Detection of main ideas and production of summaries in English, Spanish, Portuguese and Russian

60 years of research

Griselda Areli Matias Mendoza
Yulia Ledeneva
René Arnulfo García Hernández

Alfaomega

PUBLICACIONES UAEM

Detection of main ideas and production of summaries in English, Spanish, Portuguese and Russian.

60 years of research

Detection of main ideas and production of summaries in English, Spanish, Portuguese and Russian.

60 years of research

Griselda Areli Matias Mendoza

Yulia Ledeneva

René Arnulfo García Hernández

Alfaomega

Detection of main ideas and production of summaries in English, Spanish, Portuguese and Russian. 60 years of research
Griselda Areli Matias Mendoza, Yulia Ledeneva and René Arnulfo García Hernández

Luis Cejudo-Espinosa
Translator

First edition, October 2021

Catalog Datalog

Detection of main ideas and production of summaries in English, Spanish, Portuguese and Russian. 60 years of research.

First edition

Alfaomega Grupo Editor, S.A. de C.V. México
ISBN: 978-607-538-809-0

Universidad Autónoma del Estado de México
ISBN: 978-607-633-341-9

Size: 17 x 23 cm. Pages 256

Cover picture: Dreamstime

ISBN 978-607-633-342-6 (PDF UAEM)
ISBN 978-607-538-810-6 (PDF Alfaomega Grupo Editor)

This book was revised and approved by two double blind peers alien to Universidad Autónoma del Estado de México. The review was supervised by the Secretary of Research and Advanced Studies, according to file number 274/2020

This publication's content is responsibility of the authoring people.

Printed in Mexico

PROLOGUE

Detection of main ideas and production of summaries in English, Spanish, Portuguese and Russian. 60 years of research, is a book that anyone can read; however, being a text that presents a task related to Natural Language Processing (NLP) it is intended for researchers, postgraduate students, engineers and for those interested in NLP and knowledge generation.

One of the intelligent tasks performed by human beings is the production of summaries of documents; this with a view to enabling the readers to quickly learn the information contained therein. However, with the exponential growth of electronic information this task has become difficult for there is a lot of information, and accessing it is time-consuming and requires resources. The Automatic Generation of Text Summaries (AGTS) is a task that reaches 60 years of research, since Luhn's publication in 1958. Ever since, there has been great advance in AGTS enquiries, especially in the English language, which is noticed in books and scientific articles that showcase the quality of the methods and techniques in a qualitative manner by assessing them. At once, there has been a lack of qualitative tests to find out whether a machine-made summary has reached the quality to deceive a human and not to notice if such summary was made by a computer or a human. To do so, we present Turing Test carried out with machines that currently generate summaries automatically in the most written and spoken languages such as English, Spanish, Portuguese and Russian. Chapter I opens with a very interesting observation, which has been subject of debate for researchers: Can a machine be intelligent? This question is approached from the AGTS's standpoint. To answer this, two Turing Tests for AGTS tasks are

presented: one for Spanish and another for English. The goal of the tests is that humans identify which two summaries, out of six, are man-made. Moreover, the results obtained are shown and a brief introduction to AGTS's issues is given.

For the task of automatic generation of text summaries to be studied, there must be certain resources; this way, in chapter II, the main elements to study and solve AGTS's tasks such as *corpora*, heuristics and evaluation metrics are presented.

In chapter III, the two main sort of summaries are distinguished according to their condensation strategy: abstractive and extractive; there is a description of each, as well of some novel scientific methods that work abstractively. Additionally, a table presents the main characteristics used in AGTS tasks, essentially for the extractive ones.

In chapter IV, an analysis of commercial AGTS tools is made. The method used by each of them as well as the steps to make them work are described, this with a view to having a panorama of the commercial tools' quality regarding the heuristics and novel scientific methods.

In chapters V, VI, VII and VIII, AGTS is approached in English, Spanish, Portuguese and Russian, respectively. But, why in these languages? In English because it is the most studied language and there are more resources and works with which make a comparison, moreover there is still much to research. In Spanish because it is our mother language and obviously, we are interested in producing text summaries in our own language. In Portuguese because it is a Romance language, such as Spanish, and although there are research works in this language, desirable results have not been accomplished yet. And finally, Russian because there are no reported research works on this language for AGTS's tasks and also because we have a native expert to generate the resources in this language. Furthermore, in this chapter, the results of the assessments carried out with specialized *corpus* for each language are shown. The best AGTS novel scientific methods, commercial tools and main heuristics are tried. Finally, in chapter IX conclusions and discussions are displayed. The contributions distinguished in this book are a report on six Turing Test, by means of which it is evinced that a machine can deceive a human and produce a better summary than one produced by another human; the integration of and report on novel methods developed thus far; a comparison of AGTS systems, integration and

report for Spanish and Russian, as for these languages there was no formal enquiry; and, finally, the results displayed are a reference source to be aware of the stage of AGTS research in each of the four languages.

Acknowledgements

The publication of this book was made possible by funding from Secretary of Research and Advanced Studies, the Mexican Society of Artificial Intelligence and the authors.

Content

Chapter 1

Introduction

In this chapter the reader is introduced to the issues of Automatic Generation of Text Summaries (AGTS). Two Turing Test are presented, one in Spanish the other in English, with a view to finding out whether an individual is able to identify summaries generated by another human or by a machine. It is also enquired if a human is able to replicate the necessary knowledge to produce a summary in an automatic manner on a machine, particularly on a computer.

Can a machine be intelligent?

This simple yet deep question made a number of scientists debate on what intelligence is; therefore, there were questions as basic as: in order to move myself from one place to another, do I need to be intelligent? Or the one who adds a couple of numbers, can they be considered intelligent? In order to answer, Alan Turing, considered one of the fathers of computer science, presented a test that could indirectly solve the original question whether a machine, particularly a computer, is to be deemed intelligent.

The Turing Test is an imitation game performed by three people: a man (A), a woman (B) and an interrogator (C) of either sex. The interrogator remains in a room separate from the other two. The aim of the game for the interrogator is to find out which of the other two people is a man and a woman. He, the interrogator, knows them by their labels X and Y, at the end of the game he says: "X is A and Y is B or the other way". He can ask questions for A and B, all the answers are given in writing so that the voice will not help. The variant introduced by Turing is to substitute one of the respondents with a machine, then the interrogator must decide in like manner, which is A and B without him knowing about the apparatus that poses as one of the respondents. The machine will pass the Turing Test when the interrogator does not manage to identify whom he is speaking to (Turing, 1950). This is to say, when the human becomes confused, frequently the machine shows intelligence. It is also because of this test that Alan Turing is considered the father of artificial intelligence as well.

Artificial intelligence has as an objective to emulate some of the human intellectual faculties in artificial systems (Benítez *et al*., 2014). While emulating is understood as the application of theoretical models in a machine (computer) with the purpose of obtaining satisfactory results for humans. The most noticeable artificial intelligence areas of study are robotics, expert systems, perception and learning issues, Natural Language Processing (NLP), among others (Coarite Choque, 2008).

Based on the Turing Test, IBM developed systems that were able to fully compete against humans in tasks that may be considered very intelligent. To do so, the firm focused on chess: it made compete chess champion Garry Kasparov against the system developed by IBM, known as Deep Blue, which after several matches managed to defeat the most intelligent player (Hsu, 1999).

Recently, in 2011, one of the most famous applications of such test was carried out by IBM to answer questions produced in natural language. The firm made compete two human champions in Jeopardy against the Watson System, which is able to recognize voice, ask a question and produce the response in voice. Watson is an artificial intelligence system and, in the competence, it was able to emulate and surpass the human (Gliozzo *et al*., 2017).

Beyond games, nowadays it is common to hear news on how artificial intelligence (including all its subdisciplines) supports and surpasses in the performance of human activities in various spheres such as medicine, security and education, to mention a few.

1.1 Turing Test applied to generate automatic summaries

One of the intelligent tasks carried out by human beings is the production of document summaries so that readers may quickly learn the information contained in them. However, with the exponential growth of electronic information this task has become cumbersome, as there is much information and accessing it takes time, thereby resources.

The Automatic Generation of Text Summaries (AGTS) is a task that becomes 60 years of age of research since its first work (Luhn, 1958). There is great advance in AGTS research in English, noticed in specialized books and scientific articles, evinced in the quality of the methods and quantitative techniques by means of assessing them (Mani, 2001), (Mihalcea and Radev, 2011), (Ledeneva *and* García-Hernández, 2013), (Torres-Moreno, 2014), (Ledeneva and García-Hernández, 2017), etc. However, no Turing Test has been performed on the machines that produce automatic summaries at present, in the most spoken and written languages in the world such as English, Spanish, Portuguese and Russian.

In order to introduce the fundamental AGTS concepts, it is suggested that the reader carries out the following Turing Test to distinguish, from a news item in Spanish (taken from Mexican newspaper *La Crónica*[1]), which summaries were produced by humans and machines.

[1]Renowned newspaper in Mexico. It can be reached at: http://www.cronica.com.mx/noticias.php

The item presented below was taken as it appears on the website.

Diseñan casa que resbalaría vientos de huracán y tornados

Es casi una costumbre ver cada año países del sureste de Estados Unidos devastados por huracanes. ¿Por qué no hacer casas más resistentes? Se preguntó el mexicano Sergio Díaz Zubieta después de atender a su lógica como arquitecto. Así nació la idea de buscar una solución y pensó que el problema no estaba en la resistencia de los materiales de las casas, sino en su dinámica. De esta forma elaboró una solución para elaborar hogares —en el estándar de la región— que prácticamente resbalaran los ciclones: casas resistentes a tornados y huracanes. Adecuándose a elaborara un diseño económico que cumpliera requisitos de la región, como la nieve, por ejemplo, desarrolló una idea sencilla pero práctica. "Si uno parte de la base de que algo que no es plano no recibe viento a diferencia de algo vertical que lo recibe totalmente, algo que tenga 45 grados de inclinación deberá de ser sólo afectado en el 50 por ciento", explica en entrevista el arquitecto. "Si recibo un viento de 300 km por hora con la solución de techos a 45 grados, el ciclón resbalará sobre la estructura porque reduce, como mínimo, la mitad del impacto". Este diseño piramidal ha sido patentado ya por el mexicano en el país y en EU, puesto que no existe algo parecido en la industria de la construcción hasta ahora. La tarea ahora es promoverla porque el papel por sí mismo "no sirve de nada". Si las afectaciones a la población por este tipo de fenómenos naturales son una constante, por qué no se había pensado en soluciones distintas para disminuir el riesgo o los daños. "Porque hay mucho poder económico en medio de esto. Pero aún así se puede modificar y hacer el esfuerzo". CONFORTABLE. De acuerdo con el arquitecto, el diseño que propone se puede componer de dos formas: elaborando las viviendas con piezas precoladas o con colado en el mismo lugar. Lo preferente sería la segunda opción, añade. "De esta forma, la estructura sería más íntegra y resistente a tornados, pero ni huracanes ni sismos le afectarían. Tenemos una opción para evitar más fallecimientos y millones de dólares en pérdidas materiales". La estructura de concreto armado podría ser armado con aditivos que la hicieran impermeable, pero incluso resistente al fuego y también confortable, porque sería menos afectada por aumento o baja de temperaturas. Díaz Zubieta tiene, por otra

parte, una solución de este tipo para México en el contexto de los ciclones que nos afectan y la situación económica de nuestra población. "Es una más sencilla y económica porque tenemos otro contexto (también sería aplicable a países de Latinoamérica). Ésta representaría sólo el 15% del costo del diseño que propongo para EU; sería de fácil construcción y pensada para población de bajos recursos". Si bien apunta que sería totalmente diferente a su primer diseño, las modificaciones que se realizarían serían sin detrimento a su calidad y seguridad.

Then, humans and machines were asked to produce 100-word summaries of the item; this way, they had to be trimmed in order to objectively assesssing. In particular, out of the summaries presented it has to be decided which two were produced by humans.

Summary 1

Es casi una costumbre ver cada año países del sureste de Estados Unidos devastados por huracanes. De esta forma elaboró una solución para elaborar hogares —en el estándar de la región— que prácticamente resbalaran los ciclones: casas resistentes a tornados y huracanes. Adecuándose a elaborara un diseño económico que cumpliera requisitos de la región, como la nieve, por ejemplo, desarrolló una idea sencilla pero práctica. CONFORTABLE. La estructura de concreto armado podría ser armado con aditivos que la hicieran impermeable, pero incluso resistente al fuego y también confortable, porque sería menos afectada por aumento o baja de temperaturas. Díaz Zubieta…

Summary 2

¿Por qué no hacer casas más resistentes? Se preguntó el mexicano Sergio Díaz Zubieta después de atender a su lógica como arquitecto. Así nació la idea de buscar una solución y pensó que el problema no estaba en la resistencia de los materiales de las casas, sino en su dinámica. "Si recibo un viento de 300 km por hora con la solución de techos a 45 grados, el ciclón resbalará sobre la estructura porque reduce, como mínimo, la mitad del impacto". Si las afectaciones a la población por este tipo de fenómenos naturales son una constante, por qué no se…

Summary 3

Es casi una costumbre ver cada año países del sureste de Estados Unidos devastados por huracanes. ¿Por qué no hacer casas más resistentes? Se preguntó el mexicano Sergio Díaz Zubieta después de atender a su lógica como arquitecto. Así nació la idea de buscar una solución y pensó que el problema no estaba en la resistencia de los materiales de las casas, sino en su dinámica. De esta forma elaboró una solución para elaborar hogares —en el estándar de la región— que prácticamente resbalaran los ciclones: casas resistentes a tornados y huracanes. Adecuándose a elaborara un diseño económico que cumpliera…

Summary 4

Es casi una costumbre ver cada año países del sureste de Estados Unidos devastados por huracanes. Así nació la idea de buscar una solución y pensó que el problema no estaba en la resistencia de los materiales de las casas, sino en su dinámica. "Si recibo un viento de 300 km por hora con la solución de techos a 45 grados, el ciclón resbalará sobre la estructura porque reduce, como mínimo, la mitad del impacto". De acuerdo con el arquitecto, el diseño que propone se puede componer de dos formas: elaborando las viviendas con piezas precoladas o con colado en…

Summary 5

CONFORTABLE. "Es una más sencilla y económica porque tenemos otro contexto (también sería aplicable a países de Latinoamérica) Díaz Zubieta tiene, por otra parte, una solución de este tipo para México en el contexto de los ciclones que nos afectan y la situación económica de nuestra población Tenemos una opción para evitar más fallecimientos y millones de dólares en pérdidas materiales" "Si recibo un viento de 300 km por hora con la solución de techos a 45 grados, el ciclón resbalará sobre la estructura porque reduce, como mínimo, la mitad del impacto" Ésta representaría sólo el 15% del costo del…

Summary 6

Es casi una costumbre ver cada año países del sureste de Estados Unidos devastados por huracanes ¿Por qué no hacer casas más resistentes?, se preguntó el mexicano Sergio Díaz Zubieta después de atender a su lógica como arquitecto. De esta forma elaboró una solución para elaborar hogares —en el estándar de la región— que prácticamente resbalaran los ciclones: casas resistentes a tornados y huracanes. "Si recibo un viento de 300 km por hora con la solución de techos a 45 grados, el ciclón resbalará sobre la estructura porque reduce, como mínimo, la mitad del impacto". Este diseño piramidal ha sido patentado…

The previous test and other two in Spanish (see Appendix A) were applied to seventy-three undergraduate and graduate students and professors, whose native language is Spanish. Results are displayed in **table 1.1.**

Table 1.1 Results of the Turing Test for humans for the Spanish language

Pairs of summaries chosen by humans	Percentage of confusion between the selected summaries (%)
Human – Machine	56
Machine – Machine	36
Human – Human	8

In the third row in the table above, it is shown that only 8% of the times people correctly identified the two man-made summaries. Most of the confusion took place in the first column, with 56%, in which a summary made by a human and another by a machine were selected. However, interestingly, 36% of the people thought that the automatically generated summaries were man-made. With these results, not only is it possible to notice that the machine has overcome the Turing Test, but surpassed humans.

It is possible that these figures may cast doubt on the reader's election, though the confusion or doubt comes from the similarity of both summaries. Over the chapters of this book, we will disclose which summaries were made by humans and which by computers.

Motivated by the results above and since the bulk of research on AGTS has been carried out in English, this test was carried out for this language. Sixty-eight people with sufficient command to read in English were involved; likewise, the reader is invited to make the *test* with the following news item.

Hurricane Gilbert Heads Toward Dominican Coast

Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas. The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph. ``There is no need for

alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly before midnight Saturday. Cabral said residents of the province of Barahona should closely follow Gilbert's movement. An estimated 100,000 people live in the province, including 70,000 in the city of Barahona, about 125 miles west of Santo Domingo. Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night. The National Hurricane Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo. The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a ``broad area of cloudiness and heavy weather" rotating around the center of the storm. The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday. Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds and up to 12 feet feet to Puerto Rico's south coast. There were no reports of casualties. San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsided during the night. On Saturday, Hurricane Florence was downgraded to a tropical storm and its remnants pushed inland from the U.S. Gulf Coast. Residents returned home, happy to find little damage from 80 mph winds and sheets of rain. Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane. The first, Debby, reached minimal hurricane strength briefly before hitting the Mexican coast last month.

The summaries of the text above are the following; as in the case of the Spanish summaries, these have an extension of one hundred words.

Summary 1

Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas. The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph. An estimated 100,000 people live in the province, including 70,000 in the city of Barahona, about 125 miles west of Santo Domingo. Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night. The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with…

Summary 2

Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas. The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph. "There is no need for alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly before midnight Saturday. Cabral said residents of the province of Barahona should closely follow Gilbert's movement. An estimated 100,000 people live in the province, including 70,000 in the city of Barahona, about 125 miles west of Santo Domingo…

Summary 3

Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas. The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph. Cabral said residents of the province of Barahona should closely follow Gilbert's movement. Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night. The National Hurricane Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto…

Summary 4

Tropical Storm Gilbert in the eastern Caribbean strengthened into a hurricane Saturday night. The National Hurricane Center in Miami reported its position at 2 a.m. Sunday to be about 140 miles south of Puerto Rico and 200 miles southeast of Santo Domingo. It is moving westward at 15mph with a broad area of cloudiness and heavy weather with sustained winds of 75mph gusting to 92mph. The Dominican Republic's Civil Defense alerted that country's heavily populated south coast and the National Weather Service in San Juan, Puerto Rico issued a flood watch for Puerto Rico and the Virgin Islands until at..

Summary 5

The first, Debby, reached minimal hurricane strength briefly before hitting the Mexican coast last month. Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night. An estimated 100,000 people live in the province, including 70,000 in the city of Barahona, about 125 miles west of Santo Domingo. The National Hurricane Center in Miami reported its position at 2 a m Sunday at latitude 16 1 north, longitude 67 5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo. Residents returned home, happy to find little damage from 80…

Summary 6

Hurricane Gilbert is moving toward the Dominican Republic, where the residents of the south coast, especially the Barahona Province, have been alerted to prepare for heavy rains, and high winds and seas. Tropical Storm Gilbert formed in the eastern Caribbean and became a hurricane on Saturday night. By 2 a.m. Sunday it was about 200 miles southeast of Santo Domingo and moving westward at 15 mph with winds of 75 mph. Flooding is expected in Puerto Rico and the Virgin Islands. The second hurricane of the season, Florence, is now over the southern United States and downgraded to a tropical…

The results for the Turing Test for the English language are presented in **table 1.2**.

Table 1.2 Results of the Turing Test for humans for the English language

Pairs of summaries chosen by humans	Percentage of confusion between the selected summaries (%)
Human – Machine	46
Machine – Machine	41
Human – Human	13

We have that only 13% of the cases were correctly identified, i.e., Human-Human. Most of the confusion occurred in the Human-Machine elections, with 46%. However, in 41% of the cases machine-made summaries were preferred; that is to say, 3 out of 4 tests were mistaken and the summary by a machine was chosen. Once again, the Turing Test was accomplished at least for the domain[2] of the news. It is worth mentioning that there are no trials with such *test* for Portuguese or Russian yet; it is considered future work, though. For the English language two more tests were carried out; these are presented in Appendix B.

Surely, a number of observations appear after the previous tests, albeit they will be addressed over the book; moreover, which summaries were produced by a machine and which by humans will be disclosed. The first aspect deals with the way summaries are made, which can be extractive or abstractive according to the sort of condensation. It is considered that extractive summaries only subtract a set

[2]Domain refers to the context of the documents, for instance, news, scientific articles, poems, tweets, emails, among others.

of sentences (paragraphs or phrases) of the original document; while abstractive ones can modify the sentences from the original document and include ideas or opinions of the human who makes the summary. Owing to this, extractive summaries are deemed more objective, whereas abstractive, more subjective; what is important is to differentiate the intended use.

In the tests above and over the book only extractive summaries are dealt with because it is of our interest to keep the authors' original ideas. On the other hand, it is considered that in reality humans make extractive summaries, as they only copy the parts they consider relevant (Jing, 2002).

1.2 How does a human make a summary?

Humans, who start education as children, are instructed on the processes to summarize. A great deal of authors agree that a series of steps must be always followed (Vivaldi, 2000), (Maqueo, 2004), (UNE 50-103-90, 1990), (Kaufman and Perelman, 1999). However, each of them has a different approach, so there is no standardized method. Though, what is sought is to extract the most important characteristics of an original text to integrate them into a shorter text which contains the main ideas.

Among some of the steps to produce a summary, the following proposed by Ana María Maqueo (Maqueo, 2004) are presented:

1. Read the text attentively.
2. Separate the main ideas in blocks.
3. Underline these ideas.
4. Draft the summary linking the main ideas by means of appropriate words.

In a study by Kaufman and Perelman (1999), it was found that one hundred and eighty children were told that summaries are produced as follows:

1. Underline the parts of the text considered important.
2. Delete everything that is not thought to be part of the summary.
3. Write: they were given a piece of paper (with thirteen lines) and are asked to write the summary in the space given, without exceeding the space.

In Vivaldi's book (2000), it is mentioned that the steps to produce a summary are the following:

1. Find out the most important chapters and take notes on the fundamental concepts.
2. Summarize the most interesting, leaving the least relevant for the end. As regards this step, the author mentions that if there is time to read slowly, notes can be taken. However, if time is short, the text is usually underlined and presented as a part of the summary.
3. The order of the summary must reflect what the books means.

If we consider the steps proposed by the authors above, it is possible to realize that the procedures to make summaries are guided by the generation of extractive summaries. Nevertheless, it is worth mentioning that authors such as Maqueo (2004) recommend drafting and rewriting the text, which is when humans exercise abstraction.

Due to the foregoing, it comes as no surprise there exist studies in which it is verified that humans, by and large, produce summaries by selecting sentences from the original document, and only in cases when exhaustive reviews and thorough comprehension of the text are needed, rewriting and abstracting techniques are resorted to (Banko and Vanderwende, 2004). Jing's work (2002) breaks down a set of three hundred summaries analyzed by humans with the purpose of finding out how they were made. The result is the identification of six operations humans perform when making summaries:

1. Sentence simplification.
2. Sentence combination.
3. Syntax transformation.
4. Lexical paraphrasing.
5. Generalization.
6. Specialization.

The experiments showed that 81% of the sentences was identical copies of those in the original texts (Jing, 2002).

Taking the previous study as a reference, the steps proposed to produce a summary and the results of Turing Test, it may be concluded that humans make

extractive summaries because they only copy the parts they consider important; among them first sentences. In this regard, there are a number of hypotheses, namely: humans become tired and do not read the full text; the author writes the most important at the beginning of the text; the domain of the document influences in a relevant manner. Owing to these reasons, the summaries produced by humans are composed of first sentences.

As mentioned earlier, the documents' domain influence on the production of a summary; particularly, the most studied texts in AGTS tasks are news items. In the sphere of news, the so-called inverted pyramid rule is used, which consists in placing the most relevant information at the beginning of the text (**figure 1.1**). Although there are other heuristics such as the traditional structure of a narrative text, in which the most important elements usually appear at the end, or that of the truncated pyramid, also called hourglass in which the most important data are presented in the first place, the rest of the information in decreasing order, and after some time, important information is drafted once again (Briones *et al*., 2012).

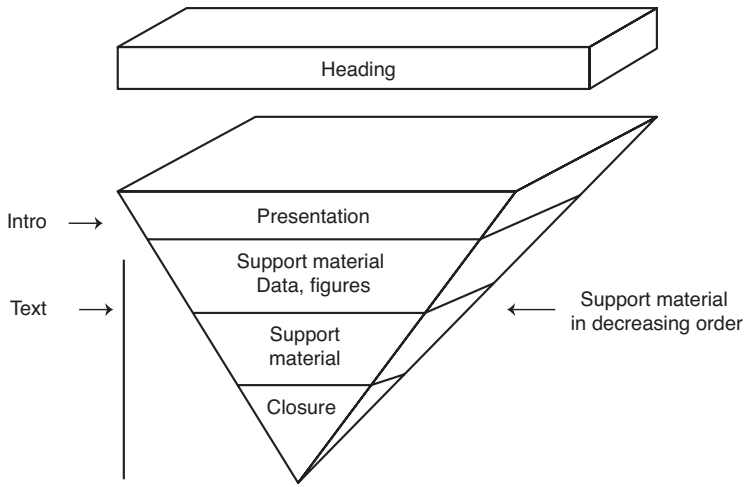


Figure 1.1 Structure of a news item (Briones *et al.*, 2012)

In Vázquez (2015), there is an analysis of a *corpus* in English (DUC02) to identify how humans select the sentences. The longest document has 177 sentences; in **figure 1.2** it is noticed that 50% of the times sentences 1 to 5 are selected; while 25%, from 6 to 10; after sentence 15, there is a selection frequency of 20%, which

descends as the number of sentences increases. For example, sentences from 61 to 177 are only selected once in the 567 documents.

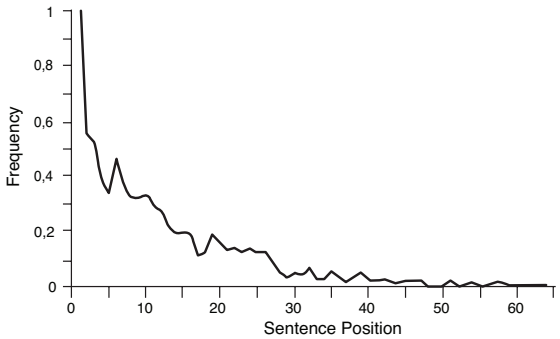


Figure 1.2 Position – Frequency relation in *corpus* DUC02

Over these sixty years of research on summaries, most of the efforts have been carried out for extractive AGTS in the English language. However, owing to the nature of the text, some of the methods created are able to work with more than one language. It is because of this that the previous tests presented four extractive summaries made by machines, tried in English, but which can work with other languages.

The results of the Turing Test clearly demonstrate that the machine managed to confuse humans in AGTS tasks; they even suggest that humans prefer machine-made summaries. With this validation it is verified that a machine can be intelligent to produce a summary in Spanish and English. However, it is important to reflect on what is being actually tried is that humans are so intelligent that they can model and reproduce the necessary knowledge for a machine may resemble them in some tasks.

This is to say, the Turing Test carried out allows noticing humans are accomplishing that a computer emulates them in AGTS. Then, if a machine can really learn or make an intelligent search for the main text's ideas over its structure and produce a summary, this can be repeated for various languages.

In order to run actual experiments in various languages, it is necessary to resort to specialized *corpora* and assessment tools. *Corpora* usually comprise a set of documents and one of two summaries of each document made by humans,

called *gold standard*,[3] which are references to assess the summaries produced by a machine (Over *et al*., 2007). For the English language one of the most utilized tools to assess is *ROUGE* (Lin, 2004), as it allows obtaining a grade for the summary made by a machine in comparison with a human. With a *corpus* and the assessment tool, the current status of the study of AGTS tasks in any language can be more easily ascertained.

In the following chapters, the problems faced by four of the most important languages in AGTS are approached. Specifically, English, Spanish, Portuguese and Russian were chosen because they are among the ten most important languages in the world. English was chosen because it is the language with the most research and also because it is the most spoken, either as a native or foreign language.[4] Spanish was picked as it is the second language in the world according to native speakers, one of the languages that has received the least attention in this regard and because it does not have Anglo-Saxon origins (Arévalo, 2017). Portuguese was chosen as it was close to Spanish and because there are some research works on this topic (Pardo and Rino, 2003), (Mihalcea and Tarau, 2005). Finally, AGTS will be studied in Russian owing to its origin, different from the other languages, and because there is no research on Russian regarding this topic.

It is worth mentioning that in this chapter, machines are referred to as the computers in which an AGTS method can be set up. Henceforward, we will refer to novel scientific methods and commercial tools as synonyms for the concept given to the machine in chapter I.

1.3 Book's structure

This book's aim is to summarize 60 years of research in AGTS. As it is known the English language is the most studied in this regard, this way, efforts in this language will be retaken and applied in Spanish, Portuguese and Russian. Plus, the results of the main heuristics (*baseline:random*, *baseline:first* and *Topline*), tools and novel scientific methods in AGTS are presented for each language.

---

[3] *Gold standard* summaries are made by humans and are also known as reference summaries.

[4] Website https://www.internetworldstats.com/stats7.htm provides the top 10 of the most spoken languages in the world.

This book comprises eight chapters. Chapter I presents two Turing Test for AGTS, one in Spanish and another in English. The goal of these tests is that humans find out which two summaries out of six were made by a person. Additionally, the results obtained are shown and there is a brief introduction to AGTS tasks.

In chapter II, the main elements to address AGTS tasks are presented, to name a few: *corpus*, heuristics and assessment metrics.

Chapter III presents the two classifications of summaries according to their condensation strategy; moreover, some novel scientific methods that work in various languages are explained.

In chapters IV to VIII, AGTS tasks are approached in English, Spanish, Portuguese and Russian, respectively. As well, the results of the assessments performed with specialized *corpora* for each language are presented; the best novel scientific methods, commercial tools and main AGTS heuristics are tried. Finally, in chapter IX discussions are developed and conclusions drawn.

Chapter II

Corpora, heuristics and assessment metrics

This chapter presents the description of *corpora*, heuristics and assessment metrics used to enrich the AGTS options for machines and humans for the English, Spanish, Portuguese and Russian languages.

The earliest studies on AGTS, which took place by the end of the 1950's, were carried out by Luhn (1953 and 1958) and Edmundson (1969). They resorted to their own *corpora* and manual assessment; hence, AGTS methods up to the year 2000 did not have a comparison reference (Luhn, 1953), (Luhn, 1958), (Edmundson and Wyllys, 1961), (Edmundson, 1969), (Kupiec *et al.*, 1995), (Mani, *et al*., 1999). In order to find out the quality of the summaries generated by machines as regards those produced by humans, it is necessary to have a *corpus*, reference heuristics and assessment metrics.

2.1 *Corpora*

According to *Diccionario Manual de la Lengua Española*[TN] [Dictionary of the Spanish Language] (*Corpus*, 2014), a *corpus* is an extensive set of various sorts of text, ordered and classified, that serves as a base for research. In the sphere of AGTS, *corpora* may comprise texts from various domains (for instance, news, scientific articles, literary texts, cookery texts, among others) and the summaries produced by humans, which together are called gold standard (in some cases there may be only one summary).

Among the earliest *corpora* created, we find those of *Document Understanding Conferences* (DUC), which started in the year 2000 and was followed by other seven, DUC01 – DUC07.

Every conference involves of a number of tasks, while the corresponding *corpora* with their own gold standards were created for each of them; the aim of these conferences was to advance AGTS research in English by means of large-scale experiments using their *corpora* and gold standards.

The Text Analysis Conference (TAC) began in 2008, organized as assessment workshops to motivate research in natural language processing and related applications. One of TAC's main goals is to gather test collections to anticipate the assessment needs of modern systems. It was in 2008, 2009, 2010, 2011 and 2014, when TAC focused on AGTS tasks, being user-oriented

[TN]According to Merriam-Webster Dictionary, "The collection of recorded utterances that is used as a basis for the descriptive analysis of a language or dialect". "*Corpus.*" Merriam-Webster's Unabridged Dictionary, Merriam-Webster, https://unabridged.merriam-webster.com/unabridged/*corpus*. Accessed 20 Apr. 2021.

multi-document summaries its main field of study. Most of the research on AGTS has been carried out in English, presently there are few works for other languages.

There has been some research for Spanish, but none of the works use a standard *corpus* or specialized for AGTS. These works make use of *corpora* adapted to extract information or simply produce their own (Acero *et al*., 2001), (Toledo-Báez, 2010), (da Cunha Fanego, 2005), (Villatoro E., 2007), (Plaza, 2011), (Venegas, 2011), (Cabral *et al.*, 2014). Therefore, research works cannot be compared and the advance of AGTS research cannot be stated.

Portuguese has one *corpus*, TeMário (Pardo and Rino, 2003), which has been utilized in most of the research (Antiqueira, 2007), (Margarido *et al*., 2008), (Leite and Rino, 2009), (Amancio *et al.*, 2012), (Cabral *et al*., 2014), (Cavalieri *et al.*, 2015), which has enabled a comparison between methods and tools for this language.

There is no standard *corpus* known for AGTS tasks in Russian. One of Braslavski and Gustelev's main works (2007) takes news items from Gazeta.ru newspaper, though it is not available. In the work by Rojas (2016), a *corpus* of the news items from the same newspaper is built for AGTS tasks; results are shown in chapter VIII.

2.2 Heuristics

Heuristics translates from Greek as finding or inventing. A heuristic is an invention of rules, procedures or techniques for a human to solve a certain problem. In scientific methodology, heuristics are applied when tasks for which there are no algorithmic solution processes appear (Polya and Zugazagoitia, 1965).

In order to find out whether a machine is intelligent enough to produce a summary, this book presents a number of heuristics intuited from the domains of the analyzed data, while the performance of methods and smart tools is analyzed by means of comparing the calculated heuristics. These AGTS heuristics take the way in which humans make summaries into consideration. For example, to produce a news item, people place the most important information in the first sentences, then they write the details of the item. Owing to this, the use an heuristic of first sentences, called *baseline:first*, has been proposed (Ledeneva,

2008); which consists in composing a summary with the first 100, 200, or 400 words, for which the number of words depends on the length of the summary to generate. Another heuristic is *baseline:random* which is considered the worst way to choose sentences for a summary (*baseline:random*) and finally, *Topline* is considered the best.

2.2.1 *Baseline:first*

Baseline:first consists in taking the first words of the text to compose a summary (Ledeneva, 2008). In the experiments presented in this book, the first words are taken. For an intelligent machine, the goal is to surpass this heuristic. Particularly, for the domain of news, the goal is too demanding, as this sort of texts contains the most important information at the beginning of the document. *Baseline:first* was included in the Turing Test run in chapter I; the results reveal that humans become confused when they read a summary produced by *baseline:first*, which displays the heavy influence of first sentences in the composition of a text. Summary 3 in Spanish and summary 2 in English were generated by *baseline:first* (section 1.2).

2.2.2 *Baseline:random*

The worst option to produce a summary would be choosing the sentences that compose it at random. This election is called *baseline:random*, and was proposed and utilized by Ledeneva (2008); when a machine produces a summary, it is expected to be more intelligent and offer better results than just random sentences. Chapter I presents the Turing Test for six summaries, two of them made by humans, two automatically, while the other two, by the heuristics.

Summaries 5 in Spanish and English were produced by *baseline:random* (section 1.2).

2.2.3 *Topline*

Topline consists in obtaining the best combination of sentences out of all the possible combinations; this allows us to find out the best result we can reach when assessing the summaries generated with a *corpus* (Rojas, 2018).

One of the main challenges for AGTS is to generate extractive summaries more similar to those produced by humans (gold standard). However, various domains' gold standard summaries are generated in an extractive manner substituting some terms and phrases of the original text. According to Verma and Lee (Verma and Lee, 2017), the gold standards of the standard *corpora* in English, DUC01 and DUC02, use about 9% of non-used words in the original documents. Consequently, the maximal level of similarity will be under 100%, moreover, if they are compared against the gold standards, the limits of maximal performance will be lower (because of the lack of concordance between humans) for any AGTS method.

For the Turing Test performed in chapter I, two gold standards made by humans were presented for each language: summaries 2 and 4 in the case of Spanish, whereas 4 and 6 in English (section 1.2).

2.3 Assessment of automatic summaries

In order to assess the summaries produced by a machine, not only is the construction of standardized sets of data (*corpora*) needed, but also the utilization of various assessment methods.

Assessment methods are classified in intrinsic and extrinsic (Sparck Jones and Galliers, 1995). The first are based on direct analysis of the automatically produced summaries; to judge quality, grammatical criteria of the text's cohesion and coherence may be used; to assess the degree of coverage, a comparison between the automatic summaries and those made by the experts was carried out.

Extrinsic assessment methods study the summary in the context of the tasks for which it was generated, intending to find out the effect on some other tasks. These tasks may include, for instance, the assessment of relevance (Berker, 2011).

The assessment metrics presented in this charter are content similarity, accuracy, Recall and F-measure; while assessment methods are ROUGE and Pyramid.

2.3.1 Content similarity

Donaway *et al.* (2000) propose assessing the informational quality of a summary, applicable to extractive and abstractive summaries. One of the measures defined to calculate such similarity is the vocabulary test, in which traditional methods to retrieve information are applied to calculate the distance between the vectoral representations of the manual and automatic summaries, using the cosine metric; this metric can be automated and use summaries produced by humans.

2.3.2 Accuracy, Recall and F-measure

Accuracy and recall measures are the traditional ones in information retrieval (Salton and McGill, 1983). They have also been used in AGTS tasks.

Accuracy (P): it accounts for the number of correct sentences extracted by the machine:

$$P = \frac{correctas}{(correctas + incorrectas)} \tag{1}$$

Recall (R): it accounts for the number of correct sentences forgotten by the system:

$$R = \frac{correct}{(correct + forgotten)} \tag{2}$$

The sentences extracted by machine and human are defined as *correct*; *incorrect* are the sentences extracted by the machine, but not the human; while forgotten, the sentences extracted by the human, not the machine.

F-measure (F): it is the harmonic measure of accuracy and recall:

$$F = 2 * \frac{accuracy * recall}{accuracy + recall} \tag{3}$$

F-measure establishes a balanced equilibrium between recall and accuracy.

2.3.3 ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004) was put forward by Lin and Hovy (Lin and Hovy, 2003), (Lin and Och, 2004), (Lin and Och, 2004). This system calculates the quality of an automatically generated summary against some others produced by humans. In short, the number of the various common units, such as word sequences, pair of words and n-grams in the summary to assess and in the ideal summaries created by humans are counted. ROUGE includes various automatic assessment measures:

- ROUGE-N (cooccurrence of n-grams): it expresses the coverage or recall of *n*-grams between a candidate summary and a set of reference summaries, and it is calculated as follows:

$$ROUGE-N = \frac{\sum_{S \in \{\text{ReferencesSummaries}\}} \sum_{gram_n \in S} Count_{coincidence}(gram_n)}{\sum_{S \in \{\text{ReferenceSummaries}\}} \sum_{gram_n \in S} Count(gram_n)} \quad (4)$$

  where *n* is the length of the *n-gram* and *count* coincidence (gram) is the maximal number of n-grams that occur in the candidate summary and in the set of reference summaries.

- ROUGE-L (longest subsequence): a sequence $S = (s_1,s_2,\ldots,s_n)$ is a subsequence of other sequence $X = (x_1,x_2,\ldots,x_m)$, if there is a strict increasing sequence $(i_1,i_2,\ldots,i_k)$ indexes X so that for every $j = 1,2,...,k$, there is $x_{ij} = s_j$. . Given two sequences X and Y, the longest Common Subsequence (LCS) of X and Y is the common subsequence with the most length, SCL is applied in the assessment of the summaries, a sentence from the summary is seen as a sequence of words. Intuitively, the LCS of two sentences is the most similar of two summaries X and Y, where X is m in length and Y of n length, supposing X is a sentence of the summary and is a sentence from the candidate summary.
- ROUGE-W (longest weighed sentence): given two sequences X and Y, SCL is called weighed if the length is calculated using a weighing function. For further details on this function, see Lin (2004).

- ROUGE-S (non-contiguous bigrams co-occurrence): a non-contiguous bigram is any pair of words in the order of the sentence that allows for an arbitrary number of spaces. The cooccurrence of non-contiguous bigrams statistically measures the coverage of noncontiguous bigrams between the candidate summary and the set of reference summaries

Lin and Hovy (2003) indicated that this sort of measures may be applied in the assessment of the quality of automatically generated summaries, as they accomplished 95% of correlation between human elections.

For each metric of the ROUGE system, indicators of accuracy, recall and F-measure are obtained.

2.3.4 Pyramid method

The assessment method based on pyramids (Nenkova and Passonneau, 2004) was developed in Columbia University; it bases upon the observation that humans, while making a summary, do not always select the same elements. In order to apply it, the automatically generated summaries are fragmented into information units called Summarization Content Units (SCU) and similar segments are identified in the texts giving various weights to each information segment, according to the number of considered reference summaries (gold standard). Each single-layer SCU is assigned a weight that depends on the number of summaries in which it appears; this way, the SCU with the most importance are placed atop the pyramid. If it is the SCU number of a summary that appears at the level, then the weight of the summary D is calculated using the equation

$$D = \sum_{i=1}^{n} i \times D_i \tag{5}$$

This way, the best summary would be the one with the most SCU in higher positions. **Figure 2.1** illustrates an example of a possible assessment using a three-level pyramid.

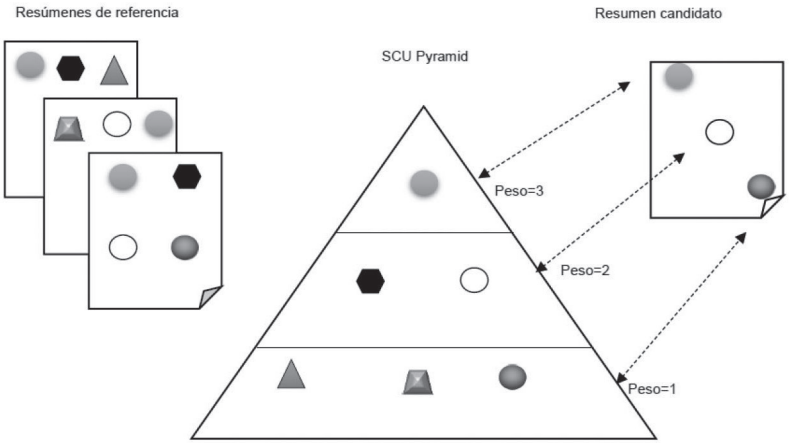


Figure 2.1 Summary assessment with the pyramid method (Nenkova and Passonneau, 2004)

Chapter III

Methods for Automatic Summary Generation

This chapter presents the two main types of summaries according to their condensation strategy, i.e., abstractive and extractive. There is a description of each, as well some novel scientific methods that work in an abstractive manner are mentioned. Moreover, a table displays the most important characteristics necessary for AGTS tasks, largely regarding extractive methods.

Among the methods proposed for AGTS we find the abstractive, which need a large number of linguistic resources (Miranda, 2013), (Lloret and Palomar, 2011), (Mateo *et al*., 2003), thereby they depend heavily on the language or need sophisticated processes in order to produce a summary. While extractive methods, which only utilize the structure and distribution of the original text are less dependent on language (Ledeneva, 2008), (Ledeneva *et al*., 2011), (Mihalcea and Tarau, 2005), (Last and Litvak, 2010), (García-Hernández and Ledeneva, 2013), (Mendoza *et al*., 2014), among others. The methods that work with a single language can produce better results than those that work with several languages. However, research on novel scientific methods has focused on the development of multi-language methods, owing to their broad application range and the exponential growth of information.

3.1 Abstractive methods

Abstractive AGTS methods base upon the comprehension and redrafting of the text. According to Plaza (2011), in the production of abstractive summaries, three stages are distinguished:

1. Construct a semantic representation of the document's sentences.
2. Perform selection, aggregation and generalization operations on these representations.
3. Finally, translate the representation into natural language.

Presently, there are few research works on this sort of summaries as they are complex and entail high computing costs required for their construction. Following, some abstractive methods are described.

3.1.1 Summons System

Summons (McKeown and Radev, 1995) is an AGTS method used in a number of documents. It creates templates for all the related articles with information from the news item; a clustering of the templates is carried out to find out the main topics, then these clusters go on to the generation stage to be combined. In this stage two steps take place:

1. A content planner generates the conceptual representation of the text meaning.
2. A linguistic component, which selects the right words to refer to the concepts contained in the information chosen; some other linguistic resources are used as well, namely: lexica, grammar books, ontologies, knowledge bases, among others. Finally, it makes use of all these resources to generate the summary.

