studies the exact certain amount of words in summaries is always generated, in order to avoid this problem.

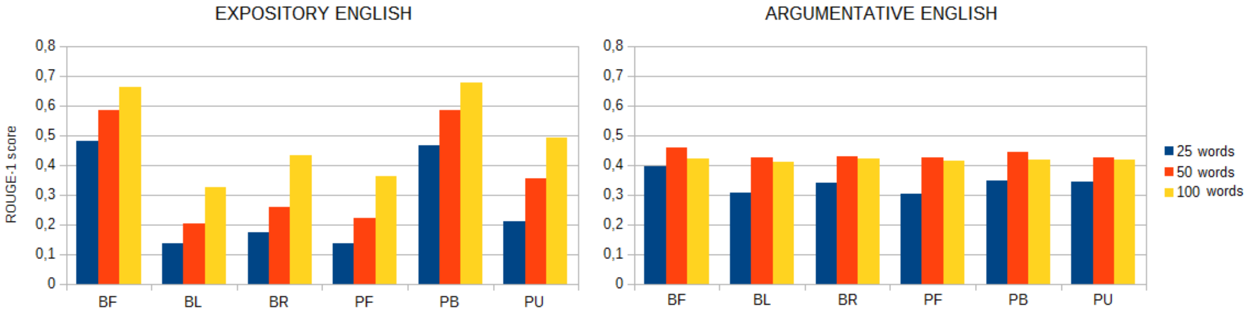

Figure 7: Graphs showing the differences in ROUGE-1 summary score for our baseline and weighted PageRank-based systems in the expository and the argumentative genre in English when taking into account a different number of words.

In Figure 7 @@, we can see how our baseline and weighted PageRank-based systems improve relatively to each other. As previously mentioned, there is an improvement in the results when more words are taken into account. However, we can see that this improvement is larger in some systems.

In the narrative genre, we can see that when around 200 words are taken into consideration when evaluating the summaries, all systems have approximately the same performance and overlap in the curve. Conversely, in the descriptive genre it can be noted that systems differ more in their results when more words are taken into account.

The above difference in behaviour can be a result of the fact that the model summaries in the narrative genre in English are much shorter (255 average words) than the model summaries in the descriptive genre (425 average words).

We must also notice that the more words that we consider in model summaries, the more likely it is that all systems picked sentences from the full document that contained those words. Because of that, it is not surprising that when considering 200 or 400-word model summaries instead of 25, 50 or 100-word summaries, the results are higher and that all systems perform almost as good as the others.

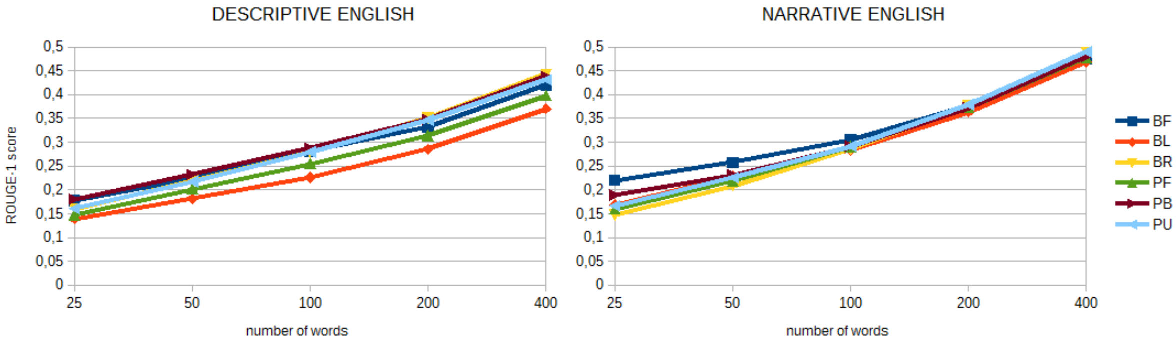

Figure 8: Graphs showing the improvement in ROUGE-1 summary scores for our baseline and weigthed PageRank-based systems in the descriptive and the narrative genre in English when taking into account a different number of words.

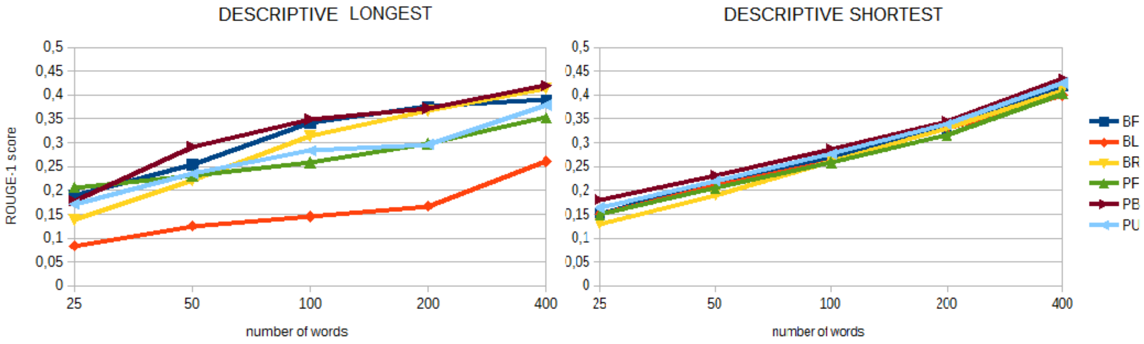

Figure 9: Graph showing the improvement in ROUGE-1 summary scores for our baseline and weigthed PageRank-based systems in the longest individual document and in the shortest individual document in the descriptive genre in English regarding different number of words.

We compare these improvements in our baseline and weighted PageRank-based systems with one of the longest and one of the shortest documents of our data set, in order to see whether the results from these documents behave differently.

We can see in Figure 8@ that in longer documents there tends to be more differences in ROUGE-1 scores across different systems. Conversely, in shorter documents the systems tend to behave similarly. However, we can see that there is a tendency that the same systems work best in all system-generated summaries.

5.2.2 Weigthed PageRank-based systems vs. Baseline systems

First of all, we compare our deterministic baselines with our weighted PageRank-based systems in Danish, English and Spanish.

We can see in Table 12@@@ (Danish), that only in the descriptive genre, our PageRank-based system PB works better (+ 0.02) than the best baseline system (BF).

However, if we look at the best systems when only taking one sentence into account, we can see that the PageRank-based system PU tends to find a better sentence than our baselines. Additionally, all PageRank-based systems have better performance than our baseline systems when only one sentence is considered.

Danish

Genre	length	Baseline	PageRank	Difference
Descriptive	full doc	BF	**PB**	+ 0.02
	1 sent	BL/BR	**PU**	+ 0.09
Narrative	full doc	**BF**	PB/PU	- 0.03
	1 sent	BL	**PU**	+ 0.08
Expository	full doc	**BF**	PB	- 0.05
	1 sent	BF	**PU**	+ 0.10
Argumentative	full doc	**BF**	PB	- 0.04
	1 sent	BL	**PU**	+ 0.16
Mixed	full doc	**BF**	PB	- 0.04
	1 sent	BL	**PU**	+ 0.10

Table 13: Comparison between the best baseline system and the best PageRank-based system for every genre in Danish. The full document or just one sentence is taken into account. Positive values indicate that the PageRank-based system surpasses the baseline system. Negative values indicate that the Baseline system has higher results than the PageRank-based system.

For English (Table 13@@), the weighted PageRank-based system PB works only better (+ 0.01) than the baseline system in the descriptive genre.

When looking at how the systems perform only taking into consideration one sentence, with the exception of the expository genre, our PageRank-based system PU works better than our baselines. In the expository genre, both PB and PU perform worse than the baseline system BF.

English

Genre	length	Baseline	PageRank	Difference
Descriptive	full doc	BF/BR	**PB**	+ 0.01
	1 sent	BL	**PU**	+ 0.13
Narrative	full doc	**BF**	PB/PU	- 0.03
	1 sent	BL	**PU**	+ 0.09
Expository	full doc	**BF**	PB	- 0.11
	1 sent	**BF**	PB/PU	- 0.01
Argumentative	full doc	**BF**	PB	- 0.02
	1 sent	BL/BR	**PU**	+ 0.11
Mixed	full doc	**BF**	PB	- 0.03
	1 sent	BL	**PU**	+ 0.07

Table 14: Comparison between the best baseline system and the best PageRank-based system for every genre in English. The full document or just one sentence is taken into account. Positive values indicate that the PageRank-based system surpasses the baseline system. Negative values indicate that the Baseline system has higher results than the PageRank-based system.

For Spanish (Table 14@@@), our weigthed PageRank-based system PF outperforms (+ 0.03) our baseline in the argumentative genre. In the descriptive genre, PB and BF and BL baseline systems gave the same result.

