



INSTITUTO POLITÉCNICO NACIONAL

CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

Multi-document Summarization using
Intra-document and Inter-document
Redundancy

TESIS

Que para obtener el grado de:
Maestro en Ciencias de la Computación.

P R E S E N T A :
Ing. Pabel Carrillo Mendoza

DIRECTORES DE TESIS.
Dr. Francisco Hiram Calvo Castro
Dr. Alexander Gelbukh



Ciudad de México, Diciembre 2016



INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

SIP-14 bis

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México siendo las 10:00 horas del día 30 del mes de noviembre de 2016 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Estudios de Posgrado e Investigación del:

Centro de Investigación en Computación

para examinar la tesis titulada:

"Multi-document Summarization using Intra-document and Inter-document Redundancy"

Presentada por el alumno:

CARRILLO

Apellido paterno

MENDOZA

Apellido materno

PABEL

Nombre(s)

Con registro:

B	1	4	0	4	5	6
---	---	---	---	---	---	---

aspirante de: **MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN**

Después de intercambiar opiniones los miembros de la Comisión manifestaron **APROBAR LA TESIS**, en virtud de que satisface los requisitos señalados por las disposiciones reglamentarias vigentes.

LA COMISIÓN REVISORA

Directores de Tesis

Dr. Alexander Gelbukh

Dr. Francisco Hiram Calvo Castro

Dr. Sergio Suárez Guerra

Dr. Grigori Sidorov

Dra. Olga Kolesnikova

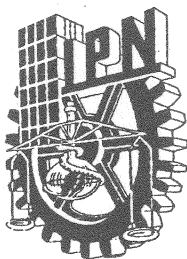
Dr. Udar Batyrshin

PRESIDENTE DEL COLEGIO DE PROFESORES

Dr. Marco Antonio Ramírez Salinas



INSTITUTO POLITÉCNICO NACIONAL
CENTRO DE INVESTIGACIÓN
EN COMPUTACIÓN
DIRECCIÓN




INSTITUTO POLITÉCNICO NACIONAL
SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA CESIÓN DE DERECHOS

En la Ciudad de México el día 07 del mes de diciembre del año 2016, el que suscribe Pabel Carrillo Mendoza alumno del Programa de Maestría en Ciencias de la Computación con número de registro B140456, adscrito al Centro de Investigación en Computación, manifiesta que es autor intelectual del presente trabajo de Tesis bajo la dirección del Dr. Alexander Gelbukh y el Dr. Francisco Hiram Calvo Castro y cede los derechos del trabajo intitulado "Multi-document Summarization using Intra-document and Inter-document Redundancy", al Instituto Politécnico Nacional para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o director del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección pabel.cm@gmail.com. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.


Pabel Carrillo Mendoza

Nombre y firma

“Science is but a perversion of itself unless it has as its ultimate goal the betterment of humanity.”

Nikola Tesla

Resumen

Al buscar información en Internet, es común enfrentarse con el problema de manejar la masiva cantidad de información que existe en la Web. Gran parte de esta información es redundante, lo que nos complica la asimilación del contenido de un grupo de textos dentro de los cuales buscamos un tópico de interés. A parte de ser un problema, se ha dicho que la gran cantidad de información es altamente valiosa. De ahí que esta información deba ser manejada cuidadosamente, tomándola en cuenta como información relevante.

Con el objetivo de atacar este problema, surge la tarea de la generación automática de resúmenes multi-documento. Esta tarea busca lidiar con esta abundancia de información buscando métodos para automáticamente extraer las ideas más importantes de varias fuentes, y presentarlas en un resumen. De entre los diferentes tipos de resúmenes que existen, nosotros estudiamos los resúmenes genéricos, los cuales tienen el propósito de mostrar un panorama general sobre lo que abordan estos textos, y así dar un rápido entendimiento sobre la información que está siendo cubierta.

En este escenario, la importancia de la información está determinada sólo con respecto al contenido de los textos. En este caso ¿qué debe considerarse como importante para mostrar al usuario? En este trabajo, mostramos cómo la cantidad de redundancia, por documento y entre documentos, puede ser usada para determinar la importancia de la información. Por ejemplo, una idea redundante entre documentos podría ser importante por su popularidad, también una idea no redundante entre documentos podría ser importante por su novedad, o también una idea redundante por documento podría ser importante al ser constantemente mencionada por el mismo autor.

Para hacer esto, desarrollamos una técnica no supervisada basada en grafos que, empleando las medidas de similitud apropiadas, puede usar la redundancia por documento y entre documentos para generar automáticamente resúmenes de múltiples documentos, dando la flexibilidad de extraer ya sea información popular o rara. La aplicación de esta técnica a los conjuntos de documentos de DUC y usando el método de evaluación ROUGE, nos permitió establecer qué hace a una idea importante para poder formar parte de un resumen en el ámbito de noticias, de acuerdo a la redundancia, y a demostrar que el uso de la redundancia ayuda a mejorar los resúmenes.

Abstract

As searchers of information, we face everyday the problem of handling massive information in the Internet. Much of this information is redundant, which complicates us the assimilation of the content of a big group of texts when looking for a specific topic. Apart from being a problem to deal with, something about the huge amount of data has been stated in the last years: it is highly valuable. Therefore, it should be carefully handled by taking it into account as relevant information.

To solve this problem, the task of multi-document summarization emerges. This task seeks to deal with this affluence of information looking for methods to automatically extract the most important ideas from multiple sources, which talk about the same topic, and present them into a final summary. Among different kinds of summaries, we studied generic summaries, which have the purpose of showing a general panorama of what the texts talk about to provide a rapid understanding of the information that is being covered.

In this setting, the importance of information is determined only with regard to the content of the input itself. So, in this case what should one consider important to show to the user? In this work, we show how the amount of redundancy per-document and across documents can be used to determine the importance of the information. For instance, an idea redundant between documents could be important because of its popularity, also an idea that is not redundant between documents could be important by its novelty, or also an idea redundant per document could be important by being constantly addressed by the same author. In this way, the massive information existent in the Internet work as a resource to find important information.

For this purpose, we developed an unsupervised graph-based technique that, based on proper similarity measures, can use intra-document and inter-document redundancy to automatically summarize multiple documents, providing the flexibility to extract either popular or rare information. Studying the application of this technique on the DUC corpora and the ROUGE evaluation method, led us to establish what makes an idea important to be part of a summary in the ambit of news, according to redundancy, and to prove that the use of redundancy helps to improve summaries.

Acknowledgements

I was fortunate to enjoy all my time during my master's degree. The research experience has been enormously exciting to me. I owe this to many persons and organizations that I would like to thank.

To my advisors Dr. Hiram Calvo and Dr. Alexander Gelbukh for their support, their trust in me, and the great ideas they provided for the development of this work. Also for the motivation and valuable teaching they have left in me to get involved in the ambit of research.

To the Computing Research Center (CIC IPN) for their efforts to ameliorate the research on Computer Science, not only in Mexico, but internationally, and to the researches of the center by their enthusiasm to share and grow the knowledge of the area. Also, to the National Council of Science and Technology (CONACYT) for the scholarship they provided me during my master's studies, giving me the opportunity to invest my full time in acquiring knowledge and applying it to the development of ideas, as the one showed in this work.

To the Mexican Society for Artificial Intelligence (SMIA) for the support to attend conferences on Artificial Intelligence. Also to the Thematic Network on Language Technologies (Red TTL, Mexico) for their initiatives on organizing and facilitating the attendance to workshops and meetings on Natural Language Processing in Mexico, and for the support they gave me to make a short research stay in the National Institute of Astrophysics, Optics and Electronics (INAOE) at Tonantzintla, Puebla.

To Dr. Luis Villaseñor-Pineda and Dr. Manuel Montes-y-Gómez, along with all the fellows of the Laboratory of Language Technologies, for receiving me at the INAOE and sharing knowledge and ideas for the improvements of our works of research.

To my parents, brother and sister for supporting me always and inspiring me with their examples of hard work. Also to my friends for feeding my soul, and for being there for the discussion of ideas.

Contents

Resumen	v
Abstract	vi
Acknowledgements	vii
List of figures	xi
List of tables	xii
1 Introduction	1
1.1 The problem	1
1.2 Hypothesis	2
1.3 Objectives	2
1.3.1 General objective	3
1.3.2 Specific objectives	3
1.4 Novelty	3
1.5 Pertinence	4
1.6 Structure of the document	4
2 Theoretical background	5
2.1 Summarization	5
2.2 Multi-document summarization	6
2.3 Categorization of summaries	6
2.3.1 By output	6
2.3.2 By language	7
2.3.3 By content	7
2.3.4 By goal	7
2.4 General extractive summarization process	8
2.5 Similarity measures	9
2.5.1 Superposition measure	9
2.5.2 Vector space model and cosine similarity	9

2.5.2.1	Bag-of-words with tf-idf	10
2.5.2.2	Paragraph vector (doc2vec)	11
2.5.2.3	Cosine similarity measure	15
2.6	Text corpora	16
2.7	Evaluation methods	17
2.7.1	Recall, precision and F-measure	18
2.7.2	ROUGE	19
2.7.2.1	ROUGE-N: n-gram co-occurrence statistics	19
2.7.2.2	ROUGE-L: longest common subsequence	20
2.7.2.3	ROUGE-W: weighted longest common subsequence	21
2.7.2.4	ROUGE-S: skip-bigram co-occurrence statistics	21
3	State of the art	23
3.1	Baseline methods	23
3.1.1	Random	24
3.1.2	Lead	24
3.1.3	Coverage	24
3.2	Traditional approaches	24
3.2.1	Feature-based approaches	25
3.2.2	Cluster-based approaches	26
3.2.3	Graph-based approaches	27
3.2.4	Knowledge-based approaches	28
3.3	Best approaches	30
3.4	How different methods deal with redundancy	33
4	Our proposal	36
4.1	Sentence extraction	37
4.2	Relevance search	38
4.3	Summary generation	43
5	Experiments and results	45
5.1	Similarity measures comparison	45
5.2	Experimental settings	47
5.3	Results	54
5.4	Comparison with existent methods	58
5.5	Discussion	59
6	Conclusions	62
6.1	Conclusions	62
6.2	Contributions	63
6.3	Publication	64
6.4	Future work	64
A	An example of the merging process	66

A.1	Cluster of texts	66
A.2	Merging process	82
A.3	Reference summaries	86
References		88

List of figures

2.1	Extractive summarization process	8
2.2	Distributed Memory Model of Paragraph Vectors (PV-DM), borrowed from [1] .	12
2.3	The projection layer of the PV-DM model	13
2.4	Distributed Bag of Words version of Paragraph Vector (PV-DBOW), borrowed from [1]	14
2.5	Cosine similarity measure	15
3.1	Cluster-based method for summarization, borrowed from [2]	27
3.2	Traditional graph representation of texts, borrowed from [3]	28
3.3	Graph representation after establishing a threshold for edge generation, borrowed from [3]	29
4.1	Complete graph representation of sentences from a cluster of texts	37
4.2	Graph representation after weighting edges	39
4.3	Strategies to define the importance of an idea according to redundancy	40
4.4	Merging process example of two nodes	43
5.1	Histograms illustrating graph edges distribution of the <i>d31033t</i> cluster from DUC 2004, built with the doc2vec-based similarity measure	48
5.2	Histograms illustrating graph edges distribution of the <i>d31033t</i> cluster from DUC 2004, built with the superposition similarity measure	49
5.3	Histograms illustrating graph edges distribution of the <i>D0741I</i> cluster from DUC 2007, built with the doc2vec-based similarity measure	51
5.4	Histograms illustrating graph edges distribution of the <i>D0741I</i> cluster from DUC 2007, built with the superposition similarity measure	52
A.1	Histograms illustrating graph edges distribution of the <i>d30027t</i> cluster from DUC 2004, built with the doc2vec-based similarity measure	83

List of tables

3.1	Best approaches for DUC 2004 task 2	31
3.2	Best approaches for DUC 2007 main task	32
5.1	Sentences from DUC 2002 used to test the similarity measures	46
5.2	Similarity values between sentences used for comparison	47
5.3	Specifications of the DUC 2004 and DUC 2007 corpora	47
5.4	Command line parameters used for execution of ROUGE	54
5.5	ROUGE results on all strategies of redundancy for DUC 2004 texts when using the doc2vec-based similarity measure	55
5.6	ROUGE results on all strategies of redundancy for DUC 2004 texts when using the superposition similarity measure	56
5.7	ROUGE results on all strategies of redundancy for DUC 2007 texts when using the doc2vec-based similarity measure	57
5.8	ROUGE results on all strategies of redundancy for DUC 2007 texts when using the superposition similarity measure	57
5.9	Comparison with baselines and best methods using DUC 2004 corpus	59
5.10	Comparison with baselines and best methods using DUC 2007 corpus	59
5.11	ROUGE-1 results on all strategies of redundancy for DUC 2004, using doc2vec, showing a 95% confidence interval	60
5.12	ROUGE-1 results on all strategies of redundancy for DUC 2007, using doc2vec, showing a 95% confidence interval	61
A.1	Merging process for the $d30027t$ cluster when $pd = 1$ and $cd = 1$	84
A.2	Merging process for the $d30027t$ cluster when $pd = -1$ and $cd = 1$	84

Chapter 1

Introduction

The purpose of this chapter is to highlight the motivation and vision of this work. We first present the problem that motivated this research. Afterwards, we pose an hypothesis with a novel idea to solve the problem, that later led us to establish the objectives that guided this work. Some facts about the novelty and pertinence of the work are also presented, to get the reader interested in this work.

1.1 The problem

A general problem that exists in the world is the difficulty of handling massive information in the Internet, which is mainly caused by its easy proliferation. As stated by different media, the growth of unstructured information in the last years has been massive. In 2013, it was reported that 1 Exabyte (10^{18}) of data was being created in the Internet daily, roughly the equivalent of data in 250 million DVDs¹. This means that by 2013, humankind was producing in two days the same amount of data it took from the dawn of civilization until 2003 to generate.

Nevertheless, much of the information found in the Internet is redundant. As searchers of information, we face this problem everyday when is necessary to digest the content on multiple sources when looking for a specific topic. This phenomenon is in part caused by the freedom to generate information on the web without *any* control on its content, which in general is a good thing. However, tools to filter this redundancy would be of great help to people who needs to digest all this information.

¹<http://www.wired.com/insights/2013/04/with-big-data-context-is-a-big-issue/>

Apart from being a problem to deal with, something about the huge amount of data has been stated in the last years: it is highly valuable. Therefore, redundancy on the Internet has two faces: it is a problem, but it is also potentially relevant; it should be carefully handled by taking it into account as valuable information.

Related with this problem of redundancy, the task of summarization, and more specifically, multi-document summarization emerge. Multi-document summarization seeks to deal with this affluence of information looking for methods to automatically extract the most important ideas from multiple sources and present them into a final summary. There are other tasks focused on understanding massive information such with visualization methods, but usually they are attached to the specific problem that the visualization method is studying. A general way to attack this problem on unstructured data such as texts is with summarization.

For this task, a problem comes ahead when methods to give relevance to an idea are explored. What makes an idea important? We could not found any generic answer, instead of that, every approach is inspired in a human idea of what is important, giving a direction on the method that every person considers appropriate. And here is where we want to study the possibility of giving importance to an idea by its level of redundancy. In this way, this massive redundancy would be of great help at assigning relevance to an idea by being popular or by being rare, given its level of redundancy in the texts.

Rare information, which can also be seen as novel, diverse, or non-redundant information, is not valuable according to the intuition of what a summary should contain. However, this does not eliminate the fact the this information could be valuable under certain areas.

1.2 Hypothesis

To solve the problem discussed above, we pose the following hypothesis: It is possible to provide a quick overview of the information that is covered by multiple texts by extracting and presenting the most popular or the most rare ideas that are addressed. In this way the user can get a rapid understanding on the general topic that is being covered.

1.3 Objectives

Motivated to prove the previous hypothesis, we established the following general and specific objectives for this work.

1.3.1 General objective

Develop an unsupervised method to automatically summarize multiple documents, providing the flexibility of extracting either popular or rare information from these documents, and present a general idea of what these multiple sources are covering.

1.3.2 Specific objectives

To achieve this goal, the following specific objectives were proposed:

- Find an appropriate graph representation to map the sentences of the texts and reflect the redundancy found in these texts.
- Develop a method to compute the relevance of the sentences based on the redundancy represented in the graph.
- Build the final summary based on the most important sentences.
- Explore similarity measures to compare sentences and find redundancy.
- Evaluate the results of the method using well-known methods and resources.

1.4 Novelty

What differentiates this work from others is the way we deal with redundancy. We first analyzed all possible cases of redundancy that can exist per document and across documents within multiple texts. This later allowed us to develop a mechanism to extract different summaries—on the same group of documents—containing ideas generated by experimenting with intra-document and inter-document redundancy, such as: ideas redundant per author, ideas redundant by different authors, etc. Studying the application and evaluation of this mechanism on the corpora and evaluation methods used by the state-of-the-art methods—which pursue to obtain the highest scores with this resources—led us to establish a general statement about what makes an idea important to belong to a summary, according to redundancy.

This method can also serve as a mechanism to eliminate redundancy in the final step of the summary generation process of any other method where a ranking of important sentences is generated. The adoption of this method can help to the improvement of the final summary by taking the redundancy as a resource to find important information.

1.5 Pertinence

The task of multi-document summarization is rich in the many existent ways to solve the problems behind this task. For giving relevance to sentences in extractive summarization, for example, simply methods such as taking the sentences with the most frequent words along the whole document, and advanced methods using supervised or unsupervised learning, genetic algorithms, graph representations, pattern recognition, or linguistic approaches, have been proposed. Also this richness exists at creating methods to assemble the final summary.

Much of what is possible to learn in the area of Computer Science could be applied in this task, and this is how it has been done. Thereby, to understand the methods of the state of the art all this knowledge is necessary, in addition to the great work that is required to get a general panorama and know how to contribute with something valuable in this area that has been studied since 1958. Behind the proposal showed in this document many ideas have been explored for each challenge of this task, however, we present only those ideas that allowed to generate this final method as a whole.

1.6 Structure of the document

This chapter introduced some information to get a general vision of the work of this thesis. Chapter 2 provides a theoretical background to understand the proposed method. Chapter 3 provides a general panorama of other works, specifically it explains how those works handle redundancy when generating multi-document summaries, along with a general explanation of the methods with the best results reported so far. Chapter 4 explains the proposed method of this work in full detail. In Chapter 5, designs of the experiments to test the functionality of the method and their results are presented. Finally, in Chapter 6 the conclusions of the results of the method and the research, together with the future work, are formulated.

Chapter 2

Theoretical background

To understand how to proceed with the automatic generation of summaries, a theoretical background is needed. This chapter presents all this background, trying to go from general to particular. First of all, we provide a definition of summary in the context of NLP and the diversity of kind of summaries that can be generated according to some need, just to know where to focus the efforts when the task of summarization is being approached.

Secondly, considering that we focused our efforts on a specific type of summaries—extractive, generic, monolingual, and informative multi-document summaries—we present the theory that is well established for this kind of summaries. As the generation of extractive summaries has been undertaken since 1958, we can give a general summary generation process.

Finally, some information is provided to understand the techniques and resources that we used to carry out this work, in specific, we present some similarity measures for sentences, text resources and evaluation methods.

2.1 Summarization

In the context of Natural Language Processing, summarization is an automatic task pursuing to take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user’s or application’s needs [4]. Thus, it should condense the source text into a shorter version preserving its information content and overall meaning [2]. The first person proposing this task was Luhn [5] in 1958, generating

the first summaries in an IBM 704 data-processing machine by choosing as important ideas those containing the most frequent words along the complete document.

2.2 Multi-document summarization

It is a special task derived from summarization, but pursuing the extraction of information from multiple texts written about the same, but unspecified topic. Is an automatic task, without any editorial touch or subjective human intervention, thus making it completely unbiased. The distinct characteristics that make multi-document summarization rather different from single-document summarization is that multi-document summarization problem involves multiple sources of information that overlap and supplement each other, being contradictory at occasions [2], and finding more frequent the problem of redundancy. Thus, the final goal is ensuring the final summary to be both coherent and complete.

The work proposed by McKeown and Radev [6] in 1995, was one of the first that undertook this task. From there, the task of multi-document summarization started to gain the attention of researchers, specially working the problem with news. Further, in 2005 McKeown et al. [7] proved that multi-document summaries of news generated automatically enable readers to more effectively complete a fact-gathering task, in comparison to having only a one-sentence summary, i.e., the first sentence of each article of the cluster, or than having no summary at all, i.e., the source documents only.

2.3 Categorization of summaries

As seen above, the importance of the ideas composing the final summary should be sensitive to a need. The importance is not a concept universally well defined, it depends always on something. In such manner, according to some criteria different kind of summaries could be generated. Let us take a look on the classical classification of summaries presented by Nenkova and McKeown [8].

2.3.1 By output

Depending on how the output summary is generated two different kind of summaries are defined:

Extractive summaries (extracts) are produced by concatenating several sentences taken exactly as they appear in the materials being summarized.

Abstractive summaries (abstracts) are written to convey the main information in the input and may reuse phrases or clauses from it, but the summaries are overall expressed in the words of the summary author.

In this work, we focus completely on generating extractive summaries. These kind summaries are enough to accomplish the goal of this research.

2.3.2 By language

Depending on the language of the texts to work with there could be two possible kinds of summaries:

Monolingual summaries are generated from text written on the same language.

Multilingual summaries are generated from text written on different languages.

We focus our work on monolingual summaries.

2.3.3 By content

Related to the content to search on the original document(s), two different summaries exist:

An **indicative** summary enables the reader to determine aboutness, this may provide characteristics such as topics, length, writing style, and so other superficial characteristics to decide if the document serves to the related need of information.

An **informative** summary can be read in place of the document, this include facts that are reported in the input document(s).

We focus our work on informative summaries.

2.3.4 By goal

Depending on what the purpose of the summary is, these are the different types:

Generic summaries are made taking few assumptions about the audience or the goal for generating the summary. Typically, it is assumed that the audience is a general one: anyone may end up reading the summary. Furthermore, no assumptions are made about the genre

or domain of the materials that need to be summarized. In this setting, importance of information is determined only with respect to the content of the input alone. It is further assumed that the summary will help the reader quickly determine what the document is about, possibly

Query-focused summaries on the other hand, takes only the information that is relevant to a specific user query—this is similar to the task of information retrieval. The summarizer tries to find information within the document that is relevant to the query or in some cases, may indicate how much information in the document relates to the query.

Update summaries are sensitive to time; it must convey the important development of an event beyond what the user has already seen.

Opinion summaries are built to give users a quick understanding of the underlying sentiments in the text.

Comparative summaries aim to summarize the differences between comparable document groups. The summary produced for each group should emphasize its differences with other groups.

For this work we pursue to generate generic summaries. In this summaries is not clear what an idea must contain to become important and be part of the summary.

2.4 General extractive summarization process

In the specific case of extractive summarization, there is a common process to construct a final summary, see Figure 2.1.

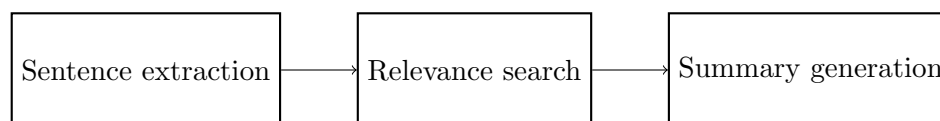


FIGURE 2.1: Extractive summarization process

The first step is to break the text into sentences. This sentences are traditionally delimited by a point, however, not always a point indicates an end of sentence and also other punctuation marks could exist such as an exclamation mark, a question mark, a quote mark, a hyphen, etc. This means that a sentence is not necessary composed by a subject, verb, predicate, or other grammatical elements, but with whatever this extract of text contains until it reaches an end

of sentence punctuation mark. After this, at the second step a method should be used to found which of this sentences are the most relevant—according to the user’s needs. Finally, a method should be used to construct the summary with the most relevant sentences considering that the final summary should be coherent and complete, for this a correct order should be found and the redundancy should be avoided.

2.5 Similarity measures

In this work we study the redundancy on multiple texts. To detect this redundancy is necessary to know if ideas are similar, hence a similarity measure is essential. Two ways of measuring similarity between texts are introduced on this section. The first one is a simple method that measures the overlapping of words between sentences, we call this *superposition*. The second one is a method which has showed good results by giving a good vector representation of a sentence, allowing to measure semantic similarities given by the context of the words on the sentence, this is called *Paragraph Vector*.

2.5.1 Superposition measure

This is a simple measure to quantify the number of overlapping words between two sentences. One way to compute this value is presented by Mihalcea and Tarau [9], see Formula 2.1. They used it to create weighted edges in a graph and then determine the relevance of each node in the graph with a modified version of the PageRank method—to take into account the weight of the edges—obtaining good results.

$$Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \ \& \ w_k \in S_j\}|}{\log|S_i| + \log|S_j|}. \quad (2.1)$$

2.5.2 Vector space model and cosine similarity

Vector space model is used in computer science to formally represent objects in a n-dimensional space. Since this makes possible to have a computational representation for an object in a vector, multiple uses of this arise. In particular, this representation allow us to compare objects with a cosine measure between any pair of vectors. The way it is possible to represent objects in a vector space is by defining the features—or characteristics—of the objects and the possible values that each feature can acquire. Each feature becomes a dimension on the space where the

object is represented. The selection of these features is a task that requires a good effort due to the complexity in finding the characteristics that best represent the object for the specific problem.

When we pass to the domain of extractive multi-document summarization, the objects that need a representation are sentences. A panorama of how to accomplish this representation is given first in the following two subsections: bag-of-words and paragraph vector. Then, the method to compare these vectors with a cosine measure in a n -dimensional space is presented.

2.5.2.1 Bag-of-words with tf-idf

This technique is one of the most well-known methods to find vector representation for documents, which text can be of variable length—from a sentence to a whole text. Bag of words consist in the use of words as features. The value for each feature of a document vector is traditionally related to the frequency of appearance of the word in the document. This representation is mainly used for comparison of documents. In information retrieval for example, this comparison is used between the given query and the collection of documents where the search is performed, in order to obtain the most similar documents to the query. Hence, the bag of words technique will generate similar vectors for documents sharing many words in common.