3.1.2 Cut and paste

The Cut and Paste method (Jing, 2001) is applied on several domains for single documents. In the first stage, the most important sentences in the text are extracted; lexical relations between words are used for identifying them and statistical measures utilized in information retrieval are incorporated among some of the document's characteristics. In the second stage, two modules are the basis to reduce the extracted phrases:

1. Sentence reduction; confusing phrases in the extracted sentences are suppressed (using resources such as WordNet, syntactic knowledge, phrases created by humans). And
2. Combination of extracted sentences; resorts to the rules identified in man-made summaries.

3.1.3 Conceptual graphs

Miranda (2013) proposes an abstractive AGTS method based on conceptual graphs, which requires semantic information (it uses WordNet), plus verbal patterns (VerbNet) to be coherent and structure the graph. By and large, the method relies on weighing and pruning processes on the conceptual graphs' nodes (synthesis), supported on generalization and union operations. The weighing process bases on the structure and semantic flows of the weighed conceptual graphs (Miranda-Jiménez *et al*., 2013) and on Kleinberg's (1999) algorithm HITS,[5] (Mihalcea,

[5]HITS: (Hyperlinked Induced Topic Search) an iterative algorithm designed to classify web pages according to their "authority" degree. Moreover, it distinguishes between "authorities" (pages with a large number of inbound links) and "hubs" (pages with a large number of outbound links).

2004) to ascertain the importance of the nodes. The pruning process considers the weighing information of HITS algorithm and uses VerbNet's verbal patterns to keep coherence in the structures over the node suppression process. The graphs which remain after the operations are considered the representation of the summary at conceptual level.

Abstractive methods are difficult to build, besides they are used in specific languages because of the use of templates or dictionaries involved.

3.2 Extractive methods

Nowadays, extractive methods are the most researched owing to their low cost and easy computational setting up, adding to their usefulness for humans when they are looking for information quickly. Because of this, they show the most relevant information without changing that of the original text.

Extractive methods for AGTS consider the structure and distribution of the sentences to be able to select the most important. In **table 3.1**, the characteristics most utilized in the generation of extractive summaries are displayed. The main goal of these methods is to find out which of the sentences has a heavier weight, which indicates if it may belong to the summary or not.

According to Ledeneva (2017), general characteristics of extractive summaries are the following stages:

- **Term selection**. Over this stage it is decided which units will be terms, for example, they may be words, n-grams, sentences.
- **Term weighing**. It is a weighing process (or estimation) of the individual terms regarding the document content.
- **Sentence weighing**. It is the process to assign the sentence a numerical measure of usefulness. For instance, one of the ways to find out a sentence usefulness is adding the usefulness weights of the individual terms that compose the sentence.
- **Sentence selection**. Sentences and other units are selected as final parts of the summary. One of the easiest ways to accomplish it is to assign the sentences a numerical measure that reflects its usefulness in the original text and only choose the best ones to produce the summary.

The methods presented in this book for each language are extractive and are explained in the corresponding section of the language for which they were tried.

3.2.1 Language-independent Methods

Currently, English language is the most studied in AGTS tasks. This is because it is the most used on the Internet, therefore, the generation of *corpora*, competences and conferences has grown over the years.

Regularly, when we search for information on the Internet, we do it in a language we command and we expect to find information in this language, however this is not always the case, usually we should have more than one option in languages we may command.

Because of this, humans need to access information in their own language, novel scientific methods have focused on other languages and have enlarged the amount of research on AGTS applied to various of them.

An AGTS method independent from language, according to Plaza (2010), is one that having a text base in a certain language, generates the summary and later translates it into others. However, authors such as Patel (*et al*., 2007), (Mihalcea and Tarau, 2005), (Wang and Cardie, 2013) and (Last and Litvak, 2010) state that an AGTS method that works with various languages is one which having a collection of multiple-language documents (written in various languages) generates the summaries with a single tool. An important requirement for any method that works with several languages is to demonstrate a similar functioning in some of them with no special adaptions such as algorithm modifications or additional data for each language.

If there is need to produce summaries in various languages the process becomes complex, since their characteristics are different. Albeit, if statistical methods (extractive) are employed, problems can be simplified. Some methods independent from language are described below.

A language-independent approach to multilingual text summarization

The method proposed by Patel (2007) is an extractive one, independent from language and for a single document. It bases on a genetic algorithm which

considers the structural and statistical factors to generate summaries in English, Hindi, Gujarati and Urdu. To weigh the sentences, the sentences' information content, reference index and the location of the characteristics are used. For English, it was verified using *corpus* DUC02. For the other languages, news items are used.

Essential Summarizer

It is a language-independent method able to work in twenty different languages. It bases on the statistical analysis of a text and to produce the summaries, it resorts to techniques such as semantic signal recognition, domain specialization and considering expressions or concepts important for the user (Lehmam, 2010).

Using a Keyness Metric for Single and Multi-Document Summarization

It is an AGTS method for one and multiple documents in English and Arabic. It bases on the frequency of words and probability calculations. The method has two initial stages: Log_Likelihood[6] calculation, and summary production following Log_Likelihood results (El-Haj and Rayson, 2013).

There are novel scientific methods that claim to be independent from language, though they only try a selection of documents, regularly in English (García-Hernández and Ledeneva, 2013), (Ledeneva *et al*., 2011), (Ledeneva and García-Hernández, 2013), (Mendoza *et al*., 2014), among others.

Table 3.2 is the result of an investigation on the main novel scientific methods; its goal is to present the most important parameters such as language they use, collections with which they verify, if it is for one or multiple documents and if it is abstractive or extractive.

[6]It is a function of the parameters of a statistical model.

Table 3.1 Text characteristics used by AGTS

No.	Text characteristics / Reference	(Mendoza *et al.*, 2014)	(Bossard *et al.*, 2008)	(Ouyang *et al.*, 2010)	(Nandhini and Balasundaram, 2014)	(Lin, 1999)	(Hirao *et al.*, 2002)	(Katragadda *et al.*, 2009)	(Uddin *and* Khan, 2007)	(Oräsan, 2003)	(Berker, 2011)	(Alfonseca and Rodríguez, 2003)	(Suanmali *et al.*, 2011)	(Qazvinian *et al.*, 2008)	(Mateo *et al.*, 2003)	(Babar *and* Patil, 2015)	(Kiyoumarsi, 2015)	Total
1	Sentences' position	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	✓	✓	14
2	Sentences' length	✓	✓		✓	✓	✓		✓		✓	✓	✓			✓	✓	11
3	Relation of the sentence to the title	✓	✓		✓	✓			✓	✓			✓	✓	✓	✓	✓	11
4	Topic (frequency) – coverage	✓				✓			✓		✓		✓		✓	✓		7
5	Proper nouns				✓	✓							✓		✓	✓	✓	6
6	Numeric data					✓			✓		✓		✓			✓		5
7	Centrality		✓		✓						✓		✓				✓	5
8	Similarity with a search		✓			✓						✓			✓			4
9	Reference phrases				✓					✓	✓						✓	4
10	Cohesion/Similarity	✓												✓			✓	3
11	Similarity with fragments		✓			✓								✓				3
12	Trigger words				✓		✓								✓			3
13	Entities' names						✓				✓							2
14	Term's weight															✓	✓	2
15	Sentiment		✓															1
16	Similarity with the first sentence		✓															1

(*continuation*)

No.	Text characteristics/ Reference	(Mendoza *et al.*, 2014)	(Bossard *et al.*, 2008)	(Ouyang *et al.*, 2010)	(Nandhini and Balasundaram, 2014)	(Lin, 1999)	(Hirao *et al.*, 2002)	(Katragadda *et al.*, 2009)	(Uddin *and* Khan, 2007)	(Orăsan, 2003)	(Berker, 2011)	(Alfonseca and Rodríguez, 2003)	(Suanmali *et al.*, 2011)	(Qazvinian *et al.*, 2008)	(Mateo *et al.*, 2003)	(Babar *and* Patil, 2015)	(Kiyoumarsi, 2015)	Total
17	Word length				✓													1
18	Polysyllable words				✓													1
19	Noun occurence				✓													1
20	Pronoun and adjective					✓												1
21	Day of the week and month					✓												1
22	Quotation					✓												1
23	Text typography														✓			1
24	Similarity of sentence with sentenceoración															✓		1
25	Indicator of main concepts																✓	1
26	Occurrence of non-essential information																✓	1

Table 3.2 Novel scientific methods for AGTS

Method	Collection	Language	Assessment	Documents	Sort
(Mihalcea, 2004)	DUC02	English	*ROUGE*	Single	Extractives
(García-Hernández and Ledeneva, 2013)	DUC02	English	*ROUGE*	Single	Extractives
(Mendoza *et al.*, 2014)	DUC02, DUC01	English	*ROUGE*	Single	Extractive
(Mendoza Becerra, 2015)	DUC01, DUC02, DUC05, DUC06	English	*ROUGE*	Single and multiple	Extractive
(Ledeneva *and* García-Hernández, 2017)	DUC02	English	*ROUGE*	Single	Extractive
(Krishna *and* Reddy, 2016)	DUC02	English	*ROUGE*	Single	Extractive
(Igave *and* Gaikwad, 2016)	DUC05	English	*ROUGE*	Multiple	Extractive
(Wang *et al.*, 2017)	DUC04	English	*ROUGE*	Multiple	Extractive
(Al Saied *et al.*, 2017)	DUC07	English	*ROUGE*	Single	Extractive
(Lynn *et al.*, 2017)	600 noticias de los periódicos CNN, BBC, UK, TechCrunch and New York Times	English	*ROUGE*	Single	Extractive
(Bhargava *et al.*, 2016)	50 documents from DUC and 51 from Amazon	English	*ROUGE*	Single	Abstractive
(Bing *et al.*, 2015)	TAC 2011	English	*ROUGE*	Multiple	Abstractive
(Miranda-Jiménez *et al.*, 2013)	DUC03	English	Accuracy, recall and *F-measure*	Single	Abstractive
(Genest *and* Lapalme, 2011)	TAC 2010	English	*Pyramid*	Multiple	Abstractive
(Khan *et al.*, 2018)	DUC03	English	*ROUGE*	Multiple	Abstractive
(Mihalcea *and* Tarau, 2005)	DUC02, TeMário	English and Portuguese	*ROUGE*	Single	Extractive
(Patel *et al.*, 2007)	DUC02, Hindi	English, Hindú, Gujarati and Urdu	Intrinsic assessment	Single	Extractive
(Villatoro E., 2007)	Desastres, CAST	Spanish, English	*ROUGE*	Multiple	Extractive
(Last *and* Litvak, 2010)	DUC02, Hebrew	English and Hebrew	*ROUGE*	Single	Extractives
(Saggion, 2011)	TAC multilingual 2011	Arabic, english, french and hindi	*ROUGE*	Multiple	Extractive
(Last *and* Litvak, 2010)	DUC02, Hebrew	English, hebrew and arabic	*ROUGE*	Single	Extractive
(El-Haj *and* Rayson, 2013)	Multiling 2013	English, Arabic	*ROUGE, AutoSumm-ENG, MeMoG* and *NPowER*	Single and multiple	Extractive
(Mingli *et al.*, 2016)	200 documents on technology	Chinese	*ROUGE*	Single	Extractive

Matias-Ledeneva-García

Chapter IV

Tools to generate automatic summaries

This chapter deals with the analysis of AGTS commercial tools. The methods used by each and the steps they carry in their functioning are described in order to have an overview of these tools' quality regarding the heuristics and novel scientific methods.

Commercial tools are those available for use either online (Internet) or downloadable to install on a computer. The method each of them follows is not commonly published since its internal functioning is not public domain and the tool has a price. Commercial tools are classified in downloadable and online. In order to work, the latter have to be installed in a computer, whereas the former can be accessed through any computer with an Internet connection.

The main objective of this chapter is to present commercial downloadable and online AGTS tools, besides their functioning is explained.

4.1 Downloadable tools

4.1.1 Copernic Summarizer

This tool was exclusively developed for AGTS, it is flexible and suitable for this task, as it offers different options for the length of the summary to produce, among which there are: five, ten, 25 and 50 percent of the number of words in the original text, which may have an extension of 100, 250 and 1000 words.

According to Copernic Summarization-Technologies White Paper (2003), Copernic Summarizer uses the following methods:

1. Statistical model (S-Model); used for finding the text's vocabulary.
2. Knowledge-intensive processes (K-Process); they consider the way humans make text summaries, going through the following stages:
   a) **Language detection**. The language of the text (english, german, french or spanish) is identified in order to apply specific processes.
   b) Sentence limit recognition and "tokenization"; it runs various heuristics such as the identification of lists with vignettes and special chains (email and scientific formulae) to isolate sentences.
   c) Concept extraction. Copernic Summarizer utilizes automatic learning techniques to extract keywords.
   d) Document segmentation. It organizes the information that can be divided into longer related segments.

e) Sentence selection. The sentences are selected following their importance (weight), discarding those that decrease legibility and coherence.

Below, the steps necessary to produce a summary with Copernic Summarizer are presented in **figure 4.1**.

1. The text to summarize is pasted.
2. Then, the desired option for summary length is selected.
3. Automatically, the tool produces the summary, which may be printed, saved and/or sent.

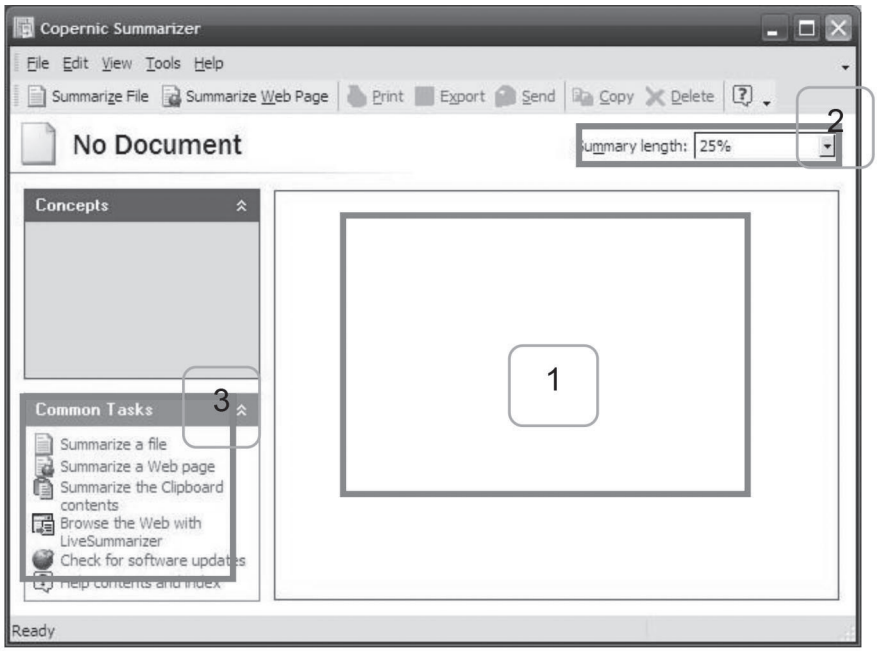


Figure 4.1 Copernic Summarizer interface

4.1.2 Microsoft Office Word Summarizer

Microsoft Office Word is a word processor that enables the creation of documents that can contain images, graphs, charts, tables and endless objects that make documents more attractive (Villar, 2005); it features an AGTS option in its versions 2003 and 2007. The summary produced by Microsoft Office Word is the result of key-word analysis and the selection of the most frequent in the document. The sentences that comprise these words are included in the summary; likewise, this program produces summaries of 10 or 20 sentences; 100 or 500 words (or fewer); or else, in percentages of 10, 25, 50 and 75 percent of words in the original text. If some of the percentages are not suitable, users may change the values according to their needs. Following, the way summaries are produced in the 2003 and 2007 Microsoft Office Word versions is described.

- ***Microsoft Office Word 2003***

Figure 4.2 displays the steps followed by Microsoft Office Word 2003 (in Spanish) for AGTS tasks

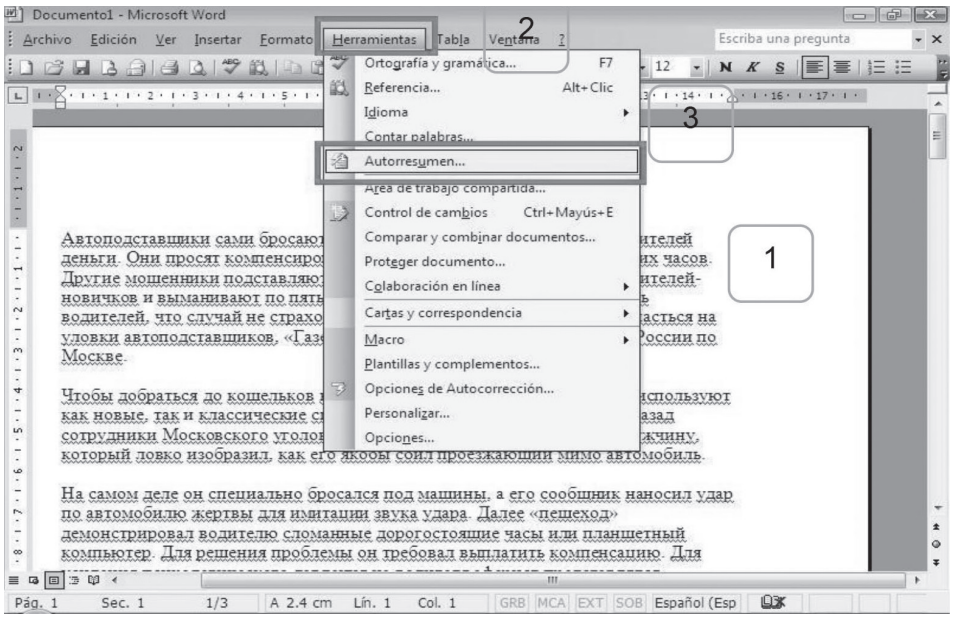


Figure 4.2 Interface to activate Autosummary (Spanish)

1. The text to summarize is pasted.
2. Click on Tools tab.
3. Select the option Autorresumen [Autosummary].

Automatically, a window presents the parameters for the user to select the summary options according to their needs.

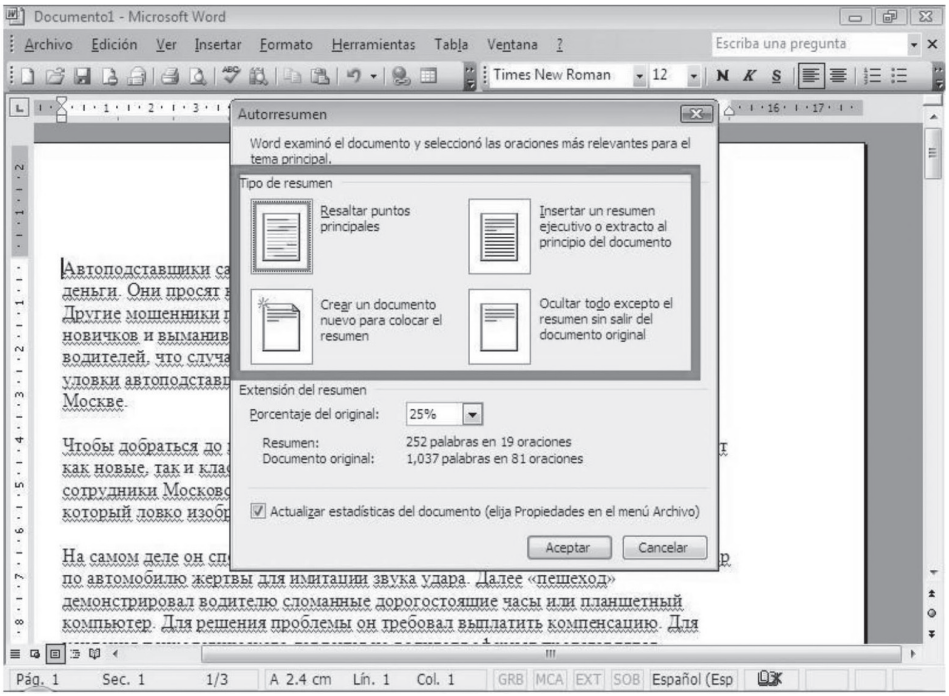


Figure 4.3 Interface to select Sort of summary (Spanish version)

As observed in **figure 4.3**, the first parameter that can be modified is the way the summary is presented: highlight the main points; insert an executive summary or an abstract at the beginning of the document; create a new document to paste the summary; and, hide everything except the summary without closing the original document.

The second parameter is the extension of the summary. The options for users to select are the required size of the text which can be number of sentences (10

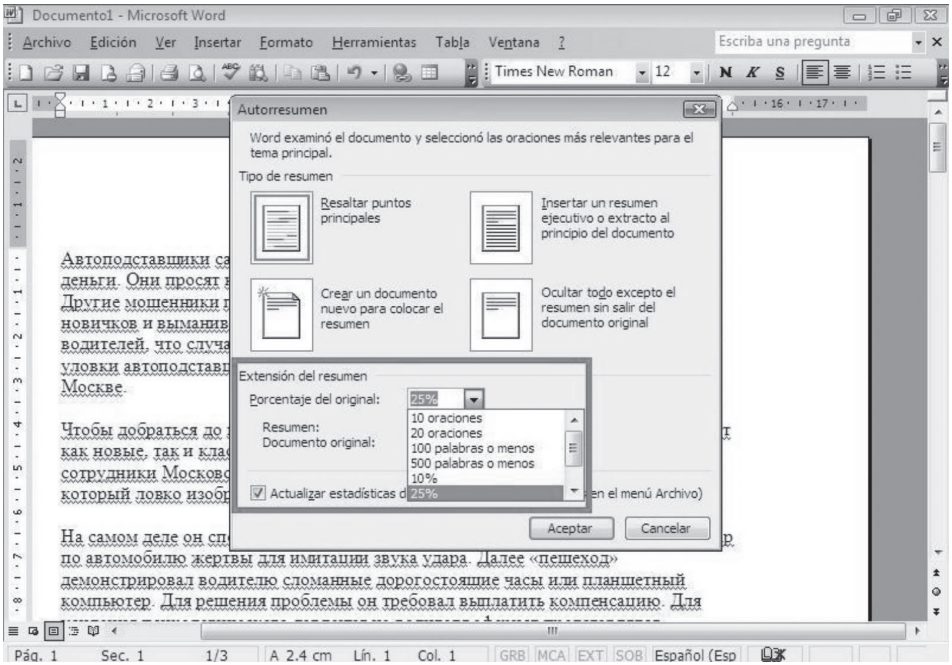


Figure 4.4 Interface to select the extension of the summary (Spanish version)

or 20), words (100 or fewer; 500 or fewer), percentage of words (10, 25 percent) (**Figure 4.4**). Finally, click on Accept to produce the automatic summary.

- ***Microsoft Office Word 2007***

AGTS in Microsoft Office Word 2007 is similar to that of version 2003. However, the Autosummary option has to be enabled. Below, the steps to enable this option are shown.

1. The Office button is selected (top left corner).
2. Options is selected (**figure 4.5**).
3. Customize is selected (upper-middle right side, **figure 4.6**).
4. Locate Command.
5. Select the option All commands.
6. Select Autosummary...
7. Add is selected >>

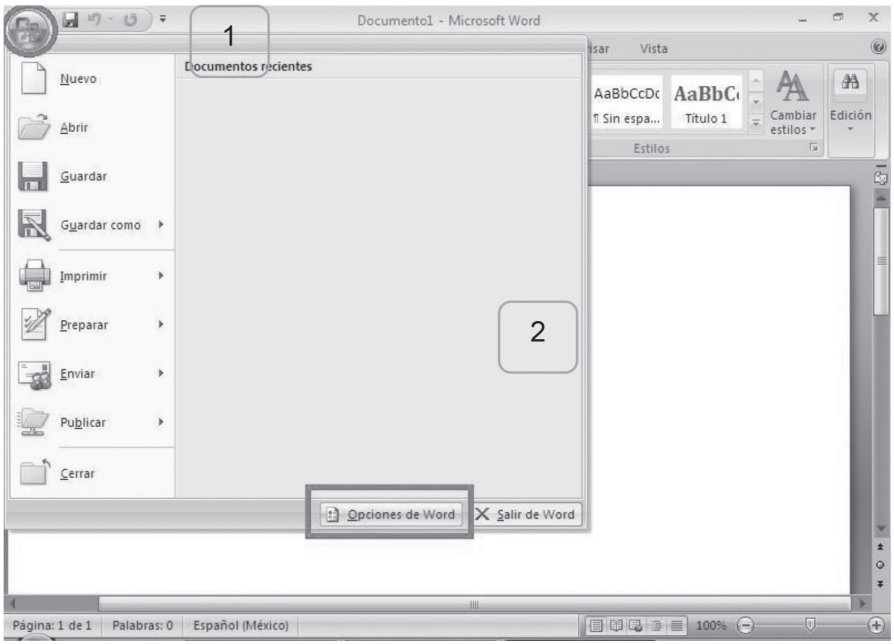


Figure 4.5 Interface to select Autosummary (Spanish version)

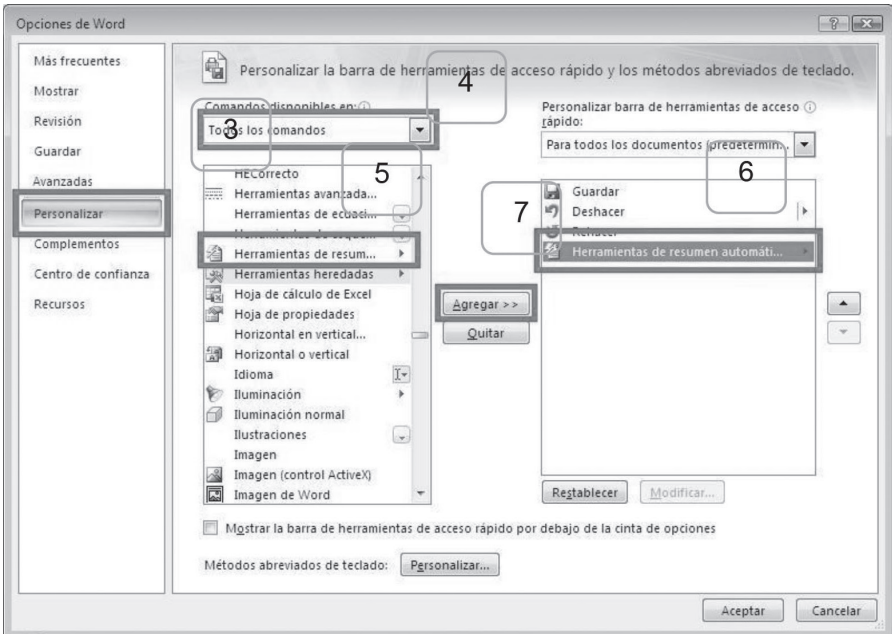


Figure 4.6 Interface to activate Autosummary (Spanish version)

Once the Autosummary option is enabled, it appears on the toolbar (**figure 4.7**), through which the option widow to produce a summary can be accessed.

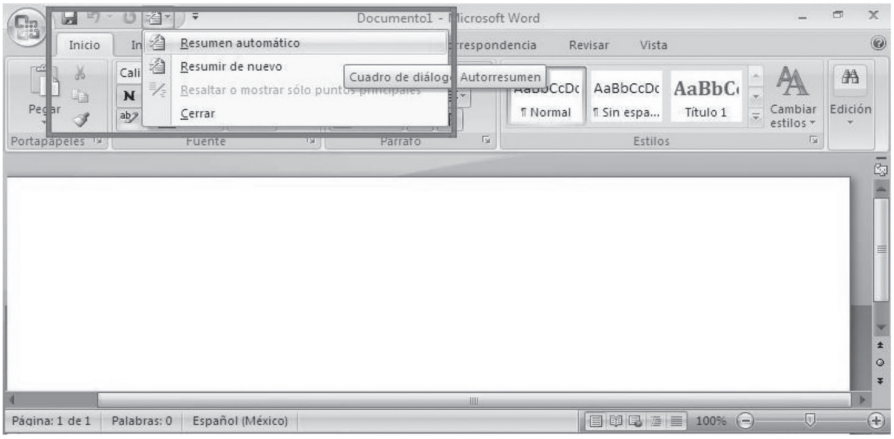


Figure 4.7 Button to activate Autosummary (Spanish version)

4.2 Online tools

4.2.1 SweSum

SweSum (Hassel and Dalianis, 2003) was the first program to summarize texts in Swedish; though at present, it works in English, Spanish, French, Norwegian, Italian, Danish, Greek, Farsi (Persian) and German. The aspects utilized to value the sentences are position and numerical values.

The SweSum process to produce summaries comprises three stages:

1. "Tokenization", text fragmentation in sentences and keyword extraction are carried out.
2. A ranking of the most frequent sentences is produced.
3. The text is summarized.

Below, an example of summary production with SweSum is presented (**figure 4.8**).

1. The text to translate is typed or pasted in the box, or else, a file is selected from the computer.
2. The sort of text is chosen (academic or newspaper).
3. Language is chosen.
4. The percentage or number of desired words in the summary are selected.
5. Finally, click on Summarize to produce the summary.

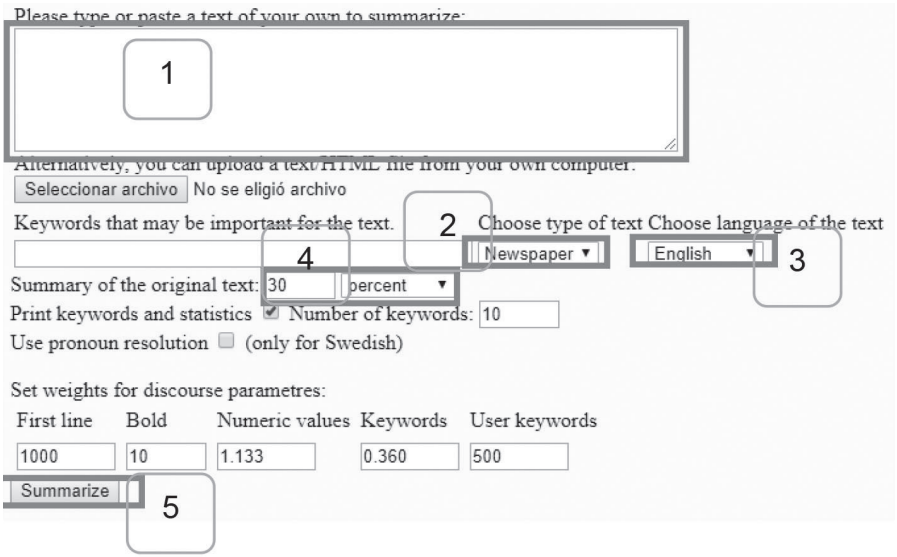


Figure 4.8 SweSum[7]

4.2.2 T-Conspectus

It is an online application to summarize articles in English, German and Russian within the news item domain.

The summary generator uses some techniques for natural language processing to automatically extract the most informative phrases from a text without format inserted into the textbox and loaded by the user or inserted from an URL.

[7]Swedish project for online summarization, available at: http://swesum.nada.k7th.se/index-eng.html

It uses a processing algorithm that contemplates a three-stage process

1. The first one is preprocessing, which comprises four main operations:
   a) Title: if the text to summarize has a title, it will be used to assign additional weights to keywords (it is advisable to introduce texts with a title).
   b) It divides the text in paragraphs: the summary generator needs to find the ends of the paragraphs to find the first and last sentences and are scored based on positions.
   c) It divides paragraphs into sentences: this operation divides into two substages; the first is the initial decomposition of the paragraph, then a correction after the separation of paragraphs into sentences.
   d) "Tokenization" of sentences: each sentence is divided into words.
2. The second stage is the scoring of the summaries by means of weighing the terms and sentences. It produces a list in a table that contains the sentences with their corresponding weights.
3. The third stage is the generation of the summary; it selects an "n" number from the first phrases in the list mentioned above. The number of sentences to be selected for the final summary is calculated in function of the user's requirements.

Once the way T-Conspectus generates summaries is explained, the steps to produce a summary with this tool are described (**figure 4.9**):

1. The text to summarize is written or pasted; though, a text from a URL or from the computer files can be summarized.
2. Using a percentage value (from 5 to 70, increasing in five-point steps), the desired summary size is defined.
3. By means of these options, the summary's keywords and statistics are displayed.
4. The button to produce the summary is Summarize, which only activates if there is text in the box.
5. The button Remove Text suppresses the text or URL; this button only activates if the corresponding boxes contain text.

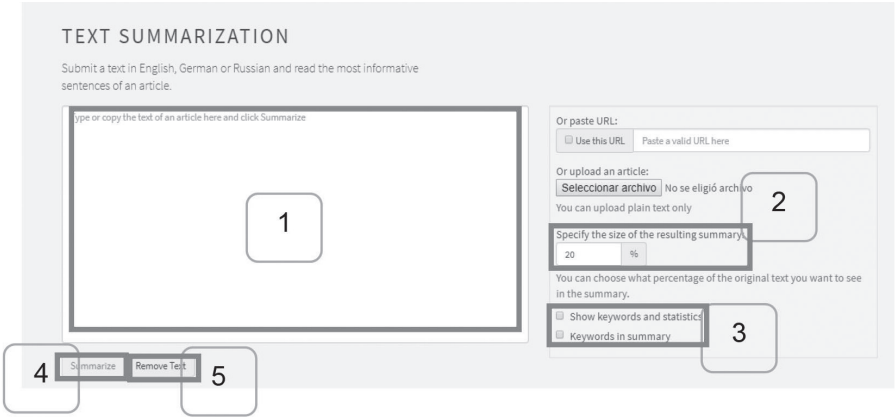


Figure 4.9 T-Conspectus [8] interface

4.2.3 Open Text Summarizer (OTS)

Open Text Summarizer[9] is an open-source application to summarize texts. It can be downloaded for free from the Internet, though it also has an online interface. OTS automatically analyzes the texts and intends to identify the most important parts; it works in various languages: English, German, Spanish, Russian, Hebrew and other twenty-five languages.

Below, the steps to produce a summary with Open Text Summarizer are described (**figure 4.10**):

1. The text to summarize is typed or pasted. Though, a text from a URL can also be summarized.
2. An output format for the information is selected, there may be two: general or keywords.
3. The output text size is chosen.
4. The language of the summary is selected.
5. Click on Send to produce the summary.

[8]A web application to summarize news articles in English, German and Russian available at: http://tconspectus. pythonanywhere.com/

[9]This is web interface to produce summaries. The tool automatically analyzes the texts in various languages and intends to identify the most important parts of the text; available at: https://www.splitbrain.org/ services/ots

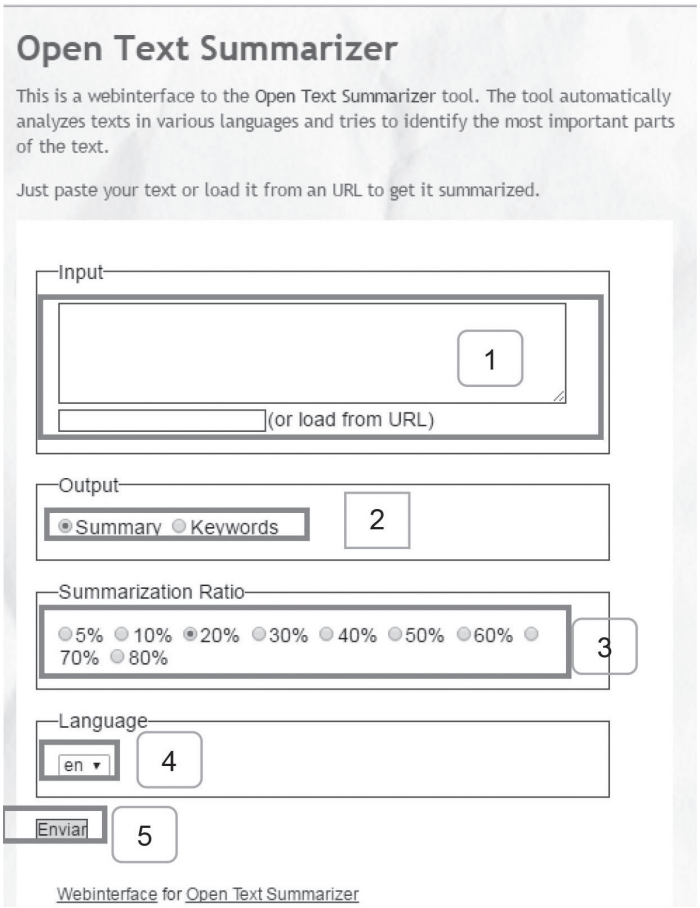


Figure 4.10 Open Text Summarizer interface

4.2.4 Text Compactor

This AGTS tool is free and available online. It was created to aid students, professors and occupations that needed to process large amounts of information. The stages of Text Compactor to produce a summary are the following:

1. The frequency of each word in the text is calculated.
2. The score is calculated for each phrase on the basis of the associated frequency of the words it contains.
3. The most important phrase is the one with the most frequency.

It is worth mentioning that this tool to generate summaries works better on textbooks and reference materials; it does not function in the same way with fiction texts (that is to say, stories on imaginary characters, places and events).

Following, each of the steps to produce a summary with *Text Compactor* are described (**figure 4.11**):

1. The text to summarize is typed or pasted.
2. The summary size is defined from 0 to 100%.
3. The summary is presented.

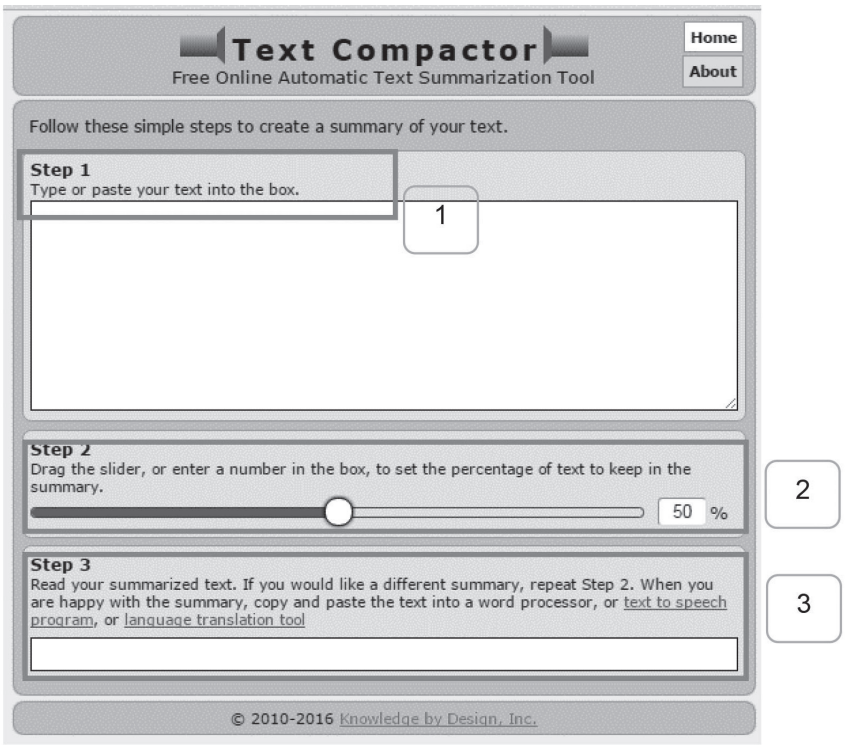


Figure 4.11 Text Compactor interface

4.2.5 Summarizing

Summarizing[10] is an AGTS online tool for articles. The stages it comprises are based on the detection of the texts' main ideas, on obtaining a description of them, which reflects the author's writing style, to finally produce the summary. This tool has parameters to produce summaries of 100, 150, 200 and 300 words. The trial version is presented.

The steps to produce a summary with this tool are presented below (**figure 4.12**).

1. The text to summarize is typed or pasted.
2. The number of words intended for the summary is chosen.
3. Click on Summarize to produce the summary.

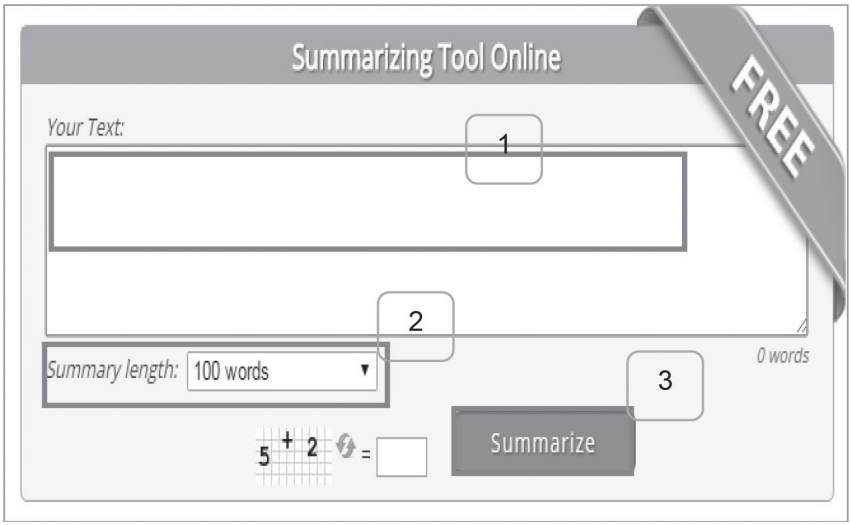


Figure 4.12 Summarizing Online Interface

[10]It is an online tool to produce summaries, available as a trial version at: https://www. summarizing.biz/best-summarizing-strategies/article-summarizer-online/. The full version can be purchased.

4.2.6 Summarizer

Summarizer[11] is a tool that enables us to produce automatic summaries. It is available as an Intellexer API component or as a desktop application. It receives an original document, extracts the text unformatted, works with the syntactic and semantic processing, extracts the information to produce the summary and assigns a score determined by sentences. This score defines the sentence importance in relation to the text. Summarizer produces summaries within a range of percentages (1 to 99).

Following, the steps to produce a summary with this online application are presented (**figure 4.13**).

1. The text or URL to summarize are pasted.
2. Then, an option is selected: percentage or paragraph; and the corresponding box is activated.
3. The percentage desired to summarize the text is entered (1-99%).
4. To finish, click on Summarize.

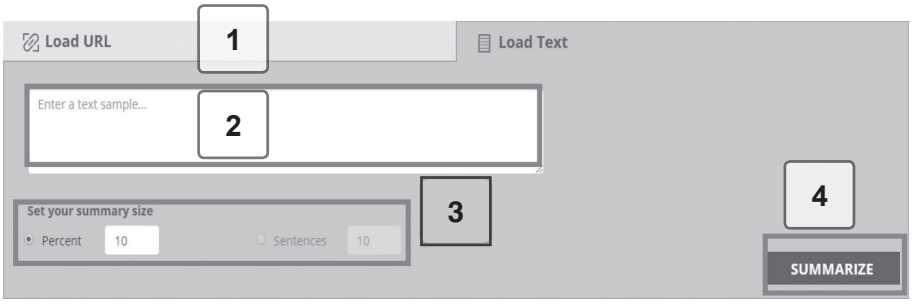


Figura 4.13 Summarizer online interface

The interface of Summarizer's downloadable version is shown below (**figure 4.14**).

[11]It is an online tool to produce summaries included in Intellexer's natural language processing tools, developed by EffectiveSoft. It can be accessed at: http://esapi.intellexer.com/Summarizer

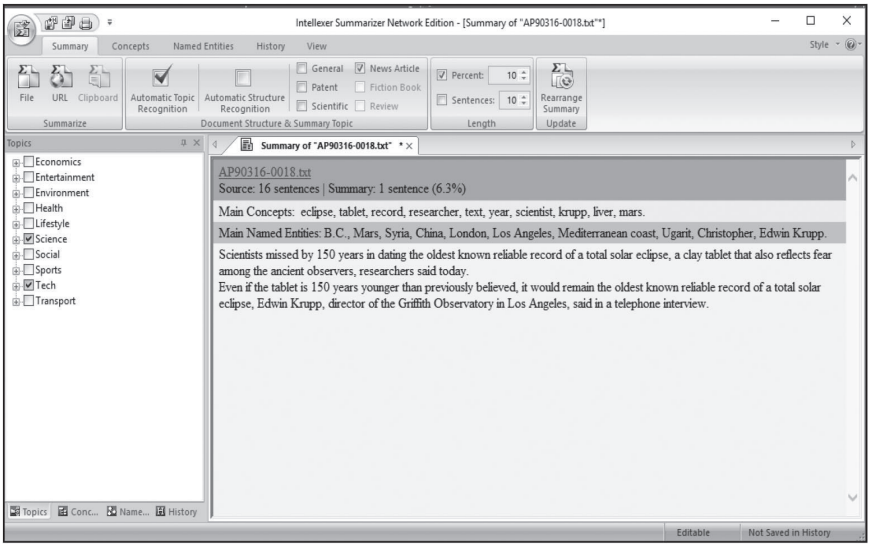


Figure 4.14 Summarizer's downloadable version interface

4.2.7 Tools4noobs

Tools4noobs[12] is an online tool that automatically creates a text summary (generally large), either pasting a text or the URL of the web page to summarize, the tool will produce a brief summary (**figure 4.15**). The process carried out by this tool comprises three stages:

a) Extract sentences.
b) Identify keywords and determine the relevance of each one.
c) Identify the phrases with most of the keywords.