When looking at only one sentence, our PageRank-based system PU works better in almost all genres with the exception of the expository genre, in which the best PageRank-based system was PB and is overpassed by BF.

Spanish

Genre	length	Baseline	PageRank	Difference
Descriptive	full doc	BF/BL	PB	0
	1 sent	BL/BR	PU	+ **0.13**
Narrative	full doc	**BF**	PB	- 0.01
	1 sent	BL	**PU**	+ 0.11
Expository	full doc	**BF**	PB	- 0.07
	1 sent	BR	**PU**	+ 0.12
Argumentative	full doc	BF/BL	**PF**	+ 0.03
	1 sent	BL	**PU**	+ 0.15
Mixed	full doc	**BF**	PB	- 0.04
	1 sent	**BF**	PB	- 0.04

Table 15: Comparison between the best baseline system and the best PageRank-based system for every genre in Spanish. The full document is taken into account or just one sentence. Positive values indicate that the PageRank-based system surpasses the baseline system. Negative values indicate that the Baseline system has higher results than the PageRank-based system.

We can see in Tables 12, 13 and 14 that the baseline BF (i.e. regarding only the first k words of a document) outperforms most of the time both the other baselines (BL and BR) and our weighted PageRank-based systems (PF, PB and PU).

However, this is not the case when taking into consideration only one sentence. In these cases, PageRank-based systems (especially PU) have better results. When checking our system-generated summaries individually, we notice that our PU systems tend to select much longer sentences than PF or PB.

The fact that PU systems tend to select longer sentences can be an explanation for why the PU system works better on very short summaries (when not taking into account the number of words that every sentence has). The more words a system catches, the more probable that it is that it catches some of the words that appeared in the model summary.

5.2.3 Performance of Personalized PageRank-based systems

The values examined in our Personalization Vectors in every genre are heuristic. Because of that, we compare the results obtained by applying those heuristic values with other values: both the similar values and the different values in order to quantify the sensitiveness of our Personalized PageRank-based system parametrization.

We check whether the changes that our Personalization Vectors caused in our results are big or whether there are no changes. For this purpose, we compare the ROUGE-1 recall scores obtained in some genres with other possible Personalization Vectors in the same genres.

First of all, we compare our Personalization Vector for the expository genre in all our Personalized PageRank-based systems (PPF, PPB and PPU). Figure 9@@ shows a vector

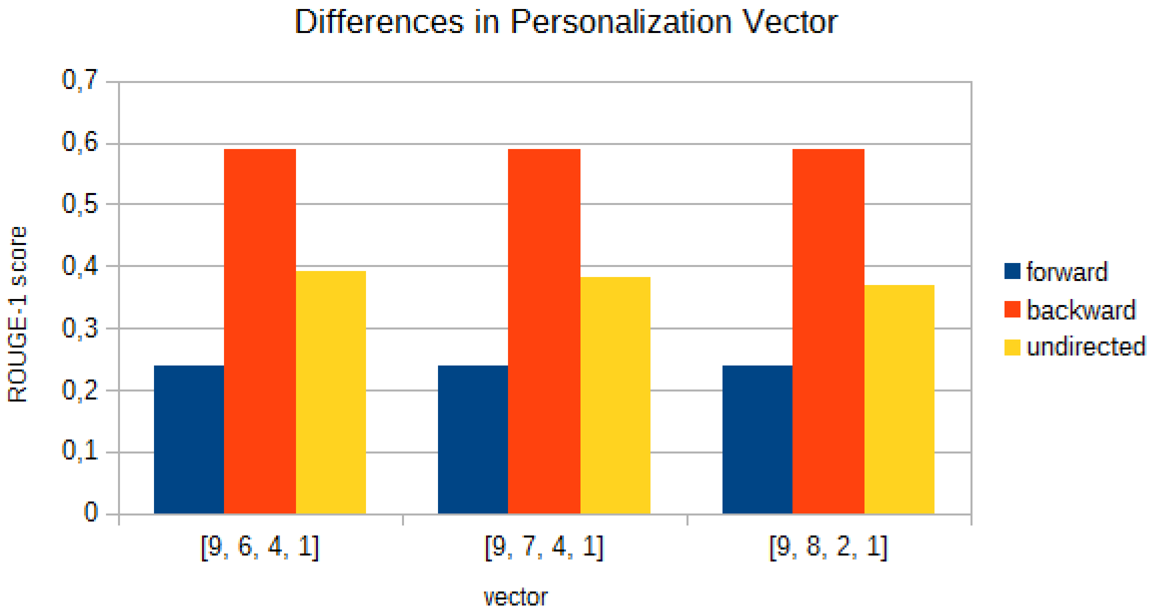

Figure 10: Graph showing differences in ROUGE-1 score with slightly different personalization vectors. The example shows the expository genre in English.

that is only different in one value ([9, 7, 4, 1] instead of [9, 6, 4, 1][30] and with a slightly different vector that keeps the same sum as the original vector (20) ([9, 8, 2, 1]).

We can see that the differences in ROUGE scores are almost imperceptible. It is actually only in the PPU system that the numbers differ. The values chosen for the expository genre [9, 6, 4, 1] seem to work a bit better than other values that are similar in the PPU system but there are no differences in the other systems (PPB and PPF).

Additionally, we compare the argumentative personalization vector [1, 4, 6, 9] with other vectors that are somewhat more different. We show the results for the three Personalized PageRank-based systems (PPF, PPB and PPU) in comparison with the results of other Personalization Vectors in Figure 10@@@@.

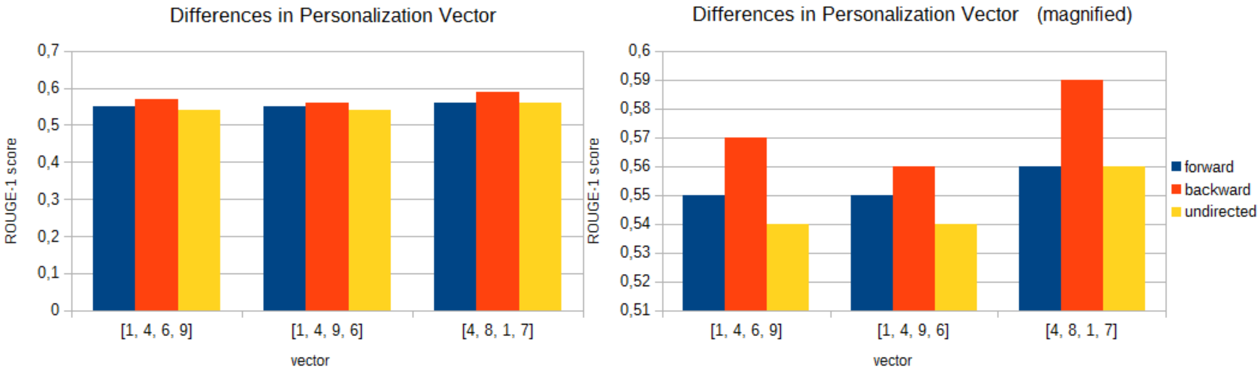

Figure 11: Graph showing differences in ROUGE-1 score applying more different personalization vectors. The argumentative genre in English is shown in the example. In the graph to the left we can see the ROUGE-1 score using the same scale as in Figure 10@@@. In the graph to the right we can see the exactly same results in a magnified version.

[30] This change causes a different normalized factor in the personalization vector.

In Figure 10@@@ we can see that the vectors are 1) the same vector but interchanging the last two values [1, 4, 9, 6] and 2) a randomly chosen vector that does not have any value in common with the original vector [4, 8, 1, 7] but keeps the same sum as the original vector (20) so that the normalized factor is the same.

Again, the changes in the three systems are almost indiscernible. It is only when zooming in on the graph that we are able to see the small differences across systems.

In the graph to the right in Figure 10@@ we can see that it is actually the randomly chosen vector that performed best. The heuristic chosen values for the argumentative genre [1, 4, 6, 9] work just a bit better than the other combination of the same values.

We have shown that the values which we fill the Personalization Vector with do influence our results in the evaluation. However, these differences are very small and sometimes imperceptible.

We have also seen that even randomly chosen vector values work better than some of our heuristic values in some cases. Therefore, the heuristic values we used to generate our Personalization Vector are not the best and there might be better combinations that would lead us to better results.

5.2.4 Personalized PageRank-based systems vs. Previous systems

We compare our Personalized PageRank-based system scores with the previously best performing systems in Danish, English and Spanish.

We can see in Table 12@@ that none of our Personalized PageRank-based systems work better than the BF or PB systems in Danish. We can observe the same results in Table 14@@@ for the Spanish data set.

We must recall that the personalization vectors for Danish and Spanish were filled following the heuristic values that best fitted the English data set. Therefore, it is not surprising that our Personalized PageRank-based system works worse than the other systems in Danish and Spanish.

However, we can claim that genres in different languages are very different in structure or that personalization vectors cannot be transposed across languages and different data sets.