Starting from a collection of documents, a dictionary of words is built by extracting all different words contained in the collection of documents, which is commonly known as the vocabulary. It is possible to do some pre-processing on the collection of documents to improve the representation, for example, removing stop words—words without important content information such as articles, prepositions, etc.—or lemmatization—words are replaced by their lemmas, the canonical or dictionary form of the word.

From here, every word of the vocabulary corresponds to a feature for all vector representation of documents, and their value is related to the frequency of each word on the document that is being represented. This frequency can be obtained simply counting the appearance of the word on the document, it is called *term frequency* (tf), denoted as

$$tf_{ij}: \text{ the frequency that word } i \text{ appears in document } j. \quad (2.2)$$

To improve this counting, another metric is used: *inverse document frequency* (idf). This is a metric obtained for each word in the dictionary and the idea of this measure is to penalize

the appearance of a word in various documents. The more common a word is in the collection of documents, the lesser it helps on differentiate a document from other. On the other hand, uncommon words can be very valuable features for a document, as they provide a unique characteristic for the document. It is computed as

$$idf_i = \log \frac{N}{DF_i}, \quad (2.3)$$

where N is the number of documents in the collection and DF_i is known as *Document Frequency*, which is equal to the number of documents where the word i appears at least once. To use the tf value in combination with idf , they are multiplied, so that the term frequency of the word is weighted with its *uniqueness* in the collection of documents. In this way they form the $tf-idf$ measure:

$$tf-idf = tf_{ij} \cdot idf_i. \quad (2.4)$$

N-grams

Apart from using words as features, it is possible to use a sequence of words. These sequence of words are called n-grams, where n indicates the quantity of words composing the sequence. In term of *grams*, a word can be called *unigram* since it has only one word, *bigrams* are sequences of two words, and *trigrams* of three words. From here, for naming all the $n > 3$ grams, n in n-gram is replaced by the number, for example: 4-gram (four-gram), 5 (five-gram), etc. In practice, unigrams and bigrams are the most used. This is because it is hard to find different sentences with a high number of sequential words in common.

2.5.2.2 Paragraph vector (doc2vec)

Paragraph vector is a technique proposed by Le and Mikolov [1] as a new way to represent variable-length pieces of text. This technique has showed to outperform bag-of-words in some tasks such as sentiment analysis and text classification. As they said, this can be because bag-of-words representations lose the ordering of the words and they also ignore semantics of the words. Paragraph vector is a technique working with simple neuronal networks that are trained to predict words in a text. It is inspired in a previous work to compute vector representation of words [10], that later was adapted for variable-length pieces of text. From here, we will refer to this variable-length pieces of text as paragraphs, even though, theses pieces can go from documents to sentences.

Two different models are proposed in *Paragraph Vector*: Distributed Memory Model of Paragraph Vectors (PV-DM) and Distributed Bag of Words version of Paragraph Vector (PV-DBOW). Both neural networks are very simple, they do not have hidden layers. The argument of the authors on the success of this method despite its simplicity, is that by training with a huge amount of data, even on a very simple neural network, it could be possible to capture semantic relations between these paragraphs given the context of their words.

In both models, every paragraph has its corresponding unique vector represented by a column in matrix D , and every word has its corresponding unique vector represented by a column in matrix W . This matrix is randomly initialized, but after the training process in the neural network, they can eventually capture semantics as an indirect result of the prediction task. As a result of this training process, vector representations of words and paragraphs are obtained and can be used as features for the paragraph—in lieu of bag-of-words.

PV-DM

The proposed neural network for this model is shown in Figure 2.2. It is a fully connected multi-layer perceptron with three layers: the input, projection and output layer. The purpose of this neural network is to predict the word that follows after a given context of words within a paragraph. In this figure for example, the purpose is to predict the fourth word of a context of three words within the given paragraph.

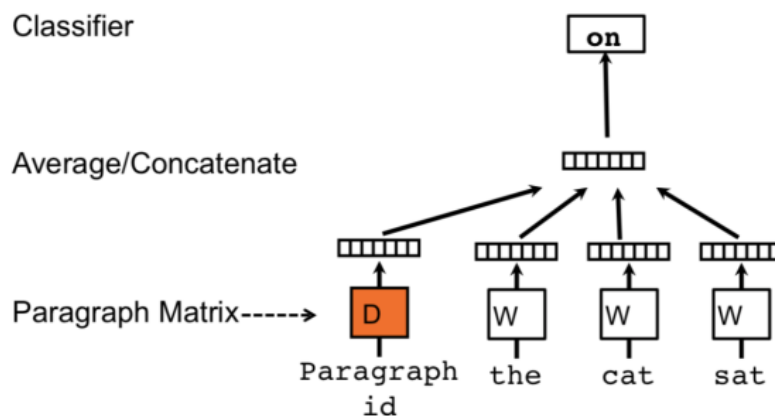


FIGURE 2.2: Distributed Memory Model of Paragraph Vectors (PV-DM), borrowed from [1]

As input for the first layer are the one-hot encoded vector of the paragraph and the one-hot encoded vectors of some words within this paragraph—the number of words is specified by a given sliding window. A one-hot encoded vector, is a 1-of-n encoding scheme where the word, or paragraph, is represented by a vector of binary valued elements in which only the i^{th} element

of the vector is set to 1 and all other values are set to 0, where i is the position of the word in the vocabulary.

Next is the projection layer. The weight matrices between the input layer and the projection layer are shared, which means that the same weights correspond to the same word independently of the contexts where this word was found. This organization increases the amount of data for training these matrices since each word of each context individually contributes to change the weight values at training. These matrices are in fact D for paragraphs and W for words, each column in these matrices represent the trained vectors of the paragraph and words respectively.

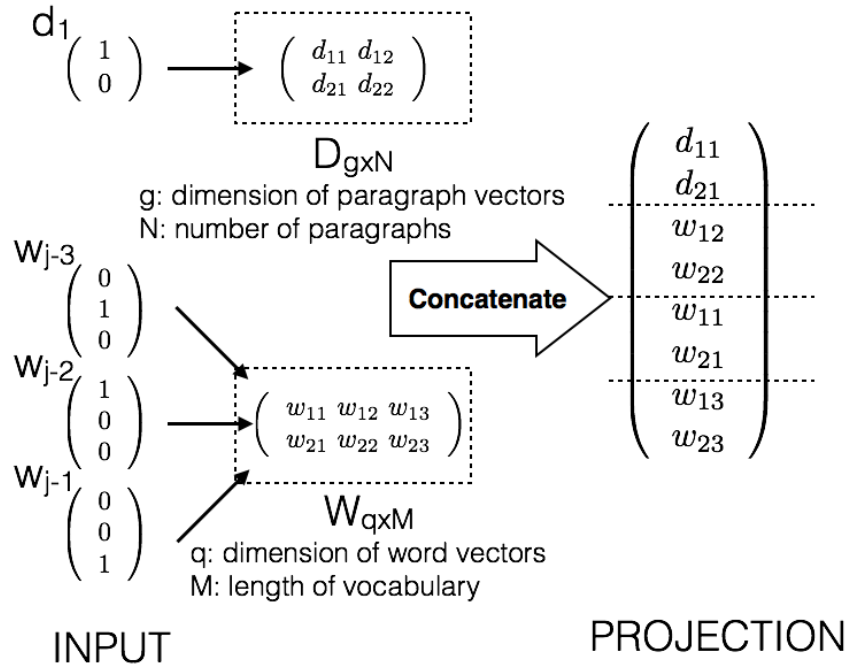


FIGURE 2.3: The projection layer of the PV-DM model

Figure 2.3 depicts how the projection layer works. The one-hot encoded vector of the paragraph is multiplied by matrix D , while the one-hot encoded vectors of the words are multiplied by matrix W , obtaining the vector representation for the paragraph and words involved. These vectors are later concatenated to form a single vector as input to the next layer.

Next layer is the output layer, where the vector representation of the word to predict—the word following the sequence of input words—is given as a reference to compare with the result of the neural network. In this way it is possible to perform the training process by minimizing the error with the stochastic gradient descent, where the gradient is obtained via backpropagation. For the output layer, a hierarchical softmax with a binary Huffman tree is used to normalize output values and optimize the training process.

The paragraph token in the input of the neural network, can be thought of as another word. It acts as a memory that remembers what is missing from the current context—or the topic of the paragraph—so it can contribute to the prediction task of the next word. For this reason, the authors call this model the Distributed Memory Model of Paragraph Vectors (PV-DM). At the end, the paragraph representation is obtained by learning from many contexts sampled from itself.

To obtain the paragraph vector for a new paragraph—which happens in most of the cases—an inference step is performed. In this step, the parameters for the rest of the model: the word vectors W , and the softmax weights and bias of the output layer, are fixed. In this way the gradient descent process is performed only to learn and obtain the vector of the new paragraph.

PV-DBOW

The proposed neural network for this model is shown in Figure 2.4. The purpose of this neural network is to ignore the context words in the input, but force the model to predict words randomly sampled from the paragraph in the output.

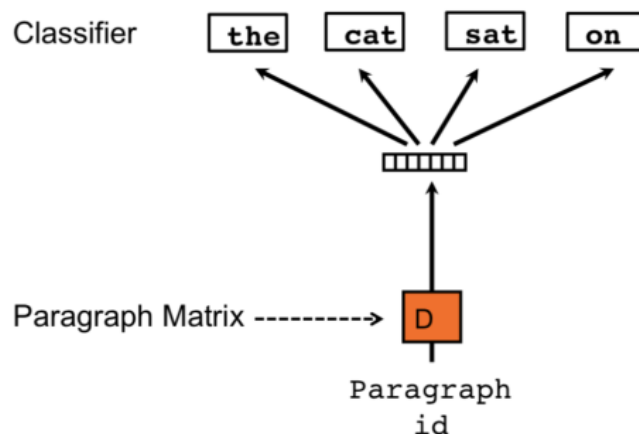


FIGURE 2.4: Distributed Bag of Words version of Paragraph Vector (PV-DBOW), borrowed from [1]

As input for the first layer, the vector for a single paragraph is extracted from matrix D . The data for comparison in the output layer changes every iteration of the stochastic gradient descent. For this, a text window is sampled from the paragraph and then a random word from this text window is given as output at each iteration to form a classification task given the paragraph vector. This technique is known as the Distributed Bag of Words version of Paragraph Vector (PV-DBOW).

For new paragraphs, a similar inference step as with PV-DM is performed to obtain their vector representation, leaving all parameters of the neural network fixed—except for the paragraph vector to be learned.

As proved by the authors, PV-DM showed better results than PV-DBOW. For this reason we used the PV-DM as our high-level similarity measure for our summarization purposes. There is also an implementation in python for this method in a library provided by the project *gensim*¹, which they call *doc2vec*.

2.5.2.3 Cosine similarity measure

Given the vector representation of sentences it is possible to compute the similarity between them. As an example in Figure 2.5 vectors of 3 sentences are represented in two dimensions, which means that only two features were taken to their representation—only two words. Of course, in real cases sentence vectors contains more than two features, but as it is impossible to visualize vectors in more than three dimensions, examples on two dimensions are given to visualize the effect of the cosine similarity measure.

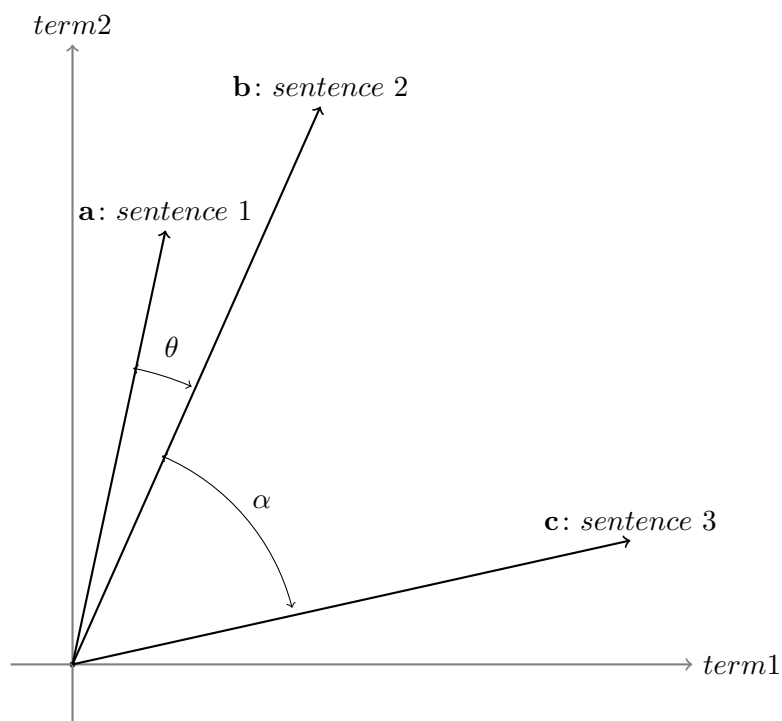


FIGURE 2.5: Cosine similarity measure

¹<https://radimrehurek.com/gensim/>

If we compare the vector of *sentence 2* against the other two, we expect the *sentence 1* to be more similar than *sentence 3* since **a** is closer to **b** than **c** to **b**. A way to measure this closeness of the vectors is by obtaining the cosine of the angle between the two pair of vectors: θ and α . The range values of cosine is $[-1, 1]$, being -1 for an angle of 180° (vectors opposed in orientation), 0 for an angle of 90° (perpendicular vectors) and 1 for angles of 0° (vector with the same orientation). It is thus a judgment of orientation and not magnitude. The cosine of the angle can be derived from the dot product of two vectors in a vector space: $\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos \theta$.

Leaving the $\cos \theta$ term alone, we obtain the formula to compute the cosine similarity measure on vectors of n dimensions:

$$\text{Similarity}(\mathbf{a}, \mathbf{b}) = \cos \theta = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} = \frac{\sum_{i=1}^n a_i b_i}{\sqrt{\sum_{i=1}^n a_i^2} \sqrt{\sum_{i=1}^n b_i^2}}. \quad (2.5)$$

2.6 Text corpora

A text corpus is a large and structured set of texts. For the area of Natural Language Processing, corpus are essential to many tasks, they are mostly used for training and testing in a massive way. These texts are usually labeled with some data of interest for the particular task. For example, for the tasks of sentiment analysis, texts as tweets are given with the information about the polarity or sentiment related to it. In the area of text classification, texts are given with the information about the topics that the text is covering, and so on. In fact, the reason why the accomplishment of this project was possible in a small period of time was the existence of this resources, which we needed to two main things:

1. To be able to execute the method and to evaluate it. For this, is necessary a big number of cluster of texts, each cluster containing various texts talking about the same topic, along with reference summaries (or also called model summaries) for each cluster to compare to.
2. To train the method used for measuring similarity between texts previously explained as *Paragraph Vector*. As it is a learning method, it needs data to be trained, and better be a huge amount a data so that it could be good at capturing all possible contexts—surrounding words—of each word within a variable-length piece of text.

For the first purpose, we use the DUC corpora. The DUC corpora is a product from a set of conferences called *Document Understanding Conferences*², which surge from 2001 to 2007

²<http://duc.nist.gov/>

organized by the government of the United States of America. Different tasks were proposed in this conferences, within which the task of summarization was proposed. The texts on which this conferences work are news provided by different news agencies. From 2002 the specific challenge of generating generic multi-document summaries was proposed until 2004. From 2005 multi-document summarization was still a challenge but summaries should have been generated responding to a specific query. We decided to work with corpora of the task 2 from DUC 2004 and the main task from DUC 2007, the last editions in the ambit of generic summarization and query-focused summarization respectively.

Reference summaries provided by DUC are gold standard as they are made by humans qualified to write good summaries. Later DUC conferences evolved to TAC conferences (*Text Analysis Conferences*) where tasks related to summarization were proposed, like: update summarization, opinion summarization, automatic evaluation of summaries, guided summarization (kind of a query-focused summarization looking for responding to different aspects specific to the category of the text, and also looking for updated information), multilingual summarization and biomedical summarization. To our knowledge, apart from this two conferences there are no more corpora to work with summarization.

For the second purpose, the Paragraph Vector algorithm only requires text divided into sentences. Many corpora exist providing huge bunch of texts. We decided to use the dataset from “One Billion Word Language Modeling Benchmark”³ that has almost 1 billion words from news. The text is already broken into sentences, giving files with a sentence per line. Furthermore, the use of this corpora is suggested by the creators of the word2vec tool⁴ which is the inspiration method from which doc2vec borrows.

2.7 Evaluation methods

Different methods to evaluate summaries exist. Most of them requires a gold standard—data generated by experts—to have a reference to compare to. Gold standard summaries are referred as *reference* summaries. On the other hand, *candidate* summaries makes reference to the summaries generated by the method that is being evaluated—the machine generated summary. Some methods have been proposed do this kind of evaluation, comparing human summaries against machine generated summaries. BLEU and ROUGE are some of these methods. From these two methods, the most used in the state of the art is ROUGE, additionally it was adopted

³<http://www.statmt.org/lm-benchmark/>

⁴<https://code.google.com/archive/p/word2vec/>

as official measure since 2003 at DUC. ROUGE stands for *Recall-Oriented Understudy for Gisting Evaluation*. As its name says it is created as a recall measure, while BLEU for example, is a precision measure. To understand these differences, we give an overview on these performance measures.

2.7.1 Recall, precision and F-measure

In the area of classification, these measures indicate how well the classification was performed—these measures are obtained for a single class at a time. Once the classification has been done with a proposed method, these measures verify how well the proposed method classified the objects in the right class. For this purpose, the classified objects should be labeled, i.e., the relation with their belonging class should be somehow specified. Lets call all reference objects belonging to the measured class the relevant objects. Relevant objects also could be seen as the objects that should be returned.

Recall measure indicates how many of the relevant objects were retrieved. It can be obtained as

$$recall = \frac{|\{relevant\ objects\} \cap \{retrieved\ objects\}|}{|\{relevant\ objects\}|}. \quad (2.6)$$

Precision measure indicates how many of the retrieved objects were relevant. It can be obtained as

$$precision = \frac{|\{relevant\ objects\} \cap \{retrieved\ objects\}|}{|\{retrieved\ objects\}|}. \quad (2.7)$$

F-measure exists in order to join precision and recall measures, and give it as a the final measure of accuracy for the classification method. It is calculated as follows

$$F_{\beta} = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}, \quad (2.8)$$

where β is a constant that can be used to weight recall when β is greater than 1, or to weight precision when β is smaller than 1. When $\beta = 1$, the F_1 score becomes the harmonic mean of the recall and precision measure, it is the most used among all the F-measures.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \quad (2.9)$$

The objects to be classified in our case of study are sentences. These measures are applied to know how well the sentences generated by our method match with the sentences on the reference summaries, i.e., to know if they are really relevant sentences or not.

2.7.2 ROUGE

There are 4 proposed ROUGE measures: ROUGE-N, ROUGE-L, ROUGE-W and ROUGE-E, with variants for each measure. The reason why this measure was born as a recall measure, is that the length on the summaries requested by the challenges on the conferences was restricted to about 250 words, i.e., the length of the retrieved documents is not variable in the denominator of the precision formula (see Formula 2.7). For this reason precision and recall are proportional when they are used for comparing different summarization methods, making the precision measure unnecessary.

To see this clearer, we can see the formula of recall in terms of precision, which will be something as

$$recall = precision \cdot \frac{|\{retrieved\ objects\}|}{|\{relevant\ objects\}|},$$

where the fraction of retrieved objects and relevant objects is constant when these values are obtained for different candidates summaries and one reference summary. This constant makes precision and recall proportional. Despite this, the 2005 release changed its formulas to get the F measures and give support to compare ROUGE scores between summaries of variable length. From the source of the author of this method, Lin [11], an explication of these measures are given.

2.7.2.1 ROUGE-N: n-gram co-occurrence statistics

This measure counts how many n-grams have in common the candidate summary and the reference summary. It is computed as

$$ROUGE-N = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}, \quad (2.10)$$

where n stands for the length of the n-gram, $Count_{match}(gram_n)$ is the number of occurrences of $gram_n$ in the candidate summary, and $Count(gram_n)$ refers to the number of occurrences of the

n-gram in the sentence S . It is clearly a recall measure since it counts how many of the relevant sentences were retrieved. BLEU is a similar measure, but it was born to evaluate machine translation as a precision measure, thereby it has in its denominator the total of n-grams in the candidate summary [12].

2.7.2.2 ROUGE-L: longest common subsequence

This way to measure summaries is by identifying subsequences in common between sentences of the reference summary and sentences of the candidate summary. Formally, a sequence $Z = [z_1, z_2, \dots, z_k]$ is a subsequence of a another sequence $X = [x_1, x_2, \dots, x_m]$, if there exists a strictly increasing sequence $[i_1, i_2, \dots, i_k]$ of indices of X such that for all $j = 1, 2, \dots, k$, we have $x_{i_j} = z_j$ [13]. That is, a subsequence is a set of elements that appear in left-to-right order on the original sequence, but not necessarily consecutively. For example, $Z = [B, C, D, B]$ is a subsequence of $X = [A, B, C, B, D, A, B]$ with corresponding index sequence $[2, 3, 5, 7]$. Given two sequences X and Y , their longest common subsequence (LCS) is a common subsequence with maximum length.

To evaluate summaries with LCS, is necessary to decompose summaries into sentences and see these sentences as sequences of words. Then, LCS are obtained at sentence level between sentences of the candidate and reference summaries. The LCS score between two sentences X of length m and Y of length n , assuming X is a reference summary sentence and Y is a candidate summary sentence, can be obtained as

$$R_{lcs} = \frac{LCS(X, Y)}{m}, \quad P_{lcs} = \frac{LCS(X, Y)}{n},$$

$$F_{lcs} = \frac{(1 + \beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}}, \quad (2.11)$$

where $LCS(X, Y)$ is the length of the longest common subsequence of X and Y , β controlling the relative importance of R_{lcs} and P_{lcs} . The intuition when applying this measure at summary-level, is that the longer the LCS of two summaries is, the more similar the two summaries are. For this idea, the score is computed from the union of LCS matches between a reference summary sentence r_i , and every candidate summary sentence c_j . Given a reference summary of u sentences containing a total of m words and a candidate summary of v sentences containing a total of n words, the summary-level LCS-based F-measure can be computed as

$$R_{lcs} = \frac{\sum_{i=1}^u LCS_{\cup}(r_i, C)}{m}, \quad P_{lcs} = \frac{\sum_{i=1}^u LCS_{\cup}(r_i, C)}{n},$$

$$F_{lcs} = \frac{(1 + \beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2P_{lcs}}, \quad (2.12)$$

where $LCS_{\cup}(r_i, C)$ is the LCS score of the union longest common subsequence between reference sentence r_i and candidate summary C . For example, if $r_i = w_1w_2w_3w_4w_5$, and C contains two sentences: $c_1 = w_1w_2w_6w_7w_8$ and $c_2 = w_1w_3w_8w_9w_5$, then the longest common subsequence of r_i and c_1 is w_1w_2 , and the longest common subsequence of r_i and c_2 is $w_1w_3w_5$. The union longest common subsequence of r_i, c_1 , and r_i, c_2 is $w_1w_2w_3w_5$. From here, we can obtain for example the recall LCS score of this union, which is: $LCS_{\cup}(r_i, C) = \frac{4}{5}$, 4 for the length of the union of all LCS $|w_1w_2w_3w_5|$ and 5 for the number of words of the reference sentence $|w_1w_2w_3w_4w_5|$.

One advantage of using LCS is that it does not require consecutive matches but in-sequence matches that reflect sentence level word order as n-grams. The other advantage is that it automatically includes longest in-sequence common n-grams, therefore no predefined n-gram length is necessary.

2.7.2.3 ROUGE-W: weighted longest common subsequence

This measure born has an improvement for ROUGE-L. For ROUGE-L little matters the spatial relations on the obtained subsequences. For example, given a reference sequence $X: [ABCDEFGF]$ and two candidate sequences $Y_1: [ABCDHIK]$ and $Y_2: [AHBKCID]$, the two candidate sequences will have the same ROUGE-L score by having the same LCS. Nevertheless, Y_1 has consecutive matches which can make it more similar to X . To solve this problem, ROUGE-W implements a regular two dimensional dynamic program table to remember the length of consecutive matches encountered so far. This weighted LCS is called WLCS.

2.7.2.4 ROUGE-S: skip-bigram co-occurrence statistics

A skip-bigram is any pair of words in their sentence order, allowing for arbitrary gaps. It does not require consecutive matches but is still sensitive to word order. This is traditionally used to measure the overlap between a candidate translation and a set of reference translations. Given translations X of length m and Y of length n , assuming X is a reference translation and Y is a candidate translation, skip-bigram-based F-measure is computed as

$$R_{skip2} = \frac{SKIP2(X, Y)}{C(m, 2)}, \quad P_{skip2} = \frac{SKIP2(X, Y)}{C(n, 2)},$$

$$F_{skip2} = \frac{(1 + \beta^2)R_{skip2}P_{skip2}}{R_{skip2} + \beta^2P_{skip2}}, \quad (2.13)$$

where $SKIP2(X, Y)$ is the number of skip-bigram matches between X and Y , β controlling the relative importance of R_{skip2} and P_{skip2} , and C is the combination function. When using a maximum skip distance, it is possible to use the same formulas only by counting in $SKIP2(X, Y)$ the skip-bigram matches within the maximum skip distance. The same modification when counting should be made on the denominators: $C(m, 2)$ and $C(n, 2)$, to take into account only the skip-bigrams with the maximum skip distance, that are possible in the reference and candidate summary respectively.

ROUGE-SU: Extension of ROUGE-S

Because ROUGE-S measures co-occurrences of word pairs, with arbitrary gaps, but also with the same order, it gives the same ROUGE-S score to: sentences with similar words but in inverse order, and sentences with complete different words, i.e., a value of zero. To differentiate similar sentences from sentences that do not have single word co-occurrence, ROUGE-S is extended with the addition of unigram as counting unit. This extended version is known as ROUGE-SU.