The steps to produce a summary with Tools4noobs are presented below (**figure 4.15**).

1. The text to summarize is typed or pasted; it also works with URLs.

[12]An online tool to produce summaries available for free at: https://www.tools4noobs.com/summarize//

2. Threshold is defined, this is to say, the summary size; moreover, the number of lines, number of characters and minimal number of words in a sentence can be defined for the summary.
3. Tools4noobs allows selecting several options to visualize the summary; among them: relevance in the sentence, stress the most relevant keywords, number of keywords, highlighting keywords, show the most distinguished phrases in the text.
4. Clicking on Summarize it! button produces the summary.

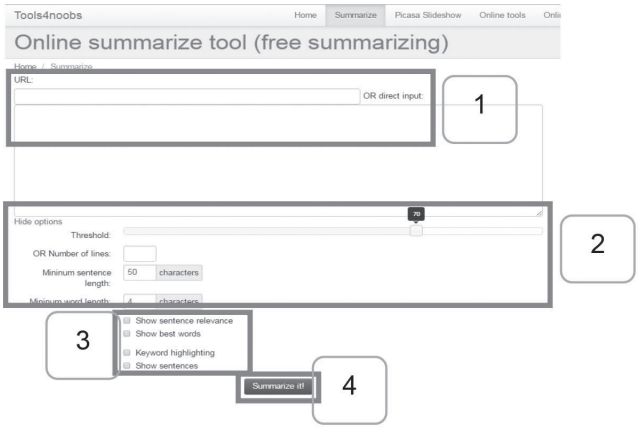


Figure 4.15 Tools4noobs online interface

4.2.8 Pertinence Summarizer

Pertinence Summarizer[13] belongs to the range of products developed with technology called KENiA© (Knowledge Extraction and Notification Architecture), which was devised by French firm *Pertinence Mining*. Pertinence Summarizer is an online tool that produces summaries in twelve languages (German, English, Arabic, Chinese, Korean, Spanish, French, Italian, Japanese, Portuguese, Russian and Dutch) from documents in various formats (html, pdf, doc, rtf and txt).

[13]An online tool that automatically generated summaries; not available anymore. However, the link to access the version tried in this work is http://pertinence. net/index_en.html

Percentages and number of words are automatically defined [1% (34 words), 5% (171 words), …and n% (n words)].

There are three ways to introduce the summary: copy-pasting the text; opening the document from its origin; introducing the URL of the webpage to summarize (**figure 4.16**).

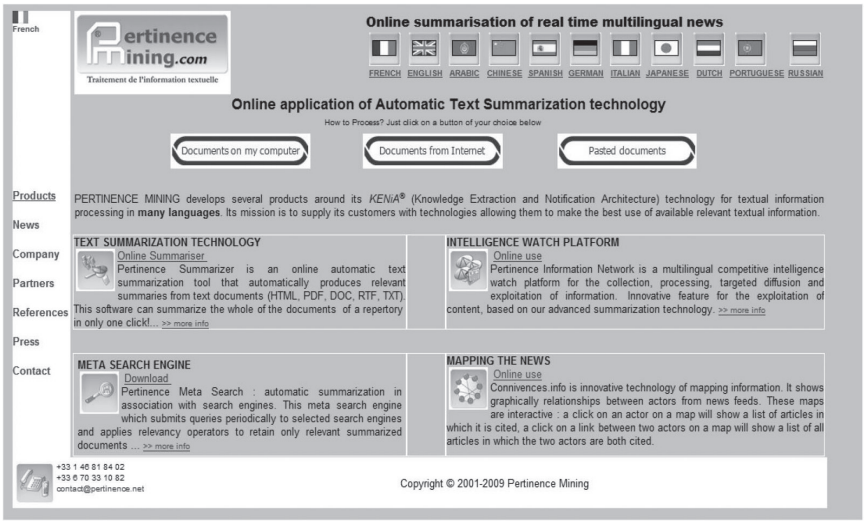


Figure 4.16 Pertinence Summarizer Interface

4.2.9 Shvoong

Shvoong[14] was created by Avi Shaked and Avner Avrahami in 2005; it is a tool that allows producing automatic summaries in twenty-one different languages. Unlike other tools, Shvoong does not deliver the summary as such, it underlines the text it considers most important in the original document though.

The steps to produce a summary with online tool Shvoong (**figure 4.17**) are:

[14]It is an online tool to produce summaries available at: http://es.shvoong.com/summarizer/

1. The text is typed or pasted.
2. Language is selected.
3. The percentage of the summary is defined.
4. Click on Summarize! to produce the summary.

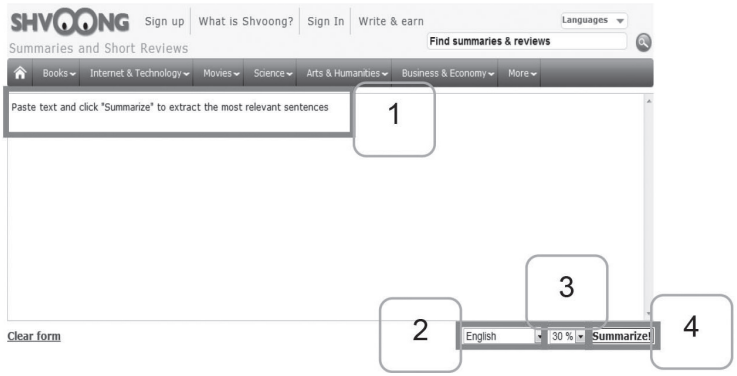


Figure 4.17 Shvoong interface

4.2.10 Resumo

Resumo[15] is a summary generator for multilanguage texts, the interface is in Portuguese; it is displayed in **figure 4.18**.

Following, the steps to produce a summary with this tool are presented

1. A text to summarize is inserted.
2. It is decided if it is a summary by number of lines or text percentage.
3. The language of the text to summarize is selected.
4. Click on "Fazer Resumo" to produce the summary.

[15]It is an online tool to generate summaries, available as a trial version; however, if large texts are to be summarized the full version must be purchased. Available at: https://www.turbinetext.com/Resumo/

Figure 4.18 Resumo Interface

4.2.11 BigdataSummarizer

BigdataSummarizer[16] is a tool that produces summaries in twenty-one languages such as Chinese, English, French, German, Italian, Russian, Spanish, etc. It works at 1, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100 percent of the original text (**figure 4.19**).

Below, the steps to produce a summary with this tool are presented:

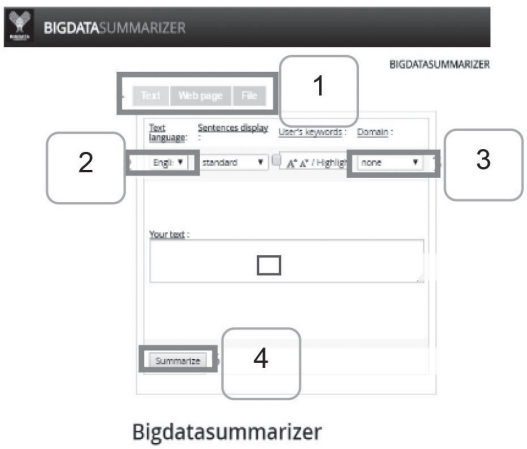


Figure 4.19 BigdataSummarizer interface

[16]It is an online tool to generate summaries especially in Russian, available at: https://bigdatasummarizer.com/summarizer/online/advanced.jsp?ui.lang=es

1. The text to summarize are typed or pasted; it is also possible to use a URL.
2. The language of the original text is chosen.
3. The text domain is set.
4. Click on Summarize to produce the summary.

4.3 Summary of tools tried in various languages

In **table 4.1** there is a listing of the downloadable and online commercial AGTS tools previously described; though, they are listed in relation to the languages in which they were tried. Results are presented in the following chapters.

Table 4.1 Tools tried in various languages

Tool	Sort	English	Spanish	Portuguese	Russian
Copernic Summarizer	Downloadable	✓	✓		
Microsoft Office Word 2003/2007	Downloadable	✓	✓		✓
SweSum	Online	✓			
T-Conspectus	Online	✓			✓
OTS	Online	✓	✓	✓	✓
Text Compactor	Online	✓	✓		✓
Summarizing	Online	✓	✓		
Summarizer	Online	✓			
Tools4noobs	Online	✓			✓
Pertinence Summarizer	Online	✓			
Shvoong	Online	✓		✓	
Resumo	Online				✓
BigdataSummarizer	Online				✓
Total		11	5	2	7

Chapter V

Automatic Summary Generation in English

This chapter presents in detail the study of AGTS for the English language. *Corpora* DUC01 and DUC02, used in the trials, are described. As well the results of the main heuristics, main commercial tools and novel scientific methods tried in this language are shown and described. Finally, there is a general comparison of heuristics, novel scientific methods and commercial tools tried with *corpora* DUC01 and DUC02.

English is the third most spoken language in the world with 339 million native speakers (Arévalo, 2017) (**table 5.1**).

Table 5.1 Languages with the most speakers in the world

No.	Language	Countries	Speakers
1	Chinese	35	1302
2	Spanish	21	427
3	English	106	339
4	Arabic	58	267
5	Hindi	4	260
6	Portuguese	12	202
7	Bengali	4	189
8	Russian	17	171
9	Japanese	2	128
10	Lahndi	8	117
11	Javanese	3	84.3
12	Korean	7	77.3
13	German	26	76.9
14	French	53	75.9
15	Telugu	2	74.2
16	Marathi	1	71.4
17	Turkish	8	71.4
18	Urdu	6	68.6
19	Vietnamese	3	68
20	Tamil	7	67.8
21	Italian	13	63.4
22	Persian	30	61

However, despite being in the third place, it is spoken in 106 countries. This way, English holds the first place in the most used on the web (**table 5.2**).

Table 5.2 10 languages most used on Internet[17]

No.	Language	Internet users
1	English	1052
2	Chinese	804
3	Spanish	337
4	Arabic	219
5	Portuguese	169
6	Hindi	168
7	French	134
8	Japanese	118
9	Russian	109
10	German	92
	Other	950

AGTS is sixty years in research, its study began in the 1950's decade with Luhn's work in 1958 (Luhn, 1958), who was the first to produce automatic extractive summaries. Later on, the study of this task continued with works by Edmundson (1969), Kupiec (*et al*., 1995), Paice (1990), Jing (*et al*., 1998), Minel (*et al*., 1997), Barzilay, Elhadad (1999), Benbrahim and Ahmad (1995), Carbonell and Goldstein (1998), Marcu (1997), McKeown and Radev (1995), Mani (*et al*., 1999) and others. AGTS research up to the year 2000 was focused on the English language because of the existing resources (*corpora* and standard assessment measures) in this language. Over the years, there has been great advance in research regarding other languages. Because of this, research in English can be taken as a basis for other languages.

Over time, research in English for summary generation has produced various heuristics for this end, which serve as references to assess the methods and tools to produce summaries. However, it was not known how much these

[17]According to a study that reveals the languages most used on the Internet https://www.internetworldstats.com/stats7.htm

heuristics influenced on both the summary construction and human preferences as regards the election of the summary. To do so, in Chapter I, a Turing Test was carried out. In this test, two summaries made by humans, two automatic and two heuristics (*baseline:first* and *baseline:random*) were included.

As previously mentioned, the heuristic *baseline:random* proposes a random selection of sentences as a summary; this is to say, it is carried out without intelligence. This summary was considered with a view to finding out whether the human is capable to distinguish this sort of texts, even though this may pose a disadvantage for computer-made summaries. **Table 1.2** displays the results of the pair of summaries chosen as regards human-machine without distinguishing the heuristics. Though, it is necessary to make such distinction to find out the influence they pose to humans.

Table 5.3 presents the Turing Test results for the English language, but with the subdivision of the pairs of summaries considering human - machine (*baseline - first*), human - machine (*baseline - random*), machine (method) - machine (method), machine (method) - machine (*baseline - first*), machine (method) - machine (*baseline - random*) and *baseline:random - baseline:first*.

Table 5.3 Turing Test results for *baseline* in English

Pairs of summaries regarding *baseline* heuristics	Confusion percentage between the selected summaries (%)
Human – *Baseline:first*	24
Human – *Baseline:random*	15
Machine – *Machine*	10
Machine – *Baseline:first*	12
Machine – *Baseline:random*	12
Baseline: random – Baseline:first	27

Table 5.3 offers a panorama of the confusion caused by *baseline* heuristics. In the first row, it is shown that 24 percent of the times humans chose *baseline:first* heuristic as the man-made summary; whereas for *baseline:random* with human only 15 percent chose this combination. However, interestingly, 27 percent of the people thought that man-made summaries were the combination of the

heuristics. It is worth mentioning that for the English language, these heuristics were troublesome for humans; and this was not the case in Spanish (**table 5.3**).

5.1 Conferences, workshops and *corpora*

Early in 2000, the assessment program Document Understanding Conferences (DUC)[18] was created with the purpose of enabling AGTS research in English to carry out large-scale experiments. Text Analysis Conference (TAC)[19] continued with the work of DUC; it comprises a series of assessment workshops organized to further research on natural language processing and related applications. One of TAC's main goals is to gather collections of tests that evolve in order to anticipate the assessment needs of modern systems. It was in 2008, 2009, 2010, 2011 and 2014 when TAC focused on AGTS tasks, being user-oriented multi-document summary its main field of study.

5.1.1 Document Understanding Conferences (DUC)

In 2000, a new program to assess summaries was launched; initially sponsored by DARPA, it was a group of researchers on summaries, experts who proposed a workplan to build a *corpus* in order for the field of text summary systems to advance (Baldwin *et al*., 2000). The above was the guideline to create DUC, with a first pilot assessment in 2001. The plan demanded the assessment of single and multiple document summaries at the specific level of reading comprehension.

DUC was created mainly for intrinsic assessment; DUC01 and DUC02 underwent intrinsic assessment as well; while, for DUC03–DUC07 extrinsic assessment was performed.

For DUC, intrinsic assessment consisted in direct observations on both well-presented linguistic information and the degree at which an automatic summary expresses the same content as one manually created (using the same set of documents to summarize).

The main idea when DUC was devised was that it should be made with generic texts. The summaries had to be different in nature, considering

[18]Document Understanding Conferences (DUC), their objective is to enable AGTS researchers in English to run large-scale experiments. https://www-nlpir.nist.gov/projects/duc/index.html

[19]Text Analysis Conference (TAC), it is organized in a series of assessment workshops to improve system assessments. https://tac.nist.gov/

a newspaper reader with a considerable schooling level to make the gold standards.

Initially, for DUC01 and DUC02 the length considered was 50, 100, 200 and 400 for multiple documents and 100 for single documents, after years of research it was noticed that the summaries' quality did not depend on size. Though for short summaries it was noticed that a reference may be <= 75 bytes, as in the generation of these summaries no grammatical or linguistic method is applied. Moreover, having summaries of such length or shorter allow users of search engines to choose a large number of them faster.

When DUC was created, it was sought to focus on the creation of extractive summaries and it was expected that they quickly change from generating extractive summaries to abstractive; however, it has not been so, mainly in the creation of very short summaries.

In DUC 02 there are summaries made by humans which are not fully extractive; this means, never will there be a summary exactly the same as those by humans when a method that automatically generates extractive summaries is applied.

Newspaper articles are part of broad-interest literature available for many people in countries around the world. News has been part of the base research on various competences such as: TREC[20] (information retrieval), MUC[21] (information extraction), TDT[22] (detection and tracking of topics) and SUMMAC[23] (summary).

To a large extent, the election of items of news of the journals in DUC followed their availability and also the fact that this domain had been previously researched. The pyramidal structure of the newspaper's articles meant that

[20]Text REtrieval Conference (TREC). Available to foster research on information retrieval from large text collections; https://trec.nist.gov/

[21]Message Understanding Conferences (MUC). Its goal is to present information extraction and show its usefulness for Internet users and natural language processing researchers; https://www-nlpir. nist.gov/related_projects/muc/

[22]Topic Detection and Tracking (TDT). An initiative to research on the state of the art in searching and tracking new events in the flow of transmitted news items; https://ciir.cs.umass.edu/tdt

[23]TIPSTER Text Summarization Evaluation Conference (SUMMAC). Large-scale assessment, independent from the developer, automatic text summary systems. https://www-nlpir.nist.gov/related_projects/tipster_summac/

baseline simple systems, which produce summaries with the first sentences of an article and even a set of them, were difficult to overcome. In spite of the advance in research and passing of time, enquiries have not changed direction and this sort of texts is still studied. However, at present, efforts are also made on short summaries applied to social media.

Table 5.4 presents a brief description of the *corpora* offered by por DUC from 2001 to 2007.

Table 5.4 Description of DUC *corpora*

Corpus	**DUC01**	**DUC02**	**DUC03**	**DUC04**	**DUC05**	**DUC06**	**DUC07**
Folders	28	59	30	114	50	50	10
Files	309	567	624	1000	1600	1250	250
Automatic assessment				*ROUGE*	*ROUGE/BE*	*ROUGE/BE*	*ROUGE/BE*
Manual assessmentl	SEE	SEE	SEE	SEE	*Pyramid*	*Pyramid*	*Pyramid*
One document	X	X	X	X			
Multiple documents	X	X	X	X	X	X	X
Size One document	100 words	100 wordss	10 words	10 words			
Size Multiple documents	50, 100, 200, 400	10, 50, 100, 200, 400	*Viewpoint* (100) *Question /topic* (100) *Event* (100)	*Event* (100) *Who is question* (100)	*Complex question* (250)	*Complex question* (250)	*Complex question* (250)
Multiple documents summary approach							

5.1.2 Text Analysis Conference (TAC)

Text Analysis Conference (TAC) comprises a series of workshops organized to encourage research on Natural Language Processing (NLP) and related applications; it contains a large number of tests, this is to say, common assessment proceedings. TAC includes sets of tasks known as "clues", each of them focuses on a particular subproblem of NLP. TAC's clues are applied to the end user's tasks, but also include assessments of components within the context of such tasks.

TAC's goals are:

- Promote NLP research based on large collections of common texts.
- Improve the methodologies and assessment measures for PNL.
- Build up a series of text collections that evolve to anticipate the assessments needs of NLP modern systems.
- Enhance communication between industry, academia and government by means an open forum to interchange research ideas.
- Accelerate the transference of technol m research laboratories to commercial products, showing substantial improvements to PNL methodologies applied to real-world problems.

TAC came on top of DUC NIST to summarize texts from the answers to the question of conference TREC. In **table 5.5** there is a description of *corpora* TAC2008 – TAC 2011 and TAC2014.

5.1.3 Corpora to assess and compare

DUC01[24] is a *corpus* that contain news items on disasters in English; it was designed to produce summaries of single and multiple documents.

[24]To access the data in *corpus* DUC01 go to https://wwwnlpir.nist.gov/projects/duc/guidelines/2001.html

Table 5.5 Description of TAC *corpora*

Corpus	2008	2009	2010	2011	2014
Multiple documents	Updating summary	A, information summaries and B, updating summaries	Semantic analysis	A, information summaries and B, updating summaries	Summary based on quotations
Languages	English	English	English	English	English
Size	100 words	100 words	100 words	100 words	250 words
Domain	News	News	News	News	Medical articles
Content	48 topics with 20 relevant documents divided into two sets, A and B. Plus 4 reference summaries for each set of articles.	44 topics, each with 1 title and their narrations in 20 relevant documents divided into two sets, A and B.	46 topics, with 20 documents for each topic.	46 groups of documents; each contains 10 news articles, classified as one topic.	20 biomedical documents. Each one has 10 documents with references from the original document.

DUC01 comprises thirty sets of reference and thirty as tests; the latter comprise three hundred and nine documents. Each set contains the original documents as well as the summaries from single and multiple documents, manually made. This *corpus* is tagged, which allows having a clear sentence separation, thereby, better management of the information it contains. Moreover, it includes the results of various *baseline* measures. The performance of commercial tools and that of *Topline* were calculated with this *corpus*. For the present book, DOC01 is used in the automatic generation of summaries of single documents; this way, the number of required words for the summary is one hundred.

DUC02[25] is a *corpus* of English-language news items on various technologic, food, politics, finance topics, among others. It was designed for AGTS in function of two tasks: multiple and single documents. It contains five hundred and sixty-seven documents. Two 100-word summaries were produced by two human experts for each of them. Besides, it has the results of various *baselines* measures. DUC02 is one of the most resorted to by AGTS researchers. Since it is tagged, it allows having a clear separation of sentences. The performance of commercial tools and *Topline* were assessed with this *corpus*.

For this book, *corpus* DUC02 is used in the automatic generation of 100-word summaries for single documents.

5.2 Heuristics

In order to run the experiments and do the calculations for heuristics, commercial tools and novel scientific methods, DUC01 and DUC02 are used because they are the collections most used in AGTS tasks for single documents in English.

5.2.1 *Baselines random*

Ledeneva (2008) proposed to calculate *baseline:random* with an average assessment by selecting ten times at random sentences for a document. Following

[25]To access DCU02 data: https://www-nlpir.nist.gov/projects/duc/guidelines/2002.html

this methodology, García (2008) reports an *F-measure* of 0.38981 for *baseline: random* for *corpus* DUC02 assessed with ROUGE. Using Ledeneva's (2008) methodology to compare *baseline:random* on DUC01, Alvarado (2017) reports a value of 0.36587.

5.2.2 *Baseline:first*

García (2008) reports the *baseline:first* value for DUC02 assessed with *ROUGE* as 0.4729. Alvarado (2017) reports a value of 0.44862 for *baseline:first* in DUC01. This heuristic works as a reference for the methods and tools that work with the news domain, though there is no evidence of good results with other domains. It is worth mentioning that this heuristic was overcome ten years ago in the English language, i.e., fifty years after the start of AGTS research.

5.2.3 *Topline*

To calculate *Topline* for *corpus* DUC01, Rojas (2018) used three hundred and eight documents to generate various sentence combinations by means of a genetic algorithm. The parameters of the algorithm above to calculate *Topline* for DUC01 are presented in **table 5.6**.

Table 5.6 Parameters of the genetic algorithm for *Topline* on DUC01

Experiment	Elite	Generations	Individuals	Selection		Cross	Mutation	
				Sort	P		Sort	P
1	Yes	20	120	Tournament	3	CX	Insertion	8

The results obtained with *Topline* for DUC01 in the assessment, using ROUGE, are presented in **table 5.7**.

Table 5.7 *Topline* results for DUC01

Measure	Recall	Accuracy	*F-measure*
ROUGE-1	0.59796	0.59046	0.59408
ROUGE-2	0.33622	0.33234	0.33422
ROUGE-SU4	0.34619	0.34202	0.34404

For the calculation of *Topline* for DUC02, Rojas (2018) utilized five hundred and seven individual documents to produce the summaries, considering combinations of sentences by means of a genetic algorithm. The parameters used in such algorithm to calculate *Topline* for DUC02 are presented in **table 5.8**.

Table 5.8 Parameters of the genetic algorithm for *Topline* in DUC02

Experiment	Elite	Generations	Individuals	Selection		Cruza	Mutation	
				Sort	P		Sort	P
1	Si	30	150	Tournament	3	CX	Insertion	8

Results obtained from *Topline* for DUC02 in the assessment using ROUGE are presented **table 5.9.**

Table 5.9 Parameters of the genetic algorithm for *Topline* in DUC02

Measure	Recall	Accuracy	*F-measure*
ROUGE-1	0.62601	0.62164	0.62367
ROUGE-2	0.35877	0.35624	0.35742
ROUGE-SU4	0.36107	0.35851	0.35970

5.3 Commercial tools

The English language assessments of the two of the main *corpora*, DUC01 and DUC02, with *ROUGE* are presented. Below, **table 5.10** shows the commercial tools tried in each *corpus*.

Table 5.10 Tools assessed with DUC01 and DUC02

Tool	Sort	English	
		DUC01	DUC02
Copernic Summarizer	Downloadable	✓	✓
Microsoft Office Word 2003/2007	Downloadable	✓	✓
SweSum	Online	✓	
T-Conspectus	Online	✓	
OTS	Online	✓	✓
Text Compactor	Online	✓	
Article *Summarizing* Online	Online	✓	
Summarizer	Online	✓	
Tools4noobs	Online		✓
Pertinence Summarizer	Online		✓
Shvoong	Online		✓

It is worth mentioning that all the tools in **table 5.10** can be applied to DUC01 and DUC 02 as they work in English. However, the results for both collections are not shown because the generation of summaries with these tools is made document by document; this way, producing the summaries is time-consuming, plus some of them are not available anymore.

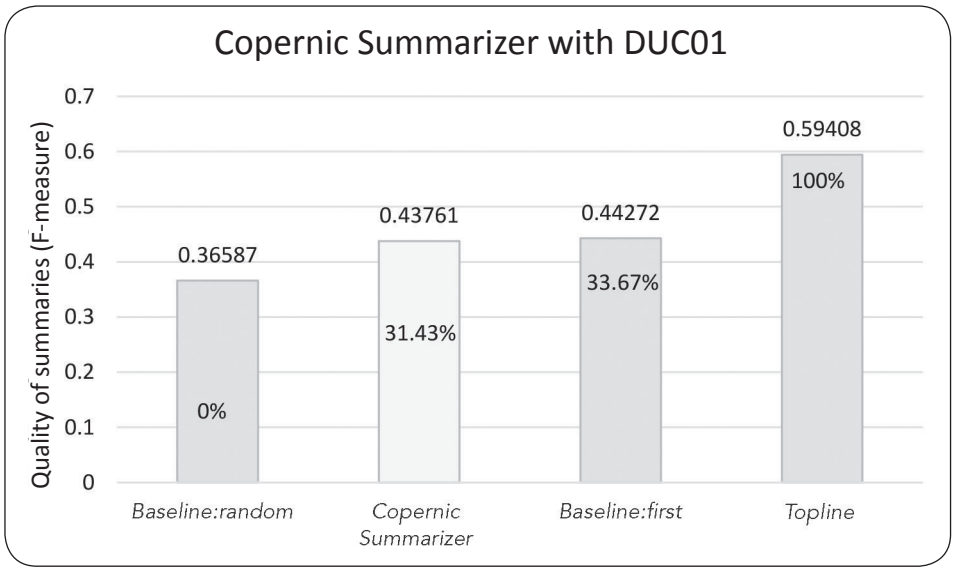
There are tools without an option for a 100-word summary, so formula 6 was applied to calculate the percentage that allows more than one hundred for each document.

$$\frac{Number\ of\ desired\ words}{Number\ of\ total\ words\ \varepsilon\ the\ document} * 100 \qquad (6)$$

5.3.1 Copernic Summarizer

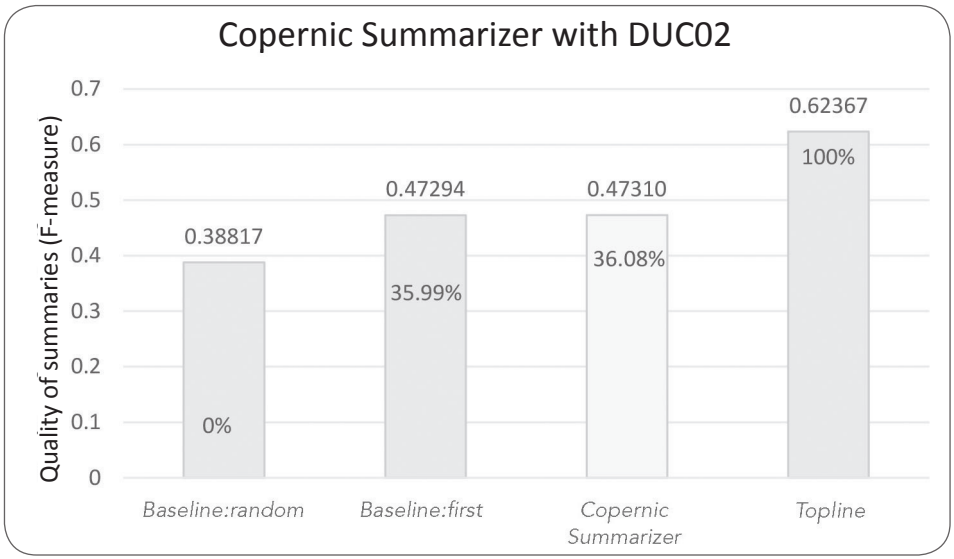
For Copernic Summarizer, DUC01 and DUC02 were assessed; it has the option to produce summaries with hundred words (length required for the two *corpus*), so this option was selected. The results obtained by this tool with DUC01 and DUC02 are presented below.

Graph 5.1 Results of Copernic Summarizer using DUC01 in the comparison with the various heuristics



Graphs 5.1 and **5.2** display the results obtained by Copernic Summarizer for DUC01 and DUC02, respectively. As it is noticed this tool does not surpass *baseline:first* in DUC01; however, it does surpass it in DUC02. Considering that *baseline:random* is the worst way to produce a summary, it is given a value of zero; whereas, *Topline* is given a maximal of one hundred, as it is considered the best. Then, taking *baseline:random* and *Topline* as references, Copernic Summarizer advances 31.43% in AGTS tasks in DUC01, while in DUC02, it advances 36.08%.

Graph 5.2 Results of Copernic Summarizer using DUC02 in comparison with the various heuristics



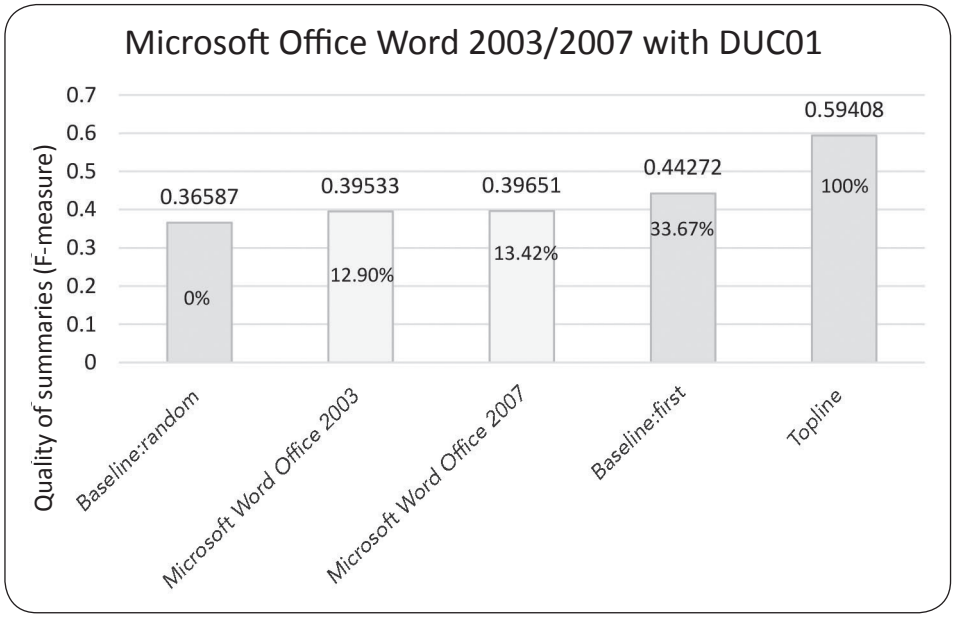
5.3.2 Microsoft Office Word

For Microsoft Office Word collections DUC01 and DUC02 were assessed; it features the option to produce summaries in its 2003 and 2007 versions, and it is possible to produce summaries with one hundred words, a requirement of the *corpora*. However, despite the tool has this option, the summaries produced by Microsoft Office Word are shorter; hence, formula 6 was utilized to calculate the percentage to use. The results obtained by Microsoft Office Word in collections DUC01 and DUC02 are displayed below.

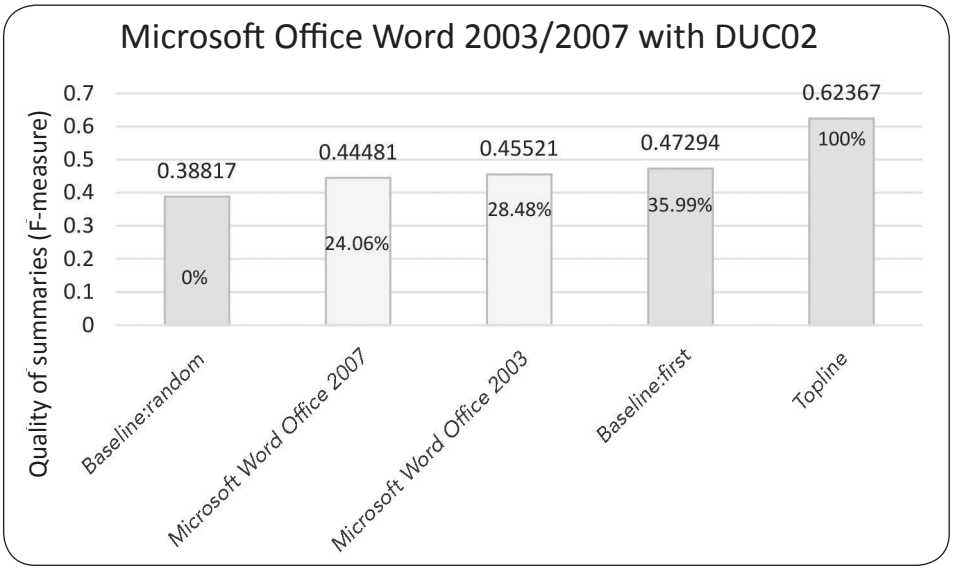
In **graph 5.3**, results for Microsoft Office Word in its versions 2003 and 2007 with DUC01 are shown. As noticed, for DUC01 the results of this tool do not surpass *baseline:first*; however, these results are better than those of *baseline:random*.

Graph 5.4 shows the results of Microsoft Office Word for in their 2003 and 2007 versions with *corpus* DUC02.

Graph 5.3 Results of Microsoft Office Word 2003/2007 using DUC01 in comparison with the various heuristics



Graph 5.4 Results of Microsoft Office Word 2003/2007 using DUC02 in comparison with the various heuristics

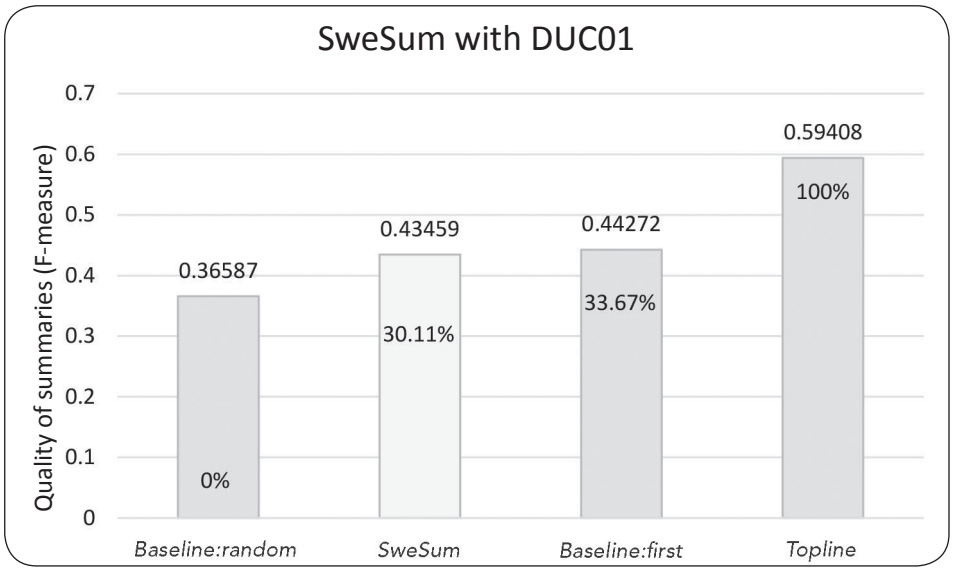


For this collection of documents, as in the case of DUC01, the tool does not surpass *baseline:first*. Though Microsoft Office Word 2003 was not expected to yield better results than the 2007 version.

5.3.3 SweSum

For *SweSum* DUC01 was assessed; it was necessary to calculate the percentage corresponding to each document with formula 6, in such manner that each has one hundred words, as indicated by the specifications of the collection itself. The results obtained by this tool in DUC01, assessed with *ROUGE*, are presented below.

Graph 5.5 Results of SweSum using DUC01 in comparison with the various heuristics

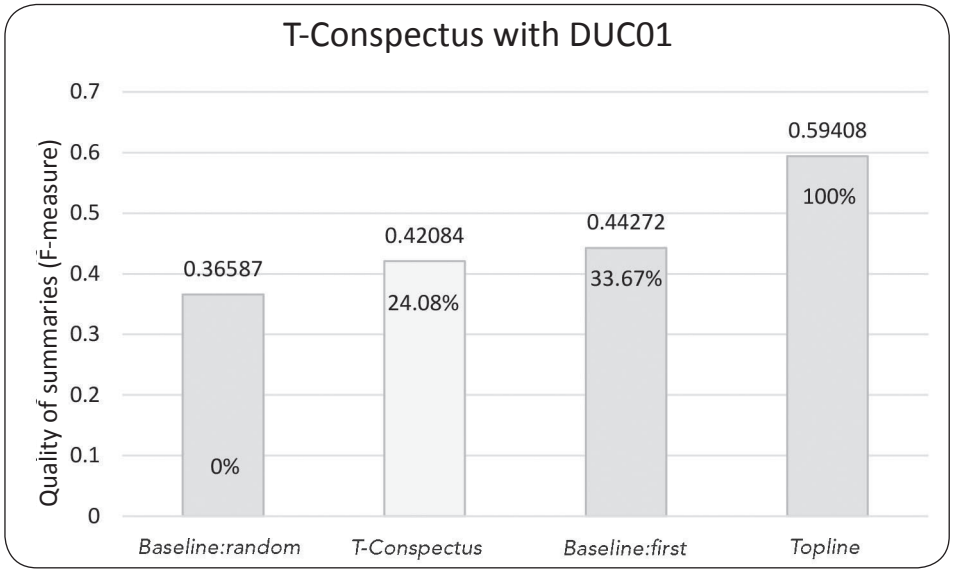


Graph 5.5 shows the results of SweSum for DUC01; it is noticed that they do not overcome *baseline:first*, though the tool advances in AGTS tasks 30.11% regarding *baseline:random* and *Topline*.

5.3.4 T-Conspectus

For T-Conspectus, DUC01 was assessed. It is necessary to calculate the percentage corresponding to each document with formula 6 so that each has one hundred words, as indicated by the specification of the collection itself. The results of this tool for DUC01, assessed by *ROUGE*, are shown below.

Graph 5.6 Results of T-Conspectus using DUC01 in comparison with the various heuristics



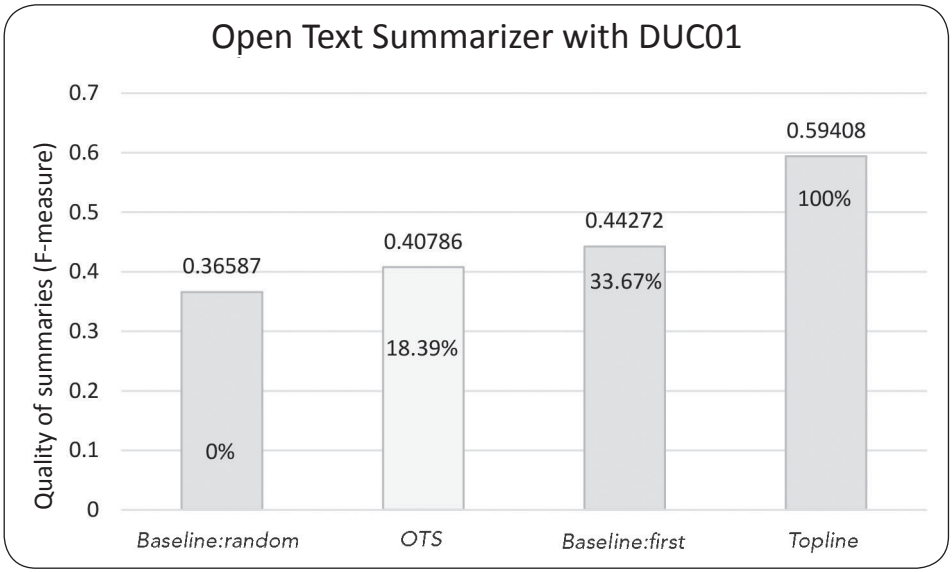
Graph 5.6 displays the results of T-Conspectus for DUC01. As noticed, the results do not surpass *baseline:first*. Albeit, there is an advance of 24.08% regarding *baseline:random* and *Topline*.

5.3.5 Open Text Summarizer (OTS)

For Open Text Summarizer (OTS) DUC01 and DUC02 were assessed. It was necessary to calculate the percentage that corresponds to each document with equation 6 so that they have one hundred words, as specified in the collections.

Following the results obtained by this tool for DUC01, assessed with *ROUGE*, are displayed.

Graph 5.7 Results of Open Text Summarizer using DUC01 in comparison with the various heuristics



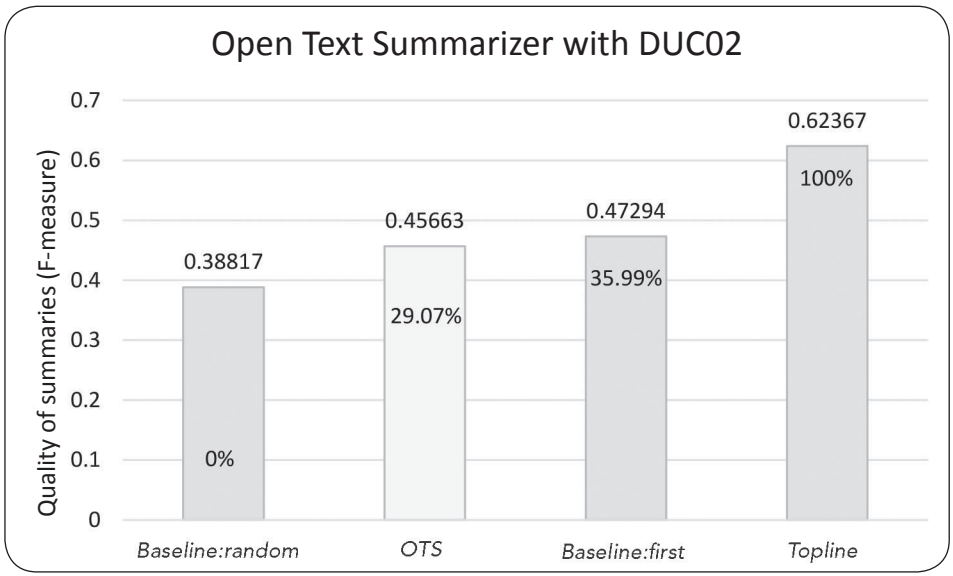
Graph 5.7 shows the results of Open Text Summarizer using DUC01. As noticed, this tool does not surpass *baseline:first*. If *baseline:random* is considered the worst way to produce a summary and *Topline* the best, then Open Text Summarizer obtains 18.39%; thus far, the lowest percentage.

For DUC02, OTS advances 29.07% regarding *baseline:random* and *Topline* (**graph 5.8**).

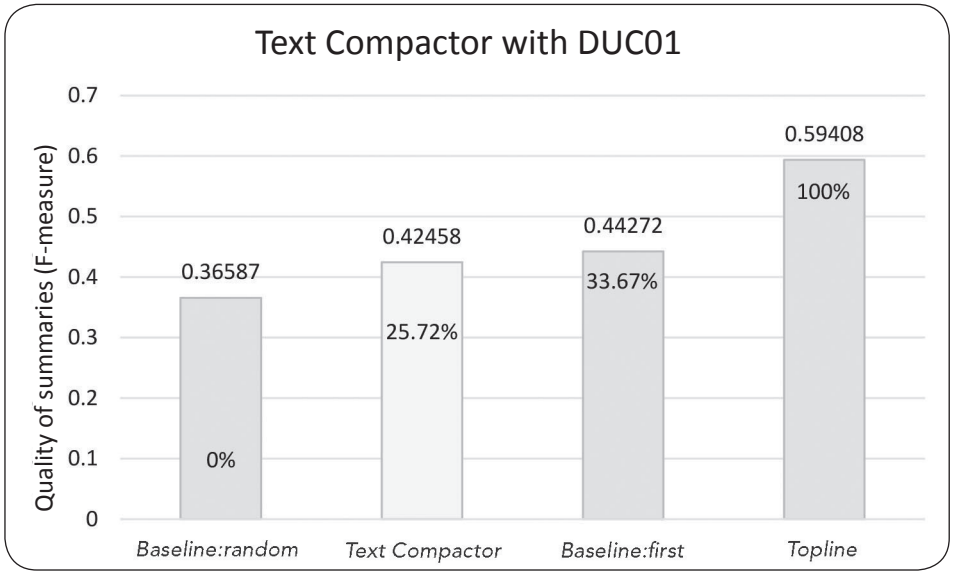
5.3.6 Text Compactor

For Text Compactor DUC01 was assessed. It is necessary to calculate the corresponding percentage with formula 6 so that each summary of the collection documents has one hundred words.