Danish

Genre	B/P	PP	Difference
Descriptive	**PB**	PPB	0
Narrative	**BF**	PPB/PPU	- 0.03
Expository	**BF**	PPB	- 0.03
Argumentative	**BF**	PPB	- 0.04
Mixed	**BF**	PPB	- 0.03

Table 16: Comparison between the best baseline/PageRank-based system and the best Personalized PageRank-based system for every genre when taking the full document into account or just one sentence.

In the English data set, we can see that we obtained better ROUGE-1 results in our Personalized PageRank-based system (PPB) than in the previous systems for the expository and argumentative genre.

We must recall that these genres are the shortest and were annotated by an annotator instead of being gathered from existing on-line resources. The length of these documents and the fact that they were collected in a different way can affect our results.

English

Genre	P/B	PP	Difference
Descriptive	**PB**	PPB	0
Narrative	**BF**	PPU	- 0.03
Expository	BF	**PPB**	+ 0.01
Argumentative	BF	**PPB**	+ 0.12
Mixed	**BF**	PPB	- 0.03

Table 17: Comparison between the best baseline/PageRank-based system and the best Personalized PageRank-based system for every genre when taking the full document into account or just one sentence.

However, in all systems we can see that the differences in scores are not big. Because of that, it is possible that these results vary a lot across different data sets even in the same language.

Spanish

Genre	B/P	PP	Difference
Descriptive	**BF/BL/PB**	PPB	- 0.01
Narrative	**BF**	PPB	- 0.02
Expository	**BF**	PPB	- 0.06
Argumentative	**PF**	PPF	- 0.01
Mixed	**PF**	PPB	- 0.04

Table 18: Comparison between the best baseline/PageRank-based system and the best Personalized PageRank-based system for every genre when taking the full document into account or just one sentence.

5.3 Testing the Hypothesis

After having presented our results and analyzed them, we show in which genres and languages our hypothesis is valid. Table 18@ shows that our hypothesis is valid for the expository and argumentative genres in English. In these two genres the sentence position information we added in our Personalized PageRank-based system was helpful and improved the results of our summaries, especially in the argumentative genre.

However, it is not valid for the other genres and languages. Thus, we have not been able to demonstrate that our hypothesis is valid in the Danish and Spanish language. The fact that our hypothesis is valid only for some genres in English is probably due to the fact that we used heuristic values based on the characteristics of our English data sets.

	Danish	English	Spanish
Descriptive	-	-	-
Narrative	-	-	-
Expository	-	+	-
Argumentative	-	+	-
Mixed	-	-	-

Table 19: Genres in which the hypothesis is valid in Danish, English and Spanish.

In the further discussion, we analyze some factors that might have influenced our generally bad results as well as our good results in the expository and argumentative genres. We also pay attention to the task of improving the automatic text summarization and, additionally, how to avoid the possible errors.

5.3.1 Significance test

There is normally not a statistical significance testing, (such as t-test) in the automatic text summarization field. Most studies just report their recall scores evaluated with ROUGE. Thus, we have only reported these recall scores.

In some studies, confidence intervals are calculated to estimate the margin of error their values have in order to show whether they systems do improve their baselines results or other studies results. They estimate these intervals by using more than one model summary.

We only have one model summary. Therefore, we cannot calculate these intervals in order to statistically check that our Personalized PageRank-based systems perform better than our baseline and PageRank-based systems.

5.4 Discussion

We further discuss qualitative aspects of the data, which are more difficult to measure objectively. We discuss where the information is placed how genres and languages differ when structuring their information. We also discuss how humans create model summaries and whether model summaries are the best tool to contrast our results with.

Furthermore, we compare our results whit the results of other studies when possible and make a mention about ROUGE-2 and ROUGE-SU4, since they have also been used in other studies. Finally, we will qualitatively evaluate our hypothesis.

5.4.1 Information placement

We have shown some figures and tables with our results after our experiments. However, after analyzing all our results, we discuss further in detail where our results point out that the information lays, and discuss the differences between genres and languages.

We analyze the differences genre by genre:

1. **Descriptive Genre**

   We have not been able to show where the most important information in the descriptive genre lays and, consequently, whether our hypothesis was valid or not. It has just been proven that the earlier the sentences appear in this type of texts, the more important they tend to be.

   There is not research in the descriptive genre that concerns Wikipedia articles. However, there is research on legal documents, which is another type of text within the descriptive genre, despite distinctive structures.

   Ceylan et al. (2010) have studied legal texts and claim that it was one of the hardest domains where TextRank could improve its performance with respect to a random baseline. In the case of Wikipedia articles, we have been able to improve our BR system. However, our results are very close to our BF system.

2. **Narrative Genre**

   Our attempt to find where most important information lays in the narrative genre has not been the most profitable. Therefore, we cannot demonstrate that our hypothesis was valid in this genre. Any of our systems could surpass the baseline BF. PB and PPB graph-based systems tend to work better than the other systems but the difference is not very big.

   In other studies on the narrative genre it has been found that sentence position is not beneficial to find the best sentences to automatically generate summaries. Following the advice of these studies, other methods should be employed in order to find out where narrative texts tend to condense their most important information.

3. **Expository Genre**

   We have shown that texts in the expository genre tend to accumulate the most imporatant information in the beginning of the documents, including their titles. In other words, titles in this genre tend to be very relevant and should definitely be included in their summaries. We have proved that our hypothesis is valid for the text of the expository genre in English.

   However, the fact that it was not demonstrated we the same heuristic values in the other languages means that our hypothesis cannot be extrapolated to other languages or other data sets in the same language.

   Most research in automatic summarization have been conducted in the newswire domain (which is a part of the expository genre) that matches the characteristics of our data set. We have found that our results match previous studies results in this genre in English.

4. **Argumentative Genre**

   We have shown that there is important information in the end of texts in the argumentative genre. We have demonstrated that our hypothesis is valid for texts in the argumentative genre in English.

   Therefore, giving more relevance to the final sentences than to the beginning of these texts is beneficial for the created summary. We have also shown that the titles of these texts tend not to be relevant. Therefore, they should not be included in their summaries.

   To our knowledge, there are no studies in the automatic text summarization field on the argumentative genre. Thus, we cannot compare our results on this genre with other studies. Moreover, studying this genre more in depth and with different data sets would be interesting.

5. **Mixed Genre**

   It has not been possible to discover where the most important information in scientific papers is located. Thus, we must say that our hypothesis is not valid for texts in the mixed genre. We have seen that in general, the title of these documents tends to be relevant and should be given importance in their summaries. It seems as well that PB and PPB graph-based methods work better than other systems.

   There are other studies that evaluated scientific papers such as Ceylan (2010) which found that it is easy to improve the results of this type of texts with TextRank.

   However, we have not been able to do as they claimed. Again, it can be due to the fact that texts in the mix genre were lengthy. It might be, just as suggested in the narrative genre, because sentence position is harmful for the results in this genre.

In conclusion, our hypothesis has only been validated in two of the five genres our of our corpus: the expository and the argumentative genres. Our hypothesis has been validate for only 13% of our data.

Nevertheless, it is only under some specific circumstances that we can say that it was proved that our heuristic values were adequate. This is due to the fact that using the same values in other data sets in other languages did not improve our baseline or weighted PageRank-based systems.

5.4.2 Direction of the relations between sentences

As mentioned earlier, PageRank was designed to rank web pages. When we have to make a judgement about which web page is forwarding another page by a link, it is clear what the direction is. However, it is not so easy to see a priori whether relations between sentences in a text are anaphorical or cataphorical.

It is not so hard to distinguish anahorical references from cataphorical references in oral speech since speech takes place in time. When people talk they do necessarily have to say one word before the next. Because of that, it is easier to find (and easier to understand) anaphorical references than cataphorical references in oral speech.

Nevertheless, texts can be written from the end to the beginning and rewritten or corrected several times. Therefore, texts do not always have a clear direction in the relations between sentences. A good example is the descriptive genre, which is rather atemporal. We can argue that when defining an object, every time the object is mentioned a backwards reference to the object is made.

Lets put the definition of a chair as an example:

Object	Chair
Sentence 1	A chair is a piece of furniture with a raised surface.
Sentence 2	A chair is used to seat a single person.

Both Sentence 1 and Sentence 2 in the example can have a backwards relation to the object that they are describing (chair), since they share the word. However, when creating a relation between Sentence 1 and Sentence 2, is creating a backwards relation then the most correct thing to do? Is a chair *primarily* a piece of furniture or used to seat a single person?

The conclusion is that since descriptions are a representation of someone or something, they are not temporal. Thus, deciding whether it is most appropriate to create a backward or forward link is difficult both for humans and machines.

Nevertheless, different genres usually have different relations between sentences. The narrative genre is mostly temporal since a narration is a telling of something that has happened (real or fictitious).