Chapter 3

State of the art

Extractive multi-document summarization is one of the Natural Language Processing tasks that has received a lot of attention by the richness in diversity to tackle the problem of this task. It has been undertaken by numerous methods, from very simple ones such as taking random sentences from the texts, to advanced methods which mainly focus on outperforming the best scores of the state of the art. In this chapter we present the simplest techniques known as baselines, four traditional approaches including some simple techniques to generate multi-document summaries, and those techniques that to our knowledge have the best scores when using texts and summaries from the Document Understanding Conferences (DUC) and using ROUGE as the evaluation method. Along with this, an overview of how redundancy is being managed by these methods is provided.

3.1 Baseline methods

Baseline methods are the simplest ways to generate an extractive summary. They were designed with the purpose of having a reference of comparison that should always be outperformed when a new method is proposed. The most well-known baselines are those required by DUC on their competitions: random, lead and coverage.

3.1.1 Random

The random baseline consists in generating summaries by extracting random sentences to build the final summary. To better represent the performance of this method, the experiment is performed many times and the averaged score of this executions is given as the final score.

3.1.2 Lead

The lead baseline consists in taking the last sentences of the last document in the cluster. In DUC corpora, this is a challenging baseline because the last document is the most recent document on the collection, conveying the most updated information—the latest news.

3.1.3 Coverage

The coverage baseline consists in taking the first sentence of the first document, then the first sentence of the second document, and so on, taking the first sentence of each document of the cluster until the maximum length of the summary is reached. This baseline is a challenging one [14], specially on news where the most relevant information is put immediately on the first paragraph to catch reader’s attention.

3.2 Traditional approaches

Different ideas to generate extractive summaries from multiple texts exist, and they are very diverse. Despite this, it is possible to group the majority of this methods. We show the four approaches stated by Kumar and Salim [2], who classify them depending on how the problem is tackled. While some methods stay out of this approaches, a general panorama of these approaches are showed in order to give an idea of how the actual techniques generate extractive summaries. It is recommended to take notice of the graph-based and by cluster-based approaches, since our method could fit on these two. All these approaches process the multiple-documents as if they were only one text; they merge all texts in only one and then the explained process is executed on this big text.

3.2.1 Feature-based approaches

In this category, methods compute the relevance of the sentence according to the values of specific features of the sentence. For this aim, it is necessary to find a way to compute a numeric value for each feature. Such features could be:

Word frequency. The importance of the sentence is given by the frequency of its words along the whole text. The common way to compute this value is to sum the value of frequency of each word in the sentence. This value of frequency is traditionally the tf value.

Title/headline word. The notion is that a sentence is important if it contains words that appears in the headline or title of the text. It is computed as

$$f_{headlineWord}(S) = \frac{\text{Number of title words in } S}{\text{Number of title words}}.$$

Sentence location. Where the sentence is located could say something about its importance. Many ideas could be valid for this feature, as stated by Suanmali et al. [15], who compute this value as

$$f_{sentenceLocation}(S) = \left\{ 1 \text{ for } 1^{st}, \frac{4}{5} \text{ for } 2^{nd}, \frac{3}{5} \text{ for } 3^{rd}, \frac{2}{5} \text{ for } 4^{th}, \frac{1}{5} \text{ for } 5^{th}, 0 \text{ for other sentences} \right\}.$$

Another idea is proposed by Kupiec et al. [16], where paragraphs are used as reference: sentences are distinguished according to whether they are paragraph-initial, paragraph-final (for paragraphs longer than one sentence) and paragraph-medial (in paragraphs greater than two sentences long).

Sentence length. What is argued in this feature is that a short sentence conveys less information, thus they are not suitable for a summary. Kupiec et al. [16] for example, the feature is True for all sentences longer than a given threshold (e.g., 5 words) and False otherwise. Another way to compute is proposed by Suanmali et al. [15]:

$$f_{sentenceLength}(S) = \frac{\text{Number of words in } S}{\text{Number of words in longest sentence}}.$$

Cue word. It is possible that certain cue words (e.g., “significantly”, “in conclusion”) indicate that the sentence is carrying an important message. The sentence acquires a True value if it contains any of this words or False otherwise.

Thematic word. The most representative words of the the text are extracted (i.e., those most representative to the topic, it could be the most frequent [15]). Then the value of the feature is computed as

$$f_{thematicWord}(S) = \frac{\text{Number of thematic words in } S}{\text{Number of thematic words}}.$$

As seen, all these characteristics can be obtained given the sentence, and maybe other elements are required such as the title, a list of cue words, etc. Also, the way to compute the value for a same feature could vary. Many other features and the way to compute their values have been proposed, however the ones presented here are the most well-known and simple to obtain.

Furthermore, a weight for each feature could be stated to give more relevance to key features. At the end the relevance value of the sentence can be computed as the following linear combination

$$Score(S) = w_1 f_1 + w_2 f_2 + w_3 f_3 + \dots + w_n f_n$$

where w_i represents the weight of the i^{th} feature, and f_i represents its value. It is possible to use supervised learning to determine the weight values for each feature.

When the scores of all sentences are computed, an algorithm is executed to choose the most relevant sentences to build the final summary, avoiding redundancy and to accomplish the other important aspects to be considered in multi-document summarization.

3.2.2 Cluster-based approaches

Clustering is a task where groups containing similar objects are formed. For this task, a similarity measure is very essential to know if objects are similar between each other, i.e., to know if they belong to the same group or not. In texts where redundancy is present, similar sentences are expected. A convenient way to deal with this redundancy, is grouping similar sentences in a cluster. By doing this with all sentences, big groups—clusters—are formed at the end in a way that each one covers a similar idea. Many similarity measures to compare sentences exist, in a section ahead of this document some of this measures are described. Typically, the cosine similarity measure between a pair of sentences represented as vectors with a bag of words, is used. After having formed different clusters with similar ideas, representative sentences for each cluster are picked to form part of the final summary. This whole process, illustrated by Kumar and Salim [2], is showed in Figure 3.1.

It is possible to categorize clustering methods in two kinds: agglomerative and partitional. Agglomerative algorithms, also called *bottom-up*, start by building a cluster for each object—a sentence in extractive summarization. Then these objects are compared in pairs and the two most similar are taken to form a new cluster. The process continues merging clusters until a

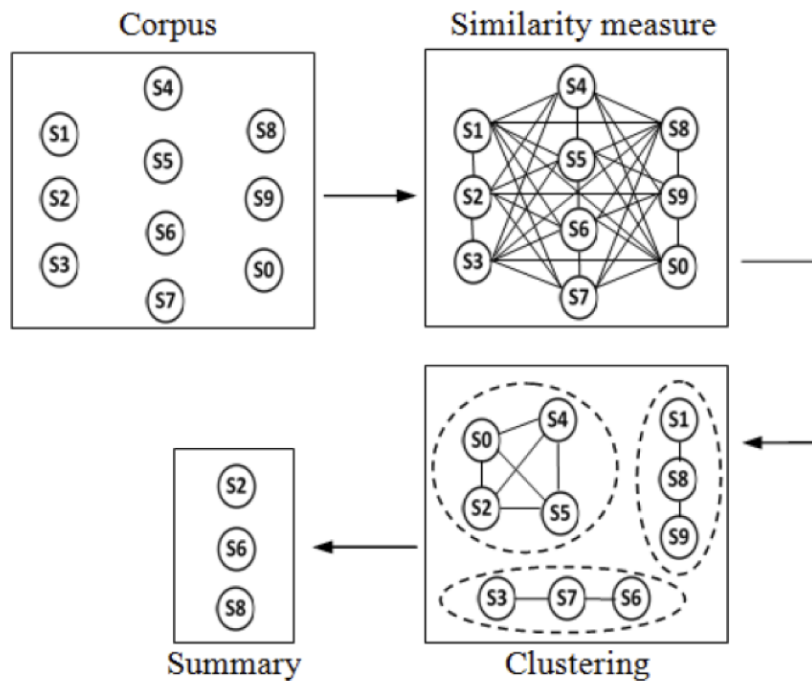


FIGURE 3.1: Cluster-based method for summarization, borrowed from [2]

stopping criteria is reached. This criteria could be when a certain quantity of cluster are built, or until a certain similarity threshold is reached, and many others.

On the other hand, partitional algorithms, also called top-down, start with a single cluster with all the objects—all the sentences in extractive summarization—and iteratively smaller clusters are created so that each sub-cluster contains objects with higher similarity; this property is usually called coherence, the more similar objects the cluster contains, more coherent the cluster is. The partition of cluster stops until a certain criteria is reached. Determining the quantity of clusters that should be created, or any other stopping criteria, is not an easy task, but many works on this subject exist.

As this approach deals with redundancy from the construction of the clusters, it is good at representing diversity. Still, careful measures have to be taken when picking representative sentences from each cluster to compose a good summary.

3.2.3 Graph-based approaches

Many graph based methods to deal with multi-document summarization exist in the state of the art. Nevertheless, we introduce a traditional method that has been a reference by many since it

was proposed. First of all, a way to map a text to a graph is required. This is done by representing sentences as nodes and edges represent the similarity between a pair of sentences. Hence, the edges are weighted with the similarity value. To see an illustration of this representation, provided by Erkan and Radev [3], see Figure 3.2.

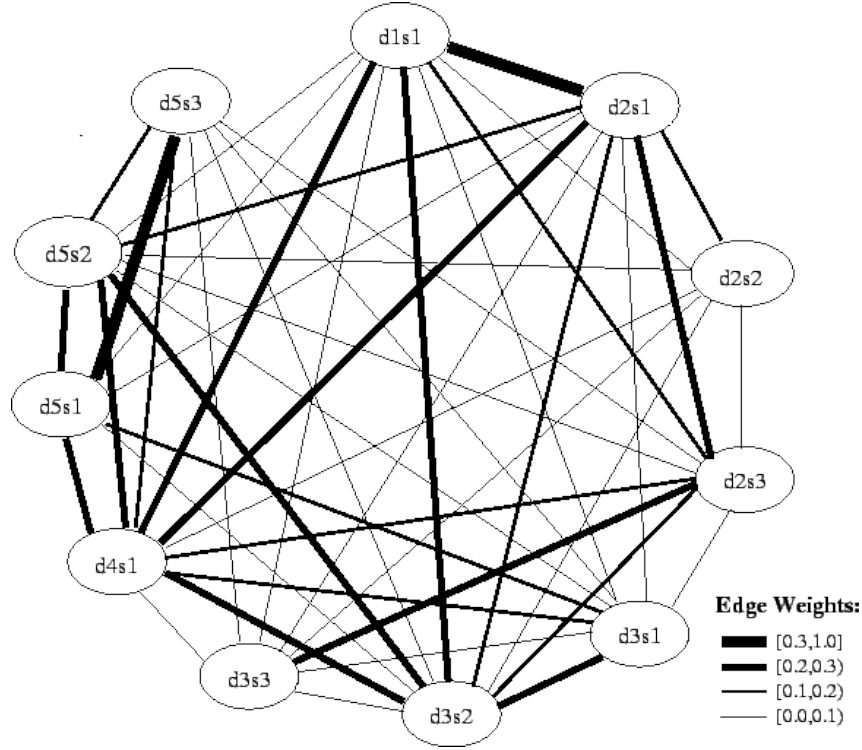


FIGURE 3.2: Traditional graph representation of texts, borrowed from [3]

It is possible to leave the edges weighted or the graph can be simplified by generating an edge only if it is above the given threshold. For example, establishing a threshold of 0.1 on the graph showed in Figure 3.2, results in a graph such the one showed in Figure 3.3.

Once the graph is built, the relevance of a node can be computed from the relationships it has to other nodes, i.e., it is said that a sentence is important if it is strongly connected to many other sentences [3]. This idea can be implemented with graph based algorithms such as HITS or PageRank. For example, Mihalcea and Tarau [9] proved the idea by successfully implementing the PageRank algorithm on the graph with weighted edges.

3.2.4 Knowledge-based approaches

These methods are characterized by using an ontology to score the importance of the sentences. An ontology is an explicit specification of a conceptualization [17]. A conceptualization is a body

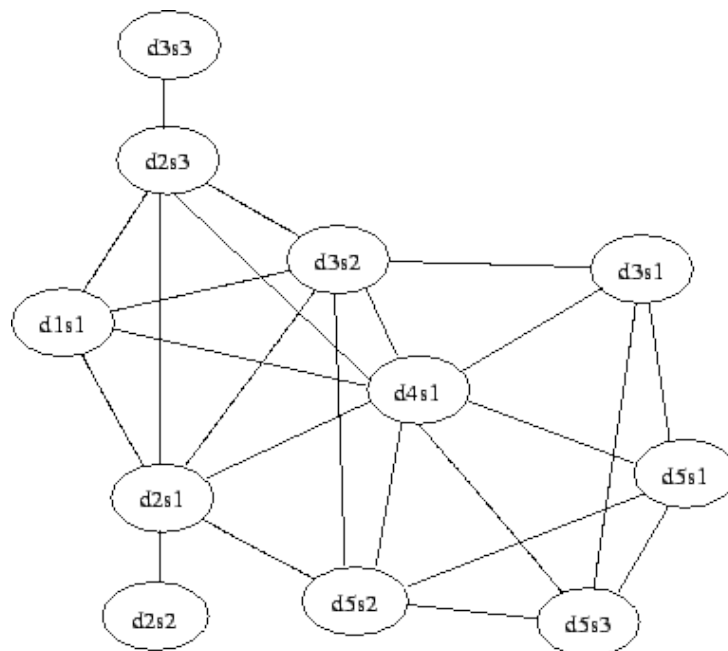


FIGURE 3.3: Graph representation after establishing a threshold for edge generation, borrowed from [3]

of formally represented knowledge, which contains the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them [18]. An ontology is mostly represented in a graph—or matrix—where nodes are the entities and edges depict the relations between entities.

To use ontologies in summarization, entities are extracted from the texts and mapped to the ontology. For example, to find a similarity between two sentences, their entities are extracted, mapped to the ontology, and the similarity between sentences can be computed by the distance of their entities in the ontology. The ontology should belong to the specific domain of the topics or events that are boarded by the texts, so that the information can be related through the shared and common understanding of the domain [19].

This approach is mainly used to generate query-focused summaries, where the similarity of a sentence to the given query is computed by extracting distances between the entities of the sentence and the query. For example, it is possible to map the query and sentences to the same ontology and then start building the summary from sentences in the sub-tree rooted at the corresponding query node [20]. Ontologies can also be employed as dictionaries to filter sentences, staying only with sentences containing words related to concepts in the ontology [21]. Additionally, the ontology has been used to find synonyms or semantically related concepts of the query so that this can be expanded [22], opening the possibility to find more results by

including more concepts on the search. At the end, this ontology can be used in different ways to take advantage of the information that this representation of knowledge contains.

3.3 Best approaches

Before presenting the best approaches, it is important to let the reader know that at the time when this work was being carried out, there was no standardized nor any kind of centralized community which focused in multi-document summarization and divulged adequate methods to solve this task. The way the best approaches were found on this work, was by tracking the methods citing and outperforming the scores of the best techniques officially reported by DUC.

When looking for results on different works, various combinations of reported results came to the game by varying the DUC version—a different corpus was produced for each edition from 2001 to 2007—and the adopted ROUGE evaluation method, depending on the convenience and judgments of the authors. For example some works were using the DUC 2002 and DUC 2004 corpora and evaluating with ROUGE-1, ROUGE-2, and ROUGE-4 measures, while some other works were using DUC 2004, DUC 2007 and evaluating with ROUGE-2 and ROUGE-SU4 measures. This led to a great combinations of results on the state of the art.

For this reason, finding which are the best methods is a difficult and uncertain task. We restricted the search to methods working with the DUC 2004 task 2—where the ROUGE measure started to be officially adopted—and DUC 2007 main task corpora, which are the last editions working on generic and query-based summarization respectively. Additionally, we keep only the methods reporting results with the ROUGE-1, ROUGE-2, or ROUGE-SU4 measures, which have been the most used ones in the state of the art given the good correlation of ROUGE-1 with human judgments [12], and the adoption of ROUGE-2 and ROUGE-SU4 as official evaluation methods for DUC 2007.

Table 3.1 shows the four best scores for each ROUGE recall measure using DUC 2004 corpus, ordered by ROUGE-1. Also the best system that performed at DUC 2004 (peer 65) is included, along with the LexPageRank method, which have been widely used for comparison in the state of the art. A short explanation of the methods that have the best score for each ROUGE measure is provided.

PatSum, a method proposed by Qiang et al. [23], reports only ROUGE-2 and ROUGE-4 results, outperforming every method they compare to, and even some human summarizers. This method represent sentences using term-weight pairs, where each term of the sentence is related to a

TABLE 3.1: Best approaches for DUC 2004 task 2

Method	Year	ROUGE-1	ROUGE-2	ROUGE-SU4
PatSum [23]	2016	–	0.10200	–
wHAASum [24]	2016	0.41670	0.09560	0.13860
AASum-W3 [25]	2014	0.41150	0.09340	0.13760
SentTopic-MultiRank [26]	2012	0.41016	0.09917	0.14324
ToPageRank [27]	2012	0.40501	0.09555	0.14034
LexPageRank (degree) [28]	2004	0.38304	0.09204	–
CLASSY (peer 65) [29]	2004	0.38220	–	–

closed-pattern—a longest frequent pattern in the document, which is discovered using closed pattern mining algorithm. The weight of each term-weight pair of the sentence is computed by accumulating the weight of its covering closed patterns with respect to this sentence, and the weight of a closed-pattern is obtained from the coverage of this closed-pattern on all the sentences and documents. Given this representation, they rank sentences giving more relevance to sentences containing more closed patterns with high weight, and being closer to the start of the document. After this, a method inspired in MMR is proposed to select sentences trading between coverage and redundancy.

Canhasi and Kononenko first proposed *AASum* [25] using archetypal analysis, which is an matrix factorization method to perform soft clustering. Initially they generate a content-graph joint model as a representation of the sentences. The content-graph joint matrix of this representation is obtained as a product between the term-sentence matrix and the sentence similarity matrix. The sentence similarity matrix describes similarities between sentences, which are obtained with the cosine similarity measure between sentences represented in vector space model using the bag of words approach. The term-sentence matrix describes the frequency of terms that occur in each sentence. After obtaining the content-graph joint matrix, the Archetypal Analysis is applied to factor it. Finally, sentences with the highest archetype membership value, from the most significant archetypes, are chosen to compose the final summary.

Later the same authors designed *wHAASum* [24], proposing an enhanced version of their previous work which they called weighted hierarchical archetypal analysis. In this new technique sub-topics are obtained—as sub-archetypes—to capture the intrinsic (low dimensional) structure of the data. This last method was adapted to generate query-focused, update and comparative summaries obtaining good results, and getting the best ROUGE-1 score until now for generic summarization.

SentTopic-MultiRank, a method proposed by Yin et al. [26], has until now the best ROUGE-SU4 score. They extract semantic topics from texts using LDA. Then they propose a way to

compute similarity using transformed radio (TR) between sentences represented by their words-topic probability distribution. Then, they rank sentences using the MultiRank [30] method and make the final selection of sentences with a modified MMR algorithm, proposed by Wan et al. [31].

For DUC 2007 corpus, the four best scores for each ROUGE recall measure are presented in Table 3.2, ordered by ROUGE-1. Also the best systems that performed at DUC 2007 are included (peers 15 and 29). A short explanation of the methods obtaining the best score for each ROUGE measure is provided.

TABLE 3.2: Best approaches for DUC 2007 main task

Method	Year	ROUGE-1	ROUGE-2	ROUGE-SU4
PLSA-JS [32]	2009	0.45843	0.11675	0.17680
HybHSum2 [33]	2010	0.45600	0.11400	0.17200
Inter w/i ND [34]	2013	0.45009	0.12410	0.17600
Bipartite graphs [35]	2014	0.44800	0.10928	0.16735
ILP2 [36]	2012	–	0.12517	0.17603
IIIT Hyderabad (peer 15) [37]	2007	–	0.12448	0.17711
Lin and Bilmes [38]	2011	–	0.12380	–
PYTHY (peer 29)[39]	2007	0.42600	0.12028	0.17074

PLSA-JS, proposed by Hennig [32], holds the best ROUGE-1 score until now. His method is based on probabilistic latent semantic analysis (PLSA), which allows him to represent sentences and queries as probability distributions over latent topics. By representing sentences and the query on the same latent topic space, it is possible to apply a similarity measure on this space. These latent topics are obtained by training PLSA with a term-sentence matrix. After obtaining this representation, the importance of sentences is computed as a weighted sum of similarity feature values. Such similarity features correspond to the comparison between the sentence and different elements of the texts such as the cluster title and the cluster narrative—which represents the query, the document title, etc. Weights are obtained experimentally. At the end, these score is recomputed with a formula inspired in MMR to avoid redundancy.

Galanis et al. [36] created *ILP2* using only ROUGE-2 and ROUGE-SU4 measures to evaluate, having achieved the best ROUGE-2 score. Unlike the other methods best performing at DUC 2004 and 2007, this is a supervised method. First they compute the importance score of each sentence with Support Vector Regression (SVR). For this, they provide feature vectors of sentences (such as sentence position, world overlap—as well as Levenshtein distance—with the given query, etc.) as training instances, and the average of their ROUGE-2 and ROUGE-SU4 scores as the target. After having this trained model—capable of predicting the averaged ROUGE-2

and ROUGE-SU4 score for new sentences, they used Integer Linear Programming to jointly maximize the importance of sentences composing the final summary, as well as their diversity. At the end, a modification on the optimization method is made to reward longer sentences.

The best ROUGE-SU4 score is held by a participant method at DUC 2007 competitions—*peer 15*—proposed by Pingali et al. [37]. They rank sentences as a trade between a query-independent and a query-dependent score. The query-independent score of a sentence is a prior value computed as a proportion between the probability of generating each of its terms from the given set of documents, and the sum of the probabilities of generating each of its terms from: the given set of documents, and a randomly picked set of documents from various topics. The terms are obtained by clustering words of the two sets of documents, according to their probability distribution on different classes. This distributional clustering method that they used, is proposed by Baker and McCallum [40] and requires training data. The query-dependent score, on the other hand, is a value of the sentence answering the given query. It is computed using co-occurrence statistics of words, by obtaining the sentences whose word probability distributions most co-occur with the word probability distributions of the query in a fixed window. At the end, entities—such as organization names, partial person names, etc.—are identified and replaced to their acronymized version to avoid redundancy.

An interesting pattern to notice in all of these best performing works in DUC 2004 and DUC 2007, is that most of them are topic-oriented, term-oriented, or clustering methods—except for the supervised approach. It seems that topics, clusters or frequent terms are good at concentrating the most important information of the texts in generic and query-oriented summaries. Therefore, an important step is to find the right way to relate this important information with sentences and pick the right ones to compose the final summary.

3.4 How different methods deal with redundancy

More than taking part of the movement of getting the best ROUGE score on the DUC corpora, we want to study this movement and see how redundancy is being taken to provide relevance to sentences.

All approaches agree that redundancy should be eliminated from the final summary. Some approximations manage redundancy at the phase of sentence selection by selecting the top n of non-redundant ranked sentences, i.e., avoiding sentences in the summary if their similarity to the previously selected sentences is greater than a given threshold [25, 33, 37].

The maximum marginal relevance (MMR) approach [41] searches novelty by penalizing the score of sentences according to their similarity with the already selected sentences for the summary. But as much as they penalize this redundancy, they benefit similarity with a given query. This method is typically applied during the process of sentence selection for query-focused summarization methods [42, 43]. So in this way, the top sentences composing the final summary are important to the query and are not redundant between themselves, i.e., are diverse.

Modifications have been proposed on the MMR technique to change the benefit in the formula. Instead of benefiting its similarity with a given query, they benefit other aspects related to the method, while the penalization on redundancy remains the same. For example Qiang et al. [23] made an adjustment to benefit the coverage of the sentence on the closed patterns of the documents. Also, Hennig [32] changed the formula to benefit the already computed score of the sentence according to its similarity with, not only the query, but other elements of the text as the title of the cluster, or the title of the document.

Furthermore, some other techniques penalize redundancy on the computation of the sentence score differently from the MMR technique. For example Radev et al. [44] penalize a sentence if it overlaps with highest ranked sentences. Or for instance, Toutanova et al. [39] use a discount factor on the computation of each feature value of the sentence, where the magnitude of the discount or penalization is related to the similarity with the first selected sentences for the summary.

Some other authors, instead of penalizing, compensate the non-redundancy or diversity. For example, Parveen and Strube [35] compensate sentences by not sharing the same entities, or as Galanis et al. [36] that compensate sentences by not sharing the same bigrams through an optimization method pursuing, between other variables, to maximize the quantity of different bigrams in the final summary.

A different approach is to deal with redundancy in previous stages. In graph-based methods this is considered from the construction of the graph. For example, Mihalcea and Tarau [45] establish a maximum similarity threshold, in this way very similar sentences are not linked with an edge and thereby they do not add importance to a node. Other graph-based approximation is presented by Shen and Li [46], where the summary is represented by the minimum dominating set on a graph built by representing nodes as sentences and edges as similarities, hence leaving out the redundant sentences by looking for the minimum set.

Clustering methods also have been successful in reducing redundancy and representing diversity by grouping sentences which are highly similar to each other into one cluster, thus generating a

number of clusters. Once sentences are clustered, sentence selection is performed by selecting a sentence from each cluster. At the end some extra measures have to be executed on the selected sentences to ensure non-redundancy on the final summary [25, 44].

As can be seen, none of these methods judges directly the importance of an idea by its absence or presence of redundancy. Their management of redundancy is mainly focused on avoiding this redundancy on their final summary. This is why this work focuses on studying redundancy handling through a flexible framework that allows us to experiment with different redundancy settings. This framework will be explained in the following section.