Graph 5.8 Results of Open Text Summarizer using DUC02 in comparison with the various heuristics



Graph 5.9 Results of Text Compactor using DUC01 in comparison with the various heuristics

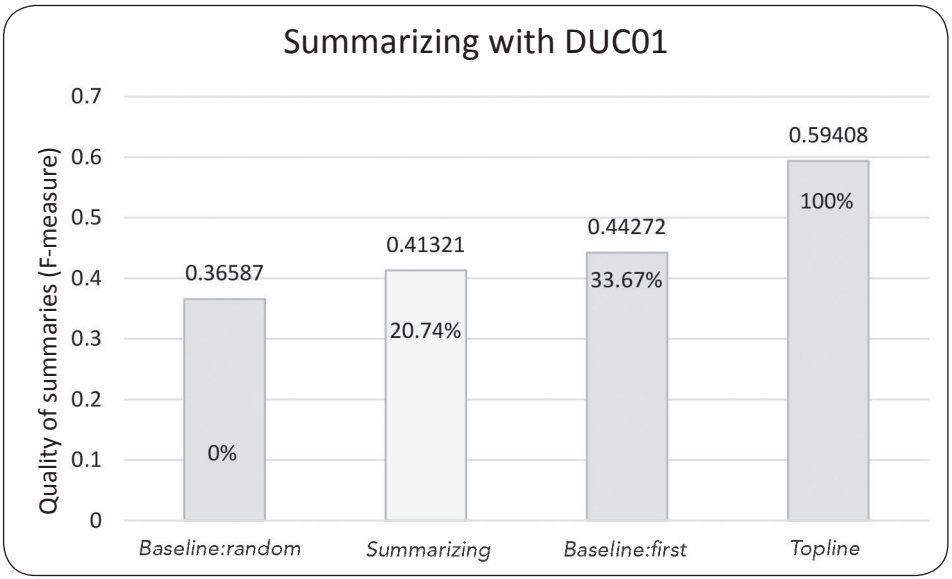


This tool's results evince that T-Compactor has an advance of 25.72% regarding *baseline:random* and *Topline* for DUC01 assessed with ROUGE.

5.3.7 Summarizing

For Summarizing DUC01 was assessed. Article Summarizing Online features an option to produce 100-word summaries (length required by the *corpus*). The results obtained by this tool for DUC01, assessed with *ROUGE*, are shown below.

Graph 5.10 Results of Summarizing using DUC01 in comparison with the various heuristics

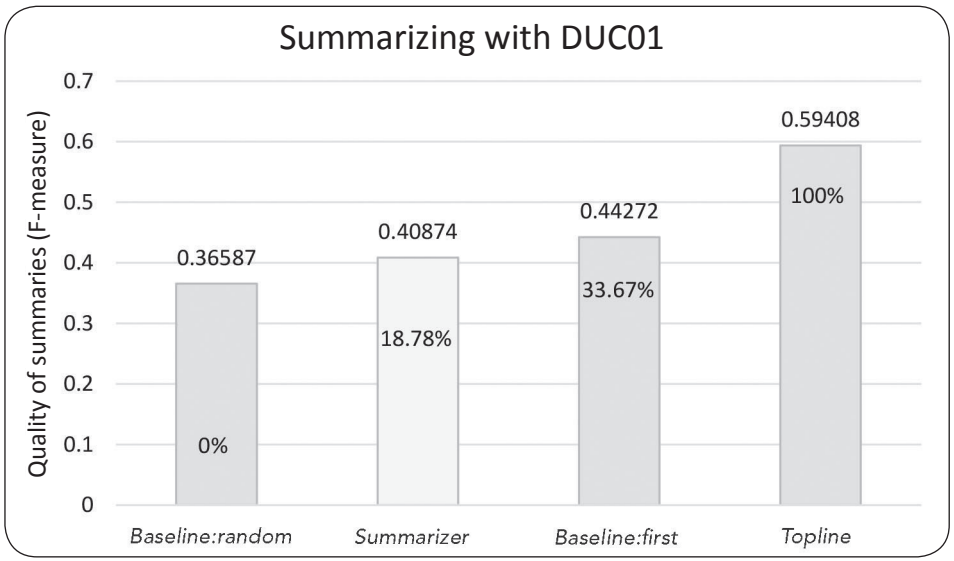


Graph 5.10 shows the results of Summarizing for DUC01. As noticed, the results obtained are not better than those of *baseline:first*. If *baseline:random* is taken as the bottom line to find out if a summary is good (0%) and *Topline* as the best result (100%), then *Summarizing* obtains 20.74%.

5.3.8 Summarizer

For Summarizer DUC01 was assessed. For this tool, the corresponding percentage so that each document has one hundred words. The advance of Summarizer is 33.67% (see Graph 5.11) regarding the worst (*baseline:random*) and the best heuristic (*Topline*).

Graph 5.11 Results of Summarizer using DUC01 in comparison with the various heuristics

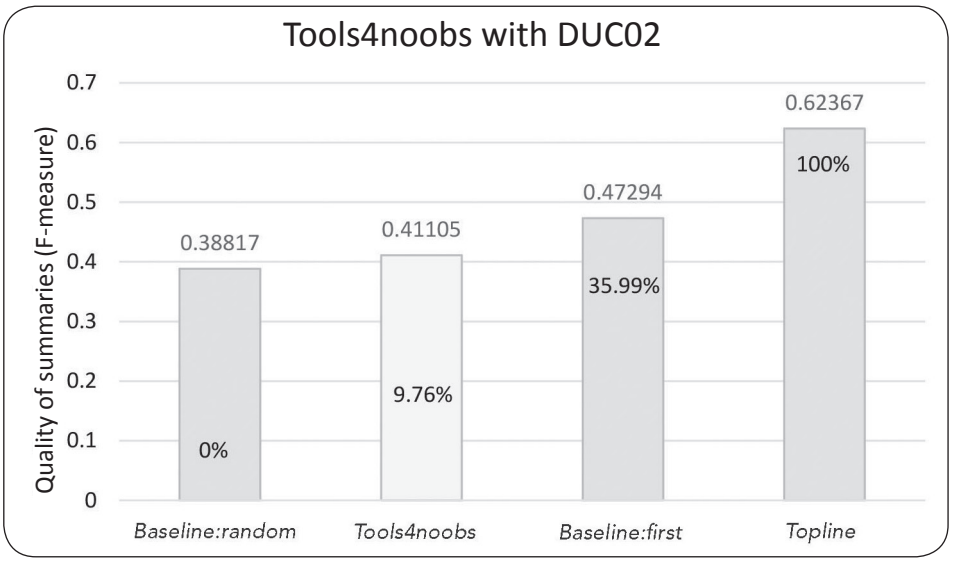


5.3.9 Tools4noobs

For Tools4noobs DUC02 was assessed. It is necessary to calculate the corresponding percentage so that each of them has a hundred words; the calculation is carried out with formula 6. Below, the results of this tool for DUC02, assessed with *ROUGE*, are presented.

Graph 5.12 Shows the results obtained with Tools4noobs using DUC02. The advance for this tool is the lowest regarding those tested for the English

Graph 5.12 Results of Tools4noobs with DUC02 in comparison with the various heuristics



language, as it only obtained 10% in comparison with *baseline:random* and *Topline*.

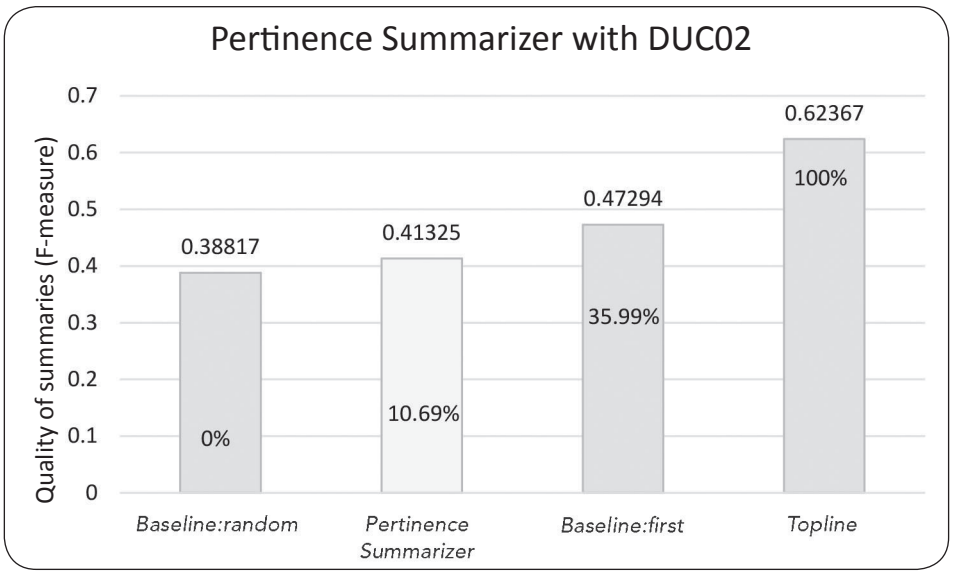
5.3.10 Pertinence Summarizer

For Pertinence Summarizer DUC02 was tried. The percentage so that all the summaries have one hundred words has to be calculated using formula 6. The advancement level is 10.69%, which ranks it at the second lowest place of commercial tools.

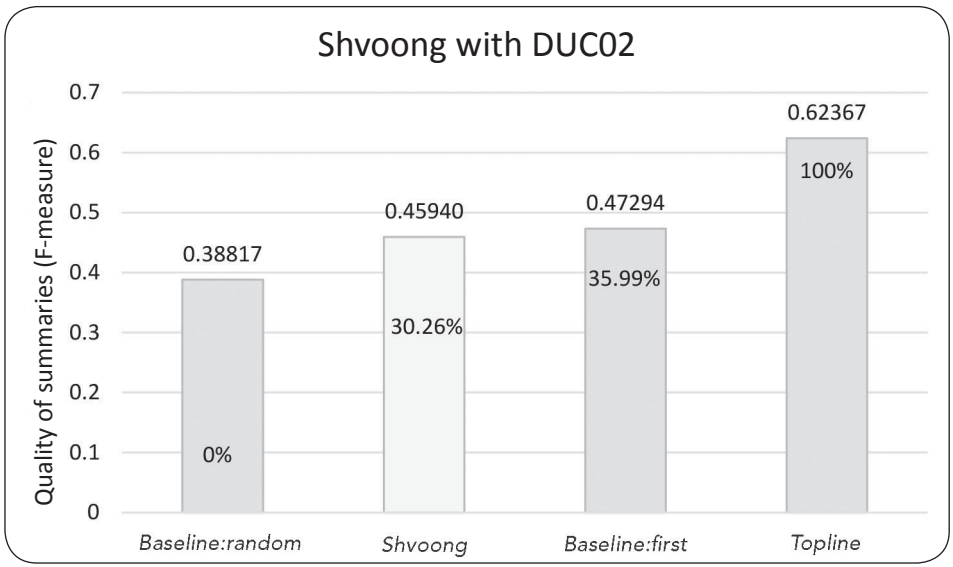
5.3.11 Shvoong

For Shvoong DUC02 was assessed. The corresponding percentage of each document is calculated with formula 6 so that each of them has one hundred words. The results obtained by this tool for DUC02, assessed with *ROUGE*, are presented below.

Graph 5.13 Results of Pertinence Summarizer using DUC02 in comparison with the various heuristics



Graph 5.14 Results of Shvoong using DUC02 in comparison with the various heuristics



Graph 5.14 shows the results of Shvoong using DUC02. As noticed, this tool does not surpass *baseline:first*. However, it advances 30.26% regarding *baseline:random* and *Topline*.

5.4 Novel scientific methods

In this section novel scientific methods tried in the English language are presented. **Table 5.11** presents the various methods reviewed in this section and the *corpora* used.

Table 5.11 Novel scientific methods assessed with DUC01 and DUC02

Method	DUC01	DUC02
AG-4feature	✓	✓
Ma-SingleDocSum	✓	✓
UnifiedRank		✓
AG-Bag-Words		✓
AG-Bigramas		✓
AG-Multi		✓
TextRank		✓
SFMs k-best		✓
SFMs (1 *best+first*)		✓
SFMs grouping		✓

5.4.1 Ma-SingleDocSum

Ma-SingleDocSum, proposed by Mendoza (Mendoza *et al*., 2014), is based upon a memetic algorithm focused on single-document summaries. In addition to using genetic operators to produce summaries, it uses local search. The parameters considered for the fitness function are sentence position, relation of the sentences

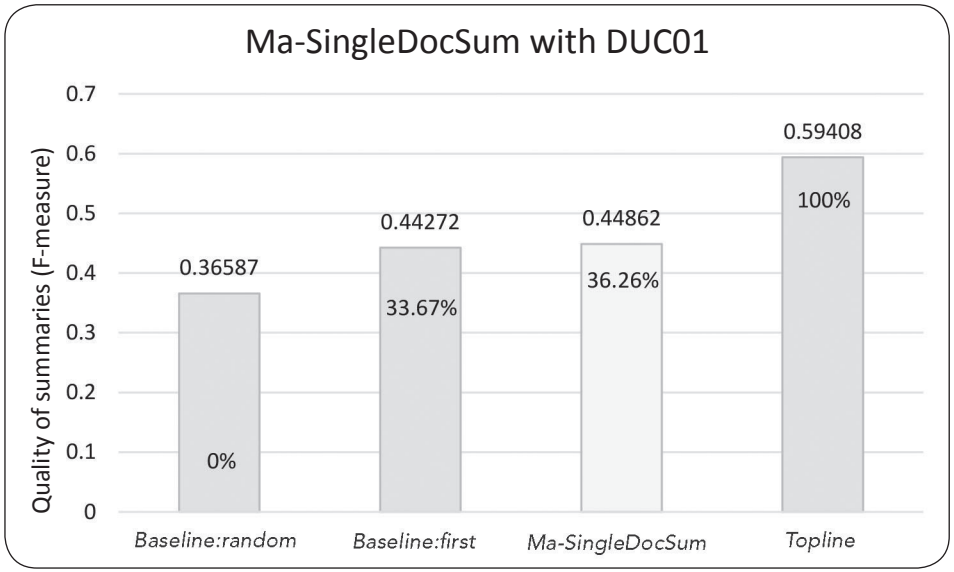
with the title, sentence length, cohesion and convergence (known as the text's topic).

It is known that the position of the sentences, their relation to title and their length are the characteristics most used in AGTS tasks (**table 3.1**, in section 3.2.1).

A comparison of Ma-SingleDocSum with the heuristics for DUC01 and DUC02, assessed with *ROUGE*, is presented below.

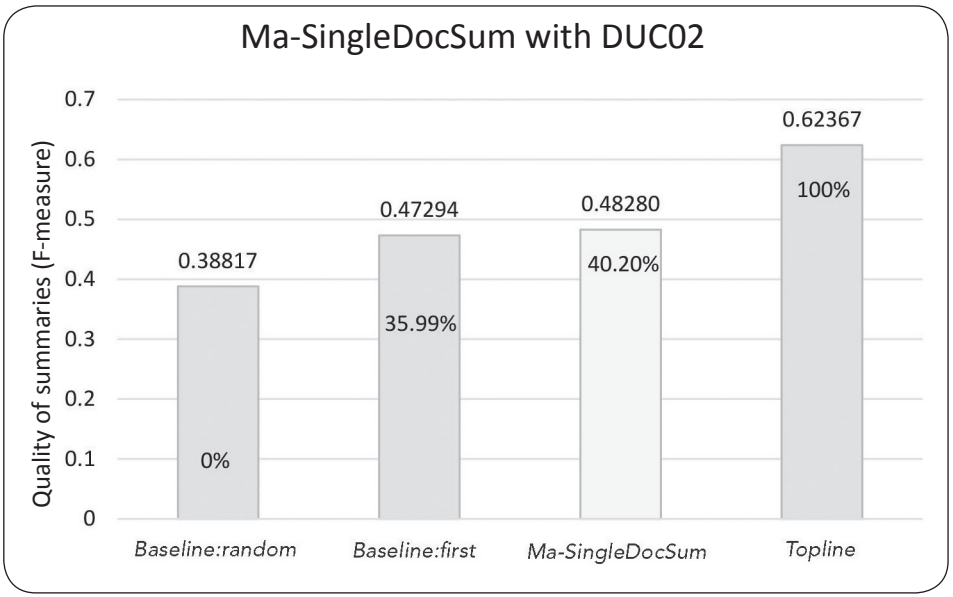
For DUC01, this method surpasses the two reference heuristics (**graph 5.15**), as it is the case for DUC02 (**graph 5.16**).

Graph 5.15 Results of Ma-SingleDocSum using DUC01 in comparison with the various heuristics



It is worth mentioning that this is one of the best AGTS methods, which owing to its characteristics can be applied to English and Spanish.

Graph 5.16 Results of Ma-SingleDocSum using DUC02 in comparison with the various heuristics



5.4.2 UnifiedRank

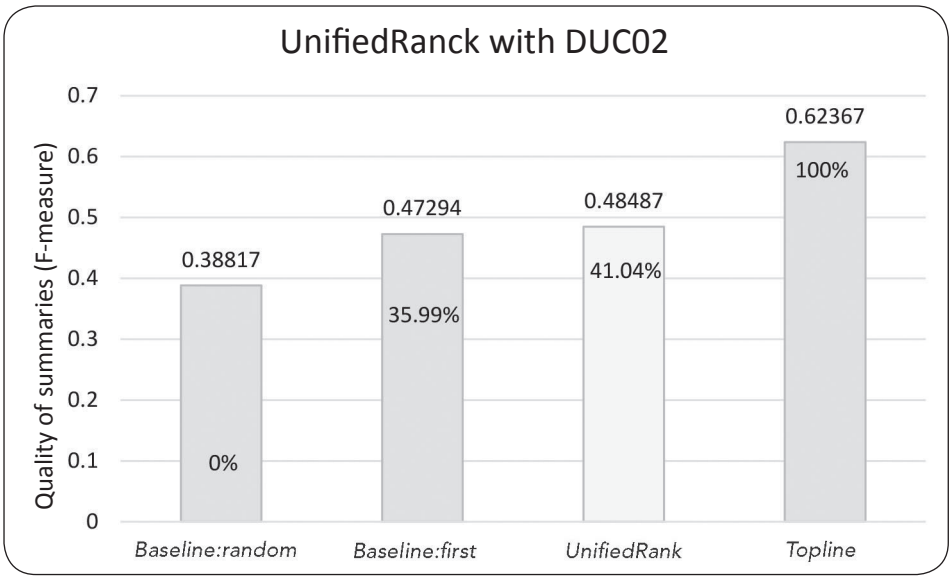
UnifiedRank (Wan, 2010) is based on graphs focused on the production of single- and multi-document summaries. This work reviews the influence between the generation of summaries for them. The *corpus* with which it works for single documents is DUC02 and it is assessed with *ROUGE*.

Thus far, UnifiedRank has produced the best results for DUC02 (**graph 5.17**).

5.4.3 AG-Bag-Words

The method proposed by García-Hernández and Ledeneva (2013) is one with the best results. It resorts to a genetic algorithm and uses the bag-of-words model. The fitness function used in this work considers two main characteristics:

Graph 5.17 Results of UnifiedRank using DUC02 in comparison with the various heuristics



- The first sentences are the most important; they are considered candidates to be part of the summary.
- Verify that the summary includes various ideas, this is to say, not repetitive, though it must contain important words as well (Accuracy-Recall).

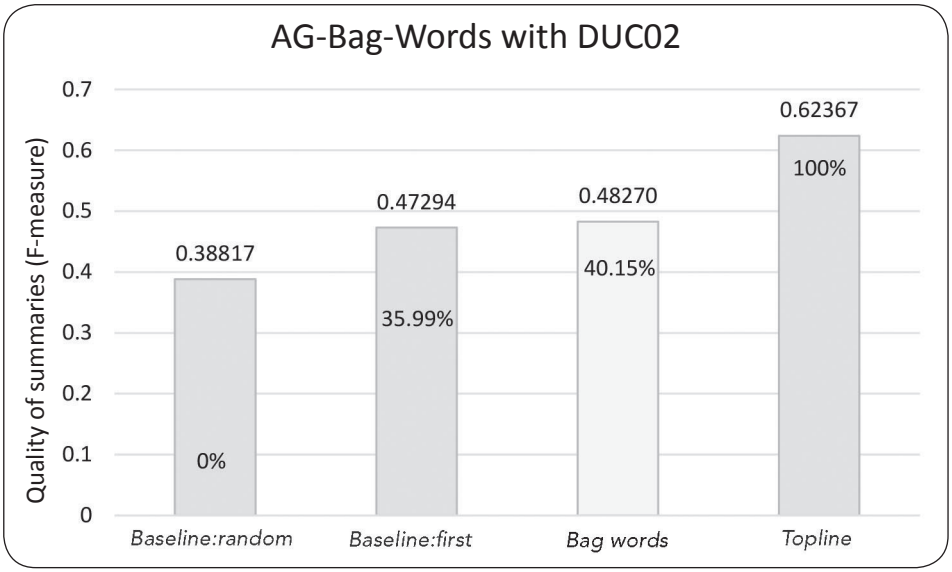
Graph 5.18 displays the comparison of this method with the various heuristics; it is noticed that bag of words surpasses *baseline:first*.

5.4.4 AG-Bigramas

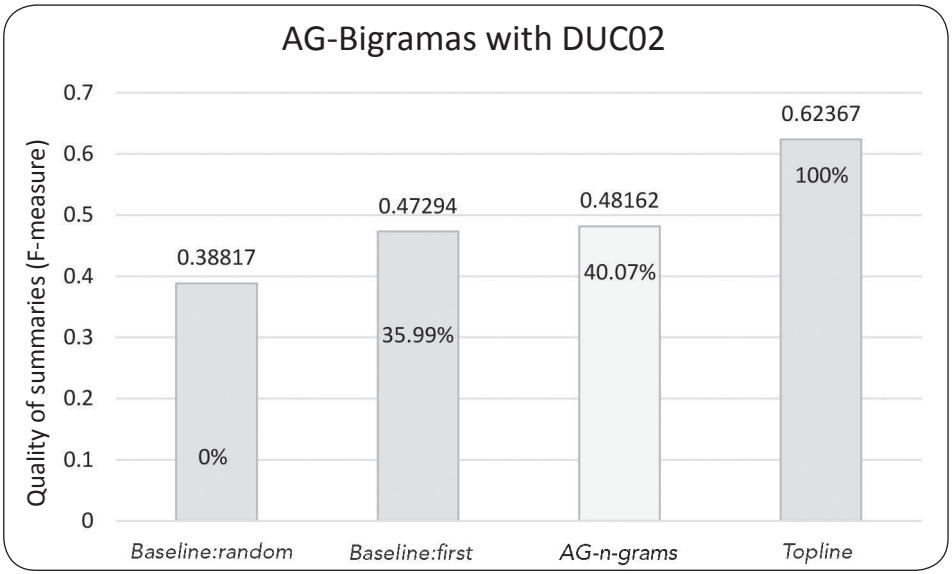
AG-Bigramas [AG-Bigrams] was proposed by Matias (2013). It is based on a genetic algorithm, specifically applying the bigram text model and works on single documents in English. The position of the sentences and frequency of terms is considered in the fitness function. Matias (2013) points out that bigrams allow managing information better with fewer losses.

AG-Bigramas is one of the models with the best results, overcoming *baseline:random* and mainly *baseline:first* (**graph 5.19**).

Graph 5.18 Results of AG-Bag-Words using DUC02 in comparison with the various heuristics



Graph 5.19 Results of AG-Bigramas for DUC02 in comparison with the various heuristics



5.4.5 AG-Multi

AG-Multi proposed by Matias (2016) is one of the most utilized in AGTS for a number of languages. It is based on a genetic algorithm and uses n-grams as a text model. For the fitness function, it considers two of the most utilized characteristics in the state of the art (García-Hernández and Ledeneva, 2013), which are term frequency and sentence position.

Term frequency

To produce a summary (*S*), the upper limit of words (*m*) has to be considered. Consequently, the number of recovery units is always limited by the threshold of the maximum number of words; therefore, the reference summary (the one made by a human) must contain, on the one side, the most relevant words of the original text (*T*), and on the other, expressivity; this is to say, it must not be redundant.

The relevance of w is represented by the frequency of each word in the original text (frequency(w,T)), and expressivity is represented only if various words the summary can contain are considered ($\{word \in S\}$).

In this sense, the best summary would have the most frequent words in the original text and each must be different. To have a normalized measure, García-Hernández (2013) proposes that the addition of the frequencies of the words in the summary must be divided by the addition of the frequency of the most frequent words in the original text.

$$\beta = \frac{\sum_{p=\{word \in S\}}^{m} frecuency(p,T)}{\sum_{q=\{word \in T\}}^{m} frecuency(q,T)} \tag{7}$$

Sentence position

By and large, this characteristic is deemed important in AGTS tasks, as displayed in **table 3.1**; fourteen out of sixteen analyzed works use it to solve the task. Starting from the heuristic that has demonstrated that first sentences are very important for AGTS. One of the ideas to give more importance *xn*;

would be to consider the first one with importance ; the second with an $xn - 1$ importance; though this may be very drastic because if it were a thirty-sentence text, it would be said that the first is thirty times more important than the last; however, this may be part of the conclusions and would not have a chance to appear in the summary.

In the work by García-Hernández and Ledeneva, it is proposed to soften the sentences' importance. To do so, it is possible to use the general equation of a line with an *m* slope. The slope indicates the importance of the sentences from first to last; if it is negative, the first sentences are more important; (-1) means that goes down to the right at a 45-degree angle; 0 means that all the sentences are as important, while positive means that the last sentences are more important; (1) means that it goes up to the right at a 45-degree angle.

For a text with *n* sentences, if the sentence is selected for the summary (this is chromosome $|c_i| = 1$), then its relevance is defined by $\mathrm{m}(i - x) + x$, where a $x = 1 + (n - 1/2)$ and *m* is the slope to ascertain. With a view to normalizing the sentence position measure (δ), the importance of the first *k* sentences, where *k* is the number of sentences chosen.

Then, the formula to calculate the importance of the first sentences is

$$\delta = \frac{\sum_{|C_i|=1}^{n} m(i-x)+\mathrm{x}}{\sum_{j=1}^{k} m(j-x)+1}, x = 1+\frac{(n-1)}{2}. \tag{8}$$

Finally, to obtain the value of fitness function the following formula is applied:

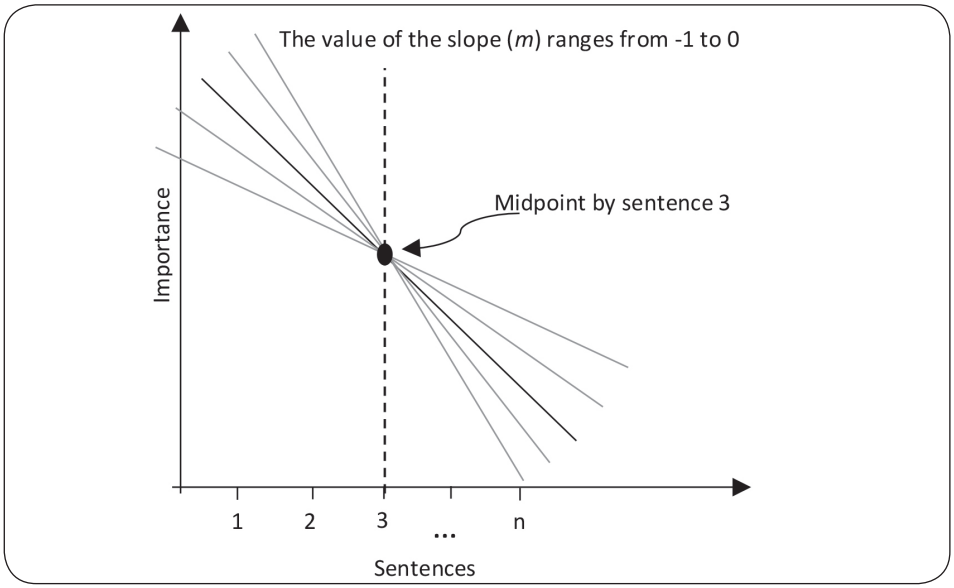
$$fitness = \beta * \delta \tag{9}$$

as previously mentioned, the slope of the line (*m)* can help us ascertain the sentences' importance. The value of *m* may vary in order to soften such importance.

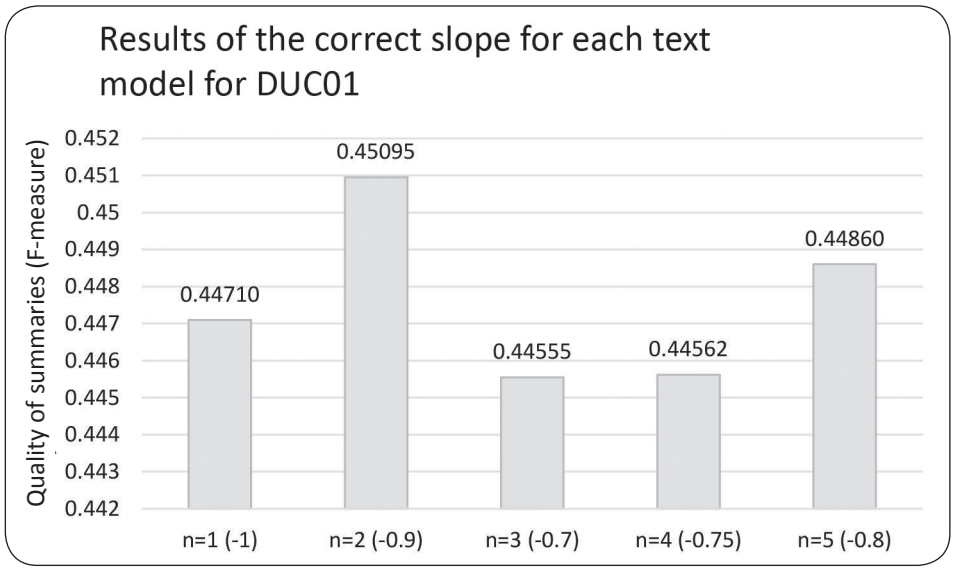
Graph 5.20 displays a presentation of a line slope. To assess the method, the following values of the slope are considered: .

$m = -0.25$, $m = -0.3$, $m = -0.375$, $m = -0.45$, $m = -0.5$, $m = -0.55$, $m = -0.6$, $m = -0.625$, $m = -0.65$, $m = -0.7$, $m = -0.75$, $m = -0.8$, $m = -0.85$, y $m = -0.9$. These values were taken at random

Graph 5.20 Graphic representation of the value of a line's slope



Graph 5.21 Results of the correct slope for each text model for DUC01

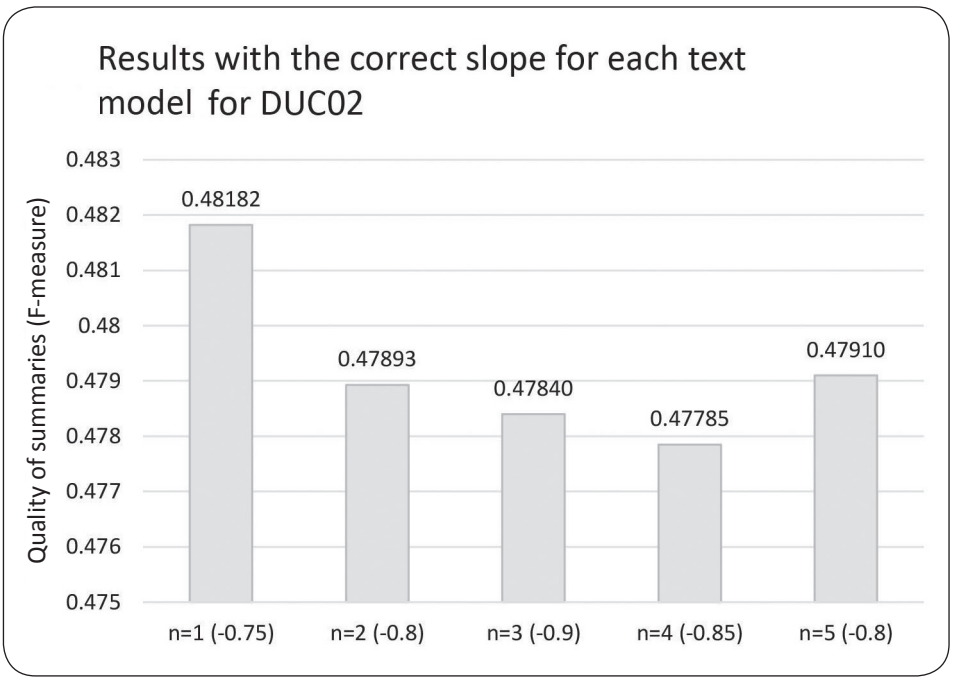


As mentioned, the text model used in Matias (2016) is n-grams. An analysis by text model was run ($n = 1$ to $n = 5$) to find out the best value for the slope (sentence position) and the best text model for DUC01 and DUC02. Graph 5.21 shows the best value obtained for each model in DUC01.

The best text model for DUC01 is Bigramas ($n = 2$) with a slope, $m = -0.9$.

The n-gram text models were also tried in DUC02 (**graph 5.22**).

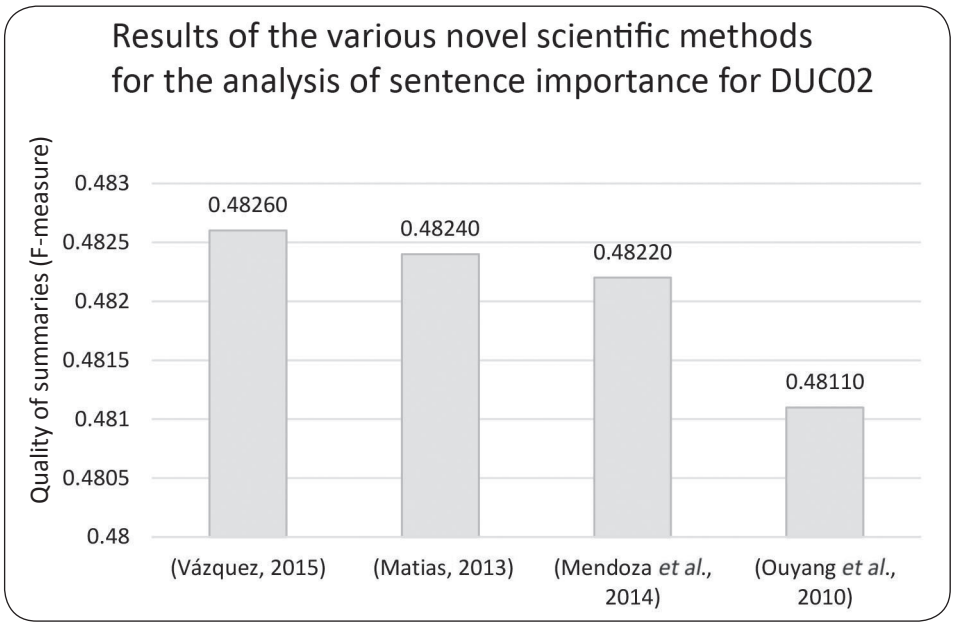
Graph 5.22 Results with the correct slope for each text model



The best text model for DUC02 was n-gramas with ("bag of words" model).

One of the characteristics most used in AGTS is sentence position. Owing to this, Matias (2016) studies the various forms to calculate such position. **Graph 5.23** shows the results obtained applying the various methods to the state of the art on DUC02.

Graph 5.23 Results of the various novel scientific methods for the analysis of sentence importance



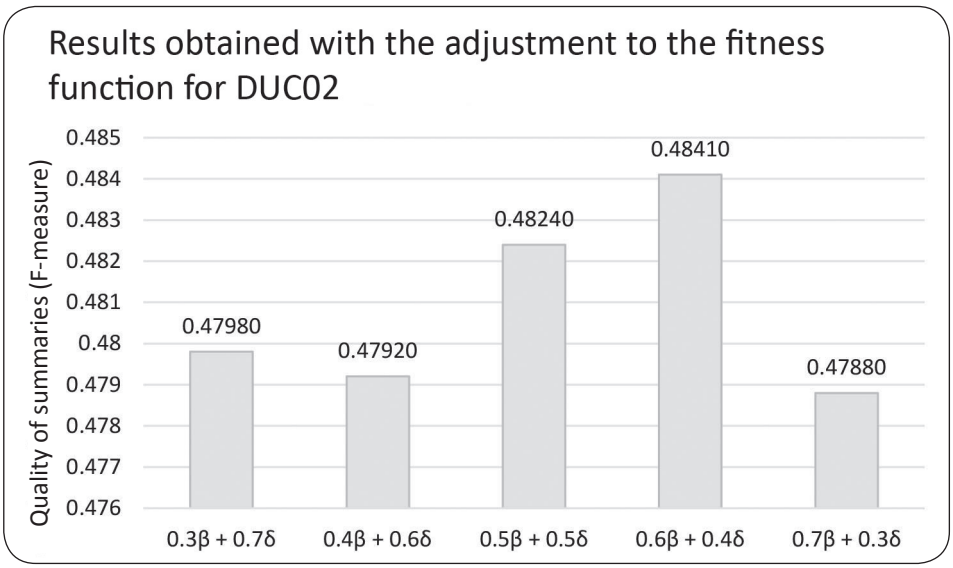
As noticed, the formula put forward by Vázquez (2015) is the one that produces the best results for DUC02.

For DUC01, tests were run in the manner proposed by Vázquez (2015). However, the results were not favorable (Alvarado B., 2017).

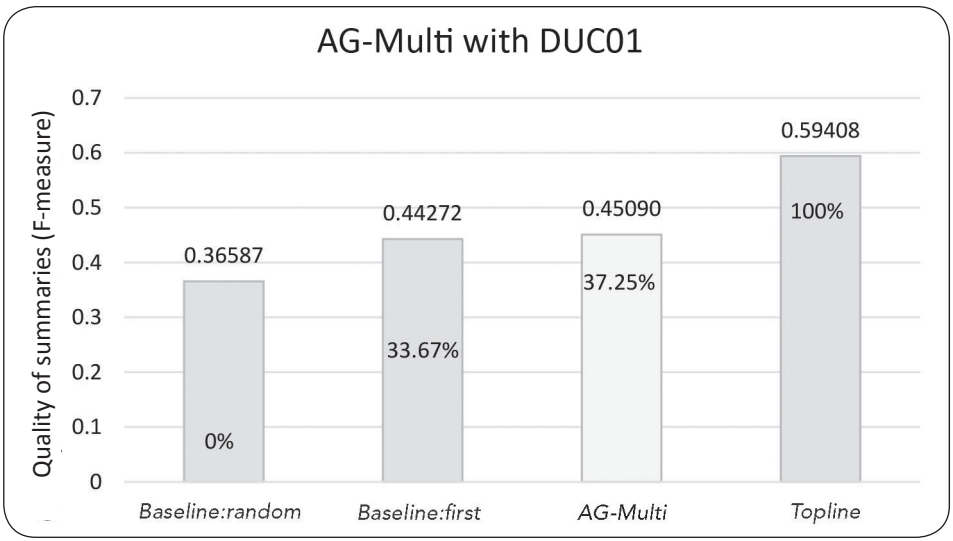
In addition to sentence position, Matias' method (2016) considers term frequency, so a weight adjustment was made on DUC02 parameters ß and ∂.

The best results for DUC01 using Matias' (2016) method are obtained with bigrams and a slope value of $m = -0.9$.

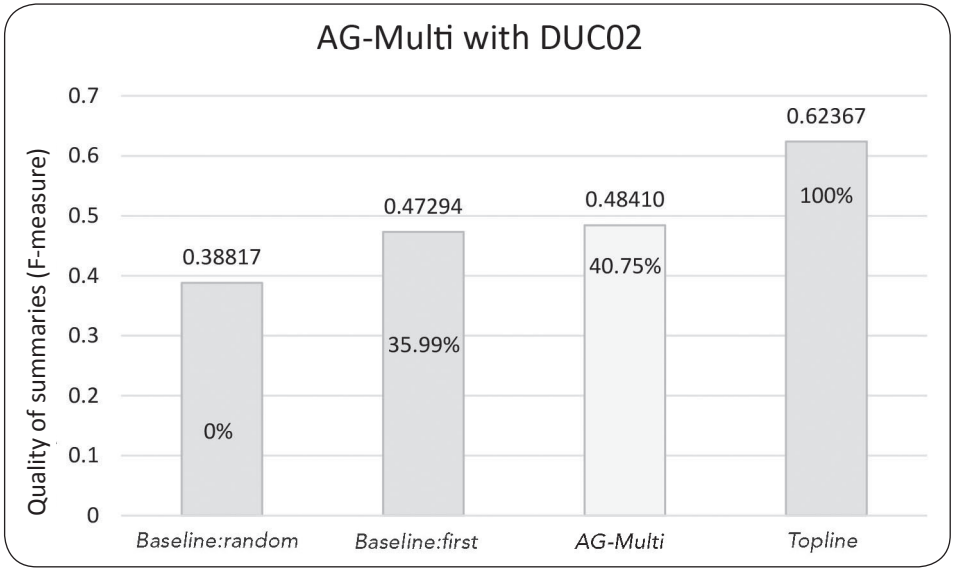
Graph 5.24 Results obtained with the adjustment to the fitness function



Graph 5.25 Results for AG-Multi using DUC01 in comparison with the various heuristics



Gráfica 5.26 Results of AG-Multi for DUC02 in comparison with the various heuristics



For DUC02, the best result is obtained with "bag of words", applying the formula put forward by Vázquez (2015) and a combination of 0.6 for term frequency and 0.4 for sentence position.

This method has been tried in English, Spanish, Portuguese and Russian; it can work with or without preprocessing, yielding acceptable results, since it has surpassed *baseline:first* in each language.

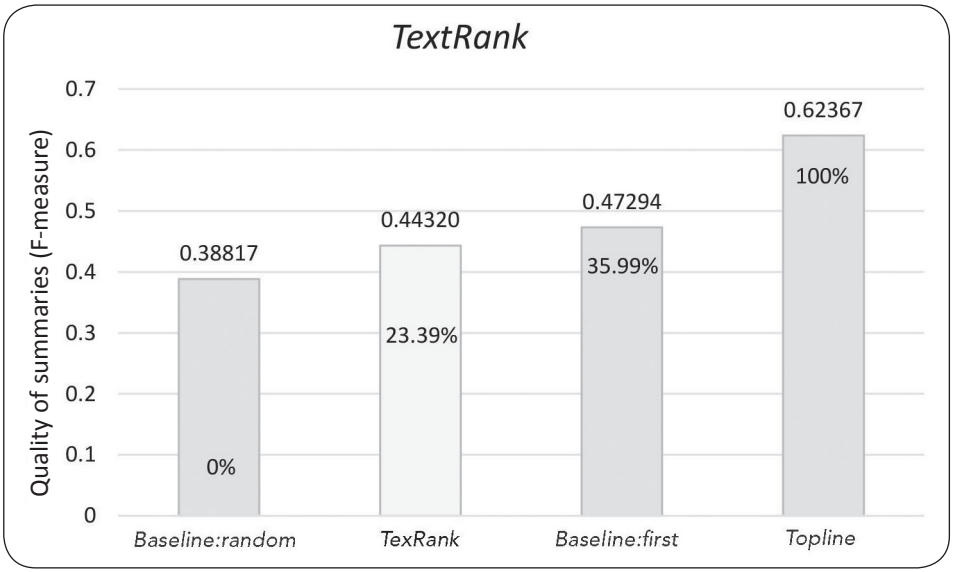
5.4.6 TextRank

This method consists in a weighing algorithm based on graphs. According to Rada Mihalcea (Mihalcea, 2004), a graph is produced to represent the text so that the nodes are words (or other text entities) interconnected by means of arches with significant relations. In order to extract the sentences, complete sentences have to be graded and classified from the most to the least important. Therefore, an arch is added to the graph by each sentence in the text.