However, when writing a story, the writer can *play* with the reader and present the information in different orders and refer to facts that have not been explained yet. This might be an extra difficulty to automatic summarization to find the correct relations between sentences and which sentence is pointing to another sentence.

In the expository genre it is more difficult to find cataphorical references. The most relevant information (as we have shown in this thesis) is always at the beginning of the document. Therefore backward links between sentences work best.

Argumentative texts can have very complex structures. Their aim is to convince the addressee of an idea. It is common that there are cataphorical references, which are pointing to the main point of the argument, if this idea is written in the end of the argument.

Finally, texts in the mixed genre are very different and cannot be analyzed as a whole. In the case of this thesis, we have analyzed scientific texts, which may also have different anaphorical and cataphorical references within them, such as references, tables or further explanations that come afterwards.

In practice, even when describing a static object, we tend to write the most important or relevant information in the beginning. Because of that, it might be better to create backward connections (as our results show) when we are not sure about what the best connection is.

5.4.3 Model summaries

One important aspect to discuss is the way humans write summaries. Two genres (expository and argumentative) in our corpus (GIC) have been annotated by annotators. These annotators were instructed to follow specific rules. However, the documents in the remaining genres (descriptive, narrative and mixed) have been collected from already existing resources.

These model summaries have been generated for different purposes. First of all, descriptive summaries are short representations of something or someone by means of language. Since all the model summaries are from Wikipedia, they tend to follow the same style, despite differing in the comprehension of the description. This is especially the case of the Spanish summaries, which are longer and more detailed than the Wikipedia summaries in Danish and English.

Secondly, narrative summaries are mainly made by students, who uploaded their own created summaries to websites so that other people can benefit from them. When analyzing these summaries, we discovered that they were made following completely different rules, especially in the Danish data set: some texts were explicitly about the beginning of the novels or tales, but did not mention the end of the story. Other texts just entail a certain amount of words and then left the summary open-ended. Other summaries include only a brief description of what the novel/tale was about.

Finally, the summaries in the mixed genre were abstracts of the full scientific papers. These summaries did not follow a strict pattern. Some summaries worked as an introduction of the paper and the topic, while others were more comprehensive and contained methodology, results and conclusions.

These differences in how people create a summary makes us reflect on the quality of our model summaries and the quality of human summarization techniques in general. Maybe we should ask ourselves: ”Does a good model summary exist?” ”How do we write a good summary?” ”Are there differences in the content of model summaries across genres and languages?”

There are many studies in this area (such as Winograd (1984), Brown (1983) and Kirkland (1991)) and they all agree on the point that some people create better summaries than others[31].

5.4.4 Comparison with other studies

If we compare the results of Ceylan et al. (2010) (see Figure 3 in section @@@) with our results in the type of texts that we both analyze, namely the literary, expository and scientific genres, we obtain different results. A comparison is shown in Table 15@@@.

In Table 15@@ it is shown that Ceylan et al. (2010) obtained better results, both in the narrative genre (their literary domain) and the mixed genre (their scientific domain). However, we obtained better results in our expository genre (their newswire domain).

Nevertheless, we cannot directly make a comparison of the results of Ceylan et al. (2010) and ours since our data sets are different. However, we can compare the improvement that they experienced from their baselines to the TextRank system of Ceylan et al. (2010) and our own.

We can see in Table 15@ that our Personalized PageRank-based system and the TextRank system of Ceylan et al. (2010) performed close to the baselines. In the expository genre, the TextRank system performed better than the Random system but worse than the

[31] They measure, among other things, 1) what parts of the text people find most important, 2) which type of content they find to be the most important, 3) how many sentences (in proportion) people selected to create the summary, 4) whether they freely paraphrased.

	Ceylan (2010)			Our results		
	Lead (=BF)	Random (=BR)	TextRank	BF	BR	PP
Literary / Narrative	.45	.45	**.46**	.26	.21	.23
Newswire / Expository	.46	.39	.44	.58	.26	**.59**
Scientific / Mixed	.47	.46	**.49**	.30	.22	.27

Table 20: Ceylan et al. (2010)'s results in comparison with our results in the narrative, expository and mixed genre in 50-word summaries. The Personalized PageRank-based system appearing in this Table is the system that had the best performance of the three available.

Lead system. With regards to our own setup, the Personalized PageRank-based system outperformed the BF system.

Finally, the results of TextRank in Ceylan et al. (2010) in their scientific domain surpassed their baselines. However, their results were not far away from their Random system either. Our Personalized PageRank-based system did surpass our BR (Random) system, but did not perform better than our baseline (BF).

One reason for not obtaining better results in the narrative and mixed genre than in our baselines can be because of what Mihalcea and Ceylan (2007) explained:

> "The position of sentences in a document seems like a pertinent heuristic for the summarization of short documents, and in particular for the newswire genre".

However, as Mihalcea and Ceylan (2007) explained in the same study, long texts do not benefit from sentence position information. Therefore, our narrative and mixed genre do not perform very well when adding this information.

As we have proved in this thesis, sentence position does not seem to be a good heuristic method to take into account when summarizing long texts. In long texts, topics change often, and therefore further research is needed to be able to find out where the most important information resides.

We can also compare our results with those from the DUC (2002) conference. In this case, the DUC (2002) conference data set was in the newswire domain and specified that summaries must contain 100 words. Because of that, we compare the best TextRank system of the DUC (2002) conference with our weighted PageRank-based system and Personalized PageRank-based systems in English when evaluating 100-word summaries.

	DUC (2002) best results	Our results	
	PageRank (weighted)	PageRank (weighted)	Personalized PageRank
Undirected	.49	.35	**.54**
Forward	**.42**	.36	.37
Backward	.50	.59	**.67**

Table 21: The best results in 100-word summaries of the DUC (2002) conference using different directed and undirected PageRank-based systems in their newswire domain compared with our results in the expository genre (using both weigthed PageRank-based systems and Personalized PageRank-based systems.

Table 16@@ shows that our Personalized PageRank-based systems PB and PU outperfom those from the DUC (2002) conference. However, our Forward Personalized PageRank-based system (PPF) does not work as well as the results in DUC's best Forward system. Nevertheless, since the data sets we used are different, it is difficult to claim that our systems are better than theirs.

5.4.5 ROUGE-2, ROUGE SU4

Since many studies consider it relevant to include other evaluation results and not only ROUGE-1, we also want to make a mention of other evaluation results: ROUGE-2 and ROUGE-SU4.

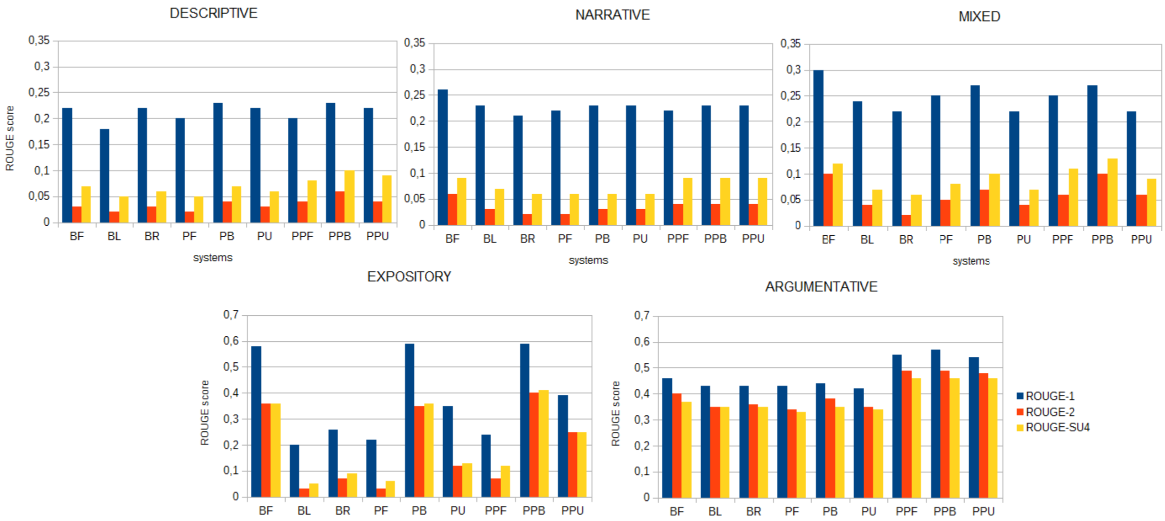

Figure 12: Graph showing scores in ROUGE-1, ROUGE-2 and ROUGE-SU4 in 50-word summaries in every genre in English.

Figure 11@ shows that different ROUGE parameters lead to very different results. However, it is a general rule that ROUGE-1 have much higher scores than the other systems, especially in the longer documents (Descriptive, Narrative and Mixed) and that ROUGE-2 has proportionally worse scores than ROUGE-SU4. Moreover, these scores seem to increase or decrease accordingly in all systems.