Chapter 4

Our proposal

In the task of generic summarization, no assumptions are made about the genre or domain of the materials that need to be summarized [8]. In this setting, the importance of information is determined only with regard to the content of the input itself. Nevertheless, the concept of importance is vague, it is still difficult to define what a generic summary must contain.

In some cases, the summarization algorithm can use some clues to determine importance, such as in query-focused, comparative or update summarization: in these approaches, a sentence is considered important if it matches a specific query, if it is the most discriminative sentence by representing the characteristics of a document group, or if it shows information different from what was previously known. However, one may need a general panorama of what the texts say, so in this case what should we consider important to show to the user? This is why we propose a different perspective to search for information in multiple documents based on redundancy.

Unlike most of the methods working on the multi-document summarization task, we want to take advantage of the presence of absence of intra-document and inter-document redundancy as relevant information. Of course at the end redundancy should be eliminated from the final summary, but this information on sentences is not ignored. This chapter presents our proposed method for this purpose. The explanation is divided in the three stages of the extractive summarization process: sentences extraction and computational representation of the texts, the process of computing relevance—from redundancy—for each sentence, and the generation of the summary from the most relevant sentences.

4.1 Sentence extraction

Starting from a cluster of texts that are about to be summarized, a representation to manipulate them is necessary. For this purpose, we represent them in a graph adopting a traditional method, by mapping all the sentences of the texts as nodes and the similarity between them as edges, i.e., edges are weighted with the similarity value. In Figure 4.1, this representation is depicted with sentences from n documents, each document with a variable number of sentences (m sentences for document 1, k sentences for document 2, l sentences for document n , etc.). The graph is complete, i.e., every node is connected to all other nodes.

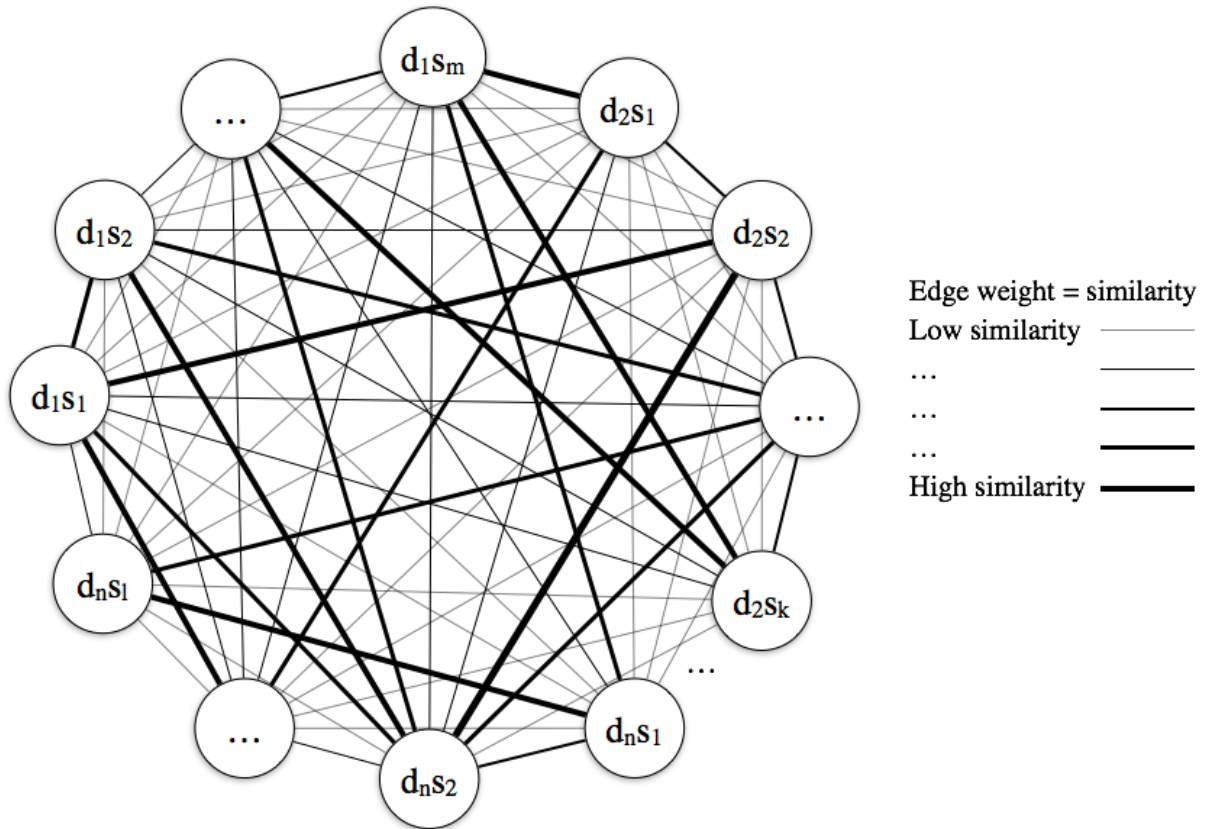


FIGURE 4.1: Complete graph representation of sentences from a cluster of texts

An important resource here is a suitable similarity measure. We decided to perform our experiments with two measures: the cosine similarity measure on sentences represented on a vector space model proposed by Le and Mikolov [1] (*Paragraph Vector*), which has shown to give good results in measuring similarity considering context; and the superposition similarity measure, which is a very simple measure to count the word overlap between sentences.

For breaking a text into sentences, we used a perl script¹ provided by NIST for evaluating summaries since DUC 2003. In this way we conform to DUC standards to determine what is a sentence and what is not. In addition, a very basic pre-processing was performed on sentences by converting them to lower case—in order to improve the overlapping and sharing of context of words—and by removing punctuation marks.

4.2 Relevance search

After extracting sentences into the graph, nodes are weighted to initially determine sentence importance. This latter is going to be useful for decision making at removing sentences. We are adopting the TextRank [9] value as the initial weight of nodes—relevance of sentences. In an undirected graph representing sentences from multiple texts, the TextRank value is somehow related to redundancy, because node weights are computed as in PageRank: the more connections the node has with other important nodes, the higher score it has. Moreover, the strength of the connection also accounts for giving it importance, i.e., the stronger the similarities of its edges with others, the more relevant the node.

Formally, the TextRank value can be computed as follows: let $G = (V, E)$ be a directed graph with the set of vertices V and the set of edges E , where E is a subset of $V \times V$. For a given vertex V_i , let $In(V_i)$ be the set of vertices that point to it (predecessors), and let $Out(V_i)$ be the set of vertices that vertex V_i points to (successors). The TextRank score of a vertex V_i is defined as follows

$$WS(V_i) = (1 - d) + d \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \quad (4.1)$$

where w_{ij} represents the strength of the connection from node V_i to node V_j , and d is a damping factor representing the probability of jumping from a given vertex to another random vertex in the graph. This factor appears in the context of Web surfing with the PageRank [47] method, where a user clicks on links at random with a probability d , and jumps to a completely new page with probability $1 - d$. We adopted the commonly used value for this d factor of 0.85 [9, 47].

Since our graph is not directed, we treated the output and input edges as equal. Initially, the TextRank value for each sentence is initialized to $\frac{1}{\text{number of sentences in the cluster}}$. Afterwards, the computation of TextRank is performed in each iteration for all sentences until convergence, i.e.,

¹<http://duc.nist.gov/duc2004/software/duc2003.breakSent.tar.gz>

until the value computed for each sentence does not have a big change with respect to the score computed in the previous iteration. After computing the initial relevance value for each node, we obtain a graph with weighted edges and weighted nodes, as depicted in Figure 4.2.

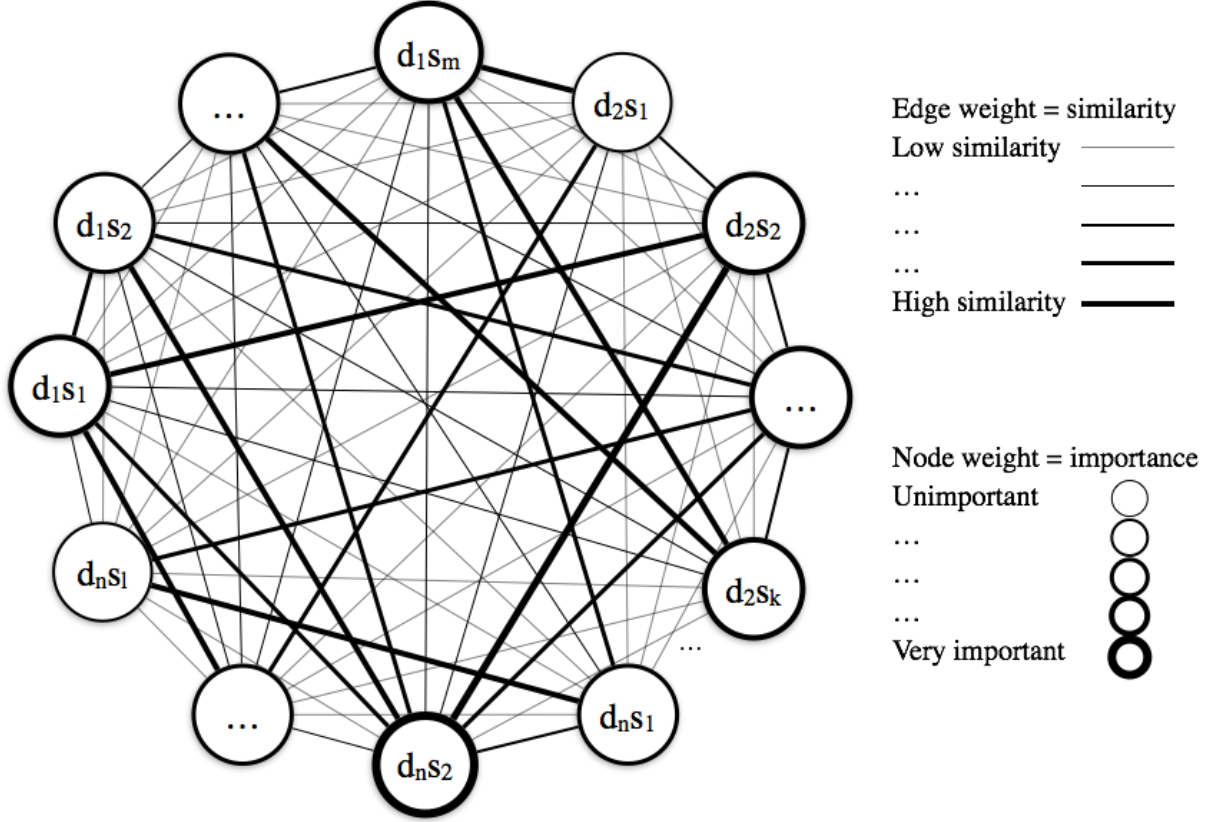
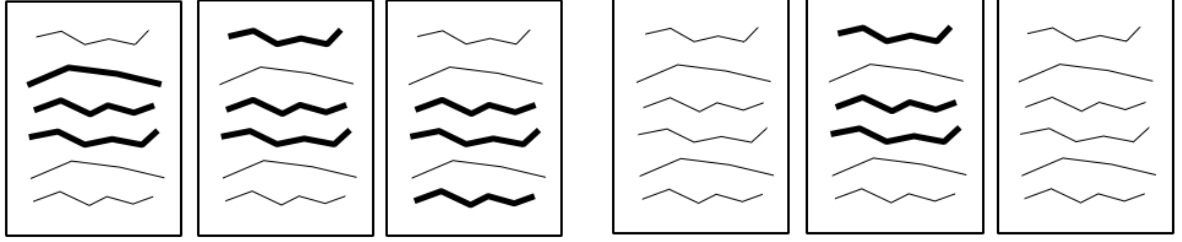


FIGURE 4.2: Graph representation after weighting edges

Once we have constructed this graph, we proceed to experiment with redundancy. The way to detect redundancy is through similarity: if two sentences are similar, then these sentences are redundant. Since this is a very common phenomenon in multi-document summarization, we can use it as a resource to measure relevance of ideas.

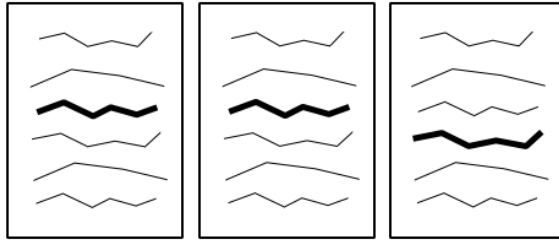
One can give relevance to an idea depending on how redundant it is across documents, i.e., inter-document redundancy, or how redundant it is within each document, i.e., intra-document redundancy. For example: one can say that an idea is important if it is redundant across all documents—such as popular information, see Figure 4.3g; or that it is important if it is not mentioned across all documents—as the concept of the inverse document frequency, that if it is mentioned by everyone then it is not relevant because everyone knows about it, see Figure 4.3h; or that an idea is important if it is redundant only in one document but not on all of them— if only one author constantly address the same idea, maybe it is important, as valuable rare

information, see Figure 4.3b. All these ideas could be valid on different areas such as medicine, news, geography, politics, etc.

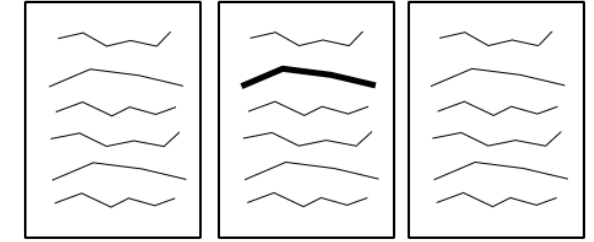


(a) 1. An idea redundant per document and redundant across documents

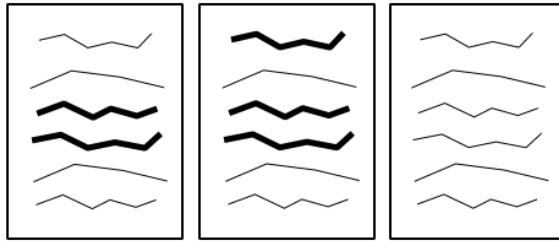
(b) 2. An idea redundant per document and non-redundant across documents



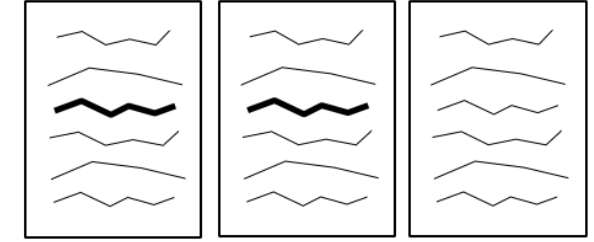
(c) 4. An idea non-redundant per document and redundant across documents



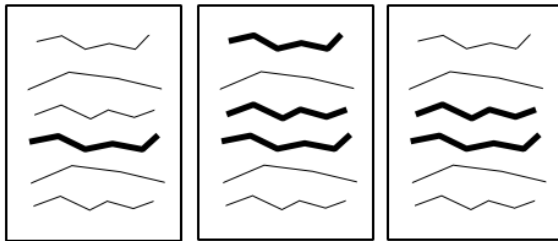
(d) 5. An idea non-redundant per document and non-redundant across documents



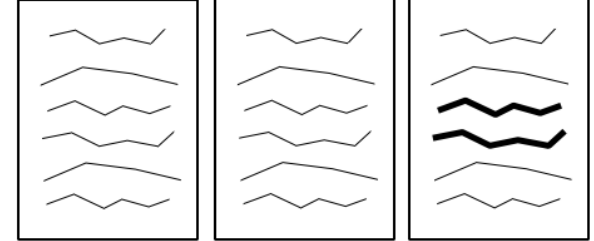
(e) 3. An idea redundant per document



(f) 6. An idea non-redundant per document



(g) 7. An idea redundant across documents



(h) 8. An idea non-redundant across documents

FIGURE 4.3: Strategies to define the importance of an idea according to redundancy

After considering these ideas, we came to nine possible strategies given the combination of intra-document and inter-document redundancy. These strategies establish that a sentence is relevant if:

1. It is redundant per document and redundant across documents.

2. It is redundant per document and not redundant across documents.
3. It is redundant per document.
4. It is not redundant per document and redundant across documents.
5. It is not redundant per document and not redundant across documents.
6. It is not redundant per document.
7. It is redundant across documents.
8. It is not redundant across documents.
9. Redundancy is not taken into account to provide relevance.

Figure 4.3 depicts the interpretation for each of these strategies, except for strategy 9, that ignores intra-document and inter-document redundancy, eliminating redundancy without taking it into account to determine the relevance of the sentence. Strategies 3 and 6 determine the relevance of a sentence based only in intra-document redundancy, i.e., inter-document redundancy is irrelevant, it could be present in any extent, see Figure 4.3e and Figure 4.3f. Something similar happens to strategies 7 and 8 but giving relevance based on inter-document redundancy, i.e., intra-document redundancy is irrelevant, it could be present in any extent, see Figure 4.3g and Figure 4.3h. The other strategies are easier to understand with the presented diagrams, see figures 4.3a, 4.3b, 4.3c, and 4.3d.

For choosing which strategy to explore, two parameters are defined:

- *pd*: per-document redundancy parameter; indicates how intra-document redundancy should be managed,

$pd = 1$	<i>if intra-document redundancy is important,</i>
$pd = 0$	<i>if intra-document redundancy is irrelevant,</i>
$pd = -1$	<i>if intra-document redundancy is undesired.</i>
- *cd*: cross-document redundancy parameter; indicates how inter-document redundancy should be managed,

$cd = 1$	<i>if inter-document redundancy is important,</i>
$cd = 0$	<i>if inter-document redundancy is irrelevant,</i>
$cd = -1$	<i>if inter-document redundancy is undesired.</i>

Having established these strategies, all that follows is the merging of similar nodes on the graph, but taking care of how the relevance of the nodes is being carried on. This merging should be done first by document and then across documents. Starting from the graph with weighted edges and weighted nodes, we present the algorithm for merging:

Step 1 Choose the most similar two nodes on the graph: N_+ and N_- . Being N_+ the node with the highest weight (the important node) and N_- the other node (the less important node).

Step 2 Merge nodes, taking into account the following aspects:

1. The text of N_+ remains, while the other is dropped.
2. All the edges of N_- must be inherited to N_+ by averaging the weight of their correspondent edges.
3. Being $R(N_+)$ the weight of N_+ , it is re-computed depending on parameters pd and cd :

```

if redundancy is important ( $pd == 1$  or  $cd == 1$ ) then
     $R(N_+) = R(N_+) + R(N_-)$ 
else if redundancy is irrelevant ( $pd == 0$  or  $cd == 0$ ) then
     $R(N_+) = R(N_+)$ 
else if redundancy is undesired (if  $pd == -1$  or  $cd == -1$ ) then
     $R(N_+) = R(N_+) - R(N_-)$ 
end if

```

Step 3 Repeat step 1 until a similarity threshold is reached.

To depict this process, a single merging between two given nodes are showed in Figure 4.4. Let us follow the merging process between these two nodes together with the presented algorithm.

At step 1, the two most similar nodes are detected in Figure 4.4a, N_+ is $d_n s_2$ and N_- is $d_2 s_2$. Next, the merging of these two nodes for step 2 are presented from Figure 4.4b to Figure 4.4d. For all cases, steps 2.1 and 2.2 of the algorithm are performed in the same way: the text of $d_n s_2$ remains, while $d_2 s_2$ disappears; before disappearing, all the edges of $d_2 s_2$ are inherited to $d_n s_2$ by averaging their weights—this is represented as thickness of the edges, and one can also see how before merging there was no edge between nodes $d_n s_2$ and $d_2 s_1$, which later appears because node $d_2 s_2$ inherited it to $d_n s_2$ before disappearing, with half of its original thickness.

Now we focus on the three possible cases at step 2.3, where the weight of N_+ is re-computed. Figure 4.4b shows the resulting relevance of N_+ when redundancy is taken as important, Figure 4.4c shows the resulting relevance of N_+ when redundancy is taken as irrelevant, and Figure 4.4d shows the resulting relevance of N_+ when redundancy is undesired. In this way, the merging process increases, ignores, or penalizes the relevance of the node when it finds a similar sentence, giving the flexibility needed to experiment with the different strategies.

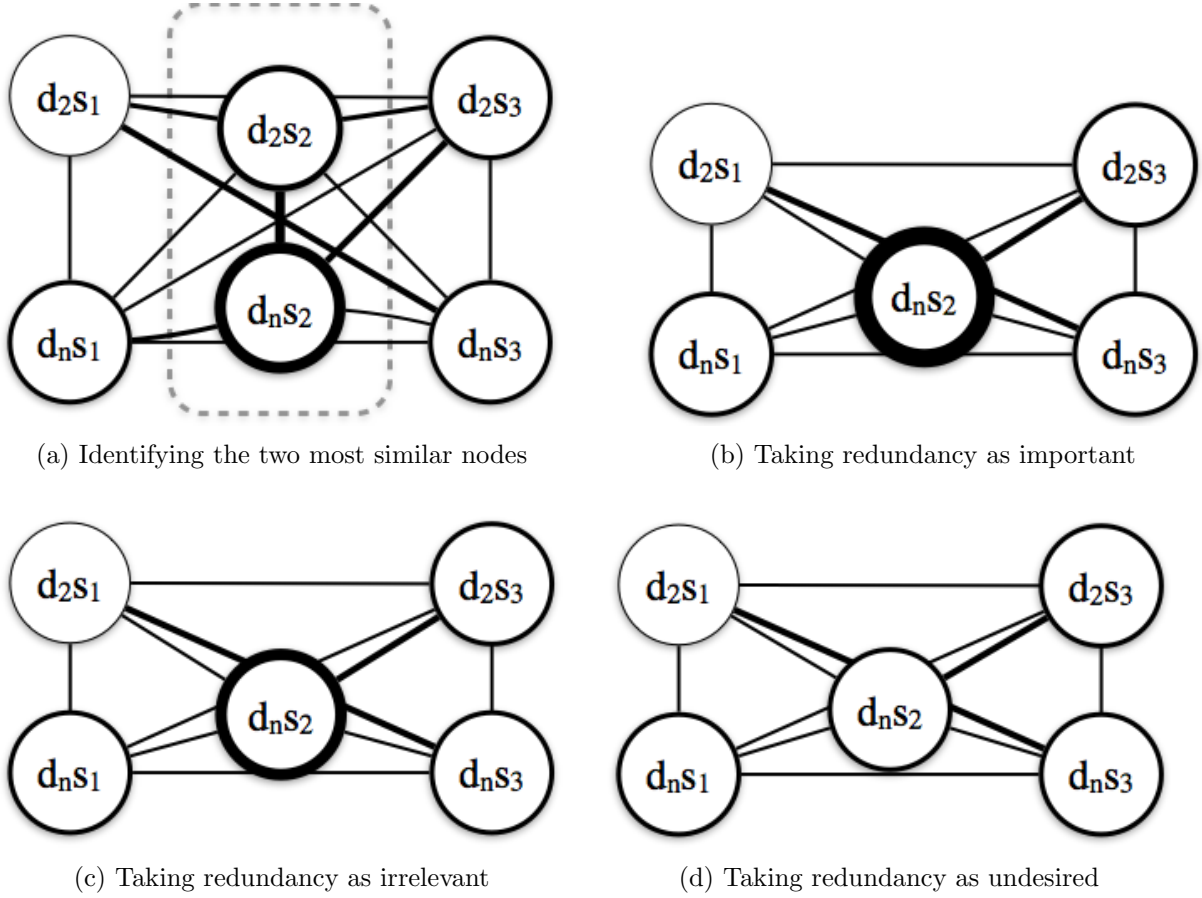


FIGURE 4.4: Merging process example of two nodes

When merging per document, at the first step the two chosen nodes must belong to the same document and the parameter taken into account at step 2.3 is *pd*. The process finishes when the established intra-document similarity threshold is reached. When merging across documents, at the first step the two chosen nodes must belong to different documents and the parameter taken into account at step 2.3 is *cd*. The process stops when the established inter-document similarity threshold is reached. Once the algorithm finishes, each sentence will have a computed relevance value according to the redundancy strategy utilized.

4.3 Summary generation

Once the assignment of relevance to the sentences has finished, a ranking of sentences will result. Usually at this stage, the most relevant sentences compose the final summary, but some measures are implemented to ensure that they are not redundant between each other.

Given that we removed redundancy together with the computing of relevance, no redundancy is expected in the ranked sentences. Therefore we can simply pick the top relevant sentences and employ them in the final summary until a maximum length is reached, ensuring non-redundancy.

Chapter 5

Experiments and results

In order to show the effectiveness of the proposed method, we present in this chapter the experiments we made along with the results obtained. First, we show a comparison on the similarity measures at measuring sentence similarity, to show how well they perform. After this, we present the experimental settings to test the proposed method, specifically, the corpora we used, and an analysis of the graphs built with these clusters of texts. Finally we present the results obtained when executing the nine proposed strategies of redundancy, along with an analysis of these results and a comparison with other methods.

5.1 Similarity measures comparison

We used two similarity measures for our experiments: the doc2vec-based similarity measure and superposition. For the doc2vec-based similarity measure, the *Paragraph Vector* method was used to create a model and generate vectors of the sentences, to later compare them using a cosine measure. Punctuation marks were removed to train the model, so the sentences to be represented afterwords in this vector space model are also processed without punctuation marks. For the superposition similarity measure, sentences were compared also without punctuation marks.

For training the doc2vec neural network, we adopted the Distributed Memory Model of Paragraph Vectors (PV-DM) model. The neural network was built with 100 neurons, a window between the predicted word and context words of 8 words, ignoring all words with total frequency lower than 5, and doing 10,000 iterations at the inference stage for computing new sentence vectors. The neural network requires big volumes of texts to be trained. For this requirement, the training was done with the dataset from “One Billion Word Language Modeling

Benchmark”¹ that has almost 1 billion words. Its texts are presented as files with a sentence per line and punctuation marks are separated from words as different tokens.

To exemplify the effectiveness of these similarity measures, we extracted and compared a few sample sentences from a cluster of news of DUC 2002, where the covered topic is the opening of the first McDonald’s in different countries. See Table 5.1 to see these sentences.