To establish the connections between sentences, a similarity relation is defined, in which the relation between two sentences may be seen as a

"recommendation" process, that is to say, a sentence that points at a certain context in the text provides the reader with a "recommendation" to refer to others that point at the same concepts and so, a link between two given sentences that share a common content may be established. To try this method, DUC02 is used.

Graph 5.27 Results of TextRank for DUC02 in comparison with the various heuristics



In spite of being one of the most recognized and used models in the state of the art, it does not surpass *baseline:first* in DUC02 (**graph 5.27**).

5.4.7 Maximal frequent sentences (MFS k-best)

This work presents a method based on statistics, independent from domain and language, to produce a single-document extractive summary. In her work Ledeneva (2008) (*et al*., 2008) experimentally shows that words parts of bigrams (two-word sequence) repeated more than once in the text are good terms to describe its content, as the so called Maximal Frequent Sentences (sequences

of words that repeat a certain number of times that are not contained in other frequent words). As well, it is evinced that the term frequency, such as term weighing, provides good results (as long as only the occurrences of one of them are counted in the repetitive bigrams).

Ledeneva applies a four-step technique to produce the summary. Such steps are term selection, term weighing, sentence weighing and selection. In term selection, MFS repetitive bigrams (they have to appear at least twice in the text) and simple words or unigrams are extracted. In term weighing, the frequency of the term is used, which is the number of times it appears in the text in an MFS. The maximal length of an MFS that contains the term is also used as weighing, as well as assigning the same weight for everyone. In sentence weighing, the weight of all the terms contained in such sentence are added.

Finally, the selection of sentences that will be part of the summary is carried out following two criteria; firstly, the best sentences are selected; i.e., those with the heaviest weight. This is repeated up to the desired length of the summary (one hundred words). The second criterion selects the best k sentences, adding to the first that appear in the document (kbest+first). This is repeated up to the desired summary length.

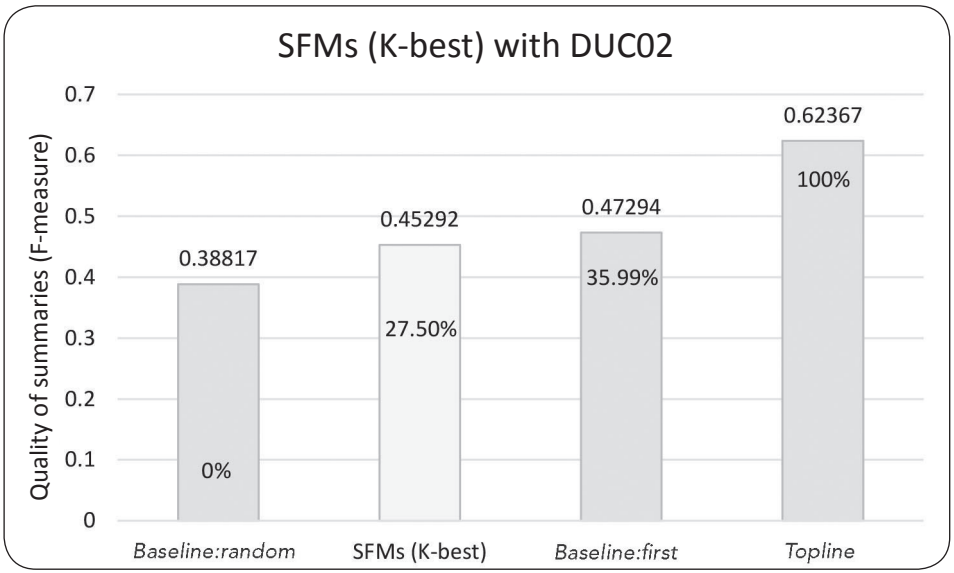
MFS (K-best) does not surpass *baseline:first* (**graph 5.28**).

5.4.8 MFS (1BEST + FIRST)

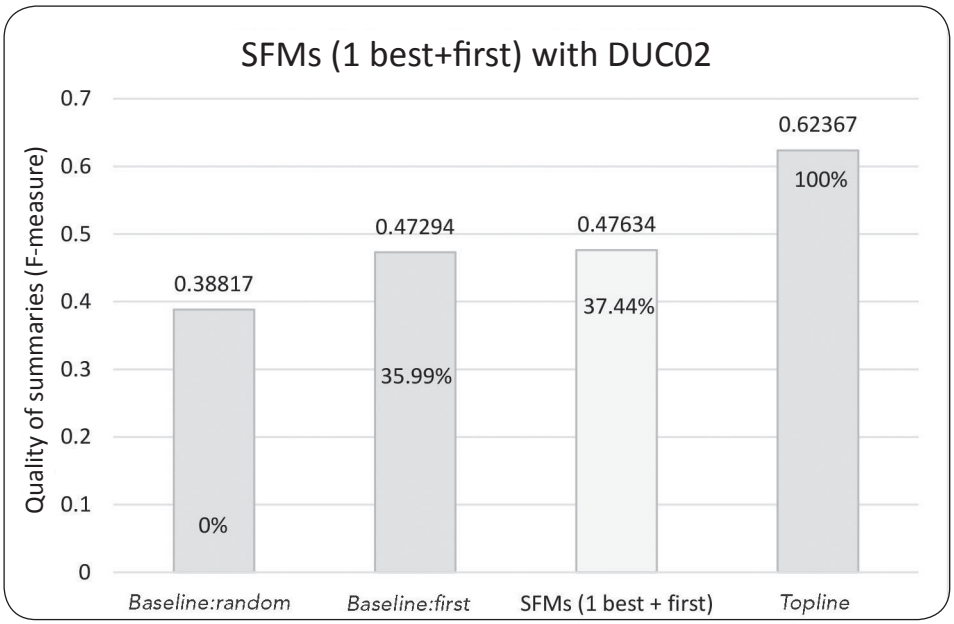
This work presents a method based on statistics, independent from domain and language, to produce of single-document extractive summaries. In her work, Ledeneva (2008) experimentally shows that words that are parts of bigrams and which repeat more than twice in the text are good terms to describe the content, as well Maximal Frequent Sentences. It is also demonstrated that term frequency produces good results as term weighing (as long as only the repetitions of one of them in repetitive bigrams are counted).

It was in 2008 that novel scientific methods started to overcome *baseline:first.* As a reference, MFS (1best + first) (**graph 5.29**).

Graph 5.28 Results of MFS (K-best) for DUC02 in comparison with the various heuristics



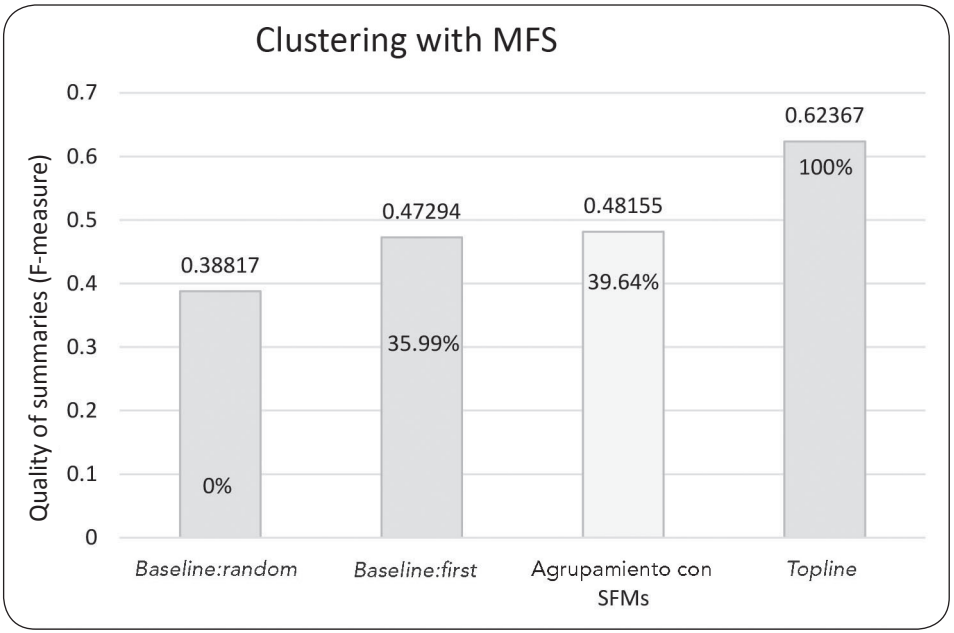
Graph 5.29 Results of MFS (1best + first) for DUC02 in comparison with the various heuristics



5.4.9 MFS CLUSTERING

In the previous method, MFS, sentences with heavier weights are selected to be part of the summary. However, if there are very similar sentences and are chosen, they do not add any information. In this grouping (clustering) of sentences with MFS and k-means (García- Hernández *et al*., 2008) separate sentences in groups, for each of which the most repetitive phrase is chosen and this becomes part of the summary. This was also carried out with EM clustering in (Ledeneva *et al*., 2011). To try this method, DUC02 is used.

Graph 5.30 Results of SFM clustering for DUC02 in comparison with the various heuristics

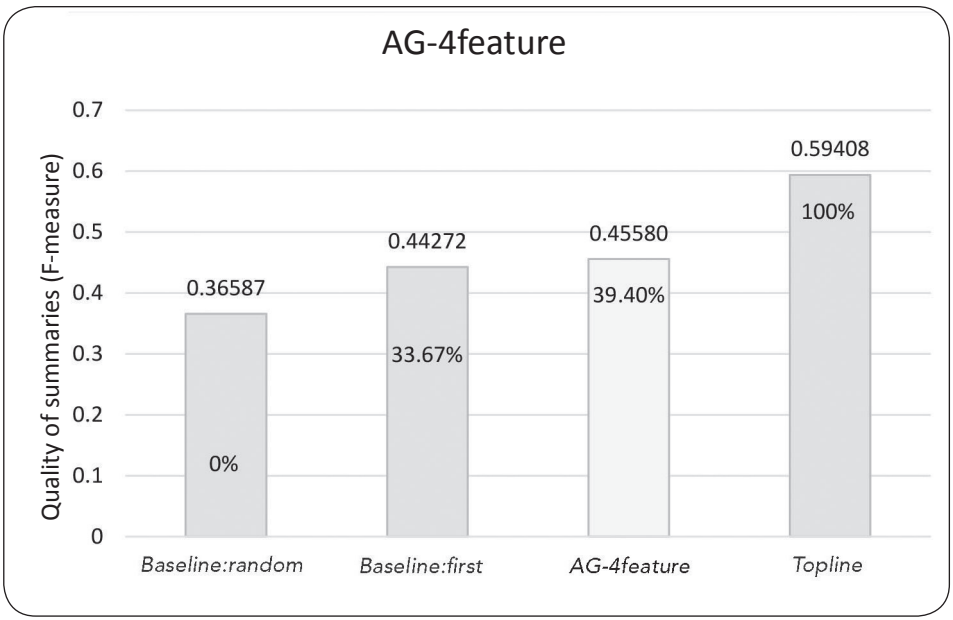


Among the English-language methods that overcome *baseline:first*, we find SMF clustering (**graph 5.30**).

5.4.10 AG-4FEATURE

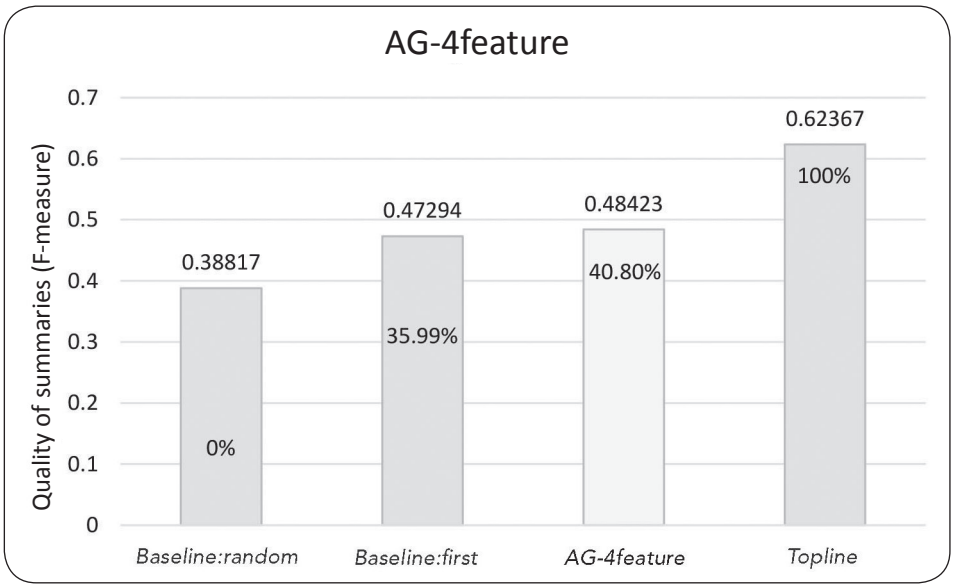
Vázquez (2018) presents a method to optimize the combination of the following characteristics on the basis of a genetic algorithm for each stage: similarity with the title (δ); sentence position (β), sentence length (γ); and, coverage (α). In the work it is concluded that the most important characteristic for the English language is: $\alpha = 0.59$, $\beta = 0.36$, $\gamma = 0.02$, $\delta = 0.03$ (Vázquez, García-Hernández and Ledeneva, 2018). Following, results for DUC01 and DUC02 are presented.

Graph 5.31 Results of AG-4feature for DUC01 in comparison with the various heuristics



Graphs 5.31 and **5.32** present the results of AG-4feature for DUC 01 and 02; it is noticed that the method is one of the best: it has an advance of 39.40% for *corpus* DUC01, and 40.80% for DUC02.

Graph 5.32 Results of AG-4feature for DUC02 in comparison with the various heuristics



5.5 Results and analysis

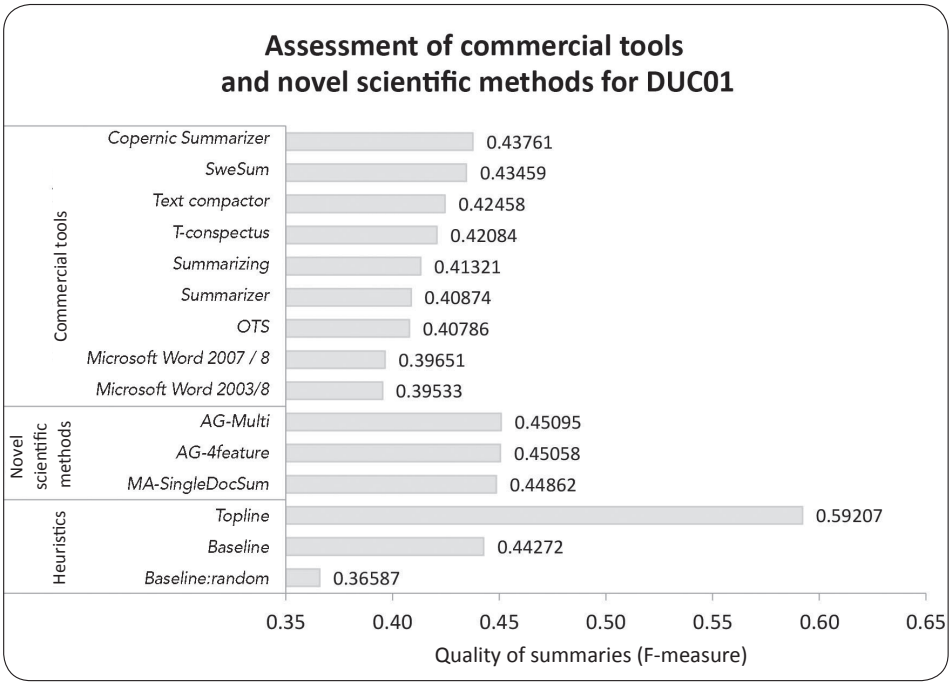
In order to assess and compare the commercial tools of AGTS novel scientific methods in English, DUC01 and 02 were assessed with *ROUGE*.

Graph 5.33 shows a comparison between novel scientific methods, commercial tools and heuristics for DUC01.

Commercial tools overcome *baseline:random*, considered the worst way to produce a summary, in all of the methods for *corpora* DUC 01 and 02. The second heuristic to surpass is *baseline:first*; it is worth mentioning that in English-language AGTS, this heuristic was overcome ten years ago.

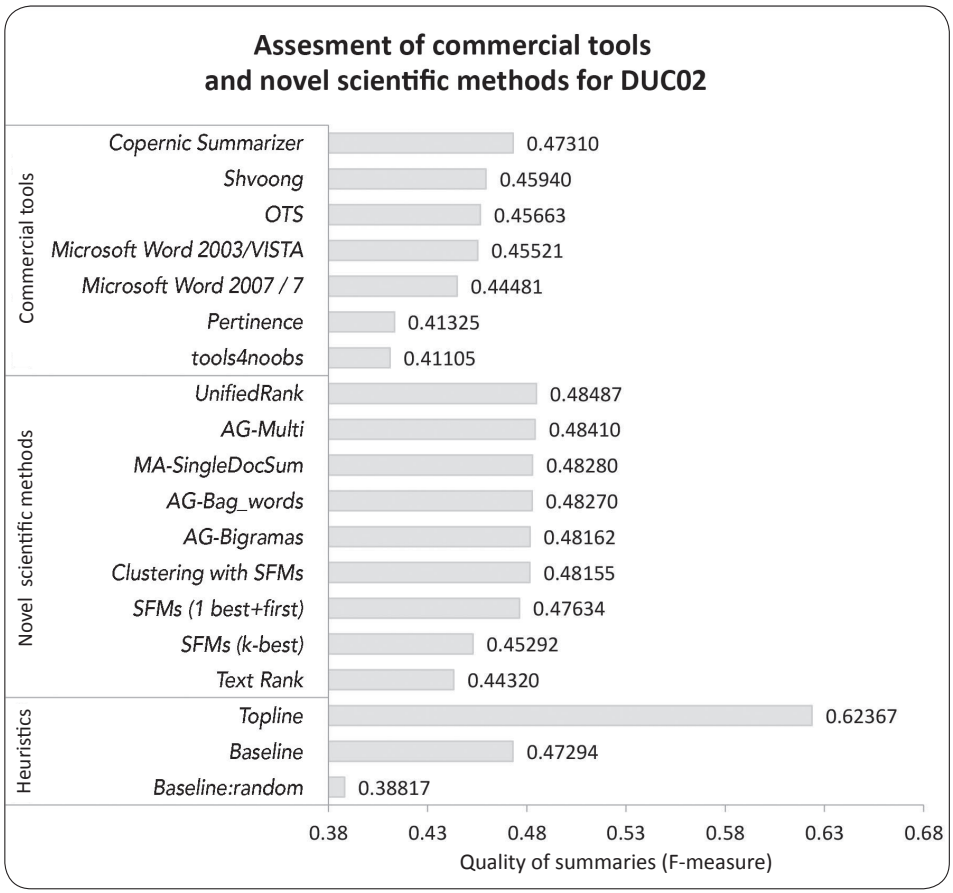
Graph 5.34 shows a comparison of novel scientific methods, commercial tools and heuristics for DUC02. The results are grouped by commercial tools, novel scientific methods and finally, heuristics.

Graph 5.33 Assessment of commercial tools and novel scientific methods for DUC01



For the Turing Test in English, Copernic Summarizer and Matias' method (2016) were chosen. Summary 1 was produced with Copernic Summarizer; while summary 3, following Matias (2016) (see, section 1.2).

Graph 5.34 Assessment of commercial tools and novel scientific methods for DUC02



Chapter VI

Automatic Summary Generation in Spanish

This chapter presents AGTS studies on the Spanish language in detail. Various conferences, workshops and corpora in this language are described. Special attention is paid to *corpus* TER, which is used to run tests in this language. As well, the results for the main heuristics, commercial tools and novel scientific methods tried in this regard are shown. Finally, there is a general comparison of these heuristics and methods, in addition to commercial tools tried with TER.

Spanish is the second most spoken language in the world, with about 477 million speakers. At present, it is spoken by 572 million people, either as native, second or foreign language. Spanish is spoken in countries such as Mexico, Colombia, Spain, Argentina, Peru, Venezuela, Chile, Ecuador, Guatemala, Cuba, Bolivia, Dominican Republic, Honduras, Paraguay, El Salvador, Nicaragua, Costa Rica, Panama, Puerto Rico, Uruguay and Equatorial Guinea. In addition to be used as

Table 6.1 Main languages spoken in the word

No.	Language	Countries	Speakers
1	Chinese	35	1302
2	Spanish	21	427
3	English	106	339
4	Arabic	58	267
5	Hindi	4	260
6	Portuguese	12	202
7	Bengali	4	189
8	Russian	17	171
9	Japanese	2	128
10	Lahndi	8	117
11	Javanese	3	84.3
12	Korean	7	77.3
13	German	26	76.9
14	French	53	75.9
15	Telugu	2	74.2
16	Marathi	1	71.4
17	Turkis	8	71.4
18	Urdu	6	68.6
19	Vietnamese	3	68
20	Tamil	7	67.8
21	Italian	13	63.4
22	Persian	30	61

an alternative language in countries such as Argelia, Australia, Brazil, Canada, United States, China, India, Israel, Japan, Norway, Russia, Switzerland, Turkey, among others; this is why, it has become the second language in communication at global level (Arévalo, 2017).

Spanish also holds an important share on the Internet and social media; at present, it is the third most used language on the web by numbers of users.

Table 6.2 Most used languages on the Internet[26]

No.	Language	Internet users
1	English	1052
2	Chinese	804
3	Spanish	337
4	Arabic	219
5	Portuguese	169
6	Hindi	168
7	French	134
8	Japonese	118
9	Russian	109
10	German	92
	Other	950

As displayed in **table 6.2**, English and Chinese surpass Spanish in numbers; though, if it is considered that, by and large, Chinese is spoken only by natives, Spanish would be the second language to communicate on the Internet (Arévalo, 2017).

Being AGTS research seventy years and Spanish one of the main languages at global level, few are the research works on this field. Particularly, the interest is to find out the present level of AGTS studies in Spanish.

[26] According to a study that reveals the languages most used on the Internet: https://www.internetworldstats.com/stats7.html

In chapter I, a Turing test was provided for the Spanish language (**table 1.1**). The results show that only 8% of the times, people correctly chose the man-made summary, while 56% confuses and choses a machine-made summary. However, in chapter I, there is only the distinction of machine- or man-made summaries, whereas among the first are those made by heuristics. To find out results for Spanish in a more specific manner, it is necessary to perform a complete and correct classification, since heuristics are utilized as references to assess AGTS tasks.

Table 6.3 displays the percentages of confusion human beings have in the selection of heuristics. In the first row it is noticed that 18% of the times people mistook *baseline:first* as the man-made summary, and 26% chose *baseline:first* and a machine-made summary as those made by humans. *Baseline:random* was also selected 15%; people chose this heuristic as the man-made summary. Moreover, 23% mistook *baseline:random* and the machine-made summary. The results obtained in these tests show that *baseline:first* has considerable correlation on humans.

Table 6.3 Results of Turing Test de Turing for *baseline* in Spanish

Pairs of summaries regarding *baseline heuristics*	Confusion percentage between selected summaries (%)
Human – *Baseline:first*	18
Human – *Baseline:random*	15
Machine – Máquina	13
Machine – *Baseline:first*	26
Machine – *Baseline:random*	23
Baseline:random – *Baseline:first*	5

6.1 Conferences, workshops and *corpora*

It is necessary to mention that for Spanish there are very few resources (conferences, workshops and *corpora*), for in spite of being one of the most spoken, works to create them for AGTS tasks in Spanish have not been furthered. Following, some *corpora* used to produce summaries in Spanish are presented, even though these have not been produced specifically for this task, but adjusted.

6.1.1 *Corpus* Desastres

The set of data Desastres [Disasters] comprises news items collected from three hundred different newspapers printed in Mexico. Each of the sentences was labeled using two basic labels, relevant and non-relevant (Téllez *et al*., 2009). This *corpus* was produced to work with information extraction systems. Albeit, it was used on an AGTS system in Villatoro (2006), in which one-hundred-word summaries were produced.

6.1.2 *Corpus* Concisus

Saggion and Szasz (2012) present a bilingual *corpus* of summary pairs, comparable in Spanish and English, about three sorts of events: air accidents, train accidents and earthquakes. It was manually made with semantic information on each event and is appropriate for experimentation in mono- and bilingual information extraction. It is worth underscoring that it is not focused on AGTS as such, but on information extraction by means of short summaries of events. This *corpus* is not labeled and does not have any *baseline* or *topline* measure.

6.1.3 *Corpus* utilized to assess and compare

The *corpus* utilized in this book to compare commercial tools and novel scientific methods was especially created for AGTS tasks in Spanish (Matias, 2016).

Textos en Español para Resúmenes, TER, [Text in Spanish for Summaries] is a collection of documents that comprises two hundred forty news items in Spanish. TER is composed of news items collected from Mexican newspaper Crónica on twelve categories. For each document in the collection, two summaries were made by two human experts.

Some of the criteria considered to build the *corpus* are:

- TER is created from items of news.
- It is specific for the Spanish language.
- It is intended for extractive generation of summaries.
- Single-document summaries.
- Digital news items.
- The length of the summaries has to be equal or greater than one hundred words.

To build the *corpus* twenty items of news were selected out of the following categories: academia, wellbeing, city, culture, sports, shows, states, national, business, editorials and society; totaling two hundred texts. One of the most important considerations to choose the items was that they were varied in length, though always more than one hundred words.

The TER categories are shown in **table 6.4**: the total of documents comprised in the *corpus*; the number of sentences per category; and, the average of words for each text. Generally, each text has 442 words and 14 sentences and an extension between one and two pages per news item.

6.2 Heuristics

To run the calculations of heuristics, commercial tools and scientific methods, *corpus* TER is used as it is the only available and specifically designed for AGTS tasks in Spanish.

Table 6.4 Parameters of the full texts in *corpus* TER

News-paper	Category	Number of texts	Number of words	Word average	No. of sentences	Sentence average
Chronicle	Academia	20	10966	548.3	382	19.1
	Wellbeing	20	11801	590.05	405	20.25
	City	20	7568	378.4	219	10.95
	Culture	20	8631	431.55	297	14.85
	Sports	20	9519	475.95	363	18.15
	Shows	20	8869	443.45	311	15.55
	States	20	7471	373.55	185	9.25
	World	20	7108	355.4	247	12.35
	Nacional	20	7533	376.65	186	9.3
	Business	20	7523	376.15	229	11.45
	Editorial	20	12716	635.8	443	22.15
	Society	20	6507	325.35	228	11.4
	Total	240	106212		3495	
	Average			442.55		14.5625

6.2.1 *BASELINE:RANDOM*

Table 6.5 se presents the results obtained by *baseline:random* for TER; it is worth mentioning ten runs were performed in order to certify the results obtained.

Table 6.5 *Baseline:random* results for TER

Measure	Recall	Accuracy	*F-measure*
ROUGE-1	0.4969	0.4980	0.4973
ROUGE-2	0.2930	0.2936	0.2933
ROUGE-SU4	0.3204	0.3208	0.3201

For *baseline:random* results tend to be low since the sentences are selected at random. For the state of the art, *baseline:random* is useful as a reference for the worst results.

6.2.2 *Baseline:first*

Table 6.6 shows the results obtained by *baseline:first* for TER.

Table 6.6 *Baseline:first* results for TER

Measure	Recall	Accuracy	*F-measure*
ROUGE-1	0.7233	0.7221	0.7226
ROUGE-2	0.6235	0.6224	0.6229
ROUGE-SU4	0.6332	0.6321	0.6326

As it is noticed, *baseline:first* results are very high, which displays that first sentences in this sort of texts are very important. For the Spanish language, this heuristic is a challenge to overcome, as its value is very high. The results shown in section 6.3 demonstrate that no commercial tool surpasses it; in section 6.4, it is shown that only a state-of-the-art method can do it.

6.2.3 *Topline*

For TER, *Topline* with *ROUGE-1* is: 0.8344 in *F-measure*. If a comparison is drawn with the value of *baseline:first*, it is noticed that the range to overcome is short. However, being Spanish a language on which AGTS is scarcely studied, it poses a challenge to overcome. *Topline* result was obtained by means of a genetic algorithm (Rojas J., 2017). **Table 6.7** shows the data obtained with various *ROUGE* combinations.

Table 6.7 *Topline* results for TER

Measure	Recall	Accuracy	*F-measure*
ROUGE-1	0.8369	0.8320	0.8344
ROUGE-2	0.7687	0.7642	0.7664
ROUGE-SU4	0.7672	0.7627	0.7649

6.3 Commercial tools

For the Spanish language, commercial tools tried with TER were also used. The length of the summaries made with this *corpus* is one hundred words, this way, those generated by the methods and tools must have as well one hundred words. To calculate the percentage that corresponds to the least number of words, the following formula is used

$$\frac{Number\ of\ desired\ words}{Total\ number\ of\ words\ in\ the\ document} * 100 \qquad (10)$$

For AGTS in Spanish the following commercial tools are used (**table 6.8**).

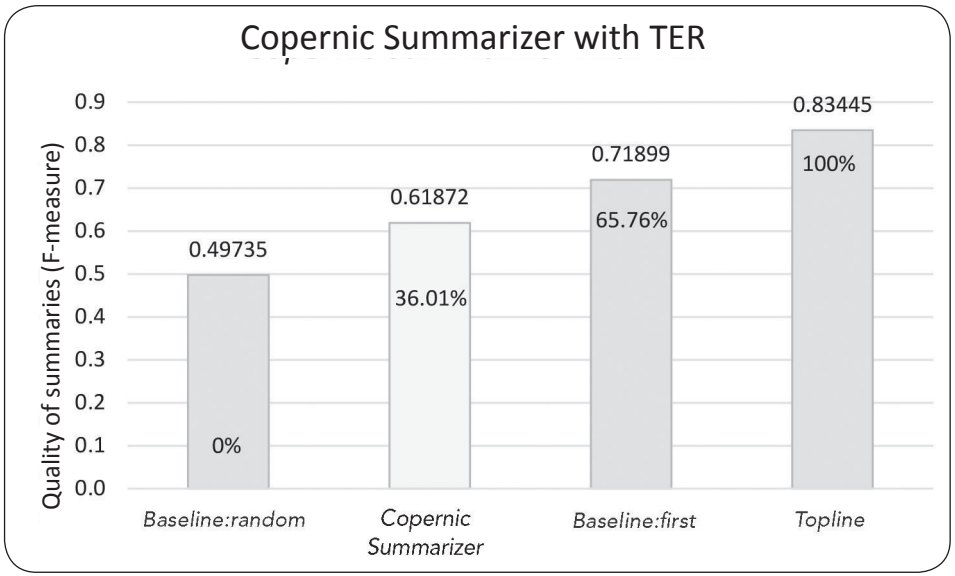
Table 6.8 List of tools tried in Spanish

Tool	Type	Spanish
Copernic Summarizer	Downloadable	✓
Microsoft Office Word 2003/2007	Downloadable	✓
OTS	Online	✓
Text Compactor	Online	✓
Summarizing	Online	✓
Total		5

6.3.1 Copernic Summarizer

Copernic Summarizer has the option to produce one-hundred-word summaries (length required by TER), so this option was selected. The results of this tool assessed by *ROUGE* are shown below.

Graph 6.1 Results of Copernic Summarizer for TER in comparison with the various heuristics



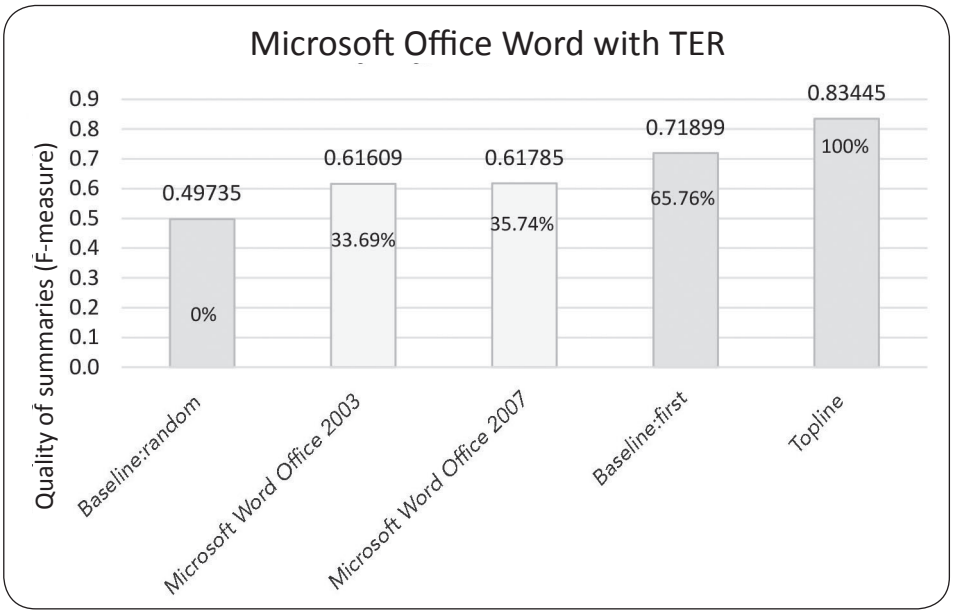
As it is noticed, Copernic results do not surpass *baseline:first* (graph 6.1).

6.3.2 Microsoft Office Word

Below the results of Microsoft Office Word assessed with *ROUGE* are presented.

In **graph 6.2**, we notice this tool does not surpass *baseline:first* and the difference between the 2003 and 2007 versions is not relevant regarding AGTS.

Graph 6.2 Results of Microsoft Office Word for TER in comparison with the various heuristics

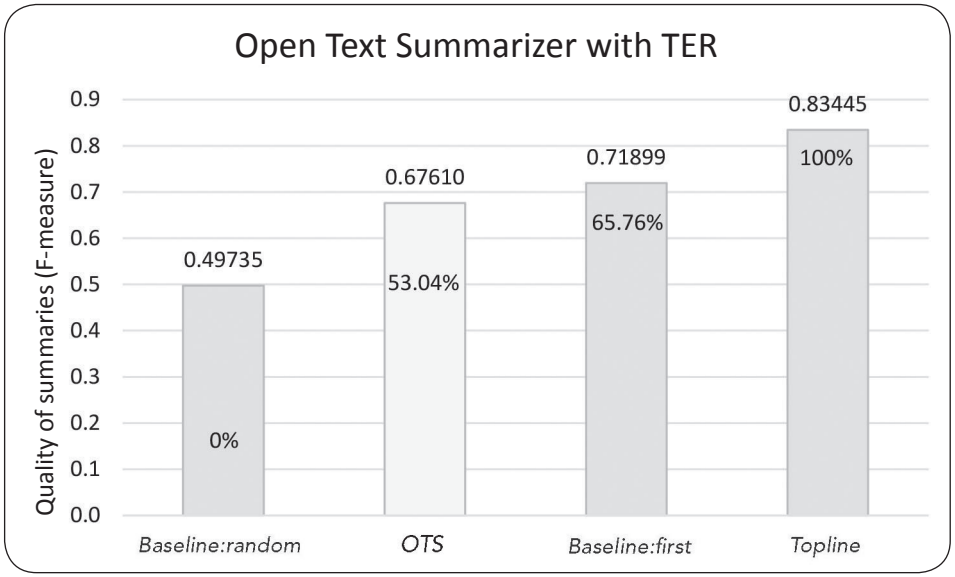


6.3.3 Open Text Summarization

For Open Text Summarizer (OTS) it is necessary to calculate the corresponding percentage so that each has one hundred words, as stated in TER specifications and to do so, formula 10 is used. Following, the results obtained by this tool assessed with *ROUGE*.

Graph 6.3 shows the results obtained with *Open Text Summarizer* with TER. As noticed, it does not surpass *baseline:first*. If *baseline:random* and *Topline* are considered the worst and best, then *Open Text Summarizer* obtains 53.04%, which so far corresponds to the lowest percentage in the novel scientific methods.

Graph 6.3 Results of Open Text Summarizer for TER in comparison with the various heuristics



6.3.4 Text Compactor

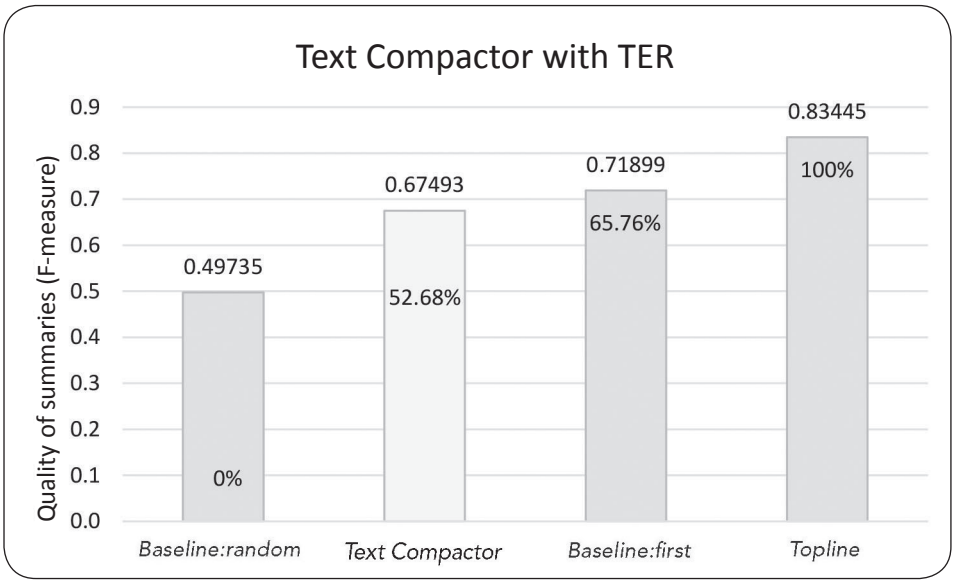
For *Text Compactor*, the corresponding percentage for each document has to be calculated so that they have one hundred words. The results of this tool assessed with *ROUGE* are displayed below.

Graph 6.4 shows the results of Text Compactor for TER; as noticed, it does not overcome *baseline:first*. However, the advance percentage regarding *baseline:random* and *Topline* is 52.68%.

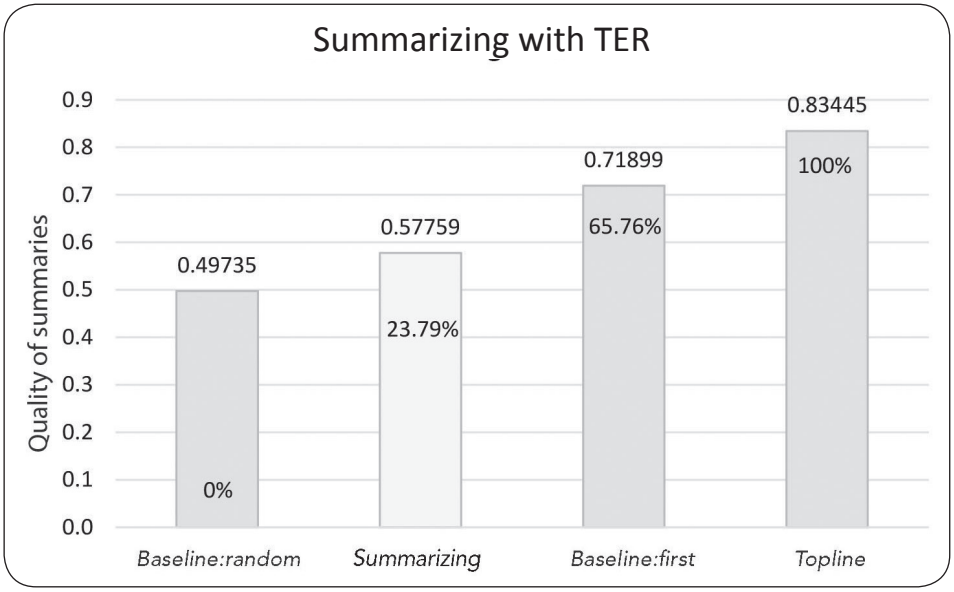
6.3.5 Summarizing

Summarizing has the option to generate one-hundred-word summaries (length required by TER), so this option was selected. The obtained results assessed with *ROUGE* are shown below.

Graph 6.4 Results of Text Compactor for TER in comparison with the various heuristics



Graph 6.5 Results of Summarizing for TER in comparison with the various heuristics



Graph 6.5 shows the results of *Summarizing* for *TER*. As noticed, the results obtained with this tool does not surpass *baseline:first*, and out of the tested tools it is the one with the lowest percentage. Considering *baseline:random* as the worst way to produce a summary, and *Topline*, the best, *Text Compactor* obtains an advance of 23.79%.

6.4 Novel scientific methods

For AGTS tasks in Spanish there are no formal research works with corpora and assessment tools that may be verifiable. However, it is important to be aware of the effort made. Some works and their own or adjusted corpora are described below.

6.4.1 Semantic graphs

Plaza's work (2011) comprises three case studies in which the design method is configured and utilized to generate various sorts of summaries of several domains and with very different structure and characteristics: biomedical scientific articles, news items and web pages of tourist information in Spanish.

The method used is based on the use of sematic graphs, which comprises the following stages:

- Preprocessing.
- Transition of sentences into concepts.
- Representation of sentences as degrees of concepts, construction of the document's graph.
- Concept clustering.
- Sentence assignation to clusters.
- Selection of sentences for the summary.
- Summary construction.

6.4.2 Automatic phrase compression

In Molina (2013), the automatic generation of summaries is proposed for the Spanish language considering the following text characteristics.

- Discursive segmentation consists in representing the document by means of a hierarchical tree that contains rhetoric/discursive information.
- The comprehension of phrases by elimination of discursive segments is based on the grammaticality of the resulting phrase, on its normativity (understood as the quality of retained important information) and on comprehension rate.
- Grammaticality consists in defining whether a phrase is correct or not.
- Normativity is based on the frequency of words.

In Molina's work (2013), two algorithms based on the characteristics above to generate automatic summaries are proposed. The first, by eliminating sections; the second, by eliminating segments with comprehension rate as argument. To experiment, Molina (2013) uses a *corpus* of his own that is not available.

6.4.3 Multiple-document summary generation

The work of Villatoro E. (2007) is based on a classifier and the use of supervised learning tools. The basic ideas with which the method works is that an inductive process automatically builds a classifier by means of observing the characteristics of a set of previously summarized documents, which gives the learning algorithm the pairs of documents; at once, these will comprise the original document or full text and the summary. In such manner that the problem of summary generation becomes a supervised learning activity.

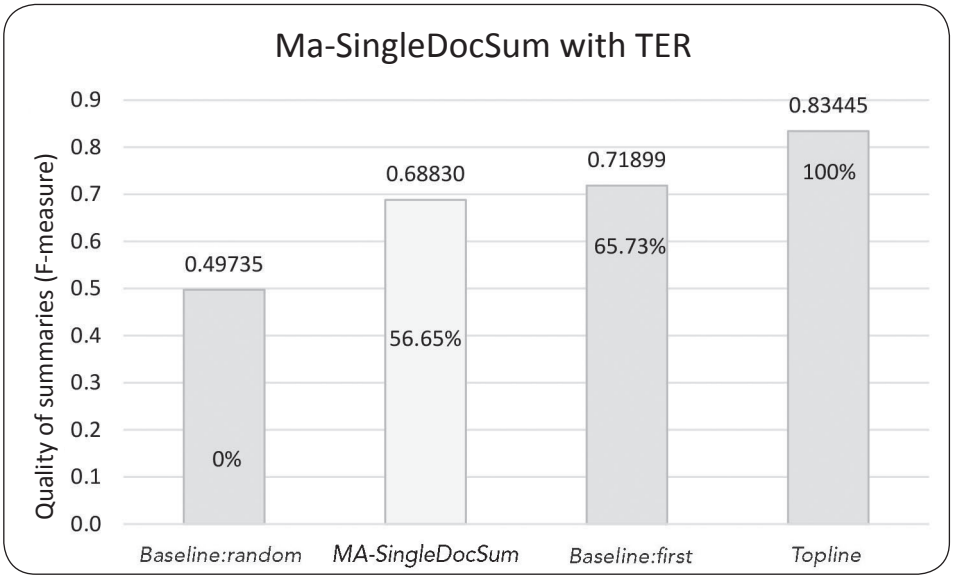
For experiments in Spanish language, *Desastres corpus* is used (Téllez *et al*., 2009); which was designed for classification and was adapted to AGTS.

To find out the state of research on AGTS, some of the best scientific methods used for the English language were tried. The results obtained are shown below.

6.4.4 Ma-SingleDocSum

It is one of the methods presented for English; the description was made in section 5.5.1.