However, in the shorter documents the behaviour of ROUGE-2 and ROUGE-SU4 is not exactly the same. We can see in Figure 11@@ that in the expository genre ROUGE-1 keeps having higher scores than ROUGE-2 and ROUGE-SU4. However, we can see that ROUGE-2 scores are equally high or higher than ROUGE-SU4 in most systems. This means that in this type of texts, we have more bigrams that are exactly the same in both model and system-generated summaries than in the longer texts. In longer texts, it seems that it is not that common to have bigrams in common but skip-unigrams (in a threshold of four words) in common.

If we look at the argumentative genre, which is even shorter in word length than the texts in the expository genre, it is clear that ROUGE-2 scores are as high as ROUGE-SU4 or higher. We can also see in this genre that ROUGE-1, ROUGE-2 and ROUGE-SU4 scores are much closer to each other than in the other genres. Again, original texts are much shorter and when writing a 50-word summary, there is not so much information to pick. Therefore, it makes sense that the information in the model summary is very similar to the information of the extractive system-generated summary.

5.4.6 Graphs connectivity

In this thesis, we analyze the graph connectivity for every genre and check whether there were differences. We checked whether the undirected graphs were connected and whether our directed graphs were weakly connected.

Due to the characteristics of our documents, when undirected graphs were connected, directed graphs were weakly connected. The reason is that the nodes in directed graphs were only connected in one direction (forward or backward) while edges between nodes were created in both directions (forward and backward) when building undirected graphs.

Table 19@@@ shows the percentage of documents whose undirected graphs were connected and whose directed graphs weakly connected. We can see that there are two genres (descriptive and mixed) where none of the documents had a connected or weakly connected graph. Conversely, the expository genre is the genre in which most graphs are connected and weakly connected.

There is a big difference between the number of documents in the argumentative genre that were connected in Danish (27%), English (7%) and Spanish (0%). While there are four documents in Danish whose graphs were connected or weakly connected, there are only 2 documents in English that were connected or weakly connected. In addition, in Spanish there were no documents that were connected or weakly connected.

We can also see that graphs of the documents in Spanish are not connected except for the expository genre, where 14 graphs were connected. In English and Danish there are some graphs that are connected in the narrative and argumentative genres.

Genre	Danish	English	Spanish
Descriptive	0%	0%	0%
Narrative	13%	13%	0%
Expository	33%	33%	47%
Argumentative	27%	7%	0%
Mixed	0%	0%	0%

Table 22: Table showing the percentage of documents that were connected or weakly connected in every genre in Danish, English and Spanish

The above-mentioned results do not fully resemble the results obtained by Mihalcea et al. (2007), since Mihalcea et al. (2007) claim that their graphs were connected. Wondering how there can be such a big difference between their work and ours, we saw that Mihalcea et al. (2007) were stemming words and not removing stopwords from their texts. That might be the reason why we find these differences in our data sets. However, we cannot fully claim that these differences make our systems work worse.

A graph with fewer connections can also have some advantages such as less noise or, as Mihalcea et al. (2007) already discussed, the density in their graphs grew so much that it was intractable in very large documents. Because of the intractableness of their graphs, they needed to find a solution that was giving a threshold of 75, so that further from that number of sentences, no more connections would be created. To create graphs can be a great idea to make the algorithms run faster in long documents.

We have shown that there are differences in the connectivity of graphs of the documents in different genres but these differences also vary across languages.

5.4.7 Example of a text and its system-generated summaries

In Appendix 4, we have attached a series of system-generated summaries of one of the documents in the argumentative genre in order to be analyzed more in depth. In these examples, we can see the differences of analyzing 25, 50 or 100 words in a summary and how different summaries are.

This example is actually one of those texts in the argumentative genre which title (”Hmm”) was not informative at all and does not appear in other sentences in the same text. However, it is not only the BF (or BR) system that takes this word in its summaries but also all weigthed PageRank-based systems include this word in their 50 and 100-word summaries.

An explanation of this behaviour might be that the graph is not connected (or weakly connected), so that there are many sentences that obtain the same ranking after running PageRank and they are just sorted out by the Counter function in python@@@@.

However, what it is most interesting in this case is the fact that all Personalized PageRank-based systems rank this sentence (”Hmm”) as the last one. The reason behind this behaviour is the initial value that we assigned to the title in the argumentative genre in the personalization vector (1), where the title was given least importance.

We can see that, especially in short and not connected graphs, a Personalized PageRank-based system is very useful assigning pre-defined values so that when there are many sentences that have the same PageRank ranking running a regular or weighted PageRank, a Personalized PageRank-based system can make a difference in the results.

We must mention that our annotators were instructed to write the title of the documents in their summaries, regardless of their importance. Because of this fact, it is not surprising that systems that were including these titles (such as BF or PB) had better results than others (such as BL or PF), as it happens in the example in the appendix.

Additionally, if we check the last sentence of the document ”Let the fun and games begin”, we can see that all weighted PageRank-based systems left this sentence out of their summaries. It is a sentence that does not have a high ranking in PageRank since does not have many relations to other sentences and, additionally, we can see in the model summary that it does not appear.

However, all Personalized PageRank-based systems include this sentence in their 50 and 100-word summaries. The reason for this ”mistake” is the same as for the previous wise choice this systems made. The sentence ”Let the fun and games begin” is the last sentence of a text in the argumentative genre, so it has a high initial value (9), which affects the results in the system-generated summaries.

With this example we can see that a Personalized PageRank-based system does include information that helps us have better summaries. However, at the same time, these types of systems make mistakes.

5.4.8 An evaluation of the hypothesis

the linguistic resource an assessment of the relevance of relevant text sections in automatic summarization @@@@@@

6 Conclusions

@@@@ y todo esto para que lo hago

6.1 Summary

@STEMING, STOPWORDS

In this thesis, we have presented a new approach to analyzing automatic text summarization by genre. We have shown that the text structure of different genres does influence the disposition of a text document and, therefore, the location of the most relevant information can be retrieved. However, this was not possible in all cases. This means that this task is not simple and further research must be carried out.

We have also shown that analyzing automatic text summarization results in sections can help to find the most important parts of a text. Other studies solved the problem of large data sets by splitting texts into sections and combining the results afterwards. However, we claim that the information that certain sections of texts possess can be valuable on their own and used to improve the task of automatic summarization. A further analysis of this already created parts, which were already created in other studies so that summaries are shorter, can show important information about text structure.

We have presented the GICorpus, which is a large, genre-informed, multilingual corpus for automatic text summarization. It contains 450 documents divided into five different genres, namely descriptive, narrative, expository, argumentative and mixed. It also contains a model summary for each document.

We have tested the following systems for every genre and language:

1. Three different baselines (BF, BL and BR).
2. Three different PageRank-based system (PF, PB and PU) summaries for every document to test the GICorpus.
3. Three different summaries for each document to check the relative relevance of each section of these documents were also created (the title of the texts, the remaining part of the text after removing the title and the two last third parts of the document).
4. Three Personalized PageRank-based systems (PPF, PPB and PPU) with a personalization vector formed by heuristic values that we deduced from the previously created summaries. These parameters beforehand determine which parts of a text that are most important, depending on the genre that the text belongs to.

We can see this distribution in Table 17 @@ (see section @@). It seems that the baseline BF and the PageRank-based system PB worked best in our data sets because they obtained higher results. We think that these systems were the best performing since, as we have explained in the discussion, anaphorical references are more common than cataphorical references. Additionally, there is a tendency in most genres to condense the most important information at the beginning.

We have seen in this thesis that texts are genre-dependent and that they have different structures. However, these structures must not be the same in all languages and might not be the same in all texts that belong to a genre. Nevertheless, we cannot prove that the idealized curves we propose in this work were the best values to represent the structure of the texts of all five genres we analyze.

We have learnt particularities of our data (the GICorpus) and generated new Personalized PageRank-based systems with that knowledge. Finding parameters that display the relative importance of the different parts of a text (i.e. introduction, development and end) has helped to improve the performance of PageRank-based summarization systems in some cases. However, we have not been able to improve the performance in all our data sets.

We tried to economize the task of human summary generation by gathering model summaries from public on-line sources. However, model summaries written by annotators were provided when gathering model summaries was not possible.

Ceylan et al. (2010) claimed that different domains have different difficulties. We have shown that our results, despite the fact that texts cannot be directly compared across type of texts or languages, differ greatly in scores in some genres. We conclude short texts actually work best. Probably, this is due to the fact that compression ratios are not as big in these genres and, consequently, it is easier for any system to pick the correct sentences (even for a random system).

In conclusion, we claim that texts can be grouped in broader categories than domains, namely genres. The reason of this classification is based on the structure that these genres tend to follow. Taking into account the structure of text is important to generate good automatic summaries.