TABLE 5.1: Sentences from DUC 2002 used to test the similarity measures

Id	Sentence
s1	John Onoda, a spokesman at McDonald’s Oak Brook, Ill., headquarters, said it was the first of the chain’s outlets in a communist country.
s2	The restaurant has 350 seats and employs 110 people capable of serving 2500 meals per hour.
s3	Normally dour citizens broke into grins as they caught the infectious cheerful mood from youthful Soviet staffers hired for their ability to smile and work hard.
s4	The world’s largest version of the landmark American fast-food chain rang up 30000 meals on 27 cash registers, breaking the opening-day record for McDonald’s worldwide, officials said.
s5	McDonald’s hamburgers, fries and golden arches came to China on Monday when the fast-food chain opened its first restaurant in a nation famed for its distinctive cuisine.
s6	With seats for 900 and 27 cash registers, it served 30000 people on opening day.
s7	The world’s largest McDonald’s, with 27 cash registers and a seating capacity of 900, brings to Moscow not only hamburgers, french fries and shakes (called ‘milk cocktails’ here), but also a living lesson in Western-style marketing.
s8	At training sessions before opening day, cashiers were taught the importance of greeting customers cheerfully, of saying ‘please’ and ‘thank you’ — all of which promises something distinctly different from the typically surly service at most of Moscow’s dingy state cafes.

Let us consider sentences $s4$ and $s3$ as a reference to be compared with the other sentences. In the table one can see that the texts of sentence $s3$ and sentence $s8$ refer to a similar idea, also sentences $s4$, $s6$ and $s7$ seem to refer to the same fact. The similarity values between them are shown in Table 5.2. Similarities of the doc2vec-based measure are given between -1 and 1 , while similarities of the superposition measure start from 0 having no upper limits—the upper limit depends on the length of the sentence.

One can see that the highest values reported by the doc2vec-based similarity measure correspond to the most similar ideas. Whereas the values reported by the superposition similarity measure performs well at comparing sentence $s4$, but does not perform well when comparing $s3$ by giving

¹<http://www.statmt.org/lm-benchmark/1-billion-word-language-modeling-benchmark-r13output.tar.gz>

TABLE 5.2: Similarity values between sentences used for comparison

	s1	s2	s3	s4	s5	s6	s7	s8
Doc2vec-based similarity values								
s3	0.012	0.071	1.000	−0.081	0.216	0.048	0.055	0.240
s4	0.149	0.131	−0.081	1.000	0.123	0.396	0.153	0.136
Superposition similarity values								
s3	0.356	0.764	9.187	0.688	1.390	0.772	0.998	0.991
s4	2.077	1.113	0.688	9.387	2.708	2.999	2.920	1.611

the highest value to an idea that is not similar. One can see that doc2vec performs better at comparing sentences that are similar but not necessarily by sharing the same words—as the case between *s3* and *s8*.

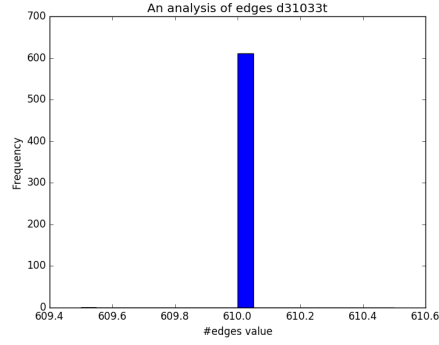
5.2 Experimental settings

For our experiments we used the DUC 2004 task 2 and DUC 2007 main task corpora. No data set to train the method was necessary since it is unsupervised. The specifications of this corpora is showed in Table 5.3. The corpus of DUC 2004 task 2 consists of 50 clusters of 10 documents each (500 documents in total) and 4 reference summaries for each cluster (200 summaries in total). For the main task of DUC 2007, 45 clusters are provided, each cluster with 25 documents (1125 documents in total) and also with 4 different summaries for each cluster (180 summaries). Summaries on DUC 2007 are query-focused, while DUC 2004 summaries are completely generic.

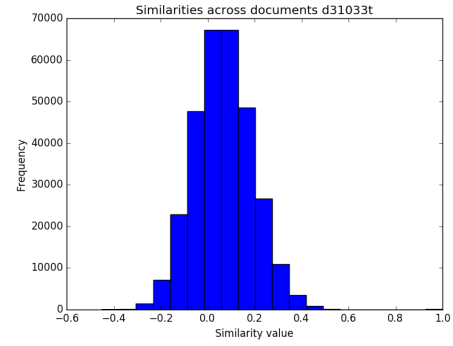
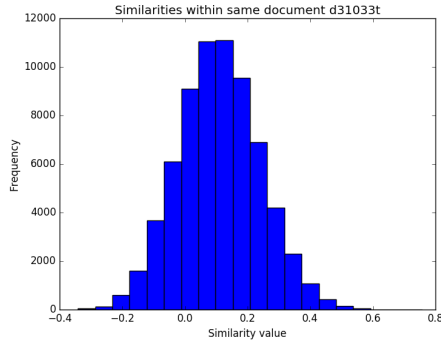
TABLE 5.3: Specifications of the DUC 2004 and DUC 2007 corpora

	DUC 2004 task 2	DUC 2007 main task
Type of summaries	Generic	Query-focused
Number of clusters	50	45
Number of documents per cluster	10	25
Number of reference summaries per cluster	4	4
Maximum length allowed in summaries	665 bytes	250 words

By using DUC corpora, we wanted to study which of the previous stated redundancy strategies are gold standard DUC summaries using. Since the majority of works involved in summarization compares their resulting summaries against those summaries provided by DUC, the study led us to state the general understanding of what makes an idea important to become part of a summary—according to redundancy.

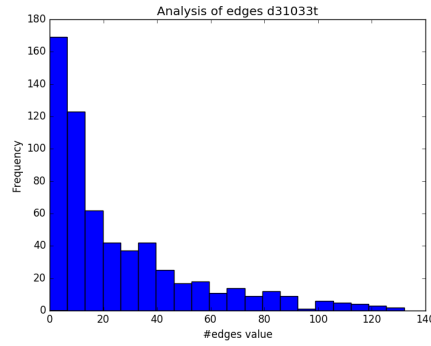


(a) Frequency of edges per node

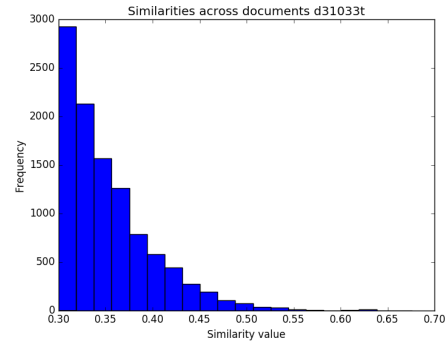
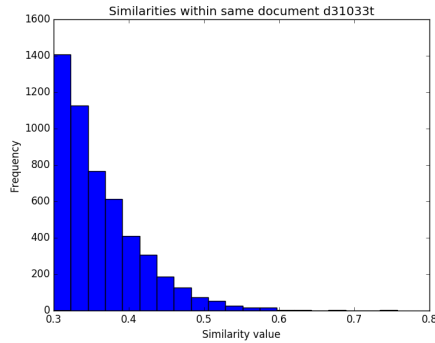


(b) Similarity distribution of sentences within the same document

(c) Similarity distribution of sentences from different documents



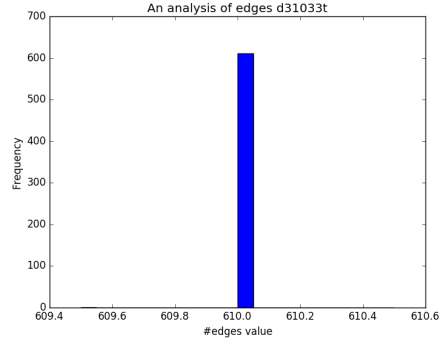
(d) Frequency of edges per node after pruning



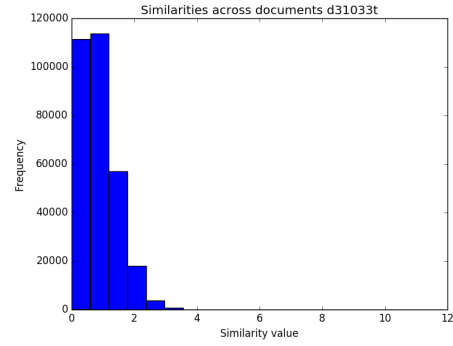
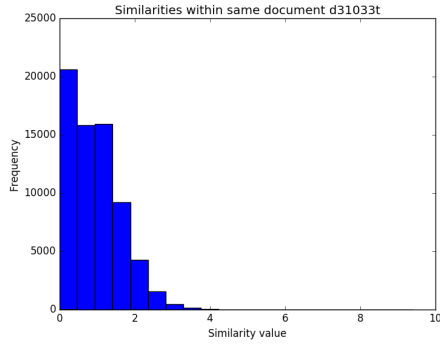
(e) Similarity distribution of sentences within the same document after pruning

(f) Similarity distribution of sentences from different documents after pruning

FIGURE 5.1: Histograms illustrating graph edges distribution of the $d31033t$ cluster from DUC 2004, built with the doc2vec-based similarity measure

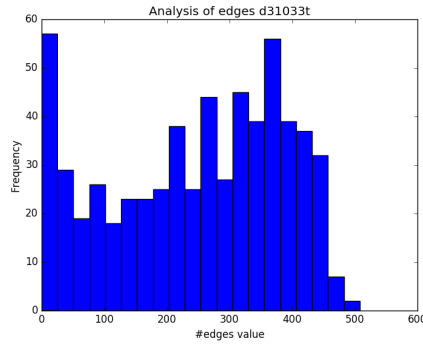


(a) Frequency of edges per node

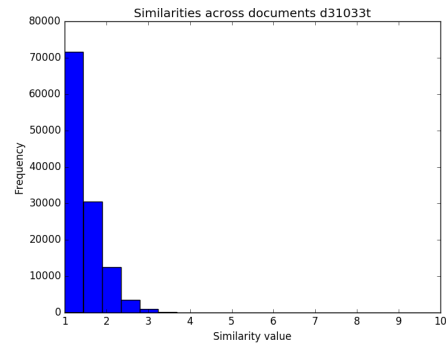
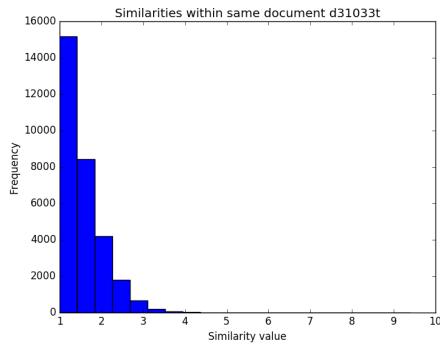


(b) Similarity distribution of sentences within the same document

(c) Similarity distribution of sentences from different documents



(d) Frequency of edges per node after pruning



(e) Similarity distribution of sentences within the same document after pruning

(f) Similarity distribution of sentences from different documents after pruning

FIGURE 5.2: Histograms illustrating graph edges distribution of the *d31033t* cluster from DUC 2004, built with the superposition similarity measure

In order to obtain general information about the constructed graphs, plots have been generated for each graph, i.e., for each cluster of texts. These plots are histograms showing the quantity of edges on the graph, and the similarity distribution over its edges. After analyzing the histograms of each cluster we observed a similar behavior for all of them.

An example of the histograms for a single cluster of texts from DUC 2004 is shown in Figure 5.1 for the graph built with the doc2vec-based similarity measure, and in Figure 5.2 for the graph built with the superposition similarity measure. These plots correspond to the cluster of texts with the greatest number of sentences.

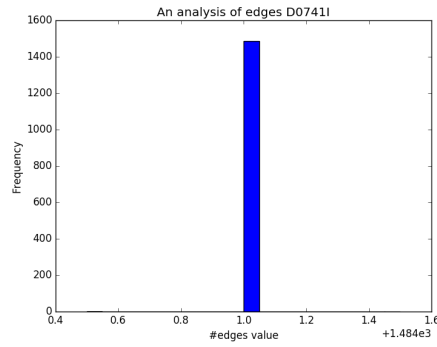
As it is possible to visualize in figures 5.1a and 5.2a, this cluster contains 611 sentences, generating a total of 372710 edges—because is a complete graph (610 edges each node). On the other hand we observed that the cluster with the least number of edges was cluster *d30055t*, having 153 sentences and 23256 edges (152 edges each node).

The obtained distribution of similarities per and across documents for all clusters, were similar to the ones showed in figures 5.1b and 5.1c for the doc2vec-based similarity measure, and to the ones showed in figures 5.2b and 5.2c for the superposition similarity measure. For the doc2vec-based similarity measure, it is a normal distribution, which mean is close to 0.1. One can see also that strong similarities exist at big values on similarities across documents.

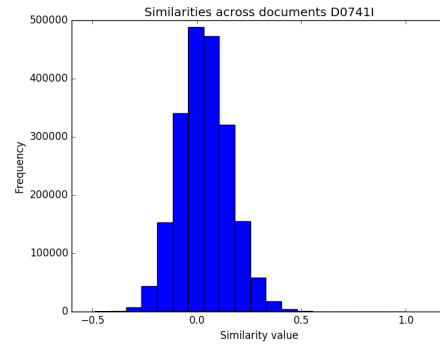
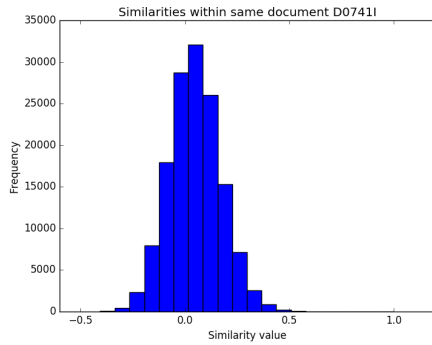
Additionally, an example of the histograms for a single cluster of texts from DUC 2007 is shown in Figure 5.3 for the graph built with the doc2vec-based similarity measure, and in Figure 5.4 for the graph built with the superposition similarity measure. These plots correspond to the cluster of texts with the greatest number of sentences.

As showed in figures 5.3a and 5.4a, this cluster contains 1485 sentences, generating a total of 2203740 edges—because is a complete graph (1484 edges each node). On the other hand we observed that the cluster with the least number of edges was cluster *D0732H*, having 160 sentences and 25440 edges (159 edges each node).

An interesting fact to notice in these plots is how the effect of redundancy across documents predominates over redundancy per document. One can see this by noticing that the number of edges between sentences from different documents is much bigger than the number of edges between sentences within the same document. For example, see the *y* axis of figures 5.3b and 5.3c, the counting of edges in the first case is made in tens of thousands, whereas the counting of edges in the second case is made in hundreds of thousands.

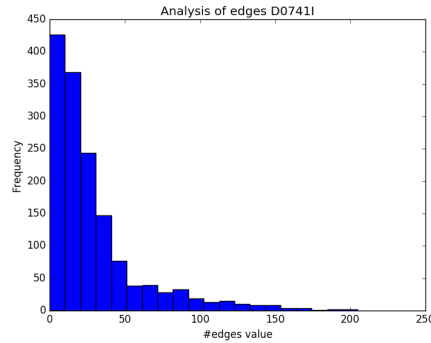


(a) Frequency of edges per node

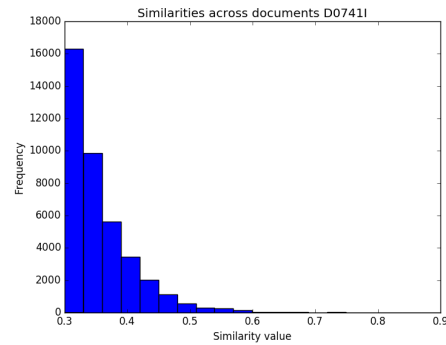
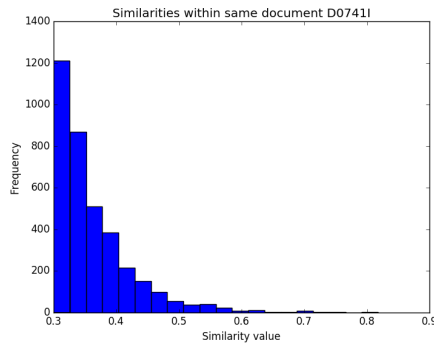


(b) Similarity distribution of sentences within the same document

(c) Similarity distribution of sentences from different documents



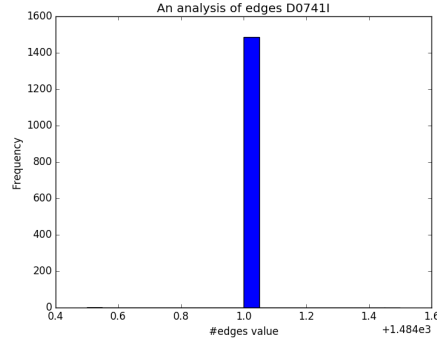
(d) Frequency of edges per node after pruning



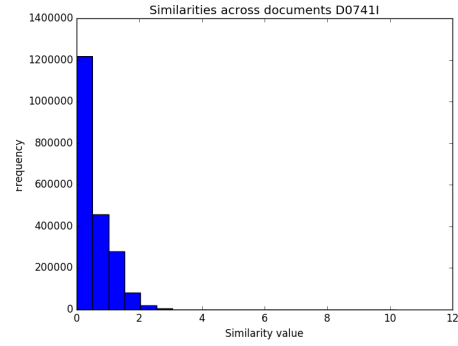
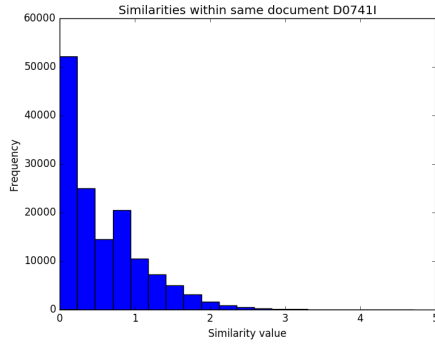
(e) Similarity distribution of sentences within the same document after pruning

(f) Similarity distribution of sentences from different documents after pruning

FIGURE 5.3: Histograms illustrating graph edges distribution of the *D0741I* cluster from DUC 2007, built with the doc2vec-based similarity measure

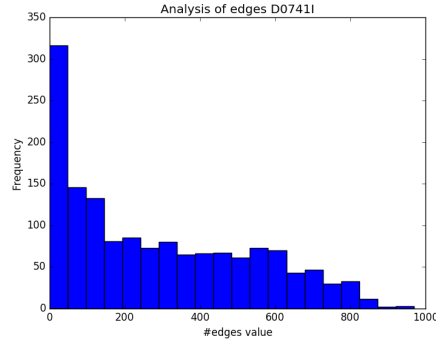


(a) Frequency of edges per node

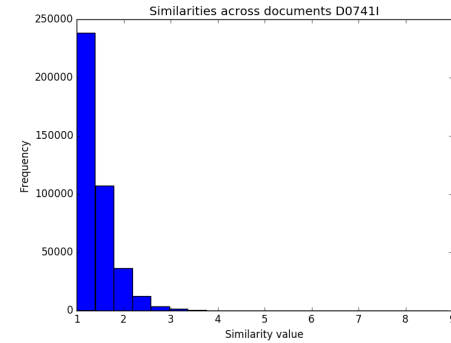
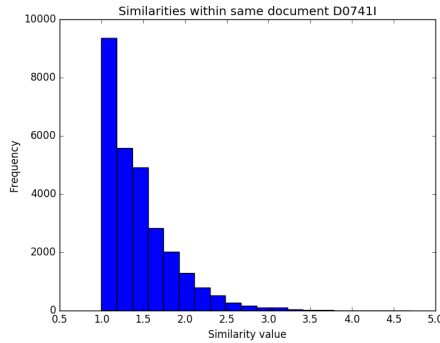


(b) Similarity distribution of sentences within the same document

(c) Similarity distribution of sentences from different documents



(d) Frequency of edges per node after pruning



(e) Similarity distribution of sentences within the same document after pruning

(f) Similarity distribution of sentences from different documents after pruning

FIGURE 5.4: Histograms illustrating graph edges distribution of the *D0741I* cluster from DUC 2007, built with the superposition similarity measure

Analyzing this information, we decided to establish an edge threshold on the graph to determine if generate or not an edge. With this threshold we could avoid having a very complex graph with useless information, i.e., deleting small similarities with which do not provide information about redundancy.

The thresholds we used for this purpose on the doc2vec-based graph were 0.3 as the minimum and 0.9 as the maximum similarity. The upper threshold allows us to eliminate similar ideas whose text is almost the same—with the same words but missing or changing one or two of these words. For valuable redundancy, we do not look for precisely identical sentences but similar ones, expressed in a different way; being this similarity usually unconscious to the writers.

To see an example of the resulting doc2vec graphs after the implementation of the edge threshold, see again figures 5.1 and 5.3. From Figure 5.1d one can see that the graph for the biggest cluster of documents in DUC 2004 went from having 610 edges per node to having no more than 140 edges per node, varying now the quantity of edges per node. A bigger reduction is saw in Figure 5.3d for the biggest cluster of documents in DUC 2007, it went from having 1484 edges per node to have no more than 250 edges.

On the superposition graph, we established an edge threshold of 1 as the minimum and 10 as the maximum. For this case we could not establish an appropriate maximum threshold to avoid edges between sentences with identical words, because this value vary depending on the length of the sentences in spite of being exactly alike.

To get an idea of the resulting superposition graphs after the implementation of the edge threshold, see again figures 5.2 and 5.4. The reduction of edges in this case was not as big as with the doc2vec graphs. One can see in Figure 5.2d that there are tens of nodes with a number between 300 and 400 of edges. A similar case of reduction can be observed for the cluster of texts of the DUC 2007 in Figure 5.4d.

Besides the edge threshold, this analysis also allowed us to establish an appropriate threshold to stop the merging process per and across documents. Since the histograms do not show a big difference in the similarity distribution over similarities per document and across documents, we established the same threshold value for both merging processes.

For the doc2vec graph, a merging threshold of 0.5 was established. With a similarity as strictly high as 0.5, we increase the possibilities to merge nodes with good similarities. Take a look at Table 5.2 were some high similarities values do not guarantee a real similarity, for example, see the Doc2vec-based similarity between sentence $s5$ and sentence $s3$. One can see in figures 5.1e

and 5.1f, for DUC 2004, and in figures 5.3e and 5.3f, for DUC 2007, that in spite of the high threshold, a high number of edges still exist to perform the merging process.

For the superposition graph, a merging threshold of 2 was established. This is not as high as the one established for doc2vec, one can see in figures 5.2e and 5.2f, for DUC 2004, and in figures 5.4e and 5.4f, for DUC 2007, that a lot more of edges remain on the graph by establishing this threshold, in comparison with the doc2vec threshold. We expected that by decreasing the threshold, we would give a chance to this measure to perform as well as doc2vec by taking into account more similarities—more information—in merging process.

With all these threshold values set, we proceeded to build the graphs ignoring edges out of the range between the maximum and minimum edge thresholds, and we executed the merging process per document and across documents until the merging threshold was reached.

5.3 Results

We tested the merging method with all nine redundancy strategies, on the graphs built by using the two chosen similarity measures on the cluster of texts of DUC 2004 and DUC 2007 corpora. The resulting summaries were evaluated against summaries provided by DUC with the ROUGE method.

The ROUGE script version we used was ROUGE 1.5.5, Table 5.4 shows the command line parameters used for its execution.

TABLE 5.4: Command line parameters used for execution of ROUGE

-n 4	to generate the ROUGE-N scores from unigrams to 4-grams,
-b 665	to use the first 665 bytes in the candidate summary for the evaluation of DUC 2004 summaries,
-l 250	to use the first 250 words in the candidate summary for the evaluation of DUC 2007 summaries,
-m	to stem both model and system summaries using Porter stemmer before computing various statistics,
-2 4	to compute skip bigram (ROUGE-S) co-occurrence with a maximum gap length between two words of 4,
-u	to include the unigram counting to the computed skip bigrams (ROUGE-SU),
-a	to evaluate all systems specified in the ROUGE-eval-config-file,
-d	to print per evaluation average score for each system.

The execution line was:

```
./ROUGE-1.5.5.pl -n 4 [-b 665, -l 250] -m -2 4 -u -a -d <ROUGE-eval-config-file>
```

The specifications of how to configure ROUGE, along with the details for each parameter, can be consulted on the README text file that comes with the whole package of the official software².

Tables 5.5 and 5.6 present the results of the experiments performed on the DUC 2004 corpus, while Tables 5.7 and 5.8 present the results of the experiments performed on the DUC 2007 corpus. Only recall measures are presented for these experiments, and results are ordered by the ROUGE-1 in descending order,

As one can see, the results when using the doc2vec-based similarity measure have some differences when using the superposition similarity measure. As we presented before, doc2vec performs better than superposition by giving good similarities on sentences that do not use the exact same words. Also, the ROUGE scores obtained with doc2vec were predominantly better than those obtained with superposition. For these reasons, we provide the analysis of the results mainly based on the doc2vec-based similarity measure, taking the resulting scores of the superposition similarity measure to reinforce our ideas.

TABLE 5.5: ROUGE results on all strategies of redundancy for DUC 2004 texts when using the doc2vec-based similarity measure

Strategy	Parameters		ROUGE-1	ROUGE-2	ROUGE-SU4
	<i>pd</i>	<i>cd</i>			
7	0	1	0.3775	0.0771	0.1283
4	-1	1	0.3763	0.0765	0.1273
1	1	1	0.3709	0.0754	0.1251
9	0	0	0.3594	0.0719	0.1210
6	-1	0	0.3590	0.0709	0.1198
8	0	-1	0.3560	0.0675	0.1171
5	-1	-1	0.3546	0.0669	0.1162
3	1	0	0.3504	0.0657	0.1137
2	1	-1	0.3471	0.0634	0.1120

One can see in Table 5.5 that for DUC 2004, when using doc2vec, the strategy that generated the best ROUGE results was number 7, i.e., giving relevance to ideas redundant across documents. Even the next two best results, along with the four best results when using the superposition measure (see Table 5.6), reinforce the same idea: giving relevance to ideas redundant across

²www.berouge.com

TABLE 5.6: ROUGE results on all strategies of redundancy for DUC 2004 texts when using the superposition similarity measure

Strategy	Parameters		ROUGE-1	ROUGE-2	ROUGE-SU4
	pd	cd			
1	1	1	0.3739	0.0736	0.1236
9	0	0	0.3714	0.0698	0.1212
7	0	1	0.3677	0.0740	0.1219
4	-1	1	0.3667	0.0718	0.1203
3	1	0	0.3581	0.0590	0.1130
2	1	-1	0.3510	0.0579	0.1102
6	-1	0	0.3490	0.0617	0.1120
8	0	-1	0.3427	0.0522	0.1056
5	-1	-1	0.3100	0.0379	0.0916

documents ($cd = 1$) is the best possible option. On the other hand, the strategies where an idea is penalized by being redundant across documents ($cd = -1$) were ranked in the four worst results for both similarity measures.