Graph 6.6 Results of Ma-SingleDocSum for TER in comparison with the various heuristics



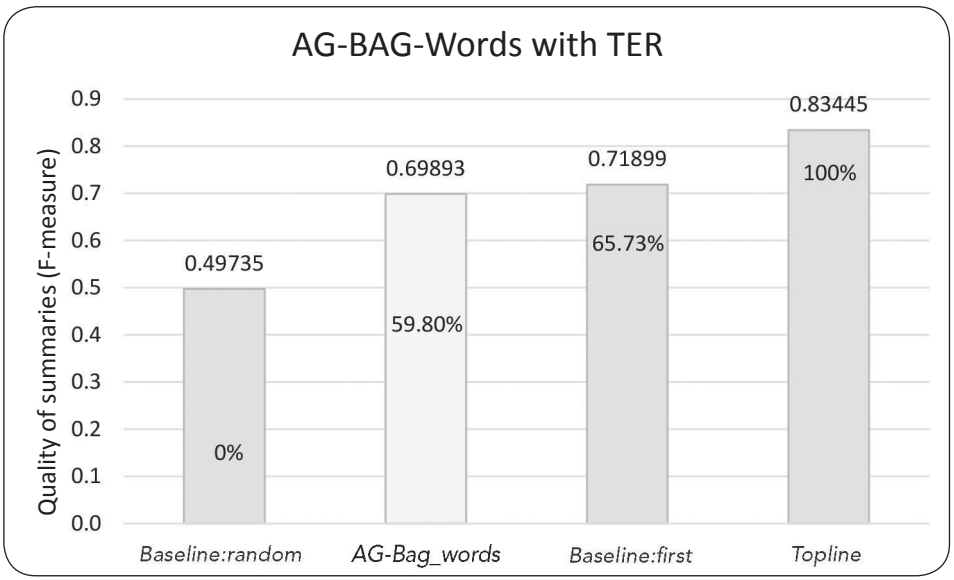
Ma-SingleDocSum for Spanish overcomes *baseline:random* (**graph 6.6**), though not better than *baseline:first* as in English.

6.4.5 AG-Bag-Words

AG-Bag-Words is a method based on a genetic algorithm only applied to English. However, due to its composition it may work with other languages, in this case Spanish. The description of this method was carried out in section 5.5.3.

AG-Bag-Words for Spanish does not surpass *baseline:first* (**graph 6.7**).

Graph 6.7 Results of AG-Bag-Words for TER in comparison with the various heuristics



6.4.6 AG-Multi

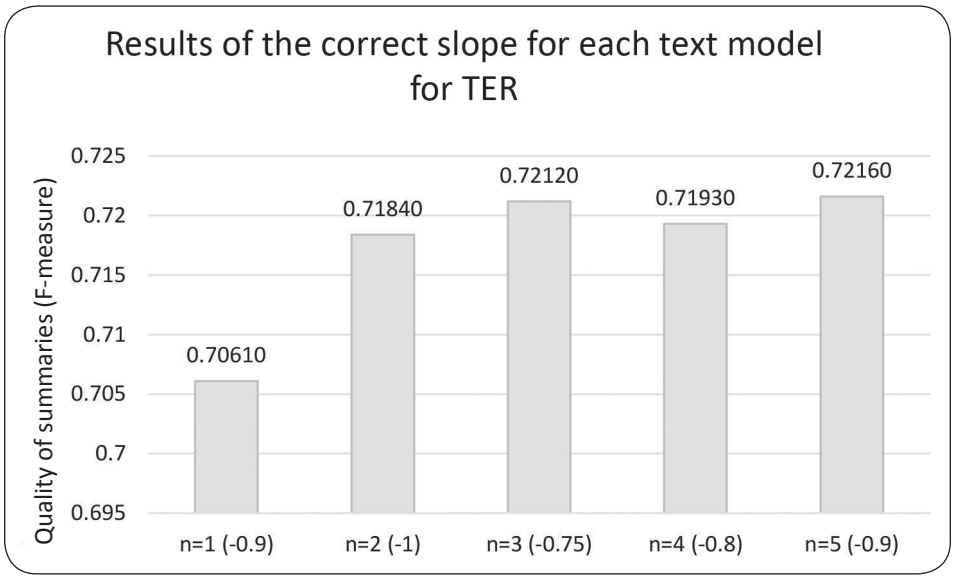
AG-Multi is a method based on a genetic algorithm applied to several languages. The description was made in section 5.5.5.

According to the stages of AG-Multi, proposed by Matias (2016), the results are the following.

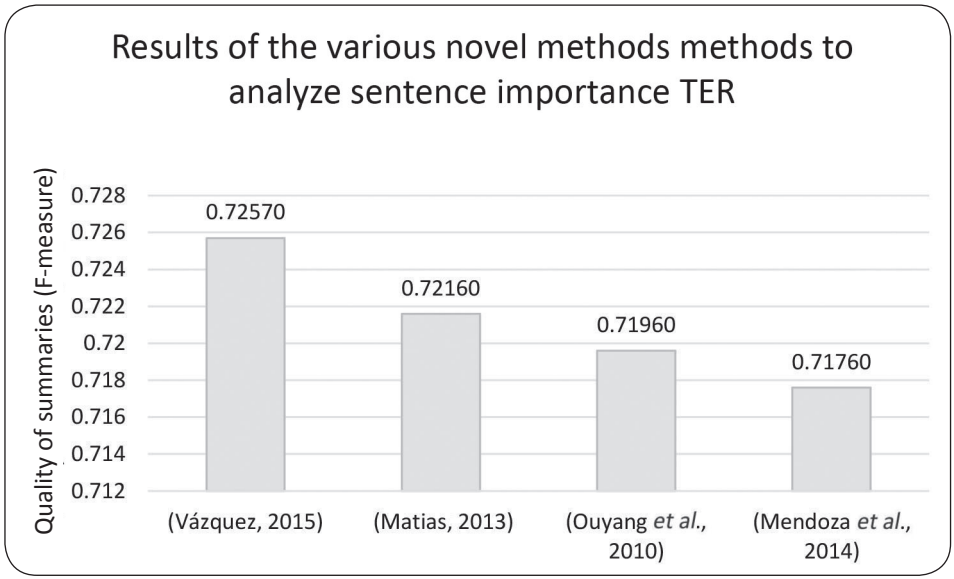
For the tests by text model, *n-grams* with $n = 5$ is the one with the best results for the Spanish languages and the best slope is $m = -5$ (**graph 6.8**). It is worth mentioning that for Spanish the best results were obtained without preprocessing and are presented below.

As in the case of the English language, an analysis of the novel scientific methods that resort to sentence position is run (**graph 6.9**).

Graph 6.8 Results of the correct slope for each text model for TER

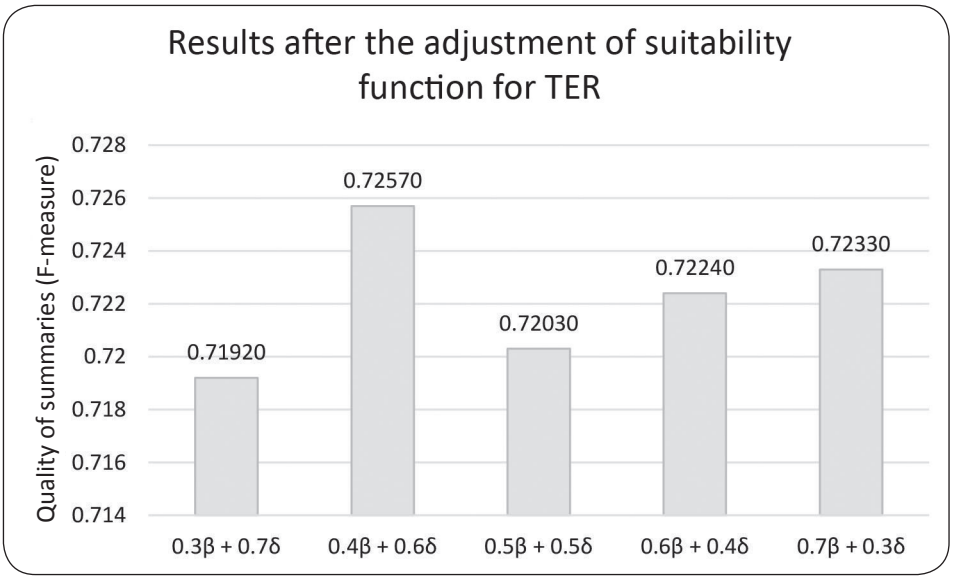


Graph 6.9 Results of the various novel methods to analyze sentence importance



For Spanish the adjustment of parameters and for TER is 0.4 for term frequency and 0.6 in sentence position (**graph 6.10**). This means that the position of sentences is more important.

Graph 6.10 Results after the adjustment of suitability function



Graph 6.11 shows the comparison of the results of AG-Multi in comparison with the various heuristics.

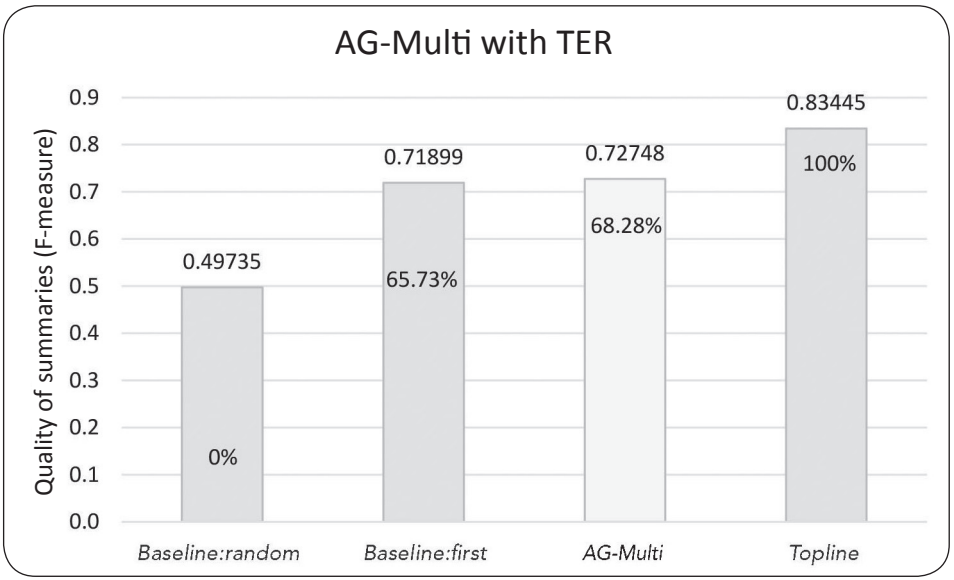
AG-Multi is the only method which thus far has surpassed the main heuristics: *baseline:random* and *baseline:first* for Spanish, with 67.75%.

6.4.7 TextRank

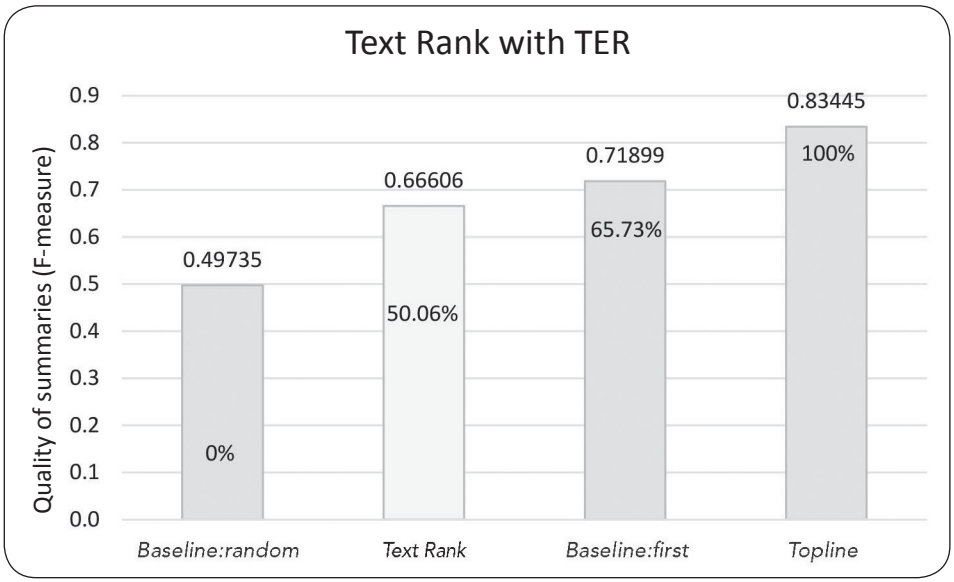
TextRank is a method based on graphs, applied to English and Portuguese. The description was made in section 5.5.6.

TextRank is one of the most used methods in AGTS for the English language. Besides being tried in English, it was used in Portuguese. This method is independent from language. For Spanish, it does not surpass *baseline:first*. Though, it does overcome *baseline:random* (**graph 6.12**).

Graph 6.11 Results of AG-Multi for TER in comparison with the various heuristics



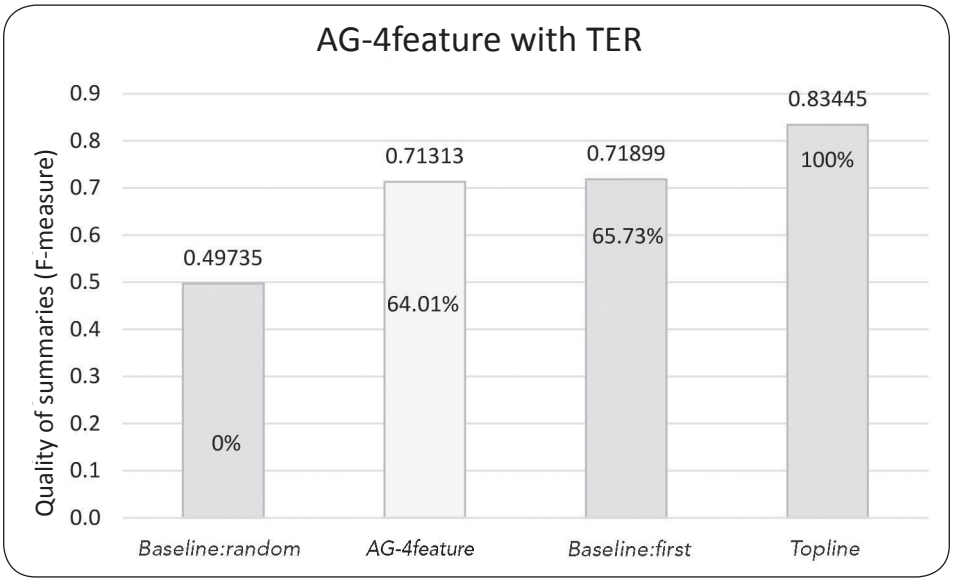
Graph 6.12 Results of Text Rank for TER in comparison with the various heuristics



6.4.8 AG-4feature

It is an AGTS method in English. The description was made in section 5.5.10.

Graph 6.13 Results of AG-4feature for TER in comparison with the various heuristics



AG-4feature does not overcome *baseline:first*. However, it is not far behind, as it has 1.72% (**graph 6.13**).

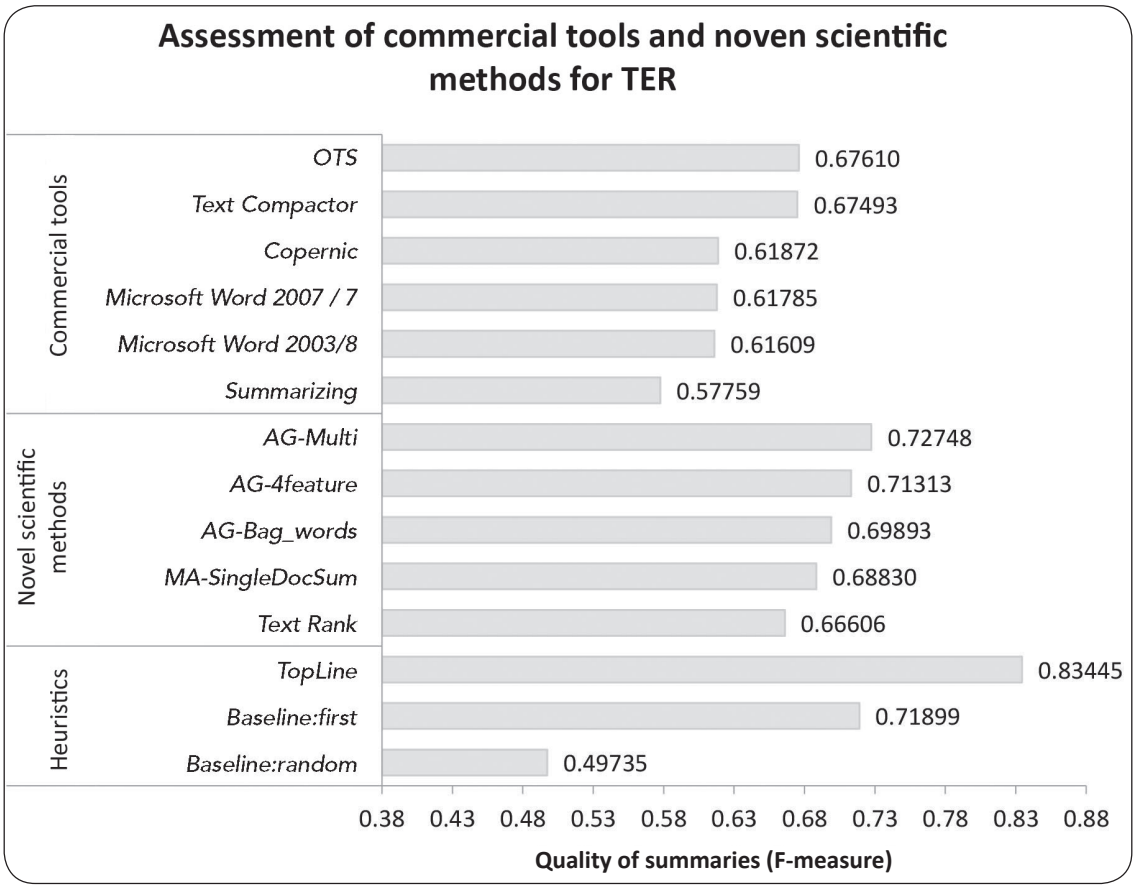
6.5 Results and analysis

As previously mentioned, resources are limited in Spanish. Though owing to TER, the available novel scientific methods can be tried with TER; those with the best results in English have been tested.

The first heuristic to be overcome by AGTS methods and tools is *baseline:random* and as noticed in **table 6.14**, all the methods and tools surpass

it. Specifically for Spanish, *baseline:first* is one of the highest. So far, only one novel scientific method surpasses this heuristic.

Graph 6.14 Assessment of commercial tools and novel scientific methods for TER



For the Turing Test carried out in Spanish, Microsoft Office Word 2007 and Matias' method (2016) were considered. Summary 1 corresponds to Microsoft Office Word, whereas Summary 6, to Matias' method (2016) (section 1.2).

Chapter VII

Automatic Summary Generation in Portuguese

This chapter deals with the thorough presentation of AGTS studies on Portuguese. The *corpus* utilized to run the tests in this language, TeMário is described, as well the main results of the main heuristics, commercial tools and novel scientific methods tried in this regard are displayed. Finally, a general comparison of the elements above worked with TeMário is provided.

Portuguese is the sixth most spoken language in the world; it has two hundred million native speakers in two hundred and two countries (**table 7.1**).

Table 7.1 Main languages spoken in the world

No.	Language	Countries	Speakers
1	Chinese	35	1302
2	Spanish	21	427
3	English	106	339
4	Arabic	58	267
5	Hindi	4	260
6	Portuguese	12	202
7	Bengali	4	189
8	Russian	17	171
9	Japanese	2	128
10	Lahndi	8	117
11	Javanese	3	84.3
12	Korean	7	77.3
13	German	26	76.9
14	French	53	75.9
15	Telegu	2	74.2
16	Marathi	1	71.4
17	Turkish	8	71.4
18	Urdu	6	68.6
19	Vietnamese	3	68
20	Tamil	7	67.8
21	Italian	13	63.4
22	Persian	30	61

At present, Portuguese holds the fifth place in Internet use, in addition to experiencing a significant growth in the use of social media such as Facebook and Twitter.

Table 7.2 Most used languages on the Internet[27]

No.	Language	Internet users
1	English	1052
2	Chinese	804
3	Spanish	337
4	Arabic	219
5	Portuguese	169
6	Hindi	168
7	French	134
8	Japanese	118
9	Russian	109
10	German	92
	other	950

When novel AGTS methods independent from language are developed and results for English and Spanish are shown, the question of how these methods would work in other languages arises; in this case we refer to Portuguese. As mentioned in previous chapters, most of the research and datasets are in English; as it is known, AGTS research tasks are about 60 years. For Portuguese, research works formally appeared by the turn of the XXI century with the work by Pardo (2003).

In this chapter we present a research on the main novel scientific methods and commercial tools applied to a *corpus* in Portuguese, setting them up independently from language with their later assessment.

[27]According to a study that ranks the languages most used on the Internet: https://www.internetworldstats.com/stats7.htm

Finally, the methods with the best results for the state of the art at international level are presented. An effort is made and the description of both the experimental and theoretical environment is presented to foster research on the methods and commercial tools for Portuguese.

7.1 Conferences, workshops and *corpora*

There are resources, projects and tools that are described at Centro Institucional de Lingüística Computacional; they can be accessed at the web (NILC, 2018).[28]

7.1.1 *Corpus* CSTNews

The *corpus* CSTNews (Aleixo and Pardo, 2008) is used in the automatic generation multiple-document summaries in Portuguese. Each of the collections is tagged in one of the following categories:

- Daily news items
- Global news items
- Sports
- Economy
- Politics
- Sciences

7.1.2 *Corpus* CSTNews-Update

CSTNews-Update (Cardoso *et al*., 2011) is a different configuration of CSTNews, which comprises fifty text collections with two or three related and which were retrieved from the main Brazilian news agencies.

[28] The Interinstitutional Center for Computational Linguistics. Research and development projects in Computational Linguistics and Natural Language Processing. Available at http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources. [Consulted on May 21st, 2018]

7.1.3 *Corpus* to assess and compare

In this book the experimentation with TeMário, whose name is a sort of compound noun for "**TE**xtos com su**MÁRIO**s". The *corpus* comprises one hundred newspaper articles and a summary for each made by the same news item writer (Pardo and Rino, 2003).

The main objective of the *corpus* is to compare the summaries generated by automatic systems with those made by humans. Moreover, it may serve for other automatic AGTS tasks, for example, the linguistic analysis of texts and summaries, the construction and composition of automatic summaries and the assessment of those made by experts with those automatically produced by the systems.

The use of the *corpus* can be extended to the areas of topic detection and information retrieval; nowadays, research is developed on how the experts recognize relevant information in a text to produce their summaries, or the identification of parameters that indicate the criteria to summarize with a view to building the model of computational systems (Pardo and Rino, 2003) and (Martins *et al*., 2001). The *corpus* is composed of news items on various topics. The length of the summaries has to be 25-30% of the source document.

Table 7.3 shows the structure of TeMário.

Table 7.3 General characteristics of TeMário

Newspapers	Sections	Number of texts	Number of words	Summaries' average words
Folha de São Paulo	Special	20	12340	617
	World	20	13739	686
	Editorial	20	10438	521
Jornal do Brasil	International	20	12098	604
	Politics	20	12797	639
	Total	100	61412	
	General averages		12282	613

7.2 Heuristics

The heuristics are calculated to provide a reference and a comparison of the novel scientific methods and commercial tools.

One of the problems in Portuguese is the flexibility in the range that can be considered for the generation of summaries, as it offers 25-30% of the original document's length for the summary. Additional to the range offered for the summaries' length, the *corpus* is neither tagged nor separated in sentences, which makes it difficult to reach an agreement in the various heuristics. In the tests carried out for the book, a 30% of the original and the *corpus*, separating the text in sentences, were utilized.

7.2.1 *Baseline:random*

To calculate this heuristic for Portuguese the sentences are taken at random and a summary is produced up to reaching 30% of each document, with a view to meeting the requirements of TeMário. For this heuristic the value reported by Matias (2016) is 0.4574.

7.2.2 *Baseline:first*

As previously mentioned, owing to the flexibility of the *corpus* to decide on the summaries' length, there are state-of-the-art works with different values for this heuristic, so the length considered in the test is supposed to be a probable cause.

For *baseline:first*, with a 30-percent length, the value obtained by Matias (2016) is 0.4846. However, in Mihalcea (2005), the value is 0.4963, due to the additional preprocessing carried out.

7.2.3 *Topline*

Topline was assessed in Rojas J. (2017) using genetic algorithms. It is calculated by means of two approaches: sentence and paragraph combination as parts of the text to combine.

The best parameters of the AG put forward by Rojas J. (2017) are presented in **table 7.4**.

Table 7.4 AG parameters to calculate *Topline* with TeMário

Experiment	Elite	Generations	Individuals	Selection		Crossing	Mutation	
				Sort	P		Sort	P
1	Si	30	150	Tour-na-ment	3	CX	Inser-tion	8

Table 7.5 shows the results of *Topline* for TeMário in Portuguese.

Table 7.5 Results of *Topline* for TeMário

Measure	Recall	Accuracy	*F-measure*
ROUGE-1	0.6450	0.6059	0.6235
ROUGE-2	0.3328	0.3141	0.3225
ROUGE-SU4	0.3274	0.3078	0.3166

7.3 Commercial tools

For AGTS in Portuguese the following commercial tools were utilized (**table 7.6**). To carry out the assessments TeMário was used at a 30-percent length.

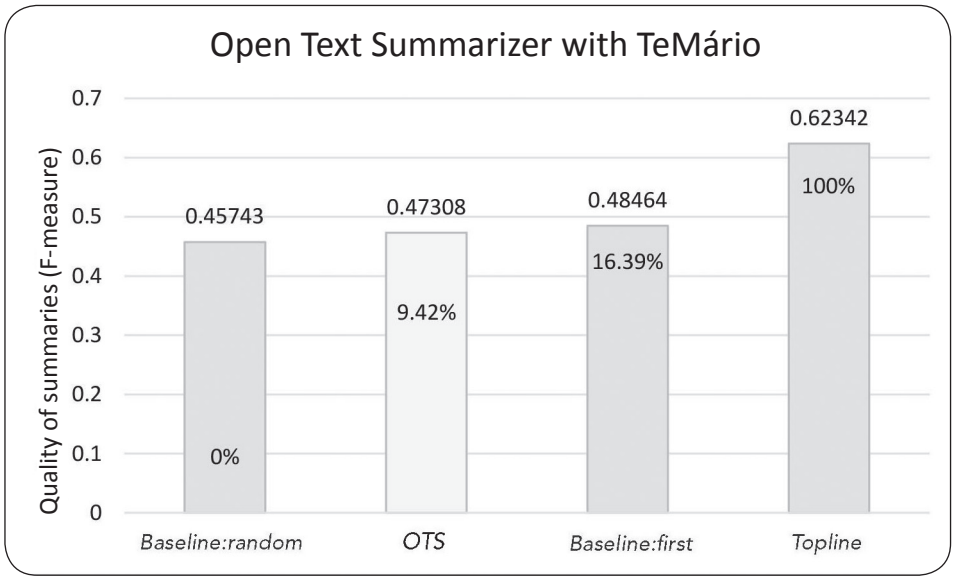
Table 7.6 Commercial tools assessed in Portuguese

Tool	Sort	Portuguese
OTS	Online	✓
Shvoong	Online	✓
Total		2

7.3.1 Text Summarizer

In this tool the percentage required for the summary can be chosen. The length selected was 30% for each document in the collection. Below, the results for TeMário assessed with *ROUGE* are displayed.

Graph 7.1 Results of Open Text Summarizer for TeMário in comparison with the various heuristics

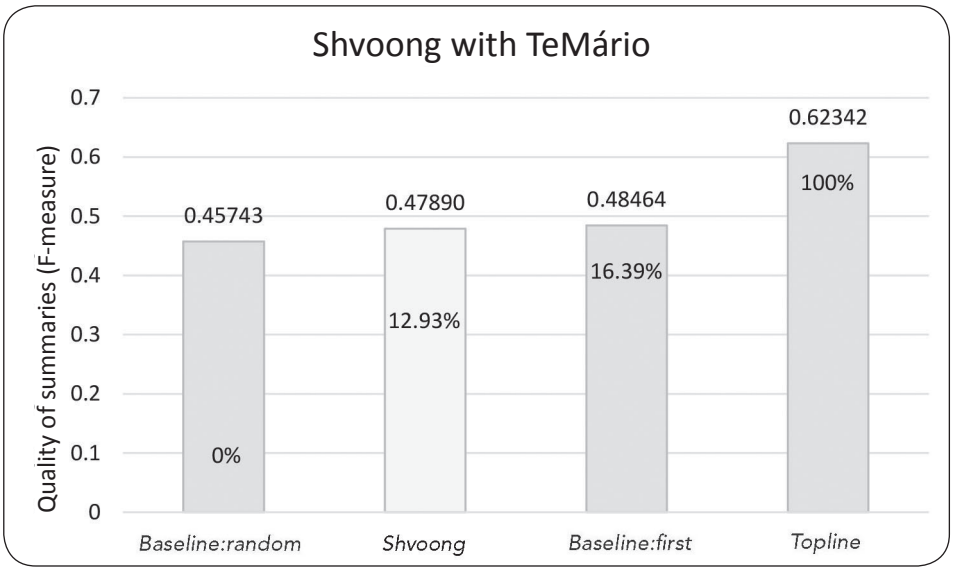


Graph 7.1 shows the results of Open Text Summarizer for TeMário. As noticed, this tool does not surpass *baseline:first*. If *baseline:random* is considered the worst way to produce a summary, while *Topline*, the best, the maximum value that can be obtained, then Pertinence Summarizer reaches 9.43%.

7.3.2 Shvoong

Following, the result obtained with Shvoong for TeMário assessed with *ROUGE* are displayed.

Graph 7.2 Results of Shvoong for TeMário in comparison with the various heuristics



Graph 7.2 shows the results of Shvoong for TeMário. As noticed, this tool does not surpass *baseline:first*. Though, Shvoong advances 12.92% in AGTS tasks regarding *baseline:random* and *Topline*.

Among the NILC projects, a number of systems have been developed to generate and assess AGTS methods such as GistSumm (Pardo *et al*., 2003), NeuralSumm (Pardo *et al*., 2003b), DMSumm (Pardo, 2002), SuPor (Modolo, 2003) and UNLSumm (Martins, 2002). Among those, the most important is the one proposed by Pardo, GistSumm; developed in 2003, it is constantly updated and available for free on the web up to the present (Pardo *et. al*., 2003a).

Following, some of the tools available for Portuguese are described; these were not tried in this work, though are available for use. Among them we find Rsumm, ViSum and NILC-WISE.

RSumm[29]

It is an online tool that solves the task of automatically generating the summaries from a specific search on Google news. It groups all the information gathered by the user with the possibility to add more items of news or take some out of the summary, considering the users' needs.

ViSum[30]

It is a system to visualize the summaries made for the task of multiple-document automatic generation.

NILC-WISE[31]

It is an application with a web interface developed in NILC with a view to provide researchers with a way and a repository to assess their automatic summaries.

7.4 Novel scientific methods

In this section novel scientific methods used in the task of automatic generation of text summaries in Portuguese are presented. The first two in **table 7.7** are not available, so they were not tried for TeMário.

[29]The extension online summarizing RSumm News was consulted at: http://conteudo.icmc.usp.br/pessoas/taspardo/sucinto/RSumm%20News%20-%20Tutorial/home.html

[30]Consulted on May 15th, 2018, from http://conteudo.icmc.usp.br/pessoas/taspardo/

[31]NILC-WISE – Web interface to assess summaries. http://nilc.icmc.usp.br/nilcwise/login

Table 7.7 Novel scientific methods assessed in Portuguese

Tool	Portuguese
SuPor	
SaBio	
GistSumm	✓
AG-Multi	✓
TextRank	✓
Total	3

7.4.1 SuPor

SUmmarizer for PORtuguese (SuPor) is a system based on a learning machine (Modolo, 2003). This way, it has two different processes. Training and extraction based on Naive-Bayes' method. This allows combining linguistic and nonlinguistic features. The characteristics considered by SuPor to generate summaries are sentence length (at least 5 words), word frequency, phrase signalization, sentence location and occurrence of proper nouns. The functioning of SuPor is described here: firstly, the set of characteristics of each sentence is extracted; secondly, the Bayesian classified is applied to each set, which defines the probability that the sentence is included in the text. Those with higher probabilities will become parts of the summary.

7.4.2 *SaBio*

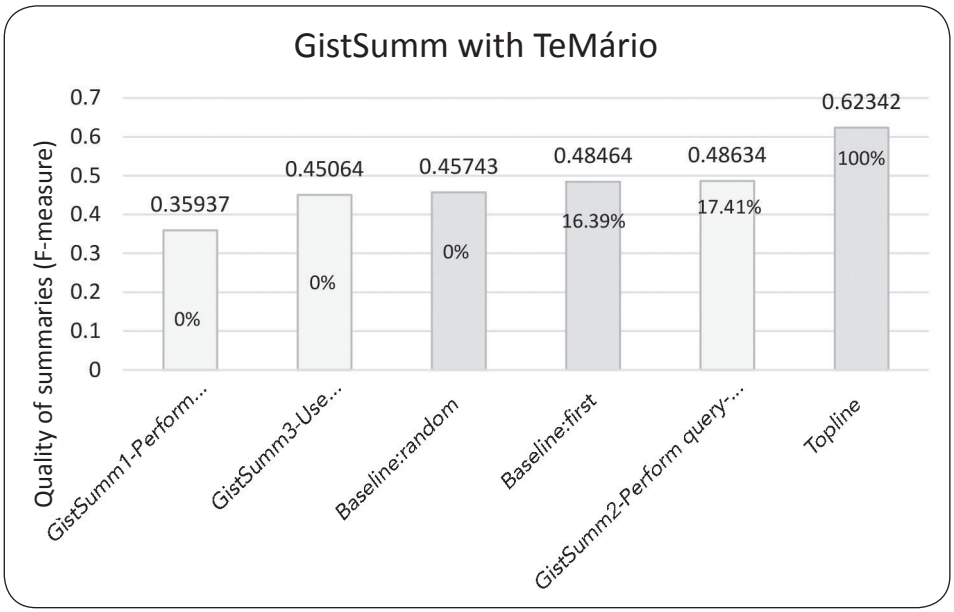
Automatic *Summarizer* for the Portuguese language with more biologically plausible connectionist architecture and learning, or SaBio, is based on a neural network trained with news items from TeMário (Orrú *et al*., 2006). This application considers the following characteristics: sentence size; position of the sentence in the text; position of the sentence in its corresponding paragraph; presence of keywords; sentence value regarding the distribution of words in the text and term frequency.

7.4.3 GistSumm

GistSumm is an automatic summarizer supported on an integration method called *gist-based* (Pardo *et al*., 2003). It comprises three stages: text segmentation; sentence ranking; and, summary generation. The ranking of sentences is based on Luhn's (1958) method, which uses keywords; it weighs each sentence of the original text resorting to the frequency of words, and the key ones have a heavier weight. The summary is produced after considering the correlation between the keywords and the relevance they have in relation to the text content.

The selected length was 30% for each document in the collection. The results of this tool for TeMário assessed with *ROUGE* are presented below.

Graph 7.3 Results of *GistSumm* for TeMário in comparison with the various heuristics



The method of GistSumm has three configurations (GistSumm1-Perform-Intrasetial summarization; GistSumm2-Performquery-based summarization and GistSumm3- Use averagekeywords ranking method) of which GistSumm1 and

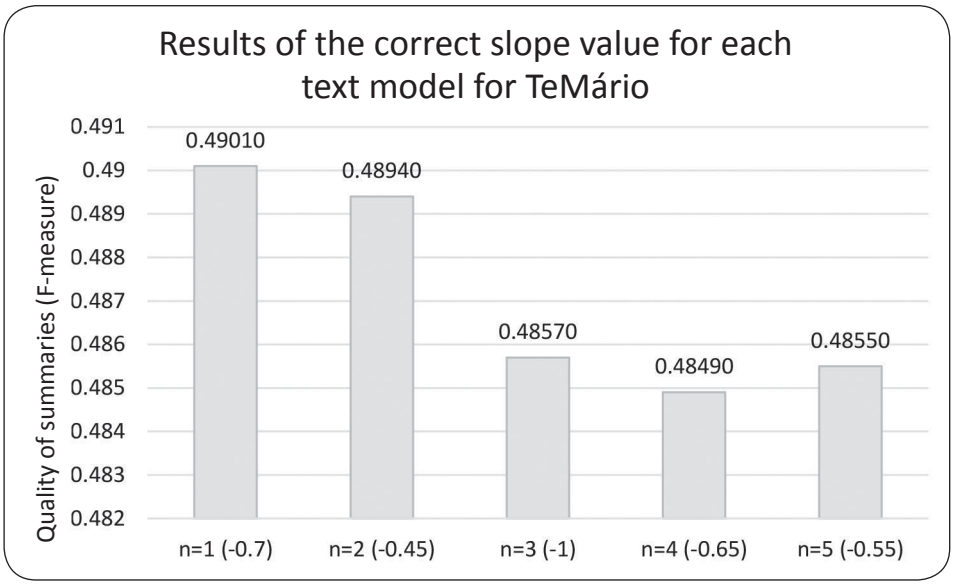
GistSumm3 do not overcome *baseline:random*. Though the configuration for GistSumm3 surpasses both *baseline:random* and *baseline:first*.

7.4.4 AG-Multi

In section 5.5.5 the method proposed by Matias (2016) is discussed. The tests carried out on TeMário using it were at a summary length of 30%. The results of this tool for TeMário assessed with *ROUGE* are displayed below.

Graph 7.4 shows the results obtained by the *n-grams* model.

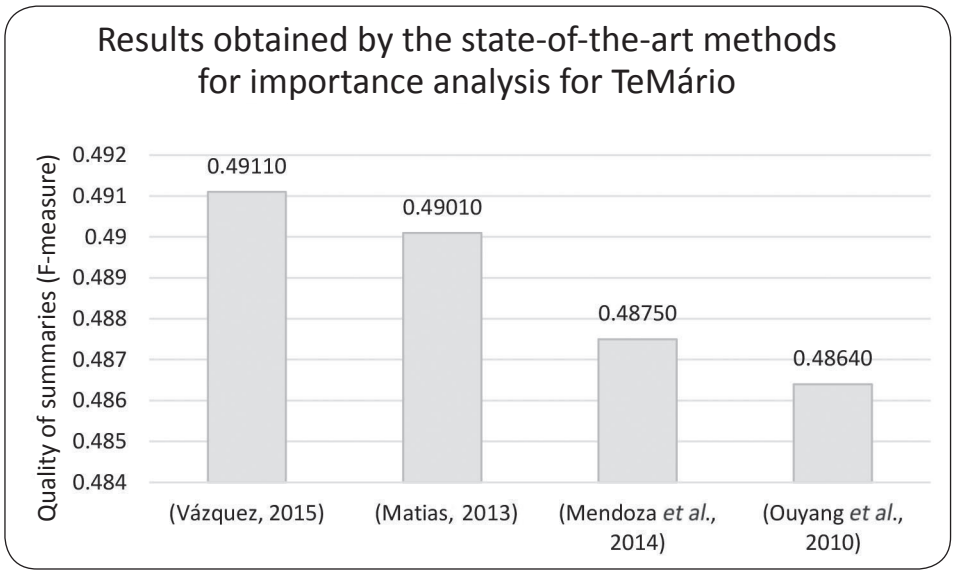
Graph 7.4 Results of the correct slope value for each text model for TeMário



The best text model for TeMário is bag of words .

Graph 7.5 shows the results of the analysis carried out on the novel scientific methods to find out the best way to calculate the characteristic of sentence position.

Graph 7.5 Results obtained by the state-of-the-art methods for importance analysis for TeMário



As noticed, the formula proposed by Vázquez (2015) is the one that yields the best results for TeMário.

If the two characteristics used by Matias' (2016) method, i.e., term frequency and sentence position, are used as fitness function, then both have the same importance in TeMário.

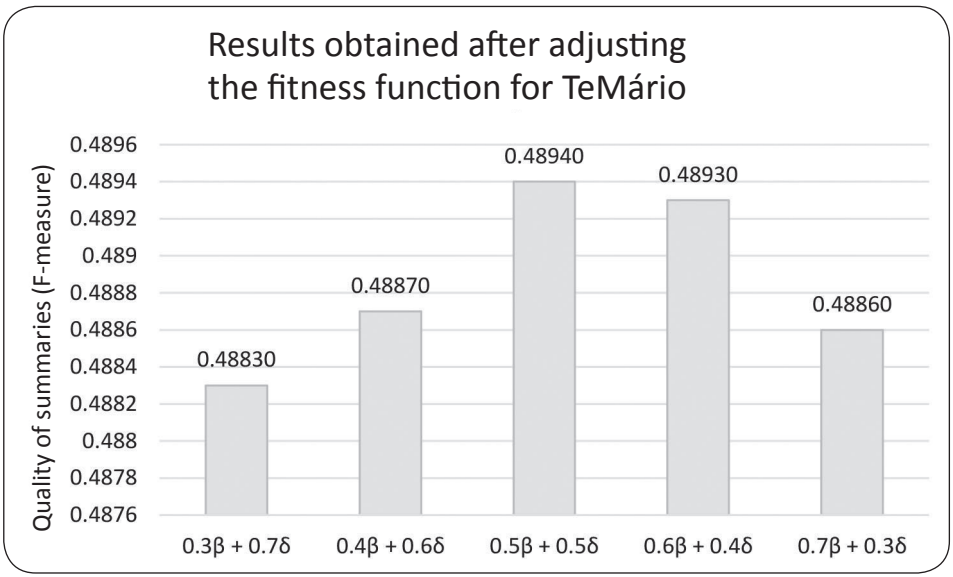
Considering the results in **graphs 7.6** and **7.7**, it may be stated that the characteristics used in the fitness function are equally important; this way, the result chosen for the comparison is that in **graph 7.5**.

As noticed, AG-Multi surpasses *baseline:random* and *baseline:first* (**graph 7.7**).

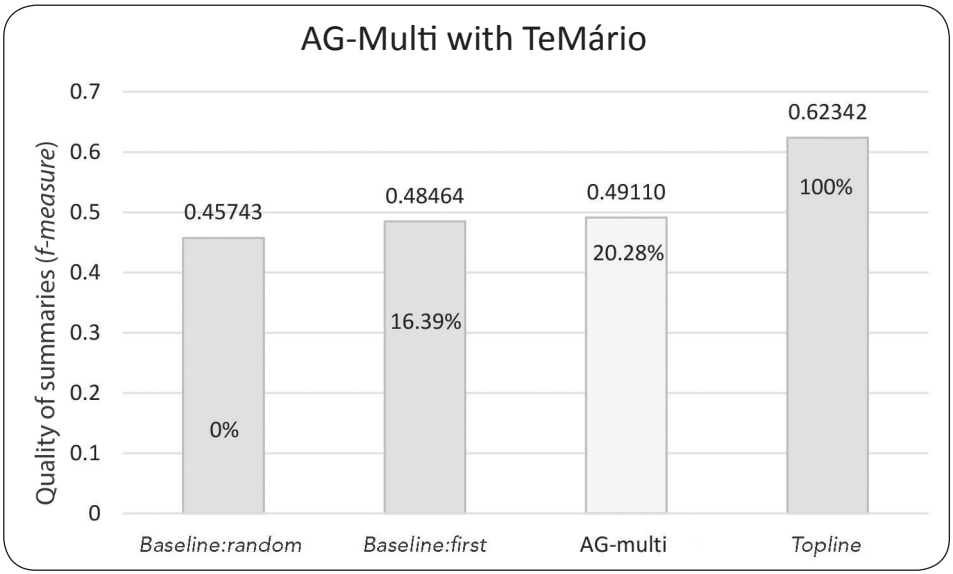
7.4.5 TextRank

It is a method based on graphs proposed by Mihalcea (2004), which uses the following algorithms.

Graph 7.6 Results obtained after adjusting the fitness function



Graph 7.7 Results of AG-Multi for TeMário in comparison with the various heuristics



PageRank

It is one of the most popular classification algorithms; it was designed as a method to analyze web links. Unlike other classification algorithms based on graphs, it integrates the inbound and outbound links in a single model, thereby it produces only one set of results (Brin and Page, 2012). PageRankW method adds a weight between the vertices of the graph produced. This way, the algorithm classification is adapted to include weighed edges.