6.2 Future Research

@@@que seria interesante hacer en el futuro, porq direccion no hay q ir, etc.

This study entailed the following limitations due to lack of time and resources: First of all, it would have been relevant to test more than one type of document inside every genre, such as editorials in the argumentative genre or legal texts in the descriptive genre. However, the task of gathering new documents and their summaries was time-consuming and sometimes impossible.

We considered it to be enough to gather just one model summary per document following the argumentation of Mihalcea et al. (2007): "the use of more than one reference summary does not influence the results". However, Mihalcea et al. (2007) used two model summaries in order to contrast one with the other and see how much agreement there were between them. If the resources were available, it would be relevant to compare a model summary written by an annotator against an on-line gathered model summary.

In a best-case scenario all documents would have approximately the same length so that our comparisons between genres would have been more accurate. However, it was a very arduous task. We discarded many documents because of their length so that every document in every genre were more similar. However, the variance across the length in words of documents within the same genre is nevertheless considerable.

We think that it is also necessary to speed up the task of automatic text summarization. In some cases, generating some summaries using PageRank-based systems took hours on our computer. Other studies, such as Mihalcea (2007) already had encountered this problem when analyzing books and found a solution in removing some edges in their graphs. It would be interesting to try this method and see whether our results remain the same or improve so that this technique can be applied in even longer texts with no loss in accuracy in our evaluation. If the task of automatic text summarization could be speeded up, it would be possible to create system-generated summaries for all the possible values in our personalization vector. Additionally, we can learn from these results which system-generated summaries are best. With this knowledge at hand, we can apply these values in different data sets and see whether they are consistent or not. It would also be interesting to know whether this parameter learning would lead us to values in our personalization vector that were close to our heuristic values or whether we were wrong in our assumptions to build these parametrization.

In the future, we would like to compare our results using a stemmer to see how graphs changed and how it influenced our results in Danish, English and Spanish.

6.3 que hacer con esto

si hay mucha literature: quitar a almeida, loza, aktas

7 References

[1] Almeida, Miguel B., et al. "Priberam Compressive Summarization Corpus: A New Multi-Document Summarization Corpus for European Portuguese." In LREC, 2014.

[2] Aktas, Mehmet S., et al. "An Application of Personalized PageRank Vectors: Personalized Search Engine".

[3] Brown, Ann L., Jeanne D. Day, and Roberta S. Jones. "The development of plans for summarizing texts." Child development (1983): 968-979.

[4] Ceylan, Hakan, et al. "Quantifying the limits and success of extractive summarization systems across domains." Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010.

[5] Conroy, John M., et al. "Left-brain/right-brain multi-document summarization." Proceedings of the Document Understanding Conference (DUC 2004), 2004.

[6] Dalianis, Hercules, and Erik strm. SweNama Swedish named entity recognizer. Technical Report. Department of Numerical Analysis and Computing Science, TRITA-NA-P0113-IPLab-189. Stockholm, Sweden, 2001.

[7] DUC. 2002. Document Understanding Conference 2002. http://www-nlpir.nist.gov/projects/duc/.

[8] Erkan, Gnes, and Dragomir R. Radev. "LexRank: Graph-based lexical centrality as salience in text summarization." Journal of Artificial Intelligence Research (2004): 457-479.

[9] Gustavsson, Pr, and Arne Jnsson. "Text summarization using random indexing and pagerank." Proceedings of the third Swedish Language Technology Conference (SLTC-2010), Linkping, Sweden, 2010.

[10] Herings, P., Gerard Van der Laan, and Dolf Talman. "Measuring the power of nodes in digraphs." Gerard and Talman, Dolf, Measuring the Power of Nodes in Digraphs (October 5, 2001).

[11] Hong, Kai, et al. A repositary of state of the art and competitive baseline summaries for generic news summarization. Proceedings of LREC, May, 2014.

[12] Hong, Kai, and Ani Nenkova. "Improving the estimation of word importance for news multi-document summarization." Proceedings of EACL. 2014.

[13] Jakobson, Roman. Closing statement: Linguistics and poetics. Style in language, 1960, 350: 377.

[14] Jakobson, Roman. "Metalanguage as a linguistic problem." Roman Jakobson. The framework of language. Ann Arbor: Michigan Studies in the Humanities (1980): 81-92.

[15] Kirkland, Margaret R., and Mary Anne P. Saunders. "Maximizing student performance in summary writing: Managing cognitive load." Tesol Quarterly 25.1 (1991): 105-121.

[16] Kazantseva, Anna, and Stan Szpakowicz. "Challenges in evaluating summaries of short stories." Proceedings of the Workshop on Task-Focused Summarization and Question Answering. Association for Computational Linguistics, 2006.

[17] Kikuchi, Yuta, et al. "Single document summarization based on nested tree structure." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Vol. 2. 2014.

[18] Kleinberg, Jon M. "Authoritative sources in a hyperlinked environment." Journal of the ACM (JACM) 46.5 (1999): 604-632.

[19] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text summarization branches out: Proceedings of the ACL-04 workshop. Vol. 8. 2004.

[20] Loza, Vanessa, et al. "Building a Dataset for Summarization and Keyword Extraction from Emails." In LREC, 2014

[21] Luhn, Hans Peter. "The automatic creation of literature abstracts." IBM Journal of research and development 2.2 (1958): 159-165.

[22] Rada Mihalcea and Ceylan, Hakan. "Explorations in Automatic Book Summarization." Association for Computational Linguistics, 2007.

[23] Mihalcea, Rada. "Language independent extractive summarization." Proceedings of the ACL 2005 on Interactive poster and demonstration sessions. Association for Computational Linguistics, 2005.

[24] Mihalcea, Rada, and Paul Tarau. "A language independent algorithm for single and multiple document summarization." (2005).

[25] Mihalcea, Rada. "Graph-based ranking algorithms for sentence extraction, applied to text summarization." Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. Association for Computational Linguistics, 2004.

[26] Mihalcea, Rada, and P. T. Textrank. "Bringing order into texts." Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2004.

[27] Page, Lawrence; Brin, Sergey; Motwani, Rajeev and Winograd, Terry (1999). "The PageRank citation ranking: Bringing order to the Web". , published as a technical report on January 29, 1998.

[28] Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.

[29] Pardo, Thiago Alexandre Salgueiro, and Lucia Helena Machado Rino. TeMario: a corpus for automatic text summarization. Technical report, NILC-TR-03-09, 2003.

[30] van Dijk, Teun Adrianus (ed) 1985. Handbook of discourse analysis. 4 vols, Academic Press, London.

[31] Van Dijk, Teun Adrianus. Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition. Lawrence Erlbaum Associates, 1980.

[32] Van Dijk, Teun Adrianus, Walter Kintsch, and Teun Adrianus Van Dijk. Strategies of discourse comprehension. New York: Academic Press, 1983.

[33] van Dijk, Teun Adrianus. Some aspects of text grammars: A study in theoretical linguistics and poetics. Vol. 63. Mouton, 1972.

[34] Wan, Stephen, and Kathy McKeown. "Generating overview summaries of ongoing email thread discussions." Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004.

[35] Winograd, Peter N. "Strategic difficulties in summarizing texts." Reading Research Quarterly (1984): 404-425.

[36] Zhou, Liang, Chin-Yew Lin, and Eduard Hovy. "Summarizing answers for complicated questions." Proceedings of the 5th International Conference on LREC, Genoa, Italy. 2006.

[37] Zhou, Liang, et al. "Paraeval: Using paraphrases to evaluate summaries automatically." Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Association for Computational Linguistics, 2006.

[38] Zhou, Liang, and Eduard Hovy. "A web-trained extraction summarization system." Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003.

Appendices

A Examples of summaries

Example of one full document in the argumentative genre and its human-generated model summary. Additionally, the three baseline systems (BF, BL and BR), the three weighted PageRank-based systems (PF, PB and PU) and the three Personalized PageRank-based systems (PPF, PPB and PPU) are shown.

In the summaries, we can see the first 25 words in blue, the next 25 words in green (so that we can see what is included in a 50-word summary evaluation) and, finally, 50 words in magenta (which summed with the previous coloured words show what a 100-word summary consists of). The letters in black were not considered in the analysis of the model summaries shown in this thesis.

Full document

Hmm

In several places in the Greater Toronto Area, traffic changes for the Pan Am Games will reduce the number of highway lanes for normal traffic by 25 to 33 per cent (Its Almost Crunch Time June 18). And yes, thats a signal to carpool, take public transit, bicycle, or drive an electric car, but those admirable results require more planning and education than making most drivers miserable for a month.

Wont road frustration increase, and high-occupancy lane compliance decrease?

Torontos Gardiner Expressway may make the notoriously slow Don Valley Parkway look like a speedway except that much of the latter also loses a lane for normal traffic. Often known as the Don Valley Parking Lot, maybe it should just hang out signs that say: Full.