Another interesting pattern to notice in Table 5.5, is when the cd parameter is left static, the pd parameter follows the same ranking: first, it is best to ignore intra-document redundancy ($pd = 0$), a less good option is to penalize it ($pd = -1$), and the worst option is to taking it into account to provide relevance ($pd = 1$). A similar case happens when the pd parameter is left static, being the same rank for the cd parameter: the best option is to take into account inter-document redundancy ($cd = 1$), then ignore this redundancy ($cd = 0$) and being the worst option to penalize it ($cd = -1$).

The more inter-document redundancy is considered as favorable, the better the result, whereas intra-document redundancy is irrelevant for the importance of the sentence.

In the case of DUC 2007, one can see in Table 5.7 that the best strategy is between strategies number 6 and number 4 just by a little difference on the scores, where both penalize the relevance of the sentence if it is redundant per document (i.e., intra-document redundancy). Strategy 6 does not take into account inter-document redundancy while strategy 4 does. Consulting the results when using superposition on Table 5.8 to decide which one is better, one can say that is strategy 4. On the other hand, the worst three strategies, for both similarity measures, resulted from penalizing inter-document redundancy ($cd = -1$).

TABLE 5.7: ROUGE results on all strategies of redundancy for DUC 2007 texts when using the doc2vec-based similarity measure

Strategy	Parameters		ROUGE-1	ROUGE-2	ROUGE-SU4
	pd	cd			
6	-1	0	0.4197	0.1008	0.1558
4	-1	1	0.4194	0.1037	0.1581
9	0	0	0.4174	0.1019	0.1566
7	0	1	0.4152	0.1026	0.1568
1	1	1	0.4122	0.1017	0.1554
3	1	0	0.4106	0.0981	0.1523
5	-1	-1	0.4105	0.0886	0.1457
8	0	-1	0.4098	0.0910	0.1474
2	1	-1	0.4081	0.0926	0.1478

TABLE 5.8: ROUGE results on all strategies of redundancy for DUC 2007 texts when using the superposition similarity measure

Strategy	Parameters		ROUGE-1	ROUGE-2	ROUGE-SU4
	pd	cd			
7	0	1	0.4263	0.1008	0.1578
4	-1	1	0.4137	0.0916	0.1492
1	1	1	0.4117	0.0898	0.1487
9	0	0	0.4085	0.0848	0.1453
3	1	0	0.4019	0.0804	0.1417
6	-1	0	0.4000	0.0842	0.1431
2	1	-1	0.3930	0.0762	0.1362
8	0	-1	0.3814	0.0678	0.1300
5	-1	-1	0.3613	0.0580	0.1206

An interesting pattern to observe in Table 5.7, is when the cd parameter is left static: the best option is to penalize intra-document redundancy ($pd = -1$), then ignore this redundancy ($pd = 0$) and the worst option is giving relevance when intra-document redundancy is present ($pd = 1$). This same pattern is not present when using superposition, see Table 5.8. In fact, there is no consistency on the results when varying the intra-document redundancy parameter. However, there is a clear consistency when varying the cd parameter: inter-document redundancy is good for a sentence.

For the case of DUC 2007, we can say that the more redundancy per document is penalized, the

better the result, whereas ideas redundant across documents (i.e., inter-document redundancy) are desirable. In other words, ideas not biased to a single source are better. Considering that DUC 2007 summaries are topic-oriented, one can think that it is more likely to solve a specific query by looking for ideas that are not continuously addressed by a single author.

The fact that some strategies got a higher score than strategy 9 ($pd = 0$, $cd = 0$), means that taking redundancy to provide relevance to sentences does help to compose a good summary. We can ensure this fact based on what a good summary is according to the gold standard of DUC corpora.

5.4 Comparison with existent methods

Despite the goal of this work is not to outperform the best techniques on multi-document summarization, we present a comparison with other techniques to show that this technique, on its best redundancy strategy for the DUC corpora, is competent when outperforming well-known baselines. We did not focus on outperforming the best results by now, to leave the method as simple as possible. In this way, it is easy to manipulate the graph and provide the flexibility needed to test different redundancy strategies in other areas different from news. However, some ideas to outperform these methods are provided later in future work.

We present a comparison between the results of our method with the best redundancy strategy using the doc2vec-based similarity measure, which was the number 7 (taking ideas redundant across documents) for DUC 2004, and number 4 (taking ideas redundant across documents, but not redundant per document) for DUC 2007, against baselines and best performing methods. For this comparisons see Tables 5.9 and 5.10.

The baselines we compare to are: random (the measure stated in the table is the median of 15 runs), lead and coverage. In addition, we compare against the best performing methods for each ROUGE measure. The way these methods work can be consulted in the *State of the art* chapter. We show only the recall measure for ROUGE-1, ROUGE-2 and ROUGE-SU4, because it is the available score for most of all these methods.

By using our best scored strategies of redundancy we did not outperform the scores of the best methods, however, we outperformed significantly—see *Discussion* section to consult confidence intervals—all baselines on DUC 2004 and DUC 2007 corpora.

TABLE 5.9: Comparison with baselines and best methods using DUC 2004 corpus

Method	ROUGE-1	ROUGE-2	ROUGE-SU4
PatSum [23]	–	0.1020	–
wHAASum [24]	0.4167	0.0956	0.1386
SentTopic-MultiRank [26]	0.4101	0.0991	0.1432
Our method ($pd = 0$, $cd = 1$)	0.3775	0.0771	0.1283
Coverage	0.3440	0.0768	0.1172
Lead	0.3299	0.0634	0.1051
Random	0.3198	0.0498	0.0968

TABLE 5.10: Comparison with baselines and best methods using DUC 2007 corpus

Method	ROUGE-1	ROUGE-2	ROUGE-SU4
ILP2 [36]	–	0.1251	0.1760
IIIT Hyderabad (peer 15) [37]	–	0.1244	0.1771
PLSA-JS [32]	0.4584	0.1167	0.1768
Our method ($pd = -1$, $cd = 1$)	0.4194	0.1037	0.1581
Random	0.3753	0.0698	0.1259
Coverage	0.3655	0.0839	0.1288
Lead	0.3091	0.0603	0.1071

5.5 Discussion

Since we are basing the analysis of the results on statistical difference, we want to make an analysis of how significant these differences are. ROUGE script provide each result with a confidence interval of 95%, with an average, minimum and maximum value. The results reported previously in the *Results* section, correspond to the average value.

The confidence interval serves as a reference for comparison to other methods. Since this comparison is made over a confidence of 95%, one can make a binary decision to say whether one is better than the another or not. For example, let A be the method with a given confidence interval, and method B the method to be compared with the average score. If the average value of method B is under the minimum of the confidence interval of method A, one can say that method B is worst than method A—with a confidence of at least 95%. On the other hand, if the average value of method B is above the maximum of the confidence interval of method A, one can say that method B is better than method A—with a confidence of at least 95%.

Taking this into account we present in Table 5.11 and Table 5.12 the reported ROUGE recall scores, using the doc2vec-based similarity measure, for DUC 2004 and DUC 2007 respectively, showing the confidence interval for each score. To decide whether one strategy is better than another, we start comparing the results by taking the minimum value of the best strategy and comparing it with the average score for all other strategies.

TABLE 5.11: ROUGE-1 results on all strategies of redundancy for DUC 2004, using doc2vec, showing a 95% confidence interval

Strategy	Parameters		average	minimum	maximum
	<i>pd</i>	<i>cd</i>			
7	0	1	0.3775	0.3651	0.3897
4	-1	1	0.3763	0.3632	0.3889
1	1	1	0.3709	0.3565	0.3851
9	0	0	0.3594	0.3461	0.3728
6	-1	0	0.3590	0.3469	0.3711
8	0	-1	0.3560	0.3423	0.3691
5	-1	-1	0.3546	0.3403	0.3680
3	1	0	0.3504	0.3357	0.3643
2	1	-1	0.3471	0.3328	0.3603

For DUC 2004, one can see that all strategies with $cd = 1$ have a minimum score greater than all other strategies average score. Therefore, we can conclude that these strategies where $cd = 1$ outperform others with a 95% confidence. And that is all that we can conclude with a high certainty, the strategies where the cd flag varies between 0 and -1 do not show significant differences, neither when the pd flag is varied among all its possible values: 1, 0, -1 .

For DUC 2007, one can see that there are no significance differences between the results. Hence, all the conclusions we have presented on the *Results* section for this corpus, are below a confidence of 95

Concluding, we can say for sure that taking inter-document redundancy as important is better than ignoring it or penalizing it for generic summaries—for DUC 2004 corpus. However, for query-focused summaries we can not ensure the same thing, since we obtained the results with a confidence below 95%. For intra-document redundancy, we saw that it could be inconvenient to take ideas redundant per document into the summary, however this affirmation is also under a confidence of 95%.

TABLE 5.12: ROUGE-1 results on all strategies of redundancy for DUC 2007, using doc2vec, showing a 95% confidence interval

Strategy	Parameters		average	minimum	maximum
	<i>pd</i>	<i>cd</i>			
6	-1	0	0.4197	0.4082	0.4313
4	-1	1	0.4194	0.4073	0.4307
9	0	0	0.4174	0.4058	0.4280
7	0	1	0.4152	0.4031	0.4263
1	1	1	0.4122	0.3996	0.4243
3	1	0	0.4106	0.3976	0.4230
5	-1	-1	0.4105	0.3996	0.4203
8	0	-1	0.4098	0.3979	0.4208
2	1	-1	0.4081	0.3949	0.4217

Analyzing some examples, we saw that intra-document redundancy is inconvenient because giving importance to this type of redundancy, reduces the possibility of taking inter-document redundancy into the final summary, which is more relevant for multi-document summarization. An analysis of this idea is provided in *Appendix A*.

Chapter 6

Conclusions

From the results of the proposed method, some valuable facts for the area of multi-document summarization were found. We also made some contributions with the development of the method, and some areas of opportunities for future work were detected. In this chapter we summarize all these observations that are worth highlighting.

6.1 Conclusions

Studying the DUC corpora led us to establish a general statement about what makes an idea important according to redundancy. We discovered that summaries should convey ideas redundant across documents, whereas taking ideas redundant per document could be inconvenient. Since the main conclusion was obtained with a confidence greater than 95%, and matches with the human intuition of what a summary is, we can define a multi-document summary as: a piece of text that contains the most redundant—popular—information between different sources.

We found also that providing relevance to ideas according to their level of redundancy does improve the summary. For all the methods were a final ranking of sentences results as a process of computing sentence relevance, this method could be implemented before eliminating redundancy to take it into account as valuable information.

As a method that is getting involved in the area of multi-document summarization, some disadvantages have to be noticed. By focusing our efforts on retrieving ideas to give a general understanding of what the multiple text are covering, we disregarded some of the challenges. From these challenges, the method manages well the phenomenon of redundancy, however, the

challenge of detecting contradictory ideas in the documents, for example, is not well managed. This is because the similarity measures acts by context or co-occurrence of words, giving the possibility of assigning a high similarity between contradictory ideas. Another aspect that we left out was the readability of the summary. Given the nature of the extractive summaries, it is difficult to produce a coherent summary because the jump between sentences can break the fluency of the reader.

In the process of this work, we observed also that good similarity measures are missing in the area of NLP. Although the doc2vec-based similarity measure performs well in some cases, it fails when giving high similarity to sentences containing words that shares similar contexts but that are not similar. The presented method in this work, can be a good tool to test similarity measures—to perform extrinsic tests on these measures. By configuring the right redundancy strategies, and implementing the proposed similarity measure, it can be used to see if the method throws the most similar or rare ideas in the texts. With a good similarity measure, this method would have great potential to exploit the redundancy present in the Internet.

6.2 Contributions

The main contributions of this work are:

1. A study of different strategies to deal with inter-document and intra-document redundancy to extract information.
2. The development and implementation of a unsupervised method to obtain generic summaries of multiple documents.

For the first contribution, we showed nine different ways of extracting information depending of how the redundancy within a document and in different documents is managed. Along with these strategies, we proposed a merging process of nodes in a graph to experiment with all these variations on intra-document and inter-document redundancy, allowing to extract popular or rare (which can be viewed also as novel or diverse) information within a cluster of documents. It can be useful for some areas where this flexibility is needed to provide a general idea of what multiple texts talk about.

For the second contribution, we implemented the merging process as an unsupervised method to summarize multiple documents. This method was capable of generating good summaries since it outperformed baselines methods. Moreover, as this method works from the computation

of an initial sentence relevance, it can be used also as a mechanism to eliminate redundancy in the final phase of the summary construction. The methods where this mechanism can be adopted is in those where a ranking of sentences is obtained with a relevance value for each sentence. With this relevance value, texts can be mapped to the graph in the same way as we showed, but changing the node weights with this given relevance instead of the one computed by TextRank. The adoption of this method can lead to the improvement of the final summary with the appropriated redundancy strategies depending on the ambit that is being studied.

Applying all the redundancy strategies using the unsupervised method on the DUC corpora, we could find which of these strategies humans experts used to generate summaries. By studying this widely used corpora, we could get to the general understanding of what makes an idea important to be part of a summary in the ambit of news, according to redundancy. Also by applying this method in summarization, we introduced a new perspective to produce generic summaries based on popularity or rarity.

6.3 Publication

Derived from this work, we participated in the 15th Mexican International Conference on Artificial Intelligence (MICA) with a publication and an oral presentation entitled “Intra-Document and Inter-Document Redundancy on Multi-Document Summarization”. In this publication, the results of the method using the doc2vec-based similarity measure on the DUC 2004 and DUC 2007 corpora were presented. The publication related with this presentation is expected to be in the Lectures Notes in Artificial Intelligence (LNAI) of Springer.

6.4 Future work

Many ideas to continue this work resulted from a continuous feedback from seminars and meetings with other colleagues in the area of AI and NLP.

One of the ideas to better this work is to enhance the readability of the final summary. Two main ideas can be implemented for this purpose. The first, is to implement a method in the phase of summary generation, to identify the correct order to present the highest ranked sentences. To find this order, one could add directed edges on the graph according to the order of appearance of the sentences on each document, indicating which sentence follows another. In this way a node would have a incoming edge from the node corresponding to the previous sentence on

the document, and an outgoing edge to the node corresponding to the next sentence on the document. These edges will remain in the process of merging, by inheriting all incoming and outgoing edges of the irrelevant node to the relevant node. The second idea is to generate from the important ideas, an abstractive summary and avoid having jumps between sentences.

Something else to better the results of this work is to continue the experiments by varying all the parameters involved. For example, it could be possible to improve the similarity measure by varying the parameters on the neural network of *Paragraph Vector* method, and adding more data. It could be possible also to experiment with more types of pre-processing on the text as stemming, or by removing or not punctuation marks or stop words. More experiments on the proposed merging method can be done in other corpora apart from DUC, to test if the same criteria of redundancy to build a good summary applies in other scopes apart from news.

Another idea came up when we analyzed the evaluation scores of each cluster of the the set of clusters from DUC. We saw that the results obtained when varying the *pd* parameter were not consistent among clusters, for some it was better to penalize intra-document redundancy, whereas for others it was better to compensate it. From these observation two main ideas could be applied: first, to implement a classifier to know when is convenient to penalize or to compensate intra-document redundancy depending on the cluster characteristics; and the other idea is to analyze the writting style of humans in charge of generating the gold standard summaries of DUC, to see if there is consistency between the judgments of how redundancy should be managed.

A last idea that can be implemented on this method is to change the way of computing the initial sentences relevance by testing different techniques from TextRank, even if this method does not use the information of the graph. As we said before, by implementing this method as a mechanism to eliminate redundancy at the end, when a previous sentence relevance was computed, it could be possible to find more relevant sentences based on redundancy. We think that applying the merging mechanism to the best performing methods, could lead to the improvement of the ROUGE scores.

Appendix A

An example of the merging process

In this appendix an example of the merging process is presented. Since one of the findings in the study was counterintuitive, i.e., that ideas redundant per document are inconvenient to be in the summary, we show a case where penalizing intra-document redundancy was better than compensating it, when using the doc2vec-based similarity measure. This was the case of the *d30027t* cluster, where independently of the *cd* parameter, i.e. how inter-document redundancy was managed, it was always better to penalize intra-document redundancy.

A.1 Cluster of texts

This section presents the texts corresponding to the cluster *d30027t* from DUC 2004 corpus, it contains a total of 348 sentences. The sentences of the texts are enumerated between square brackets, to trace it easily in the merging process.

1. NYT19981002.0250

[1] In a season of crashing banks, plunging rubles, bouncing paychecks, failing crops and rotating governments, maybe it is not the ultimate insult. [2] But the nation that bore Tolstoy and Chekhov, and still regards a well-written letter as a labor of love, is buckling a little this week, because it can no longer wish good health to Baba Anya in Omsk. [3] The Post Office is broke. [4] In 60 of the country's 89 statelike regions, more than 1,000 mail cars have been sidetracked, many stuffed with up to 18 tons of letters, newspapers and parcels. [5] The state Railway Ministry refuses to carry more mail until the Post Office makes good on some 210 million rubles

in old bills – about \$13 million in today’s dollars, or \$35 million in dollars six weeks ago. [6] Air mail, which amounts to one of every four or five letters, was also suspended at one of Moscow’s major airports until this week, when the Post Office coughed up 5 million rubles for old bills. [7] A second airport is still demanding 3 million rubles for past-due debts. [8] So much mail is backed up that post offices in Moscow and elsewhere have simply stopped accepting out-of-town mail, except for areas that can be easily reached by truck. [9] Delivery schedules have fallen weeks, and perhaps months, behind. [10] “The situation is really extraordinary,” said Vladimir Sherekhov, the deputy chief of mail service administration in the government’s Communications Committee. [11] “We’ve never had anything like this before”. [12] Maybe the Post Office has been lucky. [13] Extraordinary is the rule elsewhere in Russia. [14] Until Friday, the lower house of Parliament was preparing to sue the government for failure to provide soap, heat, toilet paper and copy-machine paper in the legislature’s monumental downtown offices. [15] It turned out that politicians had exhausted their funds by cutting short their recess and returning to address the nation’s economic crisis. [16] Earlier in the week, officials said Russia’s Arctic shipping routes may close next month because half the nation’s icebreakers are in disrepair and there is no money to fix them. [17] The oldest of the ships is so ancient that its nickname is Granny. [18] Such anecdotal evidence that Russia is losing its wheels, like one of its old, ill-maintained Volga sedans, is everywhere. [19] But oddly, real signs of public distress are not particularly common, perhaps because the system rarely seems to shed a part as big as a postal system. [20] If the U.S. Postal Service is increasingly a pipeline for sweepstakes notices and bills, the Russian Post Office still holds a special place in the national conversation. [21] Russians still write letters to each other, frequently and fervently, and many in remote regions get their news through the mail. [22] Millions use the mails to ship canned goods and other provisions to needy relatives and friends in faraway areas, an especially vital service in winter. [23] And in the last few years, the Post Office has become a vehicle for a growing mail-order trade in books, clothes and other catalog items not readily available outside big cities. [24] Exactly why all this has rumbled to a halt is in some dispute. [25] What is clear is that the Post Office and the Railway Ministry both suffer from what ails every Russian venture, private and public alike: Nobody pays his bills. [26] The state Railway Ministry complains that it is continually stiffed by customers who believe the railroads are honor-bound to carry freight whether they are paid or not. [27] The government has specified nearly 40 categories of freight which the railroads must carry for next to nothing. [28] Among the biggest deadbeats are the “power ministries” – the military and interior departments – which did not pay during Soviet times and feel little need to pay now. [29] “It’s a psychology formed during the socialist period,” said Tatiana Pashkova, the deputy spokeswoman for the ministry, “and the same situation exists with the Post Office. [30] We’ve been in dispute for a very long time. [31] Always we try to understand their differences. [32]

But we can't carry cargo for free". [33] No kidding, says the Post Office: It, too, is owed 200 million rubles by government agencies, and is barred from raising rates even though freight costs are outstripping revenues. [34] Moreover, the Post Office is also required by the government to carry some forms of mail, such as pension checks, at reduced rates. [35] Officials at the state Communications Department say they also suspect an ulterior motive in the railroads' actions: a struggle to dominate the thriving mail-order business. [36] "Here's competition between us and the railroads for the delivery of parcels, and I think decisions made by the railroads are mainly explained by this competition," Sherekhov said. [37] "They blame us for carrying some commercial cargoes instead of mail". [38] The Railways Ministry's spokeswoman, Ms. Pashkova, said such "commercial cargoes" are indeed a problem, but only because the railroads know that the Post Office has already gotten cash to transport the packages. [39] The railroads are still waiting for their share of that money, she said. [40] In the meantime, the situation has come to a boil. [41] In late September the railroads cut mail service in and out of Moscow, effectively decapitating the postal system and forcing officials to draft a fleet of trucks to move letters in and out of the city. [42] Hundreds of empty and full rail cars have clogged some local yards to the point where moving cars into position for unloading has become difficult. [43] At last count, 39 loaded cars were awaiting service at one yard. [44] And customers are getting angry. [45] "We've gotten all kinds of complaints," said Viktor Salikov, a deputy in the Communications Department's mail shipping center. [46] "People are even coming to us, searching for mail that was sent weeks ago".

2. NYT19981002.0300

[1] The United States is disappointed by the economic confusion within the new Russian government of Prime Minister Yevgeny Primakov, said Secretary of State Madeleine Albright on Friday, and she warned Russia about the dangers of an anti-Western policy. [2] In her first comprehensive review of U.S.-Russian relations since Primakov was confirmed as prime minister last month, Albright said Washington was "deeply concerned" about Russia's direction and did not think the crisis there would soon abate. [3] "We have heard a lot of talk in recent days about printing new money, indexing wages, imposing price and capital controls and restoring state management of parts of the economy," she told the U.S.-Russia Business Council in Chicago. [4] "We can only wonder if some members of Primakov's team understand the basic arithmetic of the global economy". [5] While praising Primakov, once her counterpart when he was foreign minister, as a pragmatist able to cooperate on key issues with Washington, she had harsh words of warning for him. [6] "Our initial reaction to some of the direction he's going in has not been particularly positive," she said, adding, "The question now is whether that cooperation can

continue”. [7] The United States must keep up its aid to Russia but is adjusting it to promote the building of democracy and student exchanges as well as arms control, Albright said. [8] Washington does not favor more direct aid. [9] “More big bailouts are not by themselves going to restore investor confidence in Russia,” she said. [10] “In the long run, the gap between Russia’s needs and its resources must be met not by foreign bailouts but by foreign investment”. [11] Albright sharply criticized as self-defeating the “many voices in Russia who want to shift the emphasis in Russia’s interaction with America and our allies from one of partnership to one of assertiveness, opposition and defiance for its own stake”. [12] Russia could not stand alone in the world, she said, and the United States’ ability “to help Russians help themselves will go from being merely very, very difficult to being absolutely impossible”. [13] At the same time, she said, Washington would not “assume the worst, for there are still plenty of people in Russia who will fight against turning back the clock”. [14] And she urged the world – and American critics – to “be patient with the workings of the democratic process in Russia” and “not start each day by taking a census of reformers in the Kremlin,” a census that American officials themselves promoted before President Boris Yeltsin dismissed the previous government. [15] While Moscow may continue to oppose any NATO use of force against Serbian forces in the southern province of Kosovo, Albright said, NATO must be prepared to act regardless.

3. APW19981002.0809

[1] Ukraine’s parliament on Friday refused to approve President Leonid Kuchma’s decree establishing a state fund to compensate people for savings lost in banks. [2] Deputies voted 240-47 to prepare a revised version of the decree and debate it later in the month. [3] They must consider the decree by Oct. 10, or else it automatically takes effect under the constitution. [4] Leftist factions, which voted against the proposed legislation, said it would not fully guarantee the return of savings lost during the financial instability that has recently hit Ukraine. [5] The fund Kuchma proposed to establish would have accumulated money in a special National Bank account and repay people in case the bank they kept their savings in went bankrupt or became insolvent. [6] Although the decree provided for compensation of deposits amounting only to 500 hryvna (dlrs 147 at the current exchange rate), its authors said it would cover more than 90 percent of Ukrainians who keep their money in banks. [7] The measure was meant to prevent mass withdrawals of deposits that most Ukrainian banks have already experienced as people, scared by the fall of the national currency and the turmoil in neighboring Russia, started to stock up on food, clothing and household goods. [8] Government officials say that Ukrainians have recently withdrawn at least 10 percent of the 3 billion hryvna (dlrs 882 million) they deposited in banks.