HITS

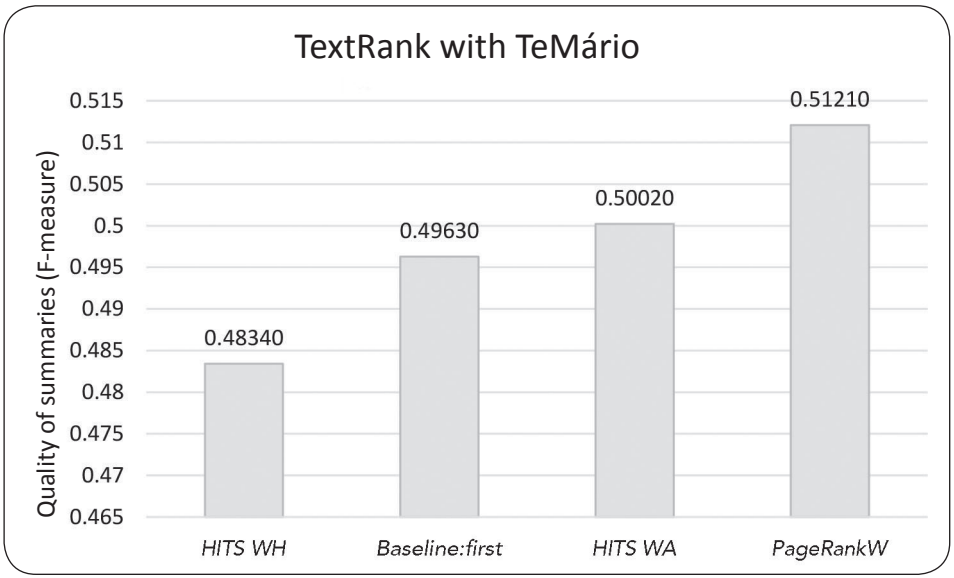
HITS (Hiperlinked Induced Topic Search) performs a search of topics of induced links. It is an iterative algorithm designed to classify web pages according to their "authority" degree. Moreover, it makes a distinction between "*authorities*" (pages with a large number of inbound links) and "hubs" (pages with a large number of outbound links) (Kleinberg, 1999). HITS assigns two values to each vertex: one of "*authority*" and the other of "*hub*". The method HITSW adds a weight between the vertices of the graph produced. This way, the algorithm classification is adapted to include weighed edges.

PageRankW and HITS were tried for TeMário by Mihalcea (2005), who consider the value of 0.4963 for *baseline:first*. Though the author does not mention the parameters used to reach the result; hence, Matias (2016) tried to repeat the experiment, but the value presented by Mihalcea could not be reached. Then, the results of the two authors cannot be compared, so they are presented separately. **Graph 7.8** shows the data obtained by Mihalcea (2005) for *baseline:first*, *PageRankW* and *HITS*.

7.5 Results and analysis

Graph 7.9 shows the results of the experiments in Portuguese using commercial tools and novel scientific methods. The experiments were carried out with TeMário and assessed with *ROUGE*. For TeMário, the possible extension of the summaries ranges from 25 to 30%. The results of the experiments displayed were produced at a 30-percent extension.

Graph 7.8 Results of TextRank using the various configurations of PageRank and HITS for TeMário

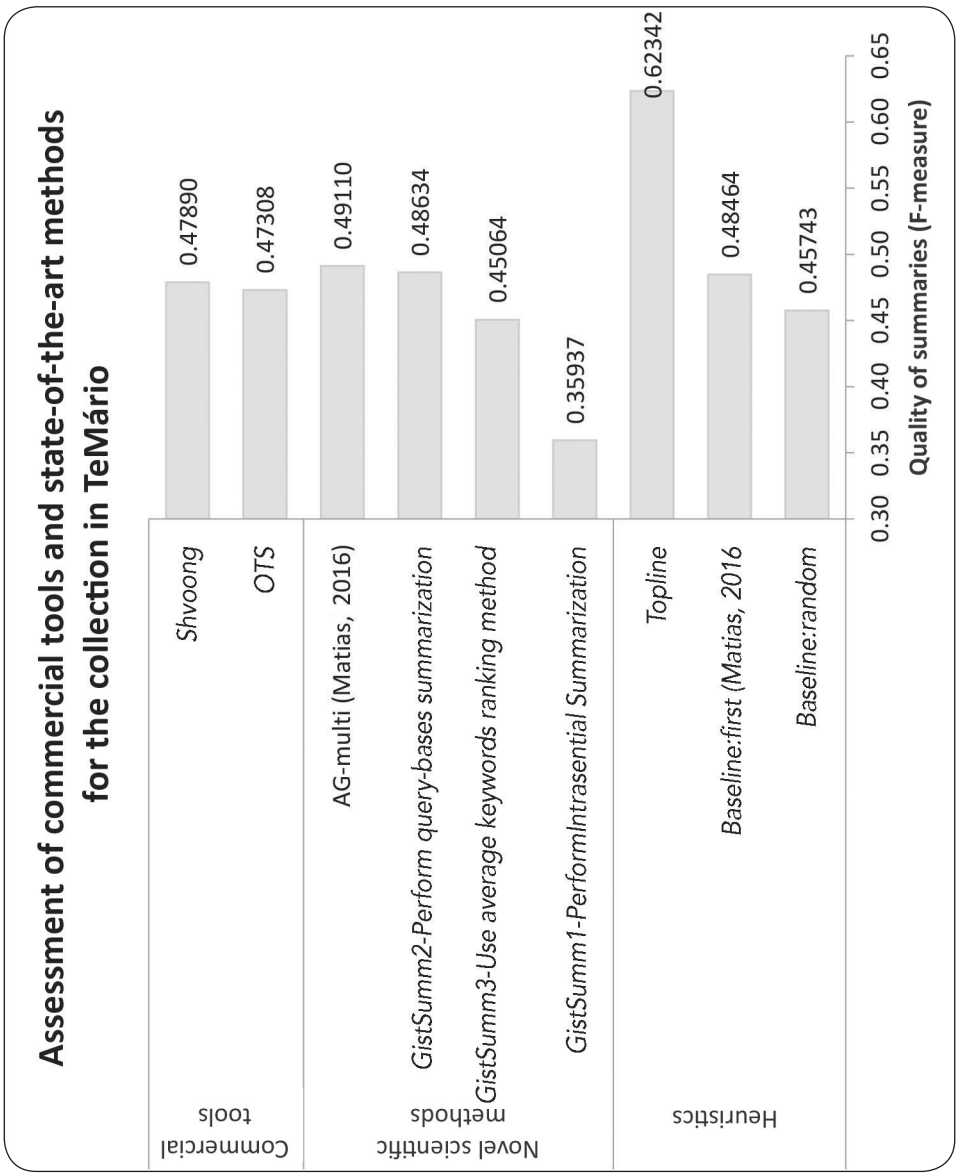


There are some state-of-the-art works that test TeMário; albeit, as it has an ample range for the summary length, it is more complicated to define a standard *baseline*, so for the present work *baseline* was obtained considering a length of 30%.

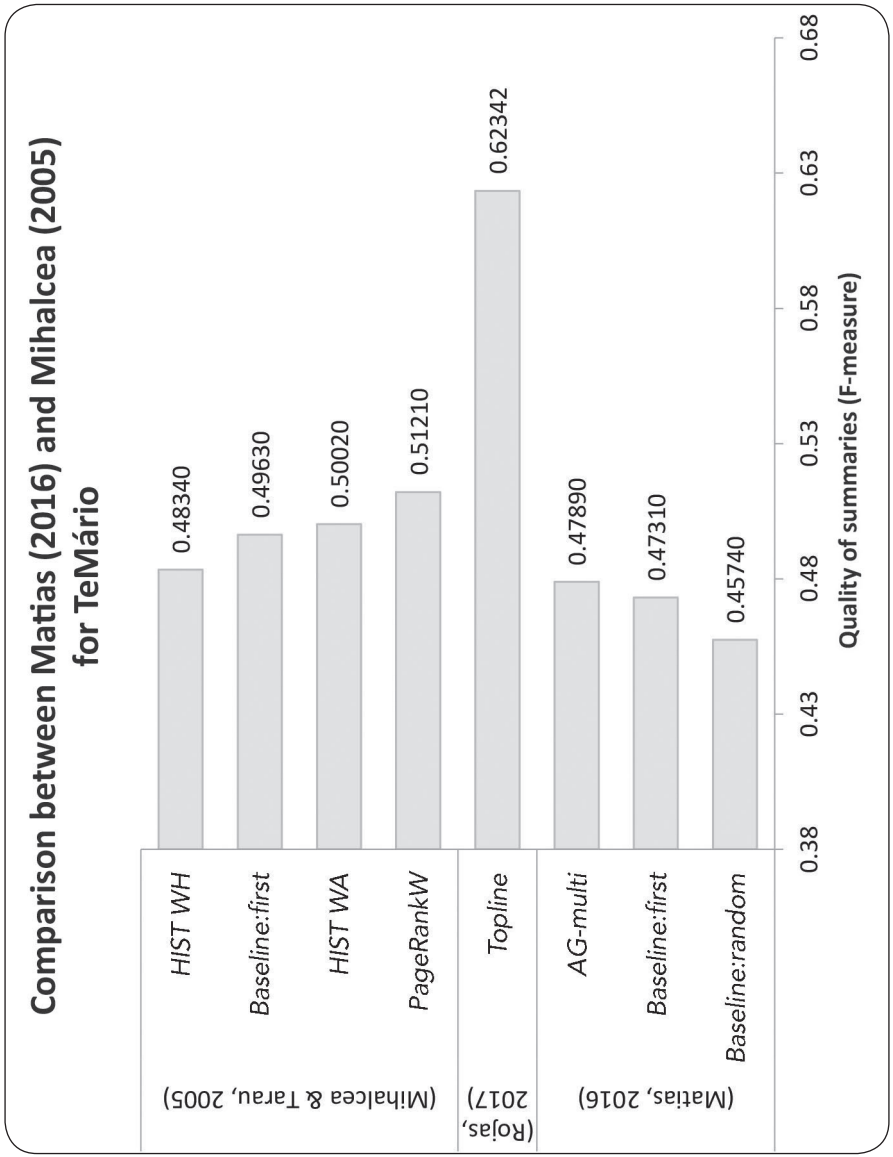
As observed in **graph 7.10**, Matias' (2016) method surpasses all online commercial tools and obtains the best results for the state of the art.

The work by Mihalcea and Tarau (2005), part of the state of the art, tries TeMário. Mihalcea's results surpass these results. However, a problem appears in the comparison because the extension range used for the collection is not well defined; thus, in order to verify the extension used by Mihalcea, the *baseline* result was obtained for TeMário in the range established by the collection (25% - 30% of the original document extension). In spite of trying all the possible extensions from 25 to 30%, Mihalcea's *baseline* result was not reached. In this regard, it is thought that the problem is the way sentences are considered.

Graph 7.9 Results of the novel scientific methods and commercial tools for the collection in Portuguese



Graph 7.10 Comparison between the present work and Mihalcea and Tarau (2005)



Graph 7.10 shows a comparison between the results of Mihalcea and Tarau (2005) and those of the methods proposed in this work (in which *baseline* was calculated at 30%).

As observed in **graph 7.10**, the results of *baseline* are different. As well, both Matias' (2016) and Mihalcea's (2005) methods surpass their proposed *baseline*.

Chapter VIII

Automatic Summary Generation in Russian

This chapter thoroughly presents the study of AGTS tasks in Russian. The *corpus* TEXTRUSS is described, as it is utilized to run the tests in this language. Also, the results of the main heuristics, commercial tools and novel scientific methods used in this regard. Finally, a general comparison of such elements tried with TEXTRUSS is presented.

Russian is an Indo-European language spoken by more than one hundred and seventy million native speakers (**table 8.1**). It holds the eighth place in the most spoken languages in the world.

Table 8.1 Main languages spoken in the world

No.	Language origin	Countries	Speakers
1	Chinese	35	1302
2	Spanish	21	427
3	English	106	339
4	Arabic	58	267
5	Hindi	4	260
6	Portuguese	12	202
7	Bengali	4	189
8	Russian	17	171
9	Japanese	2	128
10	Lahndi	8	117
11	Javanese	3	84.3
12	Korean	7	77.3
13	German	26	76.9
14	French	53	75.9
15	Telugu	2	74.2
16	Marathi	1	71.4
17	Turkish	8	71.4
18	Urdu	6	68.6
19	Vietnamese	3	68
20	Tamil	7	67.8
21	Italian	13	63.4
22	Persian	30	61

However, according to internet users, Russia holds the ninth place (**table 8.2**). Nevertheless, at present, a growth is noticed in relation to Internet use, which

places Russia second (Europe Internet Stats - Population Statistics, 2017).

Table 8.2 Languages most used on the Internet[31]

No.	Language origin	Internet user
1	English	1052
2	Chinese	804
3	Spanish	337
4	Arabic	219
5	Portuguese	169
6	Hindi	168
7	French	134
8	Japanese	118
9	Russian	109
10	German	92
	Other	950

AGTS for Russian language has not reported significant progress in the state of the art over these sixty years of research. This way, Russian is an opportunity window to approach its study.

8.1 Conferences, workshops and *corpora*

There are neither conferences nor workshops that deal with AGTS tasks. However, owing to the importance it has and the growth in the number of internet users, various works such as those by Rojas (2016) and Hernández (2018) have been produced; such studies deal with novel scientific methods and commercial tools for summary production. The *corpus* used for the summaries produced in Russian is described below.

[32]According to a study that discloses the most used languages on the Internet: https://www.internetworldstats.com/stats7.htm

8.1.1 *Corpus* utilized to assess and compare

TEXTRUSS comprises news items with their corresponding summaries, which were made by experts in Russian who, for the case of this *corpus*, were the same journalists who wrote the articles and selected the most important sentences.

The news items were downloaded from news portal gazeta.ru ("Главные новости - Газета.Ru," 2015). This *corpus* comprises various domains in eleven categories organized as follows (Name in Russian and its translation to English):

1. ПОЛИТИКА (Politics)
2. БИЗНЕС (Business)
3. ОБЩЕСТВО (Society)
4. МНЕНИЯ (Opinion)
5. КУЛЬТУРА (Culture)
6. НАУКА (Science)
7. ТЕХНОЛОГИИ (Technology)
8. НЕДВИЖИМОСТЬ (Real Estate)
9. АВТО (Automobiles)
10. СТИЛЬ ЖИЗНИ (Lifestyle)
11. СПОРТ (Sports)

Each category has twenty-two articles; this way, in total there are two hundred and two articles. The originals are called "source-texts".

The parts of each article's structure are the following (**figure 8.1**)

8.1.2 Transliteration to Russian

In order to make use of the novel scientific method presented in the book, the texts had to be transliterated.

The International Organization for Standardization defines transliteration as the action to represent the characters or signs of an alphabet with those of another, under a letter-by-letter premise (Orozco, 1989). **Figure 8.2** displays a listing of Russian letters transliterated to Latin alphabet.

Figure 8.1 Example of news item in *TEXTRUSS*

Figure 8.2 Transliteration of Cyrillic leters into Latin ones ("Traslit," 2016)

8.2 Heuristics

To calculate the heuristics, commercial tools and novel scientific methods, TEXTRUSS is used as it is the only available especially for AGTS tasks in the Russian language.

8.2.1 *Baseline:random*

An n number of sentences in the original text are selected at random (Ledeneva, 2008). **Table 8.3** shows the results of *baseline:random* for TEXTRUSS. It is worth mentioning that ten runs were made for this heuristic as a guarantee for the result displayed.

Table 8. 3 Results of *baseline:random* for TEXTRUSS

Measure	Recall	Accuracy	*F-measure*
ROUGE-1	0.8977	0.8540	0.8734
ROUGE-2	0.6221	0.5918	0.6053
ROUGE-SU4	0.7789	0.7407	0.7577

For *baseline:random* the results tend to be low, since sentences are selected at random. For the state of the art, *baseline:random* is a reference for the worst result.

8.2.2 *Baseline:first*

The first n sentences of the original text are selected up to reaching the number of desired words. This configuration produces very good results for texts in the domain of news items (Ledeneva, 2008).

Table 8.4 shows the results of *baseline:first* for TEXTRUSS.

Table 8. 4 Results of *baseline:first* for TEXTRUSS

Measure	Recall	Accuracy	*F-measure*
ROUGE-1	0.9332	0.8703	0.8994
ROUGE-2	0.7440	0.6940	0.7171
ROUGE-SU4	0.8477	0.7901	0.8168

As it is noticed, the data generated with *baseline:first* are very high, which tells us that first sentences are very important for this news item *corpus*.

8.2.3 *Topline*

For TEXTRUSS the *Topline* reached is: 1.0000. This is because there is only one gold standard and is fully extractive. Owing to this reason, a method or tool may generate the same summary as the *gold standard*, which allows defining a *Topline* of 1.0000.

8.3 Commercial tools

For the Russian language *corpus* TEXTRUSS was used, a one-hundred-letter summary extension was opted for, which was defined considering the *corpus*' characteristics. For the tools that only have the option to select the percentage, formula eleven is utilized, which aids in finding the figure to generate the summary.

$$\frac{Number\ of\ desired\ words}{Number\ of\ total\ words\ \varepsilon\ the\ document} * 100 \qquad (11)$$

The commercial tools used in the Russian language tests are described below.

8.3.1 Microsoft Office Word Summarizer

In order to run the tests with Microsoft Office Word, TEXTRUSS was utilized, to guarantee the length of the summaries at one hundred words, formula 11 was used.

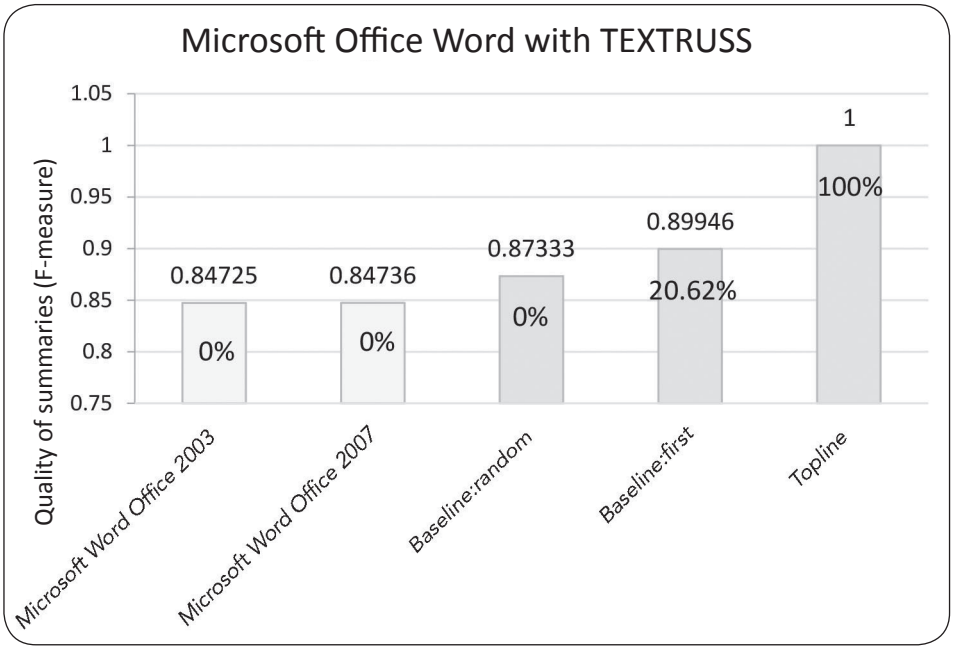
Graph 8.1 shows the results with a length of a hundred words.

Microsoft Office Word does not surpass *baseline:random*; however, this may have occurred owing to the number of tests carried out to ascertain this heuristic.

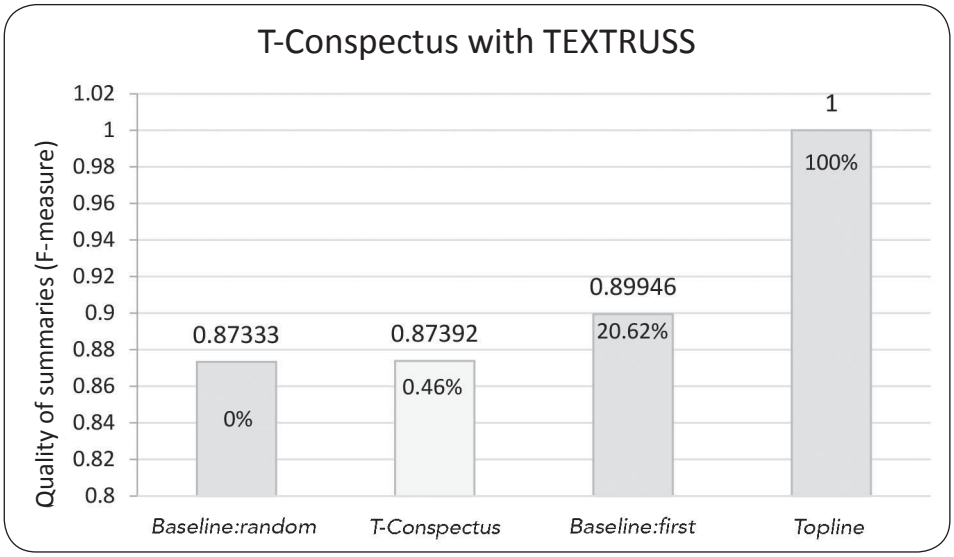
8.3.2 T-Conspectus

In order to carry out tests with T-Conspectus TEXTRUSS transliterated was utilized. Moreover, twenty percent of the document was chosen to generate the

Graph 8.1 Results of Microsoft Office Word for TEXTRUSS at one hundred words in comparison with the various heuristics



Graph 8.2 Results of T-Conspectus for TEXTRUSS in comparison with the various heuristics



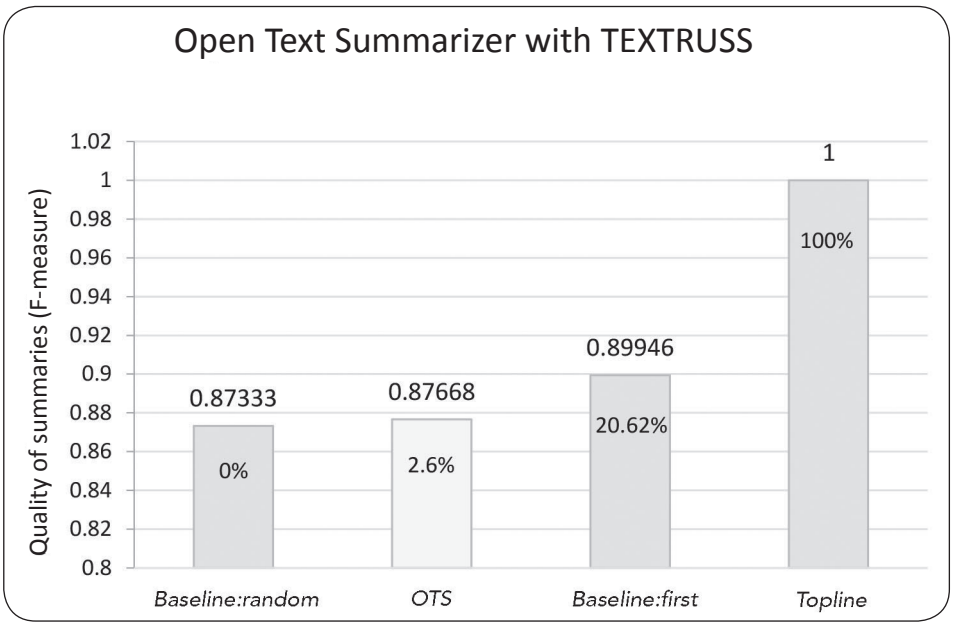
summaries, as it is the suitable percentage to accomplish texts with more than on hundred words in this tool.

T-Conspectus is one of the tools with the best results for AGTS in Russian, surpassing *baseline:random* (**graph 8.2**).

8.3.3 Open Text Summarizer (OTS)

For the tests with OTS, TEXTRUSS transliterated was used. In order to guarantee the length of the summaries at a hundred words, formula 11 was resorted to; with this it was possible to decide on the summary percentage.

Graph 8.3 Results of Open Text Summarizer for TEXTRUSS in comparison with the various heuristics

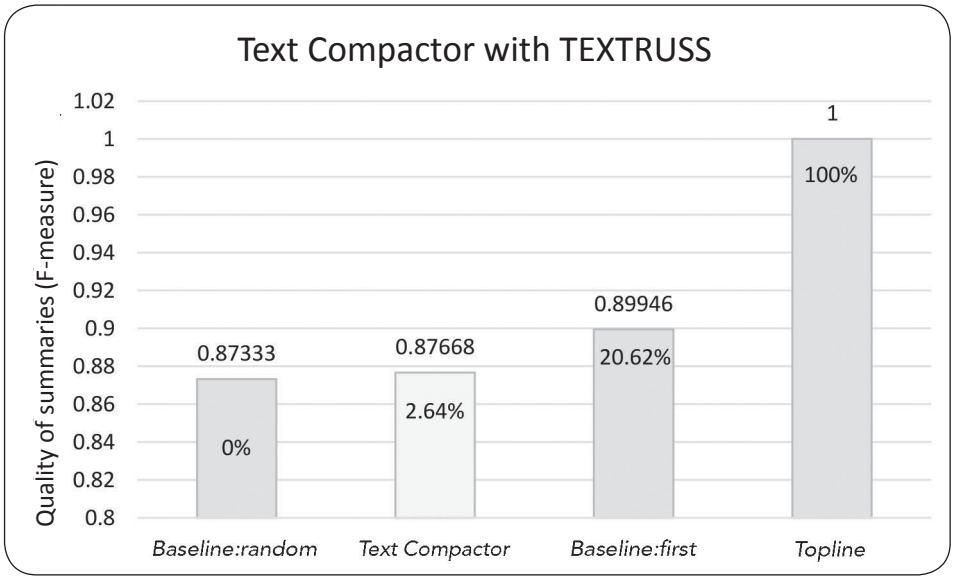


OTS surpasses *baseline:random*, showing an advance of 2.6%, if *baseline:random* is considered 0 and *Topline* 100% (**graph 8.3**). However, as noticed, the results obtained with OTS are very low in comparison with the 20.62% of *baseline:first*.

8.3.4 Text Compactor

For the tests with Text Compactor, TEXTRUSS transliterated was used. To guarantee one-hundred-word summaries, formula 11 was utilized to define the summary length. Graph 8.4 shows the results obtained by this tool: 2.64% regarding *baseline:random* and *Topline*.

Graph 8.4 Results of Text Compactor for TEXTRUSS in comparison with the various heuristics

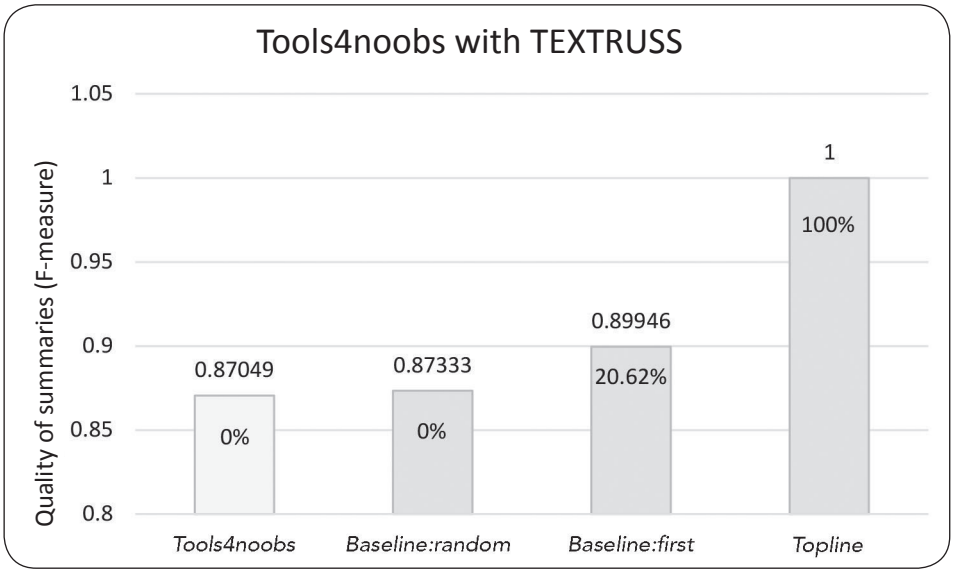


8.3.5 *Tools4noobs*

To produce one-hundred-word summaries with Tools4noobs formula 12 was used, because unlike other tools, Tools4noobs' threshold works inversely.

$$\frac{number\ of\ desired\ words * 100}{number\ of\ total\ words\ the\ document\ -\ 80} * -1 \qquad (12)$$

Graph 8.5 Results of Tools4noobs for TEXTRUSS in comparison with the various heuristics



In spite of not surpassing *baseline:random*, the difference with Tools4noobs is virtually null, so it may be considered they produce the same value **(graph 8.5**).

8.3.6 Resumo

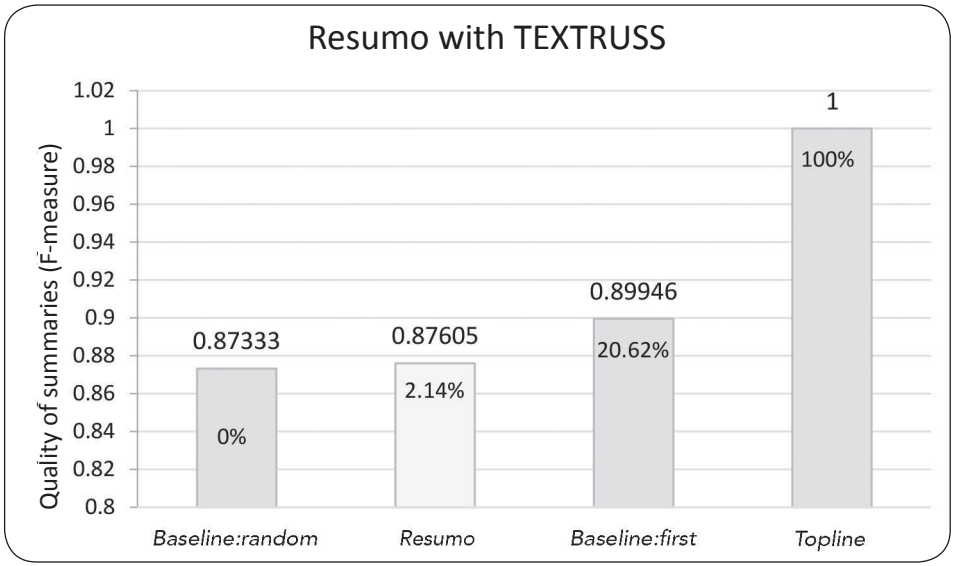
To carry out the tests with Resumo, TEXTRUSS in Russian was used. To guarantee the length of the summaries at one hundred words, formula 1 was used; with this, the length of the text percentage was reached.

Resumo allows working with text in Russian, so a transliteration is not needed; there is an advance of 2.14%, taking *baseline:random* and *Topline* as references (**graph 8.6**).

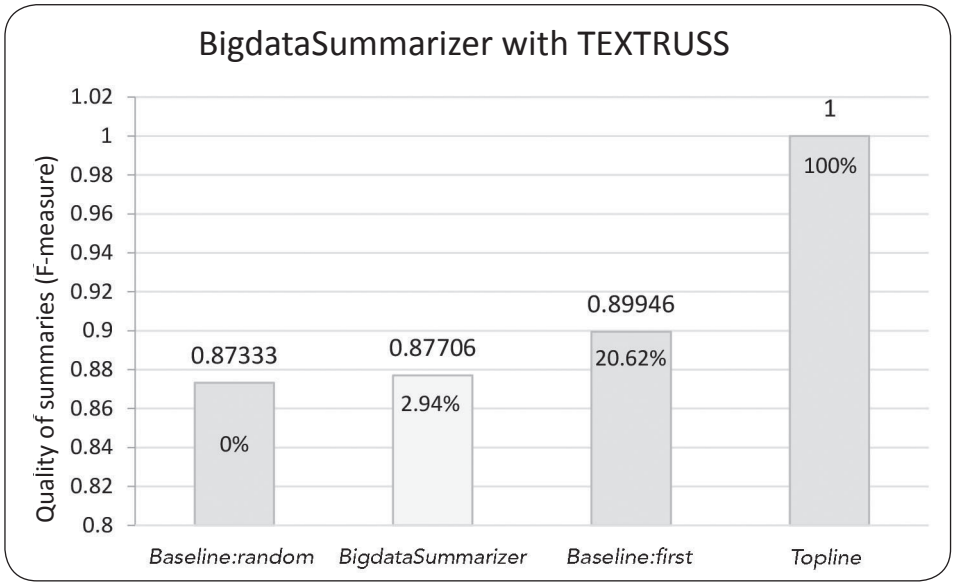
8.3.7 BigdataSummarizer

For the tests with BigdataSummarizer, TEXTRUSS in Russian was resorted to. To guarantee the length of the summaries at one hundred words, formula 8 was used; with this, the length of the text percentage was reached.

Graph 8.6 Results of Resumo for TEXTRUSS in comparison with the various heuristics



Graph 8.7 Results of BigdataSummarizer for TEXTRUSS in comparison with the various heuristics



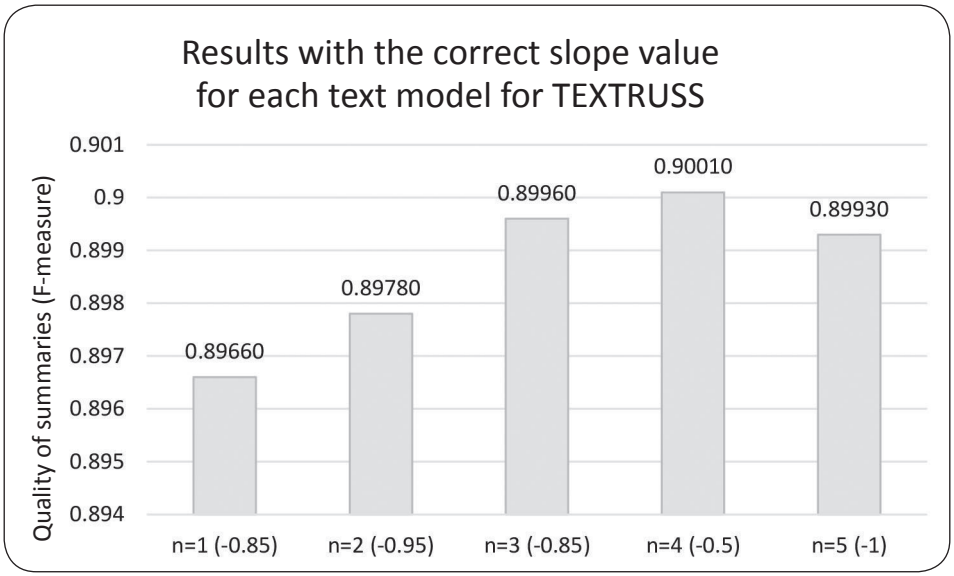
As in the case of Resumo, for BigdataSummarizer the text may be used in Cyrillic without a transliteration. This tool has an advancement of 2.94% as regards *baseline:random* (graph 8.7).

8.4 Novel scientific methods

The method by Matias (2016) has demonstrated to correctly work in various languages (English, Spanish and Portuguese); this way, it is used to try Russian. The description can be found in section 5.5.5.

Matias' (2016) method uses the n-grams model, so for the Russian language the results obtained are shown in graph 8.8.

Graph 8.8 Result with the correct slope value for each text model

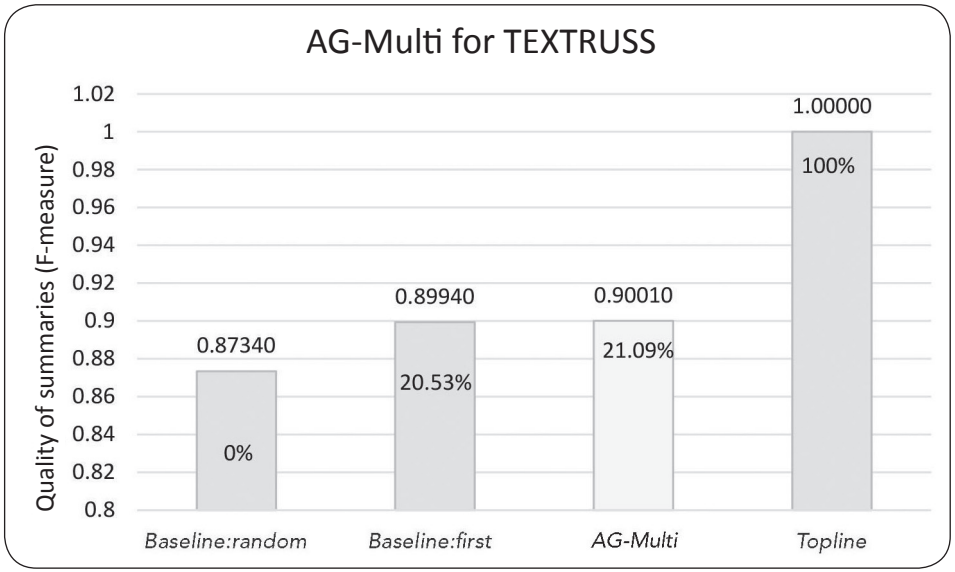


As noticed, the best results are obtained with $n = 4$ and a slope value of $m = -0.5$. It is worth mentioning that these were obtained with the *corpus* without preprocessing. Matias (2016) tried an importance adjustment of the characteristics for the fitness

function. However, for TEXTRUSS the importance of the two characteristics used by the method (term frequency and sentence position) remains the same.

Graph 8.9 shows the comparison of the result obtained with AG-Matias and the various heuristics.

Graph 8.9 Results of AG-Multi for TEXTRUSS in comparison with the various heuristics



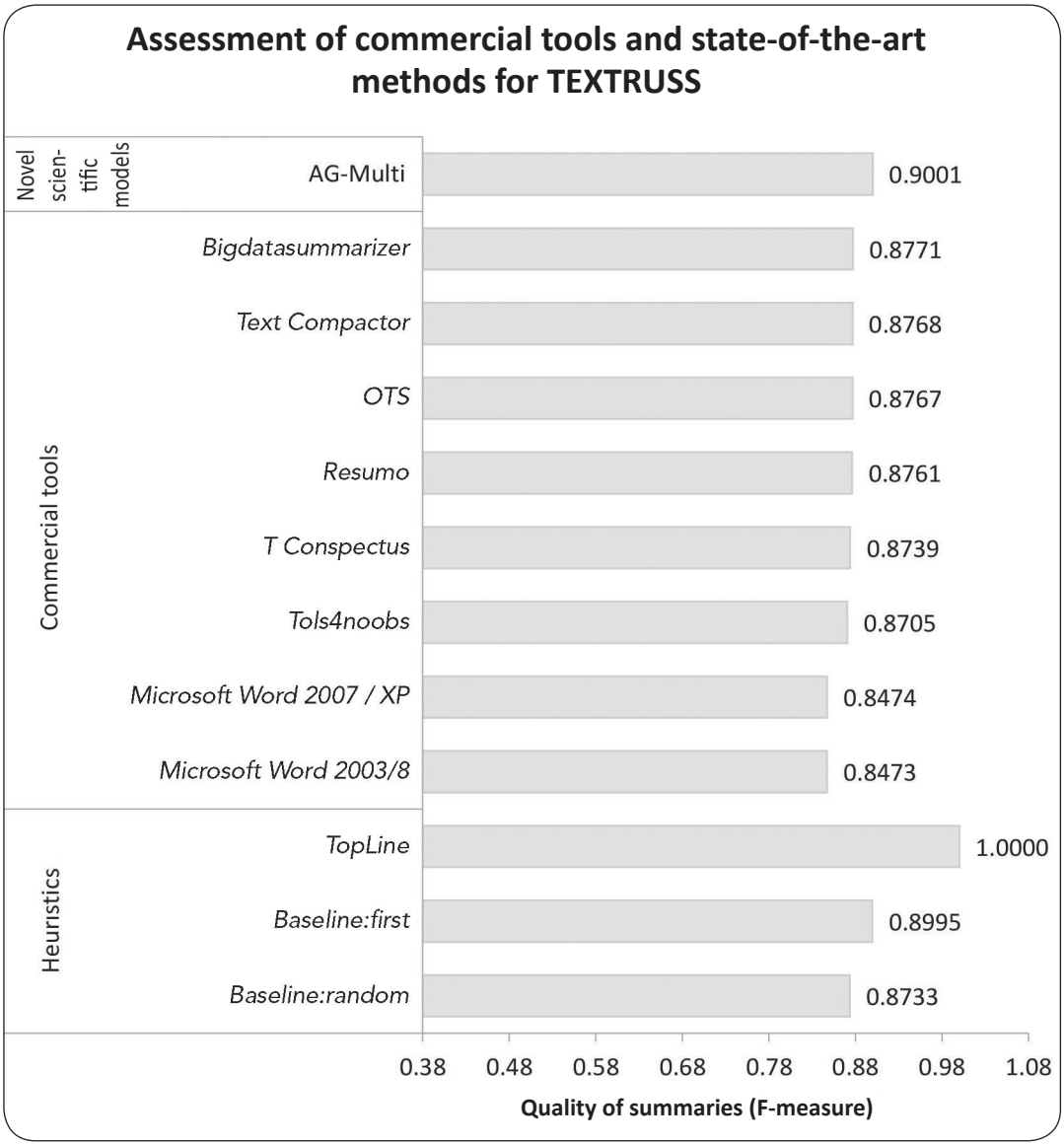
As noticed, *AG-Multi* surpasses *baseline:random* and by very little, *baseline:first*.

8.5 Results and analysis

There are no research works in Russian on AGTS tasks. In this book, the results of eight commercial tools and a state-of-the-art method are shown; additionally, the results of the main heuristics are presented. Considering those obtained with novel scientific methods and commercial tools, it may be said that sixty years of research on the automatic generation of text summaries in Russian haven

recovered. Even though there is a lot to do, as it is known that a method surpasses *baseline:first* for Russian (graph 8.10).

Graph 8.10 Results for the Russian language with novel scientific models and commercial tools



Chapter IX

Conclusions

This chapter presents the conclusions on the AGTS problems stated and developed in the book. The hypothesis considered was that humans might replicate the necessary knowledge to produce a summary automatically in a machine; the tests demonstrated that not only it imitated humans, but surpassed them. By means of the Turing Test, humans largely chose the summaries made by a machine. However, there are some pending issues that must be solved in the future.

This book presented a study on the detection of ideas and composition of summaries in English, Spanish, Portuguese and Russian with a view to recover sixty years of research in each of the languages above by means of an updating, mainly in Spanish, Portuguese and Russian.

It is said sixty years of research since the first enquiries on AGTS tasks for the English language date back to the 1950's decade (Lunh, 1953-1958), and virtually a decade back for the other languages. All the investigations have focused on the qualitative study of AGTS tasks. Although quantitative studies are important, they have been put aside as research focused on numbers and not in finding out if a machine was already able to produce summaries similar to those made by humans.

In this book, a series of Turing Tests trials in AGTS tasks was presented in order to find out if humans have been able to transmit the necessary knowledge into a machine, by means of models and methods, so that it can emulate humans. The conclusion, according to the results of Turing Tests is that a machine may be considered intelligent to generate summaries. Both in the trials made for Spanish and English there was between 56 and 46% of confusion at selecting the summaries made by people.

In like manner, a conscientious research on AGTS in English, Spanish, Portuguese, and Russian was presented, calculating for each one the values of the various heuristics (*baseline:random*, *baseline:first* and *Topline*), commercial tools and novel scientific methods. Due to foregoing, once considering the sixty years of research, the following is concluded for each language.

- English. The heuristic *baseline:first* was overcome in 2008. Progress measured versus *Topline* reaches 41.04% assessed with corpora DUC01 and DUC02. For the case of commercial tools, only Copernic Summarizer surpasses *baseline:first* in DUC02, which demonstrates a great leap of the novel scientific methods as only two of them do not overcome this heuristic in DUC02. By and large, a mayor advance regarding the various heuristics and commercial tools is noticed for the English language.
- Spanish. Research on AGTS began in 2001; over these years efforts have been made; however, they have not been compared as there was not a specialized *corpus*. In this book, the *corpus* TER is presented and the values of the heuristics *baseline:first*, *baseline:random* and *Topline*

are shown. It is worth mentioning that *baseline:first* is very high in Spanish and this poses a challenge to overcome by the novel scientific models. Commercial tools and novel scientific methods are assessed. It is concluded that Spanish has an advance of 68.25% regarding heuristics *baseline:random* and *Topline*.