Let the fun and games begin.

Model summary

Hmm

In several places in the Greater Toronto Area, traffic changes for the Pan AM Games will reduce the number of highway lanes for normal traffic considerably. It is suggested to use other means of transportation instead, but this requires planning. Car drivers will be miserable and creates frustration. Torontos Gardiner Expressway may make the notoriously slow Don Valley Parkway look like a speedway except that much of the latter also loses a lane for normal traffic.

Baseline First

Hmm

In several places in the Greater Toronto Area, traffic changes for the Pan Am Games will reduce the number of highway lanes for normal traffic by 25 to 33 per cent (Its Almost Crunch Time June 18). And yes, thats a signal to carpool, take public transit, bicycle, or drive an electric car, but those admirable results require more planning and education than making most drivers miserable for a month.

Wont road frustration increase, and high-occupancy lane compliance decrease?

Torontos Gardiner Expressway may make the notoriously slow Don Valley Parkway look like a speedway except that much of the latter also loses a lane for normal traffic. Often known as the Don Valley Parking Lot, maybe it should just hang out signs that say: Full.

Let the fun and games begin.

Baseline Last

Let the fun and games begin.

Often known as the Don Valley Parking Lot, maybe it should just hang out signs that say: Full.

Torontos Gardiner Expressway may make the notoriously slow Don Valley Parkway look like a speedway except that much of the latter also loses a lane for normal traffic.

Wont road frustration increase, and high-occupancy lane compliance decrease?

And yes, thats a signal to carpool, take public transit, bicycle, or drive an electric car, but those admirable results require more planning and education than making most drivers miserable for a month.

In several places in the Greater Toronto Area, traffic changes for the Pan Am Games will reduce the number of highway lanes for normal traffic by 25 to 33 per cent (Its Almost Crunch Time June 18).

Hmm

Baseline Random

In several places in the Greater Toronto Area, traffic changes for the Pan Am Games will reduce the number of highway lanes for normal traffic by 25 to 33 per cent (Its Almost Crunch Time June 18).

Torontos Gardiner Expressway may make the notoriously slow Don Valley Parkway look like a speedway except that much of the latter also loses a lane for normal traffic.

And yes, thats a signal to carpool, take public transit, bicycle, or drive an electric car, but those admirable results require more planning and education than making most drivers miserable for a month.

Hmm

Let the fun and games begin.

Wont road frustration increase, and high-occupancy lane compliance decrease?

Often known as the Don Valley Parking Lot, maybe it should just hang out signs that say: Full.

PageRank Forward

Often known as the Don Valley Parking Lot, maybe it should just hang out signs that say: Full.

Torontos Gardiner Expressway may make the notoriously slow Don Valley Parkway look like a speedway except that much of the latter also loses a lane for normal traffic.

Hmm

Wont road frustration increase, and high-occupancy lane compliance decrease?

In several places in the Greater Toronto Area, traffic changes for the Pan Am Games will reduce the number of highway lanes for normal traffic by 25 to 33 per cent (Its Almost Crunch Time June 18).

And yes, thats a signal to carpool, take public transit, bicycle, or drive an electric car, but those admirable results require more planning and education than making most drivers miserable for a month.

Let the fun and games begin.

PageRank Backward

In several places in the Greater Toronto Area, traffic changes for the Pan Am Games will reduce the number of highway lanes for normal traffic by 25 to 33 per cent (Its Almost Crunch Time June 18).
Torontos Gardiner Expressway may make the notoriously slow Don Valley Parkway look like a speedway except that much of the latter also loses a lane for normal traffic.
Wont road frustration increase, and high-occupancy lane compliance decrease?
Hmm
Often known as the Don Valley Parking Lot, maybe it should just hang out signs that say: Full.
And yes, thats a signal to carpool, take public transit, bicycle, or drive an electric car, but those admirable results require more planning and education than making most drivers miserable for a month.

Let the fun and games begin.

PageRank Undirected

Torontos Gardiner Expressway may make the notoriously slow Don Valley Parkway look like a speedway except that much of the latter also loses a lane for normal traffic.
In several places in the Greater Toronto Area, traffic changes for the Pan Am Games will reduce the number of highway lanes for normal traffic by 25 to 33 per cent (Its Almost Crunch Time June 18).
Hmm
And yes, thats a signal to carpool, take public transit, bicycle, or drive an electric car, but those admirable results require more planning and education than making most drivers miserable for a month.
Let the fun and games begin.
Often known as the Don Valley Parking Lot, maybe it should just hang out signs that say: Full.
Wont road frustration increase, and high-occupancy lane compliance decrease?

Personalized PageRank Forward

Often known as the Don Valley Parking Lot, maybe it should just hang out signs that say: Full.
Torontos Gardiner Expressway may make the notoriously slow Don Valley Parkway look like a speedway except that much of the latter also loses a lane for normal traffic.
Let the fun and games begin.
Wont road frustration increase, and high-occupancy lane compliance decrease?
In several places in the Greater Toronto Area, traffic changes for the Pan Am Games will reduce the number of highway lanes for normal traffic by 25 to 33 per cent (Its Almost Crunch Time June 18).

And yes, thats a signal to carpool, take public transit, bicycle, or drive an electric car, but those admirable results require more planning and education than making most drivers miserable for a month.

Hmm

Personalized PageRank Backward

Torontos Gardiner Expressway may make the notoriously slow Don Valley Parkway look like a speedway except that much of the latter also loses a lane for normal traffic.
In several places in the Greater Toronto Area, traffic changes for the Pan Am Games will

reduce the number of highway lanes for normal traffic by 25 to 33 per cent (Its Almost Crunch Time June 18).

Often known as the Don Valley Parking Lot, maybe it should just hang out signs that say: Full.

Let the fun and games begin.

Wont road frustration increase, and high-occupancy lane compliance decrease?

And yes, thats a signal to carpool, take public transit, bicycle, or drive an electric car, but those admirable results require more planning and education than making most drivers miserable for a month.

Hmm

Personalized PageRank Undirected

Let the fun and games begin.

Torontos Gardiner Expressway may make the notoriously slow Don Valley Parkway look like a speedway except that much of the latter also loses a lane for normal traffic.

Often known as the Don Valley Parking Lot, maybe it should just hang out signs that say: Full.

In several places in the Greater Toronto Area, traffic changes for the Pan Am Games will reduce the number of highway lanes for normal traffic by 25 to 33 per cent (Its Almost Crunch Time June 18).

Wont road frustration increase, and high-occupancy lane compliance decrease?

And yes, thats a signal to carpool, take public transit, bicycle, or drive an electric car, but those admirable results require more planning and education than making most drivers miserable for a month.

Hmm

B ROUGE-1 Danish Results

DANISH	length	BF	BL	BR	PF	PB	PU	PPF	PPB	PPU
	#words	25 50 100	25 50 100	25 50 100	25 50 100	25 50 100	25 50 100	25 50 100	25 50 100	25 50 100
Descriptive	full doc	**.17** .21 **.29**	.10 .15 .21	.13 .15 .21	.07 .18 .26	.16 **.23 .29**	.15 .20 .27	.13 . 19 .26	.16 **.23** .28	.14 .18 .25
	1 sent	.04 .02 .02	.07 .06 .04	.07 .06 .04	.08 .07 .05	.09 .08 .06	.15 .15 .12			
	no title	.15 .20 .28	-	-	.12 .18 .26	.16 **.23 .29**	.15 .20 .27			
	2/3	-	-	-	.12 .17 .24	.13 .18 .25	.15 .20 .26			
Narrative	full doc	**.17 .23 .29**	.15 .20 .27	.12 .18 .26	.15 .19 .28	.15 .20 .28	.14 .20 .26	.15 . 19 . 28	.14 .20 .28	.13 .20 .26
	1 sent	.05 .03 .02	.10 .10 .09	.08 .08 .07	.12 .13 .12	.12 .14 .12	.13 .18 .20			
	no title	.14 .21 .27	-	-	.15 .19 .28	.14 .20 .28	.14 .20 .26			
	2/3	-	-	-	.12 .20 .28	.15 .20 .27	.14 .20 .28			
Expository	full doc	**.47 .53 .65**	.17 .24 .39	.20 .31 .50	.18 .26 .42	.36 .48 .61	.22 .32 .47	.18 .27 .46	.39 .50 .63	.24 .34 .52
	1 sent	.19 .14 .14	.12 .11 .11	.14 .13 .12	.15 .14 .13	.20 .17 .16	.21 .24 .24			
	no title	.38 .44 .58	-	-	.18 .26 .42	.33 .43 .57	.22 .32 .46			
	2/3	-	-	-	.16 .23 .38	.18 .26 .41	.16 .25 .39			
Argument	full doc	**.37 .48 .59**	.23 .32 .47	.24 .35 .51	.24 .34 .48	.32 .44 .56	.27 .37 .51	.24 .33 .49	.30 .44 .56	.27 .28 .54
	1 sent	.09 .09 .09	.17 .16 .16	.13 .12 .16	.18 .17 .17	.18 .17 .17	.26 .25 .25			
	no title	.33 .43 .55	-	-	.23 .33 .46	.28 .41 .53	.27 .36 .50			
	2/3	-	-	-	.23 .33 .46	.25 .34 .46	.23 .32 .46			
Mixed	full doc	**.28 .30 .37**	.14 .21 .29	.12 .19 .27	.15 .22 .31	.21 .26 .36	.15 .22 .30	.15 .22 .31	.21 .27 .35	.16 .22 .31
	1 sent	.11 .08 .06	.10 .11 .09	.05 .05 .05	.14 .15 .12	.16 .14 .12	.15 .21 .22			
	no title	.22 .26 .33	-	-	.15 .22 .31	.20 .25 .35	.15 .22 .30			
	2/3	-	-	-	.15 .22 .30	.14 .19 .27	.15 .21 .28			