4. APW19981003.0292

[1] Prime Minister Yevgeny Primakov said Saturday that the economic crisis would not bring an end to the government's program of privatizing state property. [2] "Privatization will be accomplished for growth of production, growth of investment and growth of production effectiveness through renewal of major funds," Primakov said during a meeting of Western businessmen. [3] "We shall conduct privatization so that it serves the interests of the people, the state, and business," he said, according to the ITAR-Tass news agency. [4] Primakov reassured the businessmen, members of the prime minister's consultative council on foreign investment in Russia, that the government had no plans to ban the circulation of U.S. dollars in Russia. [5] But, he said, the government would take steps to staunch the flow of dollars from Russia. [6] Responding to media reports of a government economic plan that would prohibit Russians from buying U.S. dollars and other foreign currency, Primakov said the reports "absolutely do not correspond to reality," ITAR-Tass quoted him as saying. [7] He said there was no need to regulate the influx of dollars into Russia, but the government should take steps to prevent what he called "the dollar drain". [8] Russian companies stashed about dlr\$ 2.5 billion outside the country in September alone, the Interfax news agency reported Friday, quoting central bank figures. [9] Primakov said foreign investors are "the force" that will help Russia to minimize its losses from the current economic crisis. [10] "We are very much interested in foreign investments, especially in ones that go into the real production sector," Primakov said, according to Interfax. [11] "We need a continuous dialogue with foreign investors, without whom it will be difficult for us to overcome the current difficulties". [12] He said Russia wanted long-term investments. [13] "Foreign capital has been coming to the country via short-term operations," Primakov said. [14] This, he said, "is not to our liking or yours". [15] Viktor Gerashchenko, the central bank chairman, said on Friday that the government must act to stem the flow of dollars from Russia. [16] Gerashchenko and Finance Minister Mikhail Zadornov were in Washington Saturday for the annual meeting of the International Monetary Fund, where they plan to spell out the measures Russia is taking to bail out its finance system. [17] Russia wants the IMF to release the second dlr\$ 4.3 billion installment of a loan that was approved in July, a month before the country's economy crashed and the government effectively defaulted on its foreign loans. [18] IMF officials have said they want the Russian government to come up with a sound economic program before the installment is given, and have made it clear that currency controls and boosting the money supply by printing rubles are not acceptable.

5. NYT19981001.0363

[1] Russia's new prime minister picked an unusual way to reassure the nation Thursday. [2] After two weeks of deliberations he announced that he still had no plan to rescue the country from its economic crisis. [3] "I want to repeat once more - there is no program," Prime Minister Yevgeny Primakov said. [4] "It has yet to be worked out". [5] In most nations such a statement might provoke alarm. [6] But Primakov was seeking to calm an anxious public that was worried that the Kremlin's cure could be worse than the disease. [7] It is not as if Russians do not have something to be concerned about. [8] Thursday Primakov convened a meeting of top aides to try to hammer out a strategy for overcoming the economic woes. [9] A main item on the agenda was the plan drafted by Yuri Maslyukov, a Communist and the government's senior policy maker on economic issues. [10] It did not take long for Maslyukov's plan to hit the street. [11] A newspaper, Kommersant Daily, published it in full Thursday morning. [12] Not surprisingly Maslyukov's plan calls for a greater state role in the economy, including controls on hard currency. [13] A new State Bank for Reconstruction and Development would be established using the "nationalized assets" of failed commercial banks. [14] Wages and pensions would be paid in two months, and the minimum wage would be indexed to inflation. [15] There would be huge tax cuts, a combination that suggested that the government was committed to printing additional money. [16] The exchange rate of the ruble would be set by the central bank, based on changes in inflation and the balance of payments. [17] Maslyukov's plan also implies that Russians would be able to buy dollars at exchanges through the country. [18] It stipulates that hard currency could enter the country only with special authorization. [19] Exporters would be required to sell most of their hard currency reserves. [20] That is an allusion to currency controls that sent shock waves through the Russian public. [21] Russians have come to treat the dollar as a second currency, and many people have squirreled away dollars as a hedge against inflation. [22] "It is an obvious stupidity," Otto Latsis, a commentator with the newspaper Noviy Izvestiya, said. [23] "People won't give their dollars away. [24] They will go to the black market if they need to". [25] Yegor Gaidar, the former prime minister who favors free markets, said the plan was a "war against the dollar" and predicted that it would lead to a shortage of imports. [26] As the criticism grew, Primakov rushed to distance himself from the talk of currency controls. [27] "The rumors of the state's becoming a monopolist on the inflow of hard currency into the country is nonsense," he said, asserting that the document published by Kommersant Daily was just one of six possible plans. [28] Primakov may be opposed to wildly unpopular currency controls. [29] He may be worried that the plan would set off the panic buying of dollars, further depressing the ruble. [30] Or he may simply be trying to keep his distance from Maslyukov's plan while Russia tries to wrangle its next disbursement,

\$4.3 billion, from the International Monetary Fund. [31] The particulars could be modified, but many people following the maneuvering say they believe that it represents the basic thrust of the government's thinking. [32] Primakov has called for greater state regulation and an expansion of the money supply, two themes of Maslyukov's plan. [33] The plan also has many similarities with the plan presented to the government by Gorbachev-era advisers. [34] "We think it is close to being a final document," said the United Financial Group, a Russian investment business. [35] It is unclear how long Primakov can carry on without spelling out a detailed strategy. [36] The ousted tax chief, Boris Fyodorov, has argued that the IMF should not provide further aid until Primakov has taken tough measures to build a free market. [37] Primakov has, however, sought to turn that logic on its head, arguing that his government's economic strategy will depend on the fund's willingness to provide aid. [38] His aim appears to be to pressure the fund by implying that it will be the fund's fault if Russia is forced to default on its loans or take draconian measures at home. [39] Or as Primakov put it Thursday, without the fund's money, Russia will have to impose "unpopular measures".

6. APW19981002.0783

[1] President Leonid Kuchma called Friday for "corrections" to Ukraine's program of market reforms, but pledged that reforms would continue. [2] Kuchma did not elaborate in his comments to a group of Ukrainian economists, saying only that the changes were necessary because of the country's economic problems. [3] Kuchma urged the economists to come up with recommendations before a national meeting of economists in November, the Interfax news agency reported. [4] Ukraine has suffered economic problems since the 1991 collapse of the Soviet Union and it has been especially hard hit by the financial crisis in neighboring Russia, its main trading partner. [5] The Russian crisis has hurt bilateral trade, caused the Ukrainian currency to fall and led to a withdrawal of investors from Ukraine. [6] In recent months, Kuchma has implemented some economic reforms by decree, prompting the International Monetary Fund to release the first installment of a long-awaited \$2.2 billion loan. [7] Many reforms, however, remain stalled. [8] National Bank chairman Viktor Yushchenko was in Washington this week for consultations with IMF officials. [9] Yushchenko has warned that the bank would not spend its dwindling reserves to support the hryvna currency, but has avoided comments on the currency in recent days. [10] Kuchma met Friday with Yushchenko, Prime Minister Valery Pustovoitenko and other senior officials to discuss ways to stabilize the hryvna. [11] The government's press service said they focused on possible ways to keep the hryvna under 3.5 to the U.S. dollar through the end of the year. [12] The hryvna has been trading at 3.4 to the dollar in recent days but trading has been limited. [13] There are wide expectations in Ukraine that the hryvna will fall even

further, and Ukrainians have been stocking up on food, clothing and household goods to save their fast-devaluing money.

7. NYT19981001.0379

[1] When the world's finance ministers and central bankers gathered last year in Hong Kong, they nervously congratulated each other for containing – at least for the moment – a nasty financial brush fire in Asia. [2] In a year's time, many predicted in hallway chatter, the troubles in Thailand and Indonesia would look like a replay of Mexico in 1995 – a rough bump in the road for a world enjoying remarkable prosperity. [3] Talk about bad market calls. [4] Twelve months later, as the same financial mandarins clog Washington with their limousines and glide through endless receptions at the annual meeting of the International Monetary Fund and the World Bank, just about everything that could have gone wrong in the world economy has: the worst downturn in Japan since World War II, economic meltdown in Russia, a depression in Indonesia that is plunging 100 million people below the poverty line, and deep fears over what happens next in Latin America. [5] What makes this year's IMF meeting most remarkable, though, is that the harshest criticisms are directed at the monetary fund itself, and, by extension, at the U.S. Treasury, which is viewed as the power behind the IMF. [6] This year, in place of confident predictions, there are mutual recriminations. [7] Arguments are breaking out over whether the true culprits were crony capitalists and weakened leaders like Russian President Boris Yeltsin, or huge investors who poured money into the world's emerging markets with reckless abandon in the mid-1990s and panicked in the past twelve months. [8] Whatever the reason, one reality prevails: Hundreds of billions of dollars have fled from economies on four continents – seeking the safest havens possible, often in the United States – and the money is not returning anytime soon. [9] And the subtext of every seminar on capital flows and every conclave of nervous ministers will be some painfully blunt questions: Can this be stopped? [10] Or is the world headed for a global recession? [11] Fifty-three years ago the IMF was created after the Bretton Woods conference which sought to stabilize the world economy and secure the peace after World War II. [12] Now it is under attack from all sides, charged not only with worsening a bad situation by misjudging the economics, but with being politically tone-deaf in some of the most volatile capitals in the world, from Jakarta, Indonesia, to Moscow. [13] For the first time, there are disturbing questions about whether the institution itself is still capable, financially or politically, of containing the kind of economic contagion that caught the world unaware. [14] Once, the IMF's critics were largely found in Africa and South Asia, where the fund was often viewed as arrogant; today they include Wall Street's biggest players and top officials in the most powerful economies of Asia and Europe. [15] Only a few – including former Secretary of State George Schultz and members

of Congress who are increasingly suspicious of all international institutions – are talking about scrapping the IMF altogether. [16] But almost everyone is talking about creating a “new financial architecture” that can do what the old one clearly cannot: smother financial wildfires before they leap around the globe. [17] President Clinton, British Prime Minister Tony Blair and other leaders, after months of silence, have edged into the debate, in some cases wresting the issue for the first time from their finance ministers and central banks. [18] Their fear, their advisers say, is that 15 months of financial turmoil are now threatening political stability. [19] Such concerns have turned this year’s meeting into a tumbled mass of worries and a groping for short and long-term solutions. [20] The Japanese, the French, the Southeast Asians are all arriving in Washington with different diagnoses of what went wrong, and different solutions about how to set it right. [21] The United States has its own set of plans, a mix of suggestions to force more disclosure of financial data in countries around the world and to impose more American-style financial standards and regulation. [22] Meanwhile, an ideological argument is breaking out over whether the world should slow down a long march toward more free and open markets – a strategy pressed by the Clinton administration for the past six years. [23] Others argue that it is unwise to start rebuilding the hospital while the patients are still on the operating tables. [24] “Last year the standard answer that all of us were given came down to this: ‘We have the IMF and the World Bank, and they know best,’” Indonesian Foreign Minister Ali Alatas said over breakfast in Washington the other day, reflecting on how the crisis turned 30 years of astounding growth in his country into an overnight depression. [25] “Then they said everything that went wrong was our fault,” he said. [26] “But now, now I think people know that much of the problem came from the outside, and we need something better”. [27] And the IMF itself is beginning to fight back, an awkward role for an institution dominated by Ph.D. economists who are unaccustomed to being openly challenged. [28] “Every place you turn you read the same story, that we came in, that we made things worse,” said Stanley Fischer, the deputy managing director of the fund, who was born in Northern Rhodesia – now Zambia – and became chairman of the Massachusetts Institute of Technology’s economics department before taking a job that has now put him in the center of the financial storm. [29] “We frequently get the blame, some of it well-deserved,” he said. [30] “But it is politically convenient for governments around the world to cry, ‘The IMF made us do it,’ and pin their mistakes on us. [31] That’s fine. [32] We’d rather be loved, but more than that we’d like to be effective”. [33] MISCALCULATIONS, POLITICS AND SAFETY NETS On a steaming January day, Michel Camdessus, the IMF’s top official, slipped into Jakarta to the private residence of President Suharto and sat down for a four-hour meeting to tick off, line by line, the huge reforms Indonesia would have to implement in return for tens of billions of dollars in emergency aid. [34] Two previous deals had collapsed when Suharto ignored the fund’s conditions, so Camdessus insisted that he strike a deal directly with Suharto,

then Asia's longest-serving leader. [35] It was a meeting of men who knew different worlds of power politics: Suharto rose as a general in central Java, and Camdessus had detonated mines in Algeria for the French army before entering the French Treasury on his way to becoming head of France's central bank. [36] "It was all there," a senior IMF official recalled. [37] "He was told he had to dismantle the national airplane project, the clove monopoly, all the distribution monopolies". [38] At one point, Camdessus looked at the impassive Suharto and said, "You see what this means for your family," a reference to their vast investments in the country's key industries. [39] "He said, 'I called in my children, and they all understand'". [40] But within months, that exchange in Jakarta came to symbolize the IMF's twin troubles: Its inability to understand and reckon with the national politics of countries in need of radical reform, and its focus on economic stabilization rather than the social costs of its actions. [41] Suharto had no intention of fulfilling the agreement. [42] It was, one of his former Cabinet members said, "a delaying move that was obvious to everyone except Camdessus". [43] Perhaps one reason why the IMF sometimes appears tone-deaf is that its senior staff is almost entirely composed of Ph.D. economists. [44] There are few officials with deep experience in international politics, much less the complexities of Javanese culture that were at work in Indonesia. [45] Historically, experts in politics and security have gravitated to the United Nations, development experts to the World Bank, and economists to the IMF - creating dangerous gaps in a crisis like this one. [46] As a result, the fund had only a rudimentary understanding of what would happen if its demands were met and all Indonesia's state monopolies were quickly dissolved. [47] While that system lined the pockets of the Suhartos and their friends, it also distributed food, gasoline and other staples to a country that stretches for 3,000 miles over thousands of islands. [48] To help balance the budget, the fund demanded a quick end to expensive subsidies that keep the price of food and gasoline artificially low. [49] But that, combined with the huge currency devaluation that sparked the crisis, resulted in high prices and shortages that fueled riots that continue to this day, as millions of Indonesians lose their jobs. [50] The IMF - unintentionally, its officials insist - also sped Suharto's resignation, insisting on the elimination of "crony capitalism," code words for removing the Suharto family from the center of the economy. [51] Ultimately, that may prove to be Indonesia's salvation, if the new government can contain the rioting against the ethnic Chinese minority - whose money is desperately needed to save the country's fast-shrinking economy. [52] "It is worth noting," Fischer said this week, "that our programs in Asia - in Indonesia, Korea and Thailand - only took hold after there was a change in government". [53] Nonetheless, the Indonesia experience has revived the argument that the IMF is so focused on stabilizing banks and currencies, on preventing capital flight and freeing up markets, that it is blind to the social costs of its actions. [54] Among the toughest critics has been its sister institution, the World Bank, whose main charge is alleviating poverty. [55] "You've seen the tension almost every day,"

one senior World Bank official said recently. [56] The bank has gone to extraordinary lengths in recent months to differentiate its role from that of the fund, and to announce a tripling of aid to the poorest in the countries hit by the economic chaos. [57] Even U.N. [58] Secretary General Kofi Annan has joined the argument, warning in a speech at Harvard recently that “if globalization is to succeed, it must succeed for poor and rich alike. [59] It must deliver rights no less than riches. [60] It must be harnessed to the cause not of capital alone, but of development and prosperity for the poorest of the world”. [61] IMF officials say they are changing strategies when they see they are exacting too great a social cost. [62] “It’s a very difficult formula to get exactly right,” Fischer said in August, as Russia was teetering and the IMF was sending in \$4.8 billion in aid that was rapidly wasted. [63] “You need enough discipline to send the right message to the markets and keep investors from fleeing. [64] But you need enough leeway to keep people from suffering more than they otherwise would”. [65] In recent months, he noted, the IMF has allowed more spending to sustain subsidies for basic goods for longer periods in Indonesia, Korea and elsewhere. [66] “There is a new flexibility at the IMF” a senior Indonesian official concluded recently. [67] “It is a lot better”. [68] A U.S. PAWN, OR A RUNAWAY AGENCY? [69] The Clinton administration admits that the IMF has many failings, many of them on display this year. [70] But it insists that the world has gone through global financial crises without an IMF once before in this century – and the result was the 1930s. [71] “I have no doubt the situation over the past year would have been much worse – with greater devaluations, more defaults, more contagion, and greater trade dislocations – without the program agreed with the IMF and the finance it has provided,” Deputy Treasury Secretary Lawrence Summers told Congress last week. [72] Many Republicans and some Democrats are unconvinced. [73] Even though the Senate has overwhelmingly approved an \$18 billion contribution to the fund to help it fight new crises, the House defeated that measure two weeks ago. [74] The fund’s last hope of getting the money, which will free up nearly \$100 billion in contributions from other nations waiting for the United States to act, will come when the House and Senate try to resolve their budget differences in a conference committee in the next 10 days. [75] A rejection, Treasury Secretary Robert Rubin insists, would send a message around the world that the United States is turning its back on the one institution charged with restoring economic stability. [76] Everywhere else in the world, though, politicians and businessmen insist that one of the biggest problems with the IMF is that, contrary to the view of Congress, it acts as the U.S. Treasury’s lap dog. [77] Ask in Jakarta or Moscow, and the response is the same: The IMF never ventures far without looking back for the approving nod of its master. [78] When the United States weighs in, however, is when the IMF is called on to rescue a country in deep trouble. [79] Only then does the IMF – and the U.S. Treasury – have the leverage to extract commitments in return for billions in aid. [80] In theory, the U.S. influence is limited: It has an 18.5 percent vote in the fund. [81] Germany, Japan,

France and Britain have about 5 percent each. [82] But in practice the United States usually gets its way, exercising its influence behind the scenes, often in interactions between Fischer and Summers. [83] The two met when Fischer was on the MIT faculty and Summers was a graduate student taking one of his classes, later becoming a colleague at MIT. [84] Each served as chief economist of the World Bank. [85] It was Summers who was instrumental in placing Fischer in the fund's no. 2 job, and these days they talk constantly. [86] "It's usually a warm relationship," Fischer said this summer. [87] "Remember, this is a job where you cannot turn to outsiders for advice – you can't call the chief economist at a Wall Street firm, or even many of your academic friends, because so many of the issues are confidential". [88] The Treasury's relations with Camdessus are often more strained as he plays the role of world diplomat, traveling the globe and trying to coax along political leaders. [89] The tensions were obvious from the start of the Asia crisis. [90] The fund made little secret of its displeasure that the United States was not offering direct aid to Thailand, a major U.S. ally, as a sign of support and confidence. [91] Mindful of the backlash in Congress when Mexico was bailed out with U.S. money, that was the last thing the Treasury planned to do. [92] Summers, in turn, thought the fund was not forcing the Thais to implement its reform commitments rigorously enough or disclose their true financial picture. [93] Within the U.S. government there was other dissension: The State and Defense Departments felt the United States should do more for Thailand, but backed off when the Treasury asked if they would like to pony up some aid out of their own budgets. [94] There were other conflicts. [95] When Japan used the last IMF meeting to propose setting up a \$100 billion "Asia Fund" – one that would exclude the United States and would probably offer aid under much more relaxed conditions than the IMF does. [96] Rubin called up Camdessus at breakfast one morning and told him that the Japanese proposal would undercut the IMF's authority. [97] "We've just had a dispute with Michel," Rubin reported to his aides as he returned to his orange juice and croissant. [98] One of them shot back: "And it's only 8 a.m". [99] Camdessus backed down at Rubin's insistence and walked away from money that Asia could have used. [100] Japan says it will be back with a similar proposal this weekend, this time for a \$30 billion fund. [101] Camdessus has also rankled U.S. officials with statements that amounted to cheerleading to reassure the markets – sometimes in the face of the facts. [102] In June, with Russia on its way to collapse, Camdessus declared that "contrary to what markets and commentators are imagining" about the slow collapse of Russia's economy, "this is not a crisis. [103] This is not a major development". [104] The bailouts of Russia and South Korea were prime examples of how Washington muscles into the IMF's turf as soon as major U.S. strategic interests are involved. [105] Last Christmas, as South Korea slipped within days of running out of hard currency to pay its debts in December, it sent a secret envoy, Kim Kihwan, to work out a rescue package. [106] "I didn't bother going to the IMF," Kim recalled recently. [107] "I called Summers' office at the

Treasury from my home in Seoul, flew to Washington and went directly there. [108] I knew that was how this would get done". [109] Within days the Treasury dispatched David Lipton, its most experienced veteran of emergency bailouts, who is leaving his post as undersecretary for international affairs this month, to shadow the IMF staff's negotiations with the government in Seoul. [110] Fischer was displeased. [111] "To make a negotiation effective, it has to be clear who has the authority to do the negotiating," he said. [112] WHO LOST RUSSIA? [113] The pattern was repeated this summer, when the United States raced to put together a \$17 billion package for Russia. [114] The IMF's staff in Moscow declared that Russia needed no money at all - it just needed to enact policies that would restore confidence in investors. [115] The Americans and Germans came to a different conclusion. [116] Soon after, U.S. officials gathered in the White House situation room to consider what might happen to Russia if the ruble was devalued and market reforms collapsed and to push the IMF to come up with emergency money. [117] So the fund began assembling a last-ditch program to prop up a country that had resisted its reform plans for seven years. [118] Camdessus, though, was still hesitant, questioning whether the IMF should risk its scarce resources in Russia. [119] "We had to pull Michel along," a senior Treasury official recalled. [120] As it turned out, Camdessus' instincts were right while the approach championed by Rubin and Summers proved disastrously wrong. [121] The first installment of that payment - \$4.8 billion - was wasted, propping up the currency long enough, in the words of one IMF official, "to let the oligarchs get their money out of the country". [122] Then Yeltsin reversed his commitments, let the ruble devalue anyway, began printing money with abandon and fired virtually every reformer in his government - resulting in a collapse of the IMF agreements and the indefinite suspension of its aid program. [123] Now, inside the IMF and on Capitol Hill, there are recriminations over "who lost Russia". [124] Publicly, Fischer argues that "there are no apologies owed for what we attempted in Russia". [125] But some IMF officials complain privately that they let Rubin and Summers run roughshod over them, striking a deal that fell apart within weeks as the Russian parliament rebelled and Yeltsin backed away from his commitments. [126] Summers responds that the United States "took a calculated risk" because "it was vastly better that Russia succeed than not succeed". [127] The Russian collapse touched off new rounds of economic contagion, with investors fleeing Latin America, and triggering huge losses in hedge funds like Long Term Capital, the Greenwich, Conn., investment firm that needed to be rescued by Wall Street powerhouses whose money it had invested. [128] "Russia was a turning point," said Robert Hormats, the vice chairman of Goldman, Sachs & Co. [129] "It made the world realize that some countries can fail, even if the IMF and the Treasury intercede. [130] And that changed the perception of risk". [131] Now, as the countries meet to face a future that the IMF has warned could be very bleak, they need to reverse those perceptions, or watch countries slowly starve for lack of capital. [132] The emerging markets are

calling for controls on short term investments. [133] The French want a stronger IMF. [134] The Americans say the answer is more disclosure, so that investors are better warned, and tougher regulation. [135] “These are usually nice, quiet meetings; everyone very polite,” a top U.S. official said earlier this week. [136] “Not this year”.

8. NYT19981004.0132

[1] If the Communist Party has its way – and it has been planning for months – millions of Russians will take to the streets on Wednesday for some of the biggest demonstrations in years. [2] But now that Communists or politicians with the Communist stamp of approval are running the country and its economy the question is what the marchers will be demonstrating against. [3] President Boris Yeltsin and his economic advisers were easy targets two months ago. [4] But Yeltsin seems but a shadow of himself today, and his advisers are gone. [5] The Communists, who have undergone a sort of resurgence by playing on the discontent, are working hard to cast themselves as outsiders in the government that they help run – and to keep the focus of the protests on Wednesday on Yeltsin and his policies. [6] Even that strategy may backfire, however, because discontent over Yeltsin does not necessarily translate into support for his predecessors. [7] An adviser to the Communist Party leader, Gennadi Zyuganov, made headlines not long ago when he called the new government bourgeois. [8] The characterization was striking, because the new prime minister, Yevgeny Primakov, was one of Zyuganov’s top choices for the post. [9] The deputy prime minister in charge of the economy, Yuri Maslyukov, is a Communist Party figure who once headed Gosplan, the infamous central-planning program that helped bring the Soviet Union to ruin. [10] “We don’t have slogans that are aimed against the government,” the first secretary of the Moscow Communist Party committee, Alexander Kuvayev, said in a recent interview. [11] “All the slogans are aimed against the president and the economic course of the country. [12] Only the president can fully be blamed for the course that has brought the country to this situation”. [13] Anger over Russia’s fate, and their own, is drawing some Russians back to Soviet-style slogans and old-style hostility toward capitalism. [14] The ardently pro-Communist newspaper *Sovietskaya Rossiya* devoted most of its front page on Thursday to what it said were the results of a contest among readers for protest slogans. [15] The entries ranged from catchy to kitschy, from, “Legislator, official, banker – study the Constitution; the exam is Oct. 7,” to, “Imperialist! [16] Help Russia return the exported capital, and we will pay the debts at once”. [17] Russia has seen nothing approaching this sort of economic chaos and despair since 1993, when hyperinflation swept the economy. [18] There is no real way of knowing whether the latest travails will produce huge and ugly crowds or small, peaceful ones. [19] A deputy editor and political analyst at a newsweekly, *Itogi*, Masha Lipman, said

the situation was not unlike that of 1993, when Yeltsin sought, and won, a popular mandate in a referendum. [20] He used it to dissolve the Communist-dominated Parliament and increase his own power. [21] That led to an autumn showdown, the shelling of Parliament by tanks and the total defeat, for the time being, of Yeltsin's legislative opponents. [22] "You can interpret this demonstration as a Communist mandate, but I don't think they will act like Yeltsin," Ms. Lipman said. [23] "Support for the Communists is at its smallest level, maybe 20 percent". [24] In a survey of 1,714 people that was released last week the All-Russian Public Opinion Center said that nearly half the population supported the idea of demonstrations against Yeltsin, but that barely one-tenth were likely to participate in any way. [25] "Nobody is interested in any sort of struggle," Yuri Levada, who heads the center, said in an interview. [26] "This is mainly a general expression of a great wave of distrust of authorities, mainly the president. [27] He's the great scapegoat for all our sins". [28] There seems to be little enthusiasm for demonstrations of any sort. [29] Only a few thousand people turned out Sunday for the fifth anniversary of the shelling of Parliament and Yeltsin's subsequent triumph over the Communists. [30] The Communist Party and the Federation of Independent Russian Unions, a leftist organization that says it is the main sponsor of the protests, predict that 9 million Russians will participate on Wednesday. [31] The government has vowed to keep order without resorting to force. [32] But it is concerned enough that it has summoned 11,000 police officers to patrol Moscow and 6,000 military troops to intervene if violence erupts. [33] Protests since 1993 have generally been tepid, a deep-seated feeling here. [34] Some demonstrations last week, more or less practice runs, hinted at similarly dampened marches this year.

9. APW19981001.1177

[1] Wall Street extended a global stock selloff Thursday with the Dow industrials tumbling more than 200 points for a second straight day. [2] The Dow Jones industrial average, which plunged 237.90 points on Wednesday, fell an additional 210.09, or by 2.7 percent, to close at 7,632.53. [3] Broader market indicators also sank sharply in heavy trading. [4] Bank and technology stocks were particularly hard hit. [5] The selling spree, amid worries about shrinking corporate profits and fears of new financial crises, left the Dow 18.3 percent, or more than 1,700 points, below the all-time high of 9,337.97 reached on July 17. [6] It was getting closer to the low of 7,400 that was reached during trading Sept. 1, before Wall Street's best-known indicator began a comeback bid that brought it above 8,100 as recently as Monday. [7] The Dow's 12.4 percent slide in the third quarter, which ended Wednesday, was its worst quarterly performance in eight years. [8] It now is 3.5 percent below where it began this year. [9] Stock prices earlier plunged in Asia, with Tokyo shares falling 1.6 percent to a new 12-year low, and shares were falling sharply in

Europe, where Germany's central bank left interest rates unchanged. [10] Blue chips in London sank 3.1 percent to close at new lows for the year, while the key index in Frankfurt, Germany, closed down 5.5 percent and the main indicator in Paris was off 5 percent. [11] The selloff in stocks has sent a flood of money into U.S. Treasury securities, a traditional haven in times of uncertainty. [12] Yields on 30-year Treasury bonds fell further below 5 percent Thursday, reaching levels unseen for long-term government bonds since 1967. [13] Traders were alarmed to see prices on the New York Stock Exchange nosedive 2.9 percent on Wednesday, even though the Federal Reserve had lowered a key interest rate one-quarter percentage point on Tuesday. [14] Some traders were disappointed that the cut was not deeper amid fears a go-slow approach would not do enough to counter the economic crises that have swept through Asia and Russia and are threatening Latin America. [15] "The smaller-than-expected lowering of interest rates by the U.S. Federal Reserve has a chain reaction," said Lee Won-ho, an analyst at Samsung Securities Co. in Seoul, South Korea, where the main stock index fell by 1.5 percent. [16] "It is affecting Wall Street, the Japanese market, ours and others". [17] The managing director of the International Monetary Fund, Michel Camdessus, on Thursday said the Fed made the right decision in cutting rates and that global powers now must push for stronger growth to offset steep recessions in Asia and Russia. [18] Asked why stock markets, particularly in the United States, have reacted so negatively to the Fed rate cut, Camdessus said he believed confidence will soon be restored, especially if financial leaders show resolve in their discussions over the next week. [19] There also are worries about where the next financial market crisis may erupt after last week's \$3.6 billion private bailout of Long-Term Capital Management LP of Greenwich, Connecticut. [20] In addition, investors worldwide worry that the downturn on Wall Street could signal a possible slowdown in economic growth - a bad omen for the many foreign companies dependent on exports to the huge U.S. market. [21] "There's a psychological impact overall, but there's also a direct impact on companies like Sony and TDK which derive a high percentage of their earnings from overseas markets," said Pelham Smithers, a stock strategist in Tokyo at ING Baring Securities (Japan) Ltd. A new survey in Japan said confidence among small- and medium-sized businesses about the economy plunged to its worst level since the Bank of Japan began the quarterly samplings in 1967.

10. APW19981002.0778

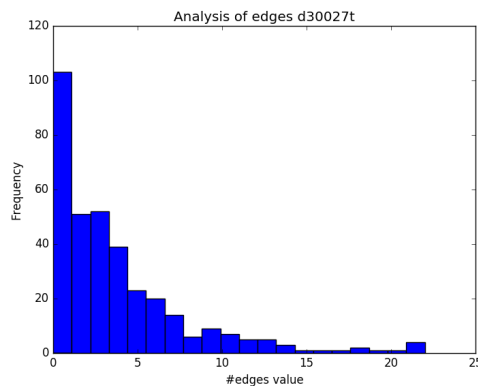
[1] President Boris Yeltsin would respond strongly to any effort to prohibit Russians from buying foreign currencies, believing the move would be like bringing another Iron Curtain down on the country, his spokesman said Friday. [2] "The president clearly understands that such a ban would be a clear violation of our rights... [3] that would mean a return to the Iron Curtain

in everyday life,” said presidential spokesman Dmitry Yakushin. [4] The remarks came after media reports of a government economic plan in the works that would prohibit Russians from buying U.S. dollars and other foreign currencies, and institute other strict economic controls, rolling back seven years of reforms. [5] Though government officials say such a plan is only one of six possibilities, Prime Minister Yevgeny Primakov warned Thursday that he might be forced to take “unpopular” measures to rescue the Russian economy if it does not receive the next installment in a dlrs 22.6 billion loan from the International Monetary Fund. [6] Central Bank Chairman Viktor Gerashchenko and Finance Minister Mikhail Zadornov flew to Washington on Friday for negotiations with the IMF over the dlrs 4.3 billion installment. [7] IMF officials have said they want the Russian government to come up with a sound economic program before the loan is given, and have made it clear that currency controls and boosting the money supply by printing rubles are not acceptable. [8] But Primakov said that the country’s plan would be dependent on the IMF loan, not the other way around. [9] No short-term economic plan will be known for another three weeks, long after the IMF is set to make its decision, said First Deputy Prime Minister Vadim Gustov. [10] A draft version of the government’s fourth-quarter budget would rely heavily on the IMF loan, the Kommersant newspaper said. [11] Without the loan, the government would have to engage in major deficit spending. [12] Calls for Soviet-style controls – part of the plan drafted by First Deputy Prime Minister Yuri Maslyukov and leaked to the press – have apparently created a rift within the Cabinet. [13] The plan drew heavy criticism from Zadornov, and Gerashchenko called the idea of banning foreign currency sales “a mad idea”. [14] Meanwhile, Russia’s stock market dropped by 5 percent to 38.8 Friday, but trading was so light the downward move meant little. [15] The stock market has continued to fall since the crisis hit Aug. 17. [16] And tax collection was down in September, the result of the collapse in Russia’s banking system, said Alexander Pochinok, head of the government’s finance and credit department. [17] Revenues totaled 14.3 billion rubles (dlrs 875 million), 700 million rubles (dlrs 43 million) less than expected. [18] The government wants to reverse this trend by raising taxes for oil and gas companies, but the idea is expected to meet sharp resistance in the lower house of Russia’s parliament, which wants to cut tax rates, Pochinok said.

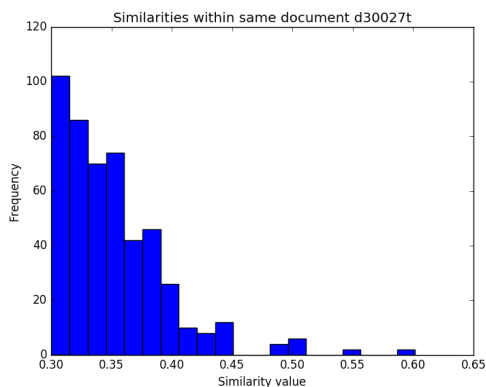
A.2 Merging process

Before passing to the merging process, information about the edges of this cluster is depicted in Figure A.1. In comparison with other clusters, the merging process with the one presented here was not very extensive. As one can see in Figure A.1b and Figure A.1c there are few edges—three edges connecting sentences from the same document, and two edges connecting

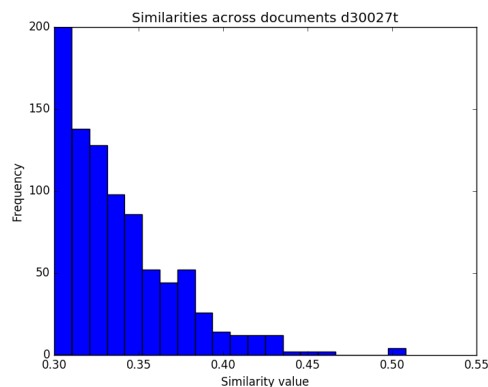
sentences from different documents—with a similarity above 0.5, which was the value of the chosen threshold to stop the merging process per document and across documents.



(a) Frequency of edges per node after pruning



(b) Similarity distribution of sentences within the same document after pruning



(c) Similarity distribution of sentences from different documents after pruning

FIGURE A.1: Histograms illustrating graph edges distribution of the *d30027t* cluster from DUC 2004, built with the doc2vec-based similarity measure

We present a comparison of how the merging process was performed when penalizing and compensating intra-document redundancy. For both settings, we analyze the case when inter-document redundancy is compensated ($cd = 1$). For a quick reference, a list of the involved sentences is provided:

d4s5. But, he said, the government would take steps to staunch the flow of dollars from Russia.

d4s7. He said there was no need to regulate the influx of dollars into Russia, but the government should take steps to prevent what he called “the dollar drain”.

d4s12. He said Russia wanted long-term investments.

d4s15. Viktor Gerashchenko, the central bank chairman, said on Friday that the government must act to stem the flow of dollars from Russia.

d4s16. Gerashchenko and Finance Minister Mikhail Zadornov were in Washington Saturday for the annual meeting of the International Monetary Fund, where they plan to spell out the measures Russia is taking to bail out its finance system.

d7s112. WHO LOST RUSSIA?

d7s123. Now, inside the IMF and on Capitol Hill, there are recriminations over “who lost Russia”.

d7s132. The emerging markets are calling for controls on short term investments.

d10s6. Central Bank Chairman Viktor Gerashchenko and Finance Minister Mikhail Zadornov flew to Washington on Friday for negotiations with the IMF over the dlrs 4.3 billion installment.

Table A.1 shows how the merging of nodes was performed when $pd = 1$ and $cd = 1$, and Table A.2 shows the case when $pd = -1$ and $cd = 1$. These tables present the step number as a reference, the pair of nodes to be merged, the relevance of each node before merging, and the similarity found between these nodes which is normalized between 0 and 1, 0 for the minimum value of the doc2vec-based similarity measure (-1), and 1 for the maximum value (1).

TABLE A.1: Merging process for the *d30027t* cluster when $pd = 1$ and $cd = 1$

Step	Pair of nodes		Node relevances		Similarity
Merging sentences per document.					
1	d4s15 ,	d4s5	2.0451 ,	1.6742	0.8008
2	d7s123,	d7s112	3.4622,	1.7753	0.7764
3	d4s15 ,	d4s7	3.7193 ,	1.8415	0.7511
Merging sentences across documents.					
4	d4s12,	d7s132	3.8749,	1.0344	0.7540
5	d4s16,	d10s6	1.6705,	1.4733	0.7528

TABLE A.2: Merging process for the *d30027t* cluster when $pd = -1$ and $cd = 1$

Step	Pair of nodes		Relevances		Similarity
Merging sentences per document.					
1	d4s15 ,	d4s5	2.0451 ,	1.6742	0.8008
2	d7s123,	d7s112	3.4622,	1.7753	0.7764
3	d4s15 ,	d4s7	0.3708 ,	1.8415	0.7511
Merging sentences across documents.					
4	d4s12,	d7s132	3.8749,	1.0344	0.7540
5	d4s16,	d10s6	1.6705,	1.4733	0.7528

As one can see, the same nodes are merged, both tables are quite similar. However, the difference is that the relevance of the remaining node after merging per document increases for Table A.1—when intra-document redundancy is being compensated—and decreases for Table A.2—when

inter-document redundancy is being penalized. One can see this effect at *Step 3*, the relevance of sentence *d4s15* is updated after being merged with another node at *Step 1*.

As a result, the best evaluated summary was when intra-document redundancy was penalized ($pd = -1$). Even when the cd parameter was changed to 0 or -1 , penalizing intra-document redundancy was the best option. These are the generated summaries.

Summary when $pd = 1$ and $cd = 1$:

Viktor Gerashchenko, the central bank chairman, said on Friday that the government must act to stem the flow of dollars from Russia [*d4s15*]. Now, inside the IMF and on Capitol Hill, there are recriminations over “who lost Russia” [*d7s123*]. He said Russia wanted long-term investments [*d4s12*]. Russia wants the IMF to release the second dlrs 4.3 billion installment of a loan that was approved in July, a month before the country’s economy crashed and the government effectively defaulted on its foreign loans [*d4s17*]. Primakov reassured the businessmen, members of the prime minister’s consultive council on foreign investment in Russia, that the government had no plans to ban the circulation of U.S. dollars in Russia [*d4s4*].

Summary when $pd = -1$ and $cd = 1$:

He said Russia wanted long-term investments [*d4s12*]. Russia wants the IMF to release the second dlrs 4.3 billion installment of a loan that was approved in July, a month before the country’s economy crashed and the government effectively defaulted on its foreign loans [*d4s17*]. Primakov reassured the businessmen, members of the prime minister’s consultive council on foreign investment in Russia, that the government had no plans to ban the circulation of U.S. dollars in Russia [*d4s4*]. Gerashchenko and Finance Minister Mikhail Zadornov were in Washington Saturday for the annual meeting of the International Monetary Fund, where they plan to spell out the measures Russia is taking to bail out its finance system [*d4s16*].

In terms of the sentence id, they will be:

Generated summary when $pd = 1$ and $cd = 1$:

d4s15 d7s123 d4s12 d4s17 d4s4.

Generated summary when $pd = -1$ and $cd = 1$:

d4s12 d4s17 d4s4 d4s16.

Sentences that are different in each summary are highlighted. Sentences in common are *d4s12*, *d4s17*, and *d4s4*. Sentence *d4s12* resulted important as a result of merging nodes across documents, see Step 4 in tables A.1 and A.2. The other two resulted important as a result of the TextRank method, since no merging was performed with these nodes. For the first summary,

sentences *d4s15* and *d7s123* resulted important as a result of merging nodes per document, see steps 1 to 3 in Table A.1. For the second summary, sentence *d4s16* resulted important as a result of merging nodes across documents, see Step 5 in Table A.2.

Penalizing intra-document redundancy in this cluster was better because the penalization, in the second summary, caused that sentences redundant per document were left out, letting in the sentence redundant across documents. Hence, intra-document redundancy is undesired because giving importance to this type of redundancy, reduces the possibility of taking inter-document redundancy into the final summary, which is more relevant for multi-document summarization.

A.3 Reference summaries

Finally, we include the four reference summaries provided by DUC 2004, for the *d30027t* cluster. The summaries obtained with our method were compared with these summaries, which by the way are abstractive summaries.

D30027.M.100.T.A

In October 1998 amid worldwide financial crises, particular concern focused on Russia where economic meltdown was exacerbated by conflicted politics. President Yeltsin's latest Prime Minister, Primakov, was supported by Communists and when word leaked out that a Communist economic program was under consideration, Yeltsin denounced it. Primakov then assured the public that "there is no program," suggesting that there would not be until the International Monetary Fund (IMF) came forth with a massive loan. IMF demanded a sound economic program before approving loan payment. Meanwhile neighboring Ukraine felt economic effects of the IMF-Primakov standoff.

D30027.M.100.T.C

As world finance and banking representatives met in Washington, the economic news continued to be bleak. IMF officials had predicted a banner year, but stocks continued to slide worldwide and the DOW probably would record its worst third quarter loss in eight years. Russia and Ukraine have been especially hard hit by the crisis. In Russia, Prime Minister Primakov had no plan to solve the problem as the economy continued to suffer. Postal service

was threatened as the Post Office could not pay its bills. President Kuchma of Ukraine called for changes in market reform even as the Parliament turned down a bill to restore lost savings.

D30027.M.100.T.E

Fifteen months of world economic turmoil are threatening political stability. Lowering Federal Reserve interest rates is not countering the crisis. IMF is worried about the turndown in Japan, economic meltdown in Russia, depression in Indonesia, and anxiety about Latin America where investors are pulling out. IMF critics say it needs to understand national politics better and focus on social issues. Russia's economic confusion is upsetting the US. Russia is considering hard currency controls, demanding IMF loans and will not end government privatization. Ukraine, affected by Russia, is trying to save its fast-developing money system and keep investors.

D30027.M.100.T.G

Early October was fraught with economic woes as the International Monetary Fund prepared for its annual meeting. The IMF faces criticism for ignoring the social costs of its actions and being a pawn to Western demands. A small cut in US interest rates lowered markets worldwide. Russia, whose economy collapsed in August, was looking for a cure—possibly instituting Soviet-style measures. Key issues were stopping dollars from leaving the country and getting foreign investment and IMF loans. The postal service was in chaos, owing everyone. Demonstrations were expected. The Ukraine also struggled, especially to keep banks working. An IMF loan was on the way.

References

- [1] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of ICML 2014, 31st International Conference on Machine Learning*, pages 1188–1196, Beijing, China, 2014.
- [2] Yogan Jaya Kumar and Naomie Salim. Automatic multi document summarization approaches. *Journal of Computer Science*, 8(1):133–140, 2012.
- [3] Güneş Erkan and Dragomir R. Radev. Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479, 2004.
- [4] Inderjeet Mani and Mark T. Maybury. *Advances in Automatic Text Summarization*. The MIT Press, 1999.
- [5] Hans Peter Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958.
- [6] Kathleen McKeown and Dragomir R. Radev. Generating summaries of multiple news articles. In *Proceedings of SIGIR 1995, 18th annual international ACM SIGIR conference on research and development in information retrieval*, volume 3, pages 74–82, Seattle, USA, 1995.
- [7] Kathleen McKeown, Rebecca J. Passonneau, David K. Elson, Ani Nenkova, and Julia Hirschberg. Do summaries help? a task-based evaluation of multi-document summarization. In *Proceedings of SIGIR 2005, 28th annual international ACM SIGIR conference on research and development in information retrieval*, pages 210–217, Salvador, Brazil, 2005.
- [8] Ani Nenkova and Kathleen McKeown. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2–3):103–233, 2011.
- [9] Rada Mihalcea and Paul Tarau. TextRank: Bringing order into texts. In *Proceedings of EMNLP 2004, 9th conference on Empirical Methods in Natural Language Processing*, volume 4, pages 404–411, Barcelona, Spain, 2004.

- [10] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of ICLR 2013, 1st International Conference on Learning Representations*, pages 1–12, Scottsdale, USA, 2013.
- [11] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of Text Summarization Branches Out, Workshop at ACL 2004*, pages 74–81, Barcelona, Spain, 2004.
- [12] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT-NAACL 2003, 1st Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 71–78, Edmonton, Canada, 2003.
- [13] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Third Edition*. The MIT Press, 3rd edition, 2009.
- [14] Ronald Brandow, Karl Mitze, and Lisa F. Rau. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685, 1995.
- [15] Ladda Suanmali, Naomie Salim, and Mohammed Salem Binwahlan. Fuzzy logic based method for improving text summarization. *International Journal of Computer Science and Information Security (IJCSIS)*, 2(1), 2009.
- [16] Julian Kupiec, Jan Pedersen, and Francine Chen. A trainable document summarizer. In *Proceedings of SIGIR 1995, 18th annual international ACM SIGIR conference on research and development in information retrieval*, pages 68–73, Seattle, USA, 1995.
- [17] Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43(5-6):907–928, 1995.
- [18] Michael R. Genesereth and Nils J. Nilsson. *Logical Foundations of Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 1987.
- [19] Khaled Khelif, Rose Dieng-Kuntz, and Pascal Barbry. An ontology-based approach to support text mining and information retrieval in the biological domain. *Journal of Universal Computer Science*, 13(12):1881–1907, 2007.
- [20] Lei Li, Dingding Wang, Chao Shen, and Tao Li. Ontology-enriched multi-document summarization in disaster management. In *Proceedings of SIGIR 2010, 33rd annual international ACM SIGIR conference on research and development in information retrieval*, pages 819–820, Geneva, Switzerland, 2010.

- [21] Rakesh Verma, Ping Chen, and Wei Lu. A semantic free-text summarization system using ontology knowledge. In *Proceedings of DUC 2007, 7th Document Understanding Conference*, Rochester, USA, 2007.
- [22] A. A. Kogilavani and B. Dr. P. Balasubramanie. Ontology enhanced clustering based summarization of medical documents. *International Journal of Recent Trends in Engineering*, 1(1):546–549, 2009.
- [23] Ji-Peng Qiang, Ping Chen, Wei Ding, Fei Xie, and Xindong Wu. Multi-document summarization using closed patterns. *Knowledge-Based Systems*, 99:28–38, 2016.
- [24] Ercan Canhasi and Igor Kononenko. Weighted hierarchical archetypal analysis for multi-document summarization. *Computer Speech and Language*, 37:24–46, 2016.
- [25] Ercan Canhasi and Igor Kononenko. Multi-document summarization via archetypal analysis of the content-graph joint model. *Knowledge and Information Systems*, 41(3):821–842, 2014.
- [26] Wenpeng Yin, Yulong Pei, Fan Zhang, and Lian'en Huang. SentTopic-MultiRank: a novel ranking model for multi-document summarization. In *Proceedings of COLING 2012, 24th International Conference on Computational Linguistics*, pages 2977–2992, Mumbai, India, 2012.
- [27] Yulong Pei, Wenpeng Yin, and Lian'en Huang. Generic multi-document summarization using topic-oriented information. In *Proceedings of PRICAI 2012, 12th Pacific Rim International Conference on Artificial Intelligence*, volume 7458, pages 435–446, Kuching, Malaysia, 2012.
- [28] Güneş Erkan and Dragomir R. Radev. LexPageRank: Prestige in multi-document text summarization. In *Proceedings of EMNLP 2004, 9th conference on Empirical Methods in Natural Language Processing*, pages 365–371, Barcelona, Spain, 2004.
- [29] John M. Conroy, Judith D. Schlesinger, Jade Goldstein, and Dianne P. O’Leary. Left-brain/right-brain multi-document summarization. In *Proceedings of DUC 2004, 4th Document Understanding Conference*, Boston, USA, 2004.
- [30] Michael Kwok-Po Ng, Xutao Li, and Yunming Ye. Multirank: Co-ranking for objects and relations in multi-relational data. In *Proceedings of KDD 2011, 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1217–1225, San Diego, USA, 2011.

- [31] Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. Manifold-ranking based topic-focused multi-document summarization. In *Proceedings of IJCAI 2007, 20th International Joint Conference on Artificial Intelligence*, pages 2903–2908, Hyderabad, India, 2007.
- [32] Leonhard Hennig. Topic-based multi-document summarization with probabilistic latent semantic analysis. In *Proceedings of RANLP 2009, 7th conference on Recent Advances in Natural Language Processing*, pages 144–149, Borovets, Bulgaria, 2009.
- [33] Asli Celikyilmaz and Dilek Hakkani-Tur. A hybrid hierarchical model for multi-document summarization. In *Proceedings of ACL 2010, 48th Annual Meeting of the Association for Computational Linguistics*, pages 815–824, Uppsala, Sweden, 2010.
- [34] Xiaoyan Cai, Wenjie Li, and Renxian Zhang. Combining co-clustering with noise detection for theme-based summarization. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(4), 2013.
- [35] Daraksha Parveen and Michael Strube. Multi-document summarization using bipartite graphs. In *Proceedings of TextGraphs-9: Graph-based Methods for Natural Language Processing, Workshop at EMNLP 2014*, pages 15–24, Doha, Qatar, 2014.
- [36] Dimitrios Galanis, Gerasimos Lampouras, and Ion Androutsopoulos. Extractive multi-document summarization with integer linear programming and support vector regression. In *Proceedings of COLING 2012, 24th International Conference on Computational Linguistics*, pages 911–926, Mumbai, India, 2012.
- [37] Prasad Pingali, Rahul K, and Vasudeva Varma. IIIT hyderabad at DUC 2007. In *Proceedings of DUC 2007, 7th Document Understanding Conference*, Rochester, USA, 2007.
- [38] Hui Lin and Jeff Bilmes. A class of submodular functions for document summarization. In *Proceedings of ACL-HLT 2011, 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 510–520, Portland, USA, 2011.
- [39] Kristina Toutanova, Chris Brockett, Michael Gamon, Jagadeesh Jagarlamudi, Hisami Suzuki, and Lucy Vanderwende. The PYPHY summarization system: Microsoft research at DUC 2007. In *Proceedings of DUC 2007, 7th Document Understanding Conference*, Rochester, USA.
- [40] Douglas L. Baker and Andrew Kachites McCallum. Distributional clustering of words for text classification. In *Proceedings of SIGIR 1998, 21st annual international ACM SIGIR conference on research and development in information retrieval*, pages 96–103, Melbourne, Australia, 1998.

- [41] Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for re-ordering documents and producing summaries. In *Proceedings of SIGIR 1998, 21st annual international ACM SIGIR conference on research and development in information retrieval*, pages 335–336, Melbourne, Australia, 1998.
- [42] Yanran Li and Sujian Li. Query-focused multi-document summarization: Combining a topic model with graph-based semi-supervised learning. In *Proceedings of COLING 2014, 25th International Conference on Computational Linguistics*, pages 1197–1207, Dublin, Ireland, 2014.
- [43] You Ouyang, Wenjie Li, Sujian Li, and Qin Lu. Applying regression models to query-focused multi-document summarization. *Information Processing and Management*, 47(2): 227–237, 2011.
- [44] Dragomir R. Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40(6): 919–938, 2004.
- [45] Rada Mihalcea and Paul Tarau. A language independent algorithm for single and multiple document summarization. In *Proceedings of IJCNLP 2005, 2nd International Joint Conference on Natural Language Processing*, pages 19–24, Jeju Island, Korea, 2005.
- [46] Chao Shen and Tao Li. Multi-document summarization via the minimum dominating set. In *Proceedings of COLING 2010, 23rd International Conference on Computational Linguistics*, volume 2, pages 984–992, Beijing, China, 2010.
- [47] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.