Table 23: Average ROUGE-1 25, 50 and 100-word summary recall scores for every genre in Danish. In this table we can see the three baselines (BF, BL and BR) and the three Weighted PageRank-based systems. Additionally, we can see the ROUGE-scores for the full document, only taking into consideration one sentence (the first one, the last one or a random one), the first 25, 50 and 100-words without the title and the PageRank ROUGE-scores for the last two thirds of the document.

C ROUGE-1 English Results

ENGLISH	length	BF	BL	BR	PF	PB	PU	PPF	PPB	PPU
	#words	25 50 100	25 50 100	25 50 100	25 50 100	25 50 100	25 50 100	25 50 100	25 50 100	25 50 100
Descriptive	full doc	**.18** .22 .28	.14 .18 .23	.16 .22 .28	.15 .20 .26	**.18 .23 .29**	.16 .22 .28	.15 .20 .26	**.18 .23 .29**	.17 .22 .28
	1 sent	.05 .03 .02	.13 .11 .10	.10 .09 .06	.16 .15 .14	.16 .15 .11	.21 **.24** .26			
	no title	**.23** .27 **.34**	-	-	**.18 .24 .30**	**.23 .28 .35**	.21 **.25 .32**			
	2/3	-	-	-	.19 **.24 .30**	.19 **.25 .32**	.20 **.25 .32**			
Narrative	full doc	**.22 .26 .30**	.17 .23 .28	.15 .21 .29	.16 .22 .29	19 .23 .29	.17 .23 .29	.16 .22 .29	.19 .23 .29	.16 .23 .29
	1 sent	.10 .08 .05	.11 .10 .08	.09 .08 .06	.14 .14 .11	.14 .13 .10	.16 .19 .18			
	no title	.18 .23 .29	-	-	.16 .22 .29	.28 .23 .29	.16 .22 .29			
	2/3	-	-	-	.15 .22 **.36**	.16 .22 **.38**	.17 .23 **.30**			
Expository	full doc	**.48 .58 .66**	.14 .20 .32	.17 .26 .43	.14 .22 .36	.22 .47 .59	.19 .21 .35	.15 .24 .37	.46 **.59 .67**	.27 .38 .54
	1 sent	.24 .21 .20	.09 .08 .08	.11 .11 .09	.10 .10 .09	.22 .20 .19	.19 .20 .18			
	no title	.41 .49 .58	-	-	.14 .22 .36	.41 .50 .59	.18 .33 .48			
	2/3	-	-	-	.15 .23 .36	.16 .25 .38	.17 .25 .37			
Argument	full doc	**.39 .46 .42**	.26 .31 .41	.34 .43 .41	.30 .43 .41	.35 .44 .42	.34 .42 **.42**	.35 **.56 .73**	.38 **.58 .74**	.37 **.57 .73**
	1 sent	.07 .06 .06	.22 .26 .26	.22 .26 .26	.20 .25 .24	.22 .25 .24	.31 .37 .37			
	no title	**.39 .57 .71**	-	-	.32 **.55 .71**	.35 **.57 .70**	.35 **.54 .71**			
	2/3	-	-	-	.31 **.54 .68**	.34 **.54 .68**	.36 **.55 .67**			
Mixed	full doc	**.28 .30 .34**	.16 .24 .30	.15 .22 .29	.17 .19 .25	.23 .27 .33	.17 .22 .27	.19 .25 .32	.23 .27 **.34**	.18 .22 .27
	1 sent	.15 .11 .07	.15 .15 .11	.11 .11 .08	.18 .17 .12	.18 .14 .10	.17 .22 .21			
	no title	.20 .25 .32	-	-	.19 .25 .32	.21 .25 .32	.17 .22 .27			
	2/3	-	-	-	.19 .25 .32	.16 .22 .28	.16 .21 .25			

Table 24: Average ROUGE-1 25, 50 and 100-word summary recall scores for every genre in English. In this table we can see the three baselines (BF, BL and BR) and the three Weighted PageRank-based systems. Additionally, we can see the ROUGE-scores for the full document, only regarding one sentence (the first one, the last one or a random one), the first 25, 50 and 100-words without the title and the PageRank ROUGE-scores for the last two thirds of the document.

D ROUGE-1 Spanish Results

ROUGE-1	length	BF	BL	BR	PF	PB	PU	PPF	PPB	PPU
	#words	25 50 100	25 50 100	25 50 100	25 50 100	25 50 100	25 50 100	25 50 100	25 50 100	25 50 100
Descriptive	full doc	**.24 .29 .35**	.15 **.29 .35**	.18 .25 .33	.18 .25 .30	.23 **.29 .35**	.21 .25 .32	.19 .24 .31	.23 .28 **.35**	.20 .25 .33
	1 sent	.05 .03 .02	.13 .11 .10	.10 .11 .10	.16 .15 .14	.16 .15 .11	.21 .24 .26			
	no title	.23 .27 .34	-	-	.18 .24 .30	.23 .28 **.35**	.21 .25 .32			
	2/3	-	-	-	.20 .26 .33	.19 .25 .32	.20 .25 .32			
Narrative	full doc	**.26 .30 .35**	.13 .19 .32	.18 .25 .33	.18 .21 .27	.22 .29 .34	.20 .27 .34	.21 .27 .33	.23 .28 .34	.20 .26 .34
	1 sent	.05 .04 .03	.14 .13 .10	.10 .08 .06	.18 .18 .14	.17 .16 .12	.20 .24 .24			
	no title	.40 .48 .58	-	-	.21 .27 .33	.21 .28 .34	.20 .27 .34			
	2/3	-	-	-	.20 .26 .33	.19 .25 .32	.20 .27 .34			
Expository	full doc	**.45 .52 .63**	.20 .27 .35	.23 .31 .43	.21 .27 .37	.37 .45 .60	.23 .33 .47	.22 .26 .37	.38 .46 .62	.27 .36 .49
	1 sent	.21 .15 .12	.16 .15. .12	.18 .17 .14	.18 .17 .14	.26 .19 .15	.23 .29 .25			
	no title	.40 .48 .58	-	-	.21 .27 .36	.35 .44 .55	.25 .34 .45			
	2/3	-	-	-	.21 .37 .36	.25 .33 .44	.23 .30 .41			
Argument	full doc	.34 .47 .70	.28 .47 .67	.29 .46 .70	**.35 .50** .71	.33 .46 -70	.34 .48 **.72**	.33 .49 .71	.31 .44 .70	.30 .45 .70
	1 sent	.09 .07 .07	.19 .22 .22	.19 .17 .16	.24 .27 .27	.21 .21 .22	.32 .37 .38			
	no title	.30 .43 .66	-	-	.32 **.50** .68	.30 **.46** .67	.32 .47 .69			
	2/3	-	-	-	.32 **.50** .67	.21 **.46** .65	.34 **.49** .67			
Mixed	full doc	**.34 .38 .41**	.21 .28 .36	.22 .27 .36	.22 .30 .37	.29 .34 .39	.20 .26 .33	.22 .30 .37	.29 .34 .39	.22 .27 .35
	1 sent	.26 .28 .11	.19 .17 .12	.18 .20 .15	.21 .23 .18	.26 .22 .16	.20 .26 .26			
	no title	.23 .29 .36	-	-	.22 .30 .37	.23 .30 .36	.20 .26 .33			
	2/3	-	-	-	.22 .30 .37	.23 .29 .35	.20 .26 .34			

Table 25: Average ROUGE-1 25, 50 and 100-word summary recall scores for every genre in Spanish. In this table we can see the three baselines (BF, BL and BR) and the three Weighted PageRank-based systems. Additionally, we can see the ROUGE-scores for the full document, only taking into account one sentence (the first one, the last one or a random one), the first 25, 50 and 100-words without the title and the PageRank ROUGE-scores for the last two thirds of the document.