- Portuguese. Research has been carried out with the *corpus* TeMário as of 2003, however the value of heuristics to validate the degree of advance was not available. Portuguese has an advance of 20.27% regarding heuristics *baseline:random* and *Topline*. Only two novel scientific methods manage to surpass *baseline:first*, while commercial tools do not surpass it.
- Finally, Russian. There were not comparable works for AGTS. In this book a specialized *corpus*, called TEXTRUSS, is presented. Formerly, there were not comparable works owing to the use of Cyrillic alphabet, but by means of TEXTRUSS transliterated texts can be obtained to be used in methods that work with Latin characters. The measures of the heuristics *baseline:first*, *baseline:random* and *Topline* in Russian for TEXTRUSS are shown. Commercial tools and novel scientific methods are analyzed. It is concluded that Russian has an advance of 21.09% regarding *baseline:random* and *Topline*.

References

Acero, I., Alcojor, M., Díaz Esteban, A., Gómez Hidalgo, J.M., Maña López, M.J. (2001, septiembre). Generación automática de resúmenes personalizados. *Proces. Leng. Nat*. No 27. Pp. 281-290.

Al Saied, H., Dugué, N., Lamirel, J.-C. (2017). Automatic summarization of scientific publications using a feature selection approach. *Int. J. Digit. Libr*. Pp. 1-13.

Aleixo, P., Pardo, T.A.S. (2008). CSTNews: um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (cross-document structure theory). *ICMC-USP*.

Alfonseca, E., Rodríguez, P. (2003). Generating extracts with genetic algorithms. *Presented at the European Conference on Information Retrieval, Springer.* Pp. 511–519.

Alvarado B., A. (2017). *Evaluación de la calidad de las herramientas comerciales y métodos del estado del arte para la generación de resúmenes del corpus DUC-2001*, México: Universidad Autónoma del Estado de México.

Amancio, D.R., Nunes, M.G., Oliveira Jr, O.N., Costa, L. da F. (2012). Extractive summarization using complex networks and syntactic dependency. *Phys. Stat. Mech. Its Appl. 391*. Pp. 1855-1864.

Antiqueira, L. (2007). *Desenvolvimento de técnicas baseadas em redes complexas para sumarização extrativa de textos*, Brasil: Universidade de São Paulo.

Arévalo, J.A. (2017). El español una lengua viva. I*nforme 2017. Instituto Cervantes. infotra*.

Babar, S., Patil, P.D. (2015). Improving Performance of Text Summarization. *Procedia Comput. Sci*. 46. Pp. 354–363.

Baldwin, B., Donaway, R., Hovy, E., Liddy, E., Mani, I., Marcu, D., McKeown, K., Mittal, V., Moens, M., Radev, D. (2000). An evaluation road map for summarization research. *TIDES July*.

Banko, M., Vanderwende, L. (2004). Using n-grams to understand the nature of summaries. *Proceedings of HLT-NAACL 2004: Short Papers. Association for Computational Linguistics*. Pp. 1–4.

Barzilay, R., Elhadad, M. (1999). Using lexical chains for text summarization. Adv. Autom. Text Summ. Pp. 111–121.

Benbrahim, M., Ahmad, K. (1995). Text summarisation: The role of lexical cohesion analysis. New Rev. Doc. Text Manag. Pp. 321–335.

Benítez, R., Escudero, G., Kanaan, S., Rodó, D.M. (2014). *Inteligencia artificial avanzada*. Editorial UOC.

Berker, M. (2011). Using genetic algorithms with lexical chains for automatic text summarization.

Bhargava, R., Sharma, Y., Sharma, G. (2016). Atssi: Abstractive text summarization using sentiment infusion. Procedia Comput. Sci. 89. Pp. 404–411.

Bing, L., Li, P., Liao, Y., Lam, W., Guo, W., Passonneau, R.J. (2015). Abstractive multi-document summarization via phrase selection and merging. *ArXiv Prepr. ArXiv150601597*.

Bossard, A., Généreux, M., Poibeau, T. (2008). Description of the LIPN System at TAC 2008: Summarizing Information and Opinions. *Presented at the TAC 2008*. Pp. 282–291.

Braslavski, P., Gustelev, V. (2007). News Summarization System Based On Machine Learning Approach. *Digit. Libr. Adv. Methods Tecnol. Digit. Collect.* Pp. 142–147.

Brin, S., Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. *Comput. Netw*. 56. Pp. 3825–3833.

Briones, E.G., Cubino, R.L., Sobrino, B.L. (2012). *La noticia y el reportaje. Proyecto Mediascopio Prensa. La lectura de la prensa escrita en el aula*. Ministerio de Educación.

Cabral, L. de S., Lins, R.D., Mello, R.F., Freitas, F., Ávila, B., Simske, S., Riss, M. (2014). A platform for language independent summarization. *Proceedings of the 2014 ACM Symposium on Document Engineering*. ACM. Pp. 203–206.

Carbonell, J., Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. P*roceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM. Pp. 335–336.

Cardoso, P.C., Maziero, E.G., Jorge, M.L., Seno, E.M., Di Felippo, A., Rino, L.H., Nunes, M.G., Pardo, T.A. (2011). CSTnews-a discourse-annotated *corpus* for single and multi-document summarization of news texts in Brazilian Portuguese. *Proceedings of the 3rd RST Brazilian Meeting*. Pp. 88–105.

Cavalieri, D.C., Bastos-Filho, T., Palazuelos-Cagigas, S.E., Sarcinelli-Filho, M., (2015). On combining language models to improve a text-based human-machine interface. *Int. J. Adv. Robot*. Syst. 12. P. 170.

Coarite Choque, R. (2008). Areas de aplicación de la Inteligencia Artificial. *Rev. Inf. Tecnol. Soc*. Pp. 18–22.

Copernic Summarization-Technologies White Paper, 2003.

Corpus, 2014. *Corpus*. Gran Dicc. Leng. Esp.

da Cunha Fanego, I. (2005). Hacia un modelo lingüístico de resumen automático de artículos médicos en español. *Proy. Investig. Univ. Pompeu Fabra Inst. Univ. Lingüíst. Apl. Dr. En Cienc. Leng. Lingüíst. Apl. Httpwww Upf Edupdiiulairia Dacunha* 0 202. 07–04.

Donaway, R.L., Drummey, K.W., Mather, L.A. (2000). A comparison of rankings produced by summarization evaluation measures. *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization. Association for Computational Linguistics*. Pp. 69–78.

Edmundson, H.P. (1969). New methods in automatic extracting. *J. ACM JACM 16*. Pp. 264–285.

Edmundson, H.P., Wyllys, R.E. (1961). Automatic abstracting and indexing—survey and recommendations. *Commun. ACM 4*. Pp. 226–234.

El-Haj, M., Rayson, P. (2013). Using a keyness metric for single and multi document summarisation. *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-Document Summarization*. Pp. 64–71.

Europe Internet Stats - Population Statistics [WWW Document] (2017). URL https://www.internetworldstats.com/europa2.htm#ru

García-Hernández, R., Montiel, R., Ledeneva, Y., Rendón, E., Gelbukh, A., Cruz, R. (2008). Text summarization by sentence extraction using unsupervised learning. *MICAI 2008 Adv. Artif. Intell.* Pp. 133–143.

García-Hernández, R.A., Ledeneva, Y. (2013). Single extractive text summarization based on a genetic algorithm. *Presented at the Mexican Conference on Pattern Recognition, Springer*. Pp. 374–383.

Genest, P.-E., Lapalme, G. (2011). Framework for abstractive summarization using text-to-text generation. *Proceedings of the Workshop on Monolingual Text-To-Text Generation. Association for Computational Linguistics*. Pp. 64–73.

Gliozzo, D.A., Ackerson, C., Bhattacharya, R., Goering, A., Jumba, A., Kim, S.Y., Krishnamurthy, L., Lam, T., Littera, A., McIntosh, I., Murthy, S., Ribas, M. (2017).

Building Cognitive Applications with IBM Watson Services: *Volume 1 Getting Started.* P. 130.

Hassel, M., Dalianis, H. (2003). Text Summarizer (with PRM) [WWW Document]. URL http*://swesum.nada.kth.se/*index-eng-adv.html

Hernández, P.T. (2018). *Desempeño de los métodos del estado del arte para la generación automática de resúmenes extractivos para el corpus TEXTRUSS*. México: Universidad Autónoma del Estado de México, Tianguistenco.

Hirao, T., Isozaki, H., Maeda, E., Matsumoto, Y. (2002). Extracting important sentences with support vector machines. *Presented at the Proceedings of the 19th international conference on Computational linguistics-Volume 1, Association for Computational Linguistics*. Pp. 1–7.

Hsu, F. (1999). IBM's deep blue chess grandmaster chips. IEEE Micro 19. Pp. 70–81.

Igave, M.S., Gaikwad, C.M. (2016). Efficient Multi-Document Summary Generation Using Neural Network. *Int. J. Adv. Eng. Manag. Sci. IJAEMS*.

Jing, H. (2002). Using hidden Markov modeling to decompose human-written summaries. *Comput*. *Linguist. 28*. Pp. 527–543.

Jing, H. (2001). *Cut-and-paste text summarization*. EUA: Columbia University.

Jing, H., Barzilay, R., McKeown, K., Elhadad, M. (1998). Summarization evaluation methods: Experiments and analysis. *AAAI Symposium on Intelligent Summarization. Palo Alto*, CA. Pp. 51–59.

Katragadda, R., Pingali, P., Varma, V. (2009). Sentence position revisited: a robust light-weight update summarization'baseline'algorithm. *Presented at the Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies, Association for Computational Linguistics*. Pp. 46–52.

Khan, A., Salim, N., Farman, H., Khan, M., Jan, B., Ahmad, A., Ahmed, I., Paul, A. (2018). Abstractive Text Summarization based on Improved Semantic Graph Approach. *Int. J. Parallel Program.* Pp. 1–25.

Kiyoumarsi, F. (2015). Evaluation Of Automatic Text Summarizations Based On Human Summaries. *Procedia-Soc. Behav. Sci*. 192. Pp. 83–91.

Kleinberg, J.M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM JACM 46*. Pp. 604–632.

Krishna, R.M., Reddy, C.S. (2016). Extractive Text Summarization Using Lexical Association and Graph Based Text Analysis. *Computational Intelligence in Data Mining—Volume 1. Springer*. Pp. 261–272.

Kupiec, J., Pedersen, J., Chen, F. (1995). A trainable document Summarizer. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM*. Pp. 68–73.

La Crónica de Hoy | La noticia hecha diario [WWW Document] (2014). URL http://www.cronica.com.mx/noticias.php

Last, M., Litvak, M. (2010). Language-independent Techniques for Automated Text Summarization. Pp. 207–237.

Ledeneva, Y. (2008). Automatic Language-Independent Detection of Multiword Descriptions for Text Summarization, National Polytechnic Institute.

Ledeneva, Y., García-Hernández, R.A. (2017). Generación automática de resúmenes. *Retos, propuestas y experimentos, 1st ed*.

Ledeneva, Y., García-Hernández, R.A. (2013). Automatic text summarization with Maximal Frequent Sequences.

Ledeneva, Y., Gelbukh, A., García-Hernández, R.A. (2008). Terms derived from frequent sequences for extractive text summarization. *Presented at the International Conference on Intelligent Text Processing and Computational Linguistics, Springer.* Pp. 593–604.

Ledeneva, Y., Hernández, R., Soto, R., Reyes, R., Gelbukh, A. (2011). EM clustering algorithm for automatic text summarization. *Adv. Artif. Intell*. Pp. 305–315.

Lehmam, A. (2010). Essential *Summarizer*: innovative automatic text summarization software in twenty languages. *Adaptivity, Personalization and Fusion of Heterogeneous Information. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE*. Pp. 216–217.

Leite, D., Rino, L.H. (2009). A Genetic Fuzzy Automatic Text *Summarizer. Anais Do Csbc 2009*. SBC. Pp. 779–788.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *Presented at the Text summarization branches out: Proceedings of the ACL-04 workshop, Barcelona, Spain*.

Lin, C.-Y. (1999). Training a selection function for extraction. *Presented at the Proceedings of the eighth international conference on Information and knowledge management, ACM*. Pp. 55–62.

Lin, C.-Y., Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics*. Pp. 71–78.

Lin, C.-Y., Och, F.J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics*. Pp. 605.

Lloret, E., Palomar, M. (2011). COMPENDIUM: Una herramienta de generación de resúmenes modular. Proces. Leng. Nat.

Luhn, H.P. (1958). The automatic creation of literature abstracts. *IBM J. Res. Dev*. 2. Pp. 159–165.

Luhn, H.P. (1953). Distributor and method for making the same. Google Patents.

Lynn, H.M., Choi, C., Kim, P. (2017). An improved method of automatic text summarization for web contents using lexical chain with semantic-related terms. *Soft Comput*. Pp. 1–11.

Mani, I. (2001). Automatic Summarization. *Natural Language Processing, 3 (Paper)*.

Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T., Sundheim, B. (1999). The TIPSTER SUMMAC text summarization evaluation. Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics. Pp. 77–85.

Marcu, D. (1997). From discourse structures to text summaries. *Intell. Scalable Text Summ.*

Margarido, P.R., Pardo, T.A., Antonio, G.M., Fuentes, V.B., Aires, R., Aluísio, S.M.,

Fortes, R.P. (2008). Automatic summarization for text simplification: Evaluating text understanding by poor readers. *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web. ACM*. Pp. 310–315.

Martins, C. (2002). UNLSumm: *Un resumen automático de textos UNL. Departamento de Informática*. Brasil: Universidad Federal de São Carlos. Sao Carlos - SP.

Martins, C., Pardo, T., Espina, A., Rino, L. (2001). Introducción a los resúmenes automáticos. Brasil: Universidad Federal de São Carlos.

Mateo, P.L., González, J.C., Villena, J., Martínez, J.L. (2003). Un sistema para resumen automático de textos en castellano. *Proces. Leng. Nat*. 31. Pp. 29–36.

Matias, G. (2016). *Generación Automática de Resúmenes Independientes del Lenguaje.* México: Universidad Autónoma del Estado de México.

Matias, G. (2013). *Generación automática de resúmenes usando algoritmos genéticos*. México: Universidad Autónoma del Estado de México.

McKeown, K., Radev, D.R. (1995). Generating summaries of multiple news articles. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM*. Pp. 74–82.

Mendoza Becerra, M.E. (2015). Generación automática de resúmenes extractivos de múltiples documentos basada en algoritmos meméticos.

Mendoza, M., Bonilla, S., Noguera, C., Cobos, C., León, E. (2014). Extractive single-document summarization based on genetic operators and guided local search. *Expert Syst. Appl*. 41. Pp. 4158–4169.

Mihalcea, R. (2004). Graph-based ranking algorithms for sentence extraction, applied to text summarization. *Presented at the Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, Association for Computational Linguistics.* P. 20.

Mihalcea, R., Radev, D. (2011). *Graph-based natural language processing and information retrieval.* EUA: Cambridge university press.

Mihalcea, R., Tarau, P. (2005). A language independent algorithm for single and multiple document summarization.

Minel, J.-L., Nugier, S., Piat, G. (1997). How to Appreciate the Quality of Automatic Text Summarization? Examples of FAN and MLUCE Protocols and their Results on SERAPHIN1.

Mingli, L.I., Sun, L., Han, X. (2016). Combining Relevance Clustering and Graph Model Methods for Automatic Summarization. *J. Residuals Sci. Technol. 13*.

Miranda, S. (2013). *Modelo para la Generación Automática de Resúmenes Abstractivos basados en gráfos conceptuales*. México: Instituto Politécnico Nacional.

Miranda-Jiménez, S., Gelbukh, A., Sidorov, G. (2013). Summarizing conceptual graphs for automatic summarization task. *International Conference on Conceptual Structures. Springer*. Pp. 245–253.

Modolo, M. (2003). Supongamos: un entorno para la exploración de métodos de extracción de resumen automático de texto en portugués. *Disert. Masters Dep. Informática UFSCar Sao Carlos - SP*.

Molina, A. (2013). Compresión automática de frases: un estudio hacia la generación de resúmenes en espanol. *Intel. Artif. 16*. Pp. 41–62.

Nandhini, K., Balasundaram, S.R. (2014). Extracting easy to understand summary using differential evolution algorithm. *Swarm Evol. Comput.* 16. Pp. 19–27.

Nenkova, A., Passonneau, R. (2004). Evaluating content selection in summarization: The pyramid method. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: Hlt-Naacl 2004.*

Orăsan, C. (2003). An evolutionary approach for improving the quality of automatic summaries. *Proceedings of the ACL 2003 Workshop on Multilingual Summarization*

and Question Answering-Volume 12. Association for Computational Linguistics. Pp. 37–45.

Orozco, A. (1989). Las RCA2 y la transliteración de nombres de autores personales rusos.

Orrú, T., Rosa, J.L.G., de Andrade Netto, M.L. (2006). SABIO: an automatic portuguese text summarizer through artificial neural networks in a more biologically plausible model. *Presented at the International Workshop on Computational Processing of the Portuguese Language, Springer.* Pp. 11–20.

Ouyang, Y., Li, W., Lu, Q., Zhang, R. (2010). A study on position information in document summarization. *Presented at the Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics.* Pp. 919–927.

Over, P., Dang, H., Harman, D. (2007). DUC in context. *Inf. Process. Manag. 43*. Pp. 1506–1520.

Paice, C.D. (1990). Constructing literature abstracts by computer: techniques and prospects. *Inf. Process. Manag. 26*. Pp. 171–186.

Pardo, T. (2002). DMSumm: A Resúmenes de generador automático. Disertación Masters. *Dep. Informática Univ. Fed. São Carlos SaoCarlos - SP*.

Pardo, T., Rino, L., Nunes, M.G. (2003b). NeuralSumm: Un enfoque Conexionista para los Textos resumen automático. *Actas XXIII Congr. Soc. Bras. Comput*. VIII. Pp. 203–245.

Pardo, T.A.S., Rino, L.H.M. (2003). *TeMário: Um corpus para sumarização automática de textos.* Brasil: São Carlos Universidade São Carlos Relatório Téc.

Pardo, T.A.S., Rino, L.H.M., Nunes, M. das G.V. (2003). GistSumm: A summarization tool based on a new extractive method. *Presented at the International Workshop on Computational Processing of the Portuguese Language, Springer*. Pp. 210–218.

Patel, A., Siddiqui, T., Tiwary, U.S. (2007). A language independent approach to multilingual text summarization. *Presented at the Large scale semantic access to content (text, image, video, and sound)*. Pp. 123–132.

Plaza, L. (2011). *Uso de grafos semánticos en la generación automática de resúmenes y estudio de su aplicación en distintos dominios: biomedicina, periodismo y turismo*. España: Universidad Complutense de Madrid, Madrid.

Polya, G., Zugazagoitia, J. (1965). Cómo plantear y resolver problemas. Trillas.

Qazvinian, V., Hassanabadi, L.S., Halavati, R., 2008. Summarising text with a genetic algorithm-based sentence extraction. Int. J. Knowl. Manag. Stud. 2. Pp. 426–444.

Rojas, J. (2018). Calculating the Significance of Automatic Extractive Text Summarization using a Genetic Algorithm. *J. Intell. Fuzzy Syst., Applications in Engineering and Technology*.

Rojas J. (2017). *Cálculo de Topline para la generación automática de resúmenes usando algoritmos genéticos*. México: Universidad Autónoma del Estado de México.

Rojas, J.M. (2016). *Evaluación de herramientas comerciales y métodos del estado del arte para la generación de resúmenes en idioma ruso*. México: Universidad Autónoma del Estado de México.

Saggion, H. (2011). Using SUMMA for Language Independent Summarization at TAC 2011. *Presented at the TAC*.

Salton, G., McGill, M. (1983). *Introduction to modern information retrieval*. McGraw-Hill Book Company.

Sparck Jones, K., Galliers, J.R. (1995). Evaluating natural language processing systems: An analysis and review. *Springer Science & Business Media.*

Suanmali, L., Salim, N., Binwahlan, M.S. (2011). GENETIC ALGORITHM BASED SENTENCE EXTRACTION FOR TEXT SUMMARIZATION. *Int. J. Innov. Comput. 1*.

Téllez, A., Montes, M., Villaseñor-Pineda, L. (2009). Using Machine Learning for Extracting Information from Natural Disaster News Reports. *Comput. Sist. Instituto Politécnico Nacional*. Pp. 33–44.

Toledo-Báez, M.C. (2010). Aproximación al resumen automático como herramienta de ayuda a la traducción jurídica en el ámbito del Derecho turístico1. *El Español, Lenguaje de Traducción Para La Cooperación y El Diálogo.* Madrid.

Torres-Moreno, J.-M. (2014). Automatic text summarization. John Wiley & Sons.

Traslit [WWW Document] (2016). Транслитератор Translitnet Бывший Translitru. URL https://translit.net/

Turing, A.M. (1950). Computing Machinery and Intelligence. *Computing Machinery and Intelligence*. Pp. 433–460.

Uddin, M.N., Khan, S.A. (2007). A study on text summarization techniques and implement few of them for Bangla language. C*omputer and Information Technology, 2007. Iccit 2007*. 10th International Conference On. IEEE. Pp. 1–4.

UNE 50-103-90 (1990). Preparación de resúmenes.

Vázquez, E. (2015). *Modelo de relevancia de la posición de las oraciones en resúmenes de textos, mediante regresión simbólica*. México: Universidad Autónoma del Estado de México.

Venegas, R. (2011). Evaluación de resúmenes en español con Análisis Semántico Latente: Una implementación posible. *Rev. Signos 44*. Pp. 85–102.

Verma, R., Lee, D. (2017). Extractive Summarization: Limits, Compression, Generalized Model and Heuristics. *ArXiv Prepr. ArXiv170405550*.

Villar, A.M. (2005). Microsoft Word 2003. *Nociones básicas: Operaciones básicas, diseño, pruebas e impresión*, 1ra ed. ideaspropias.

Villatoro E. (2007). *Generación automática de resúmenes de múltiples documentos*. México: Instituto Nacional de Astrofísica, Óptica y Electrónica.

Wan, X. (2010). Towards a unified approach to simultaneous single-document and multi-document summarizations. *Presented at the Proceedings of the 23rd international conference on computational linguistics, Association for Computational Linguistics.* Pp. 1137–1145.

Wang, B., Zhang, J., Liu, Y., Zou, Y. (2017). Density peaks clustering based integrate framework for multi-document summarization. *CAAI Trans. Intell. Technol. 2*. Pp. 26–30.

Wang, L., Cardie, C. (2013). Domain-Independent Abstract Generation for Focused Meeting Summarization. *Presented at the ACL (1)*. Pp. 1395–1405.

Главные новости - Газета.Ru [WWW Document] (2015). URL https://www.gazeta.ru/

Appendix A

Turing test in Spanish

Appendix A presents two more Turing tests in Spanish. The texts and summaries given to the individuals for them to identify which were man-made are shown in full. Additionally, the appendix contains the tables that disclose which summaries were generated by machines and which by humans.

The second text for the Turing Test in Spanish and the summaries made by humans and by machines are presented below.

Analizan estados acciones contra el dengue

El dengue es un padecimiento que afecta a 27 estados del país, por lo cual es importante realizar acciones con un impacto positivo para reducir su trasmisión, expusieron autoridades de salud en la primera Reunión de Vectores, en Nuevo Vallarta, Nayarit. En un comunicado la Secretaría de Salud (SSA) informó que con el fin de anticiparse a los posibles daños que pueda ocasionar el dengue en la región occidente los responsables de los programas y titulares de Salud de Sinaloa, Michoacán, Colima, Jalisco y Nayarit realizaron dicha reunión. En el encuentro encabezado por el director general del Centro Nacional de Programas Preventivos y Control de Enfermedades (Cenaprece), Jesús Felipe González, se revisó la situación epidemiológica del dengue de la región occidente del país, y los programas estatales de control y prevención. González Roldán subrayó la importancia del trabajo coordinado entre la federación, estados, municipios y la población para reducir el potencial de transmisión de esta enfermedad. Indicó que este trabajo de anticipación se debe enfocar en las medidas de prevención y promoción de la salud, para la eliminación de criaderos. Llamó a no bajar la guardia, pues la incidencia de letalidad en México está por debajo del indicador de la Organización Mundial de la Salud (OMS), mientras que en otras partes del mundo esta enfermedad va a la alza. Al respecto Álvaro Martín Acosta Padilla, director de Prevención y Promoción de la Salud de Sinaloa, comentó que el trabajo anticipado en el control del mosco vector redundará en la disminución del número de casos, siempre y cuando la ciudadanía tome conciencia de participar en la eliminación de criaderos. A su vez Óscar Villaseñor Anguiano, secretario de Salud en Nayarit, agregó que el trabajo coordinado entre los estados compromete acciones como el control larvario, abatización y fumigación. "Los estados occidentales debemos desarrollar acciones conjuntas en el combate al dengue, para lograr disminuir el número de casos", mencionó.

Summary 1

Llamó a no bajar la guardia, pues la incidencia de letalidad en México está por debajo del indicador de la Organización Mundial de la Salud (OMS), mientras que en otras partes del mundo esta enfermedad va a la alza. Al respecto Álvaro Martín Acosta Padilla, director de Prevención y Promoción de la Salud de Sinaloa, comentó que el trabajo anticipado en el control del mosco vector redundará en la disminución del número de casos, siempre y cuando la ciudadanía tome conciencia de participar en la eliminación de criaderos. El dengue es un padecimiento que afecta a 27 estados del país…

Summary 2

El dengue es un padecimiento que afecta a 27 estados del país, por lo cual es importante realizar acciones con un impacto positivo para reducir su trasmisión, expusieron autoridades de salud en la primera Reunión de Vectores, en Nuevo Vallarta, Nayarit. En un comunicado la Secretaría de Salud (SSA) informó que con el fin de anticiparse a los posibles daños que pueda ocasionar el dengue en la región occidente los responsables de los programas y titulares de Salud de Sinaloa, Michoacán, Colima, Jalisco y Nayarit realizaron dicha reunión. González Roldán subrayó la importancia del trabajo coordinado entre la federación, estados,…

Summary 3

El dengue es un padecimiento que afecta a 27 estados del país, por lo cual es importante realizar acciones con un impacto positivo para reducir su trasmisión, expusieron autoridades de salud en la primera Reunión de Vectores, en Nuevo Vallarta, Nayarit. En un comunicado la Secretaría de Salud (SSA) informó que con el fin de anticiparse a los posibles daños que pueda ocasionar el dengue en la región occidente los responsables de los programas y titulares de Salud de Sinaloa, Michoacán, Colima, Jalisco y Nayarit realizaron dicha reunión. A su vez Óscar Villaseñor Anguiano, secretario de Salud en Nayarit, agregó…

Summary 4

En un comunicado la Secretaría de Salud (SSA) informó que con el fin de anticiparse a los posibles daños que pueda ocasionar el dengue en la región occidente los responsables de los programas y titulares de Salud de Sinaloa, Michoacán, Colima, Jalisco y Nayarit realizaron dicha reunión. En el encuentro encabezado por el director general del Centro Nacional de Programas Preventivos y Control de Enfermedades (Cenaprece), Jesús Felipe González, se revisó la situación epidemiológica del dengue de la región occidente del país,

y los programas estatales de control y prevención. A su vez Óscar Villaseñor Anguiano, secretario de Salud en…

Summary 5

El dengue es un padecimiento que afecta a 27 estados del país, por lo cual es importante realizar acciones con un impacto positivo para reducir su trasmisión, expusieron autoridades de salud en la primera Reunión de Vectores, en Nuevo Vallarta, Nayarit. Indicó que este trabajo de anticipación se debe enfocar en las medidas de prevención y promoción de la salud, para la eliminación de criaderos. Al respecto Álvaro Martín Acosta Padilla, director de Prevención y Promoción de la Salud de Sinaloa, comentó que el trabajo anticipado en el control del mosco vector redundará en la disminución del número de casos,…

Summary 6

El dengue es un padecimiento que afecta a 27 estados del país, por lo cual es importante realizar acciones con un impacto positivo para reducir su trasmisión, expusieron autoridades de salud en la primera Reunión de Vectores, en Nuevo Vallarta, Nayarit. En un comunicado la Secretaría de Salud (SSA) informó que con el fin de anticiparse a los posibles daños que pueda ocasionar el dengue en la región occidente los responsables de los programas y titulares de Salud de Sinaloa, Michoacán, Colima, Jalisco y Nayarit realizaron dicha reunión. En el encuentro encabezado por el director general del Centro Nacional de…

Out of the summaries above, two are made by humans (gold standard), two by heuristics and two, automatically by a machine. Below, they are identified.

- Summary 1 — *Baseline:random* (heuristic)
- Summary 2 — Matias (2016) (machine)
- Summary 3 — Human 1 (gold standard)
- Summary 4 — Microsoft Office Word (machine)
- Summary 5 — Human 2 (gold standard)
- Summary 6 — *Baseline:first* (heuristic)

Following, the third text used in the Turing Test in Spanish, the summaries made by humans and those made by machines are presented.

Público mexicano sensibiliza a la Oreja de Van Gogh en el Auditorio Nacional

Se apagaron las luces del Auditorio Nacional minutos después de las 18:00 horas del domingo, los aplausos se elevaron y el grupo español, La Oreja de Van Gogh, salió al escenario para comenzar a llenar las paredes del recinto con el canto de su vocalista, Leire Martínez. Con una introducción del tema "Rosas", la cantante comenzó a entonar la melodía con un vestido corto, negro y con tres estoperoles en forma de rombos, además de hacer juego con unas botas y una pulsera enredada del mismo color. "El último vals" se hizo sonar con Leire Martínez, quien al término de la canción expresó: "Buenas noches México, como saben nuestro último trabajo, Primera Fila, lo realizamos aquí", expresó la española antes de presentar "Cuando dices adiós", bajo luces multicolores. El ritmo de "Mi calle es Nueva York" dio paso a temas de la banda como "Vestido azul", balada de su álbum Lo que te conté mientras te hacías la dormida, misma que se ilumino con luces color pastel. Prosiguió "Inmortal", canción de su material discográfico, A las cinco en el Astoria, que fuera el disco debut de Martínez en el grupo. "Algo se nos ha quedado en tantas ocasiones que hemos visitado este país, la canción es para darles un poquito de nosotros", expresó la líder de grupo para presentar "Una y otra vez", segunda canción inédita del formato Primera Fila. Antes de que la cantante expresara: "La noche irá cargada de sorpresas, hoy no podía faltar este invitado a nuestro concierto", dijo para presentar "Mi vida sin ti" junto con Samo, quien no fue bien recibido. Aparecieron los clásicos temas de La Oreja de Van Gogh como "París" que se ligó con "Europa VII", melodía de la banda para crear conciencia sobre los problemas sociales que existen alrededor del mundo; la música de Xabi San Martín levitó en el aire mientras Leire realizó un cambio de ropa. Con leggins, blusón a rayas y botas negras, la española volvió a escena para expresar: "Queremos estar muy cerca de ustedes, esta canción habla de las sensaciones cuando te entregan por primera vez a tu bebé", aseguró para acercarse al público y cantar "Palabras para Paula". Leonel García ingresó al escenario para repetir con la agrupación el momento musical logrado en el nuevo disco y cantar "La playa"; el baladista vestido con traje gris oscuro agradeció la invitación. La velada continuó y la banda rememoró el primer sencillo de toda su carrera musical: "El 28". No se fueron las sorpresas y el tema "Adiós", poco tocado en sus conciertos, resonó en el Auditorio Nacional para seguir con "María", "Deseos de cosas imposibles" y "Jueves", tema donde bajaron las luces, el público prendió los

celulares y proyectó en todo el espacio figuras que ante la vista panorámica crearon un cielo estrellado. La respuesta de Leire fue espontánea: lágrimas que cayeron sobre sus mejillas y expresó: "Gracias por recordarnos que estos son los momentos que más valen la pena, por confiar en la música en directo", aseguró Leire con voz entrecortada y visiblemente emocionada por los gritos y los aplausos de los seguidores. La intensidad musical de "Muñeca de trapo" y "La niña que llora en tus fiestas" inundó el espacio, luego llegó "El primer día del resto de mi vida" y "Pálida luna". El final se acercó en una fiesta de globos de colores que aventaron sus seguidores en todo el recinto y que dieron vida a los temas "Cometas por el cielo", "Pop", "20 de enero" y "Puedes contar conmigo". Al término, La Oreja de Van Gogh levantó la bandera de México y España y se abrazaron para celebrar la noche y despedirse de sus fans.

Summary 1

Se apagaron las luces del Auditorio Nacional minutos después de las 18:00 horas del domingo, los aplausos se elevaron y el grupo español, La Oreja de Van Gogh, salió al escenario para comenzar a llenar las paredes del recinto con el canto de su vocalista, Leire Martínez. "El último vals" se hizo sonar con Leire Martínez, quien al término de la canción expresó: "Buenas noches México, como saben nuestro último trabajo, Primera Fila, lo realizamos aquí", expresó la española antes de presentar "Cuando dices adiós", bajo luces multicolores. El ritmo de "Mi calle es Nueva York" dio paso a temas…

Summary 2

Se apagaron las luces del Auditorio Nacional minutos después de las 18:00 horas del domingo, los aplausos se elevaron y el grupo español, La Oreja de Van Gogh, salió al escenario para comenzar a llenar las paredes del recinto con el canto de su vocalista, Leire Martínez. Con una introducción del tema "Rosas", la cantante comenzó a entonar la melodía con un vestido corto, negro y con tres estoperoles en forma de rombos, además de hacer juego con unas botas y una pulsera enredada del mismo color. Aparecieron los clásicos temas de La Oreja de Van Gogh como "París" que…

Summary 3

Se apagaron las luces del Auditorio Nacional minutos después de las 18:00 horas del domingo, los aplausos se elevaron y el grupo español, La Oreja de Van Gogh, salió al escenario para comenzar a llenar las paredes del recinto con el canto de su vocalista, Leire Martínez. Con

una introducción del tema “Rosas”, la cantante comenzó a entonar la melodía con un vestido corto, negro y con tres estoperoles en forma de rombos, además de hacer juego con unas botas y una pulsera enredada del mismo color. “El último vals” se hizo sonar con Leire Martínez, quien al término de…

Summary 4

Se apagaron las luces del Auditorio Nacional minutos después de las 18:00 horas del domingo, los aplausos se elevaron y el grupo español, La Oreja de Van Gogh, salió al escenario para comenzar a llenar las paredes del recinto con el canto de su vocalista, Leire Martínez. Aparecieron los clásicos temas de La Oreja de Van Gogh como “París” que se ligó con “Europa VII”, melodía de la banda para crear conciencia sobre los problemas sociales que existen alrededor del mundo; la música de Xabi San Martín levitó en el aire mientras Leire realizó un cambio de ropa. La velada…

Summary 5

Se apagaron las luces del Auditorio Nacional minutos después de las 18:00 horas del domingo, los aplausos se elevaron y el grupo español, La Oreja de Van Gogh, salió al escenario para comenzar a llenar las paredes del recinto con el canto de su vocalista, Leire Martínez. Con una introducción del tema “Rosas”, la cantante comenzó a entonar la melodía con un vestido corto, negro y con tres estoperoles en forma de rombos, además de hacer juego con unas botas y una pulsera enredada del mismo color. “El último vals” se hizo sonar con Leire Martínez, quien al término de…

Summary 6

La respuesta de Leire fue espontánea: lágrimas que cayeron sobre sus mejillas y expresó: “Gracias por recordarnos que estos son los momentos que más valen la pena, por confiar en la música en directo”, aseguró Leire con voz entrecortada y visiblemente emocionada por los gritos y los aplausos de los seguidores Al término, La Oreja de Van Gogh levantó la bandera de México y España y se abrazaron para celebrar la noche y despedirse de sus fans El ritmo de “Mi calle es Nueva York” dio paso a temas de la banda como “Vestido azul”, balada de su álbum Lo…

Out of the summaries of this test, two were made by humans (gold standard), two by heuristics and two automatically generated by a machine. They are as follows:

- Summary 1 — Matias (2016) (machine)
- Summary 2 — Human 1 (gold standard)
- Summary 3 — Microsoft Office Word (machine)
- Summary 4 — Human 2 (gold standard)
- Summary 5 — *Baseline:first (heuristic)*
- Summary 6 — *Baseline:random* (heuristic)

Appendix B

Turing test in English

In this appendix two more Turing test trials in English are presented. The full texts and the summaries given to people to identify the two made by humans, as well as the tables which show the texts made by machines and by humans are included.

The second text used in the Turing Test in English and the summaries made by humans and those by machines are presented below

Gilbert Reaches Jamaican Capital With 110 MPH Winds

Hurricane Gilbert, packing 110 mph winds and torrential rain, moved over this capital city today after skirting Puerto Rico, Haiti and the Dominican Republic. There were no immediate reports of casualties. Telephone communications were affected. "Right now it's actually moving over Jamaica," said Bob Sheets, director of the National Hurricane Center in Miami. "We've already had reports of 110 mph winds on the eastern tip. "It looks like the eye is going to move lengthwise across that island, and they're going to bear the full brunt of this powerful hurricane," Sheets said. Forecasters say Gilbert was expected to lash Jamaica throughout the day and was on track to later strike the Cayman Islands, a small British dependency northwest of Jamaica. Meanwhile, Havana Radio reported today that 25,000 people were evacuated from Guantanamo Province on Cuba's southeastern coast as strong winds fanning out from Gilbert began brushing the island. All Jamaica-bound flights were canceled at Miami International Airport, while flights from Grand Cayman, the main island of the three-island chain, arrived packed with frightened travelers. "People were running around in the main lobby of our hotel (on Grand Cayman) like chickens with their heads cut off," said one vacationer who was returning home to California through Miami. Hurricane warnings were posted for the Cayman Islands, Cuba and Haiti. Warnings were discontinued for the Dominican Republic. "All interests in the Western Caribbean should continue to monitor the progress of this dangerous hurricane," the service said, adding, "Little change in strength is expected for the next several hours as the hurricane moves westward over Jamaica." The Associated Press' Caribbean headquarters in San Juan, Puerto Rico, was unable to get phone calls through to Kingston, where high winds and heavy rain preceding the storm drenched the capital overnight, toppling trees, causing local flooding and littering streets with branches. Most Jamaicans stayed home, boarding up windows in preparation for the hurricane. Some companies broadcast appeals for technicians and electricians to report to work. The weather bureau predicted Gilbert's center, 140 miles southeast of Kingston before dawn, would pass south of Kingston and hit the southern parish of Clarendon. Flash flood warnings were issued for the parishes of Portland on the northeast and

St. Mary on the north. The north coast tourist region from Montego Bay on the west and Ocho Rios on the east, far from the southern impact zone and separated by mountains, was expected only to receive heavy rain. Officials urged residents in the higher risk areas along the south coast to seek higher ground. "It's certainly one of the larger systems we've seen in the Caribbean for a long time," said Hal Gerrish, forecaster at the National Hurricane Center. Forecasters at the center said the eye of Gilbert was 140 miles southeast of Kingston at dawn today. Maximum sustained winds were near 110 mph, with tropical-storm force winds extending up to 250 miles to the north and 100 miles to the south. Prime Minister Edward Seaga of Jamaica alerted all government agencies, saying Sunday night: "Hurricane Gilbert appears to be a real threat and everyone should follow the instructions and hurricane precautions issued by the Office of Disaster Preparedness in order to minimize the danger." Forecasters said the hurricane had been gaining strength as it passed over the ocean after it dumped 5 to 10 inches of rain on the Dominican Republic and Haiti, which share the island of Hispaniola. "We should know within about 72 hours whether it's going to be a major threat to the United States," said Martin Nelson, another meteorologist at the center. "It's moving at about 17 mph to the west and normally hurricanes take a northward turn after they pass central Cuba." Cuba's official Prensa Latina news agency said a state of alert was declared at midday in the Cuban provinces of Guantanamo, Holguin, Santiago de Cuba and Granma. In the report from Havana received in Mexico City, Prensa Latina said civil defense officials were broadcasting bulletins on national radio and television recommending emergency measures and providing information on the storm. Heavy rain and stiff winds downed power lines and caused flooding in the Dominican Republic on Sunday night as the hurricane's center passed just south of the Barahona peninsula, then less than 100 miles from neighboring Haiti. The storm ripped the roofs off houses and flooded coastal areas of southwestern Puerto Rico after reaching hurricane strength off the island's southeast Saturday night. Flights were canceled Sunday in the Dominican Republic, where civil defense director Eugenio Cabral reported some flooding in parts of the capital of Santo Domingo and power outages there and in other southern areas.

Summary 1

Hurricane Gilbert, packing 110 mph winds and torrential rain, moved over this capital city today after skirting Puerto Rico, Haiti and the Dominican Republic. "We've already had reports of 110 mph winds on the eastern tip. "It looks like the eye is going to move lengthwise across that island, and they're going to bear the full brunt of this powerful hurricane," Sheets said. Forecasters say Gilbert was expected to lash Jamaica throughout the day and was on track to later strike the Cayman Islands, a small British dependency northwest of Jamaica. The weather bureau predicted Gilbert's center, 140 miles southeast…

Summary 2

National radio and television recommending emergency measures and the weather bureau predicted Gilbert's center, 140 miles was declared at midday in the Cuban provinces of Guantanamo, strong winds fanning out from Gilbert began brushing the island the northeast and St Mary on the north. The north coast tourist In the report from Havana received in Mexico City, Prensa Latina vacationer who was returning home to California through Miami "Right now it's actually moving over Jamaica," said Bob. The storm ripped the roofs off houses and flooded coastal areas "We should know within about 72 hours whether it's going to be…

Summary 3

Hurricane Gilbert, packing 110 mph winds and torrential rain, moved over this capital city today after skirting Puerto Rico, Haiti and the Dominican Republic. There were no immediate reports of casualties. Telephone communications were affected. "Right now it's actually moving over Jamaica," said Bob Sheets, director of the National Hurricane Center in Miami. "We've already had reports of 110 mph winds on the eastern tip. All Jamaica-bound flights were canceled at Miami International Airport, while flights from Grand Cayman, the main island of the three-island chain, arrived packed with frightened travelers. Hurricane warnings were posted for the Cayman Islands, Cuba…

Summary 4

Hurricane Gilbert hit Jamaica today with 110 mph winds and torrential rain, causing serious damage in Kingston overnight. The storm center is expected to hit land at Clarendon parish, then move lengthwise across the island. The government is preparing for the worst with government agencies on alert and coastal residents directed to move to higher ground. Communications have already been affected. Gilbert, described as one of the larger systems, has already caused some damage in

Puerto Rico, the Dominican Republic, Haiti, and Cuba. Fears are high on the Cayman Islands, the next target on its track…

Summary 5

Gilbert Reaches Jamaican Capital With 110 Mph Winds Hurricane Gilbert, packing 110 mph winds and torrential rain, moved over this capital city today after skirting Puerto Rico, Haiti and the Dominican Republic. There were no immediate reports of casualties. Telephone communications were affected. "Right now it's actually moving over Jamaica," said Bob Sheets, director of the National Hurricane Center in Miami. "We've already had reports of 110 mph winds on the eastern tip. "It looks like the eye is going to move lengthwise across that island, and they're going to bear the full brunt of this powerful hurricane," Sheets said…

Summary 6

Hurricane Gilbert, packing 110mph winds and torrential rain, moved over the Jamaican capital city of Kingston today after skirting Puerto Rico, Haiti and the Dominican Republic. It's tropical-storm force winds extend up to 250 miles to the north and 100 miles to the south. Hal Gerrish, a forecaster with the National Hurricane Center said it is one of the larger systems seen in the Caribbean for a long time. Warnings were posted for the Cayman Islands, Haiti and Cuba but discontinued for the Dominican Republic. Jamaicans are expecting to bear the brunt of Gilbert as its eye moves lengthwise across…

Out of the summaries for this trial, two were made by humans (gold standard), two by heuristics and two automatically by a machine. Their correspondence is as follows:

- Summary 1 — Copernic (machine)
- Summary 2 — *Baseline:random* (heuristic)
- Summary 3 — Matias (2016) (machine)
- Summary 4 — Human 1 (gold standard)
- Summary 5 — *Baseline:first* (heuristic)
- Summary 6 — Human 2 (gold standard)

Following, the third text used in the Turing test in English and the summaries made by humans and those made by machines are presented.

Hurricane Hits Jamaica With 115 mph Winds; Communications Disrupted

Hurricane Gilbert slammed into Kingston on Monday with torrential rains and 115 mph winds that ripped roofs off homes and buildings, uprooted trees and downed power lines. No serious injuries were immediately reported in the city of 750,000 people, which was hit by the full force of the hurricane around noon. For half an hour, the hurricane lashed the city, tearing branches from trees, blowing down fences and whipping paper through the air. The National Weather Service reported heavy damage to Kingston's airport and aircraft parked on its fields. The first shock let up as the eye of the storm moved across the city. Skies brightened, the winds died down and people waited for an hour before the second blow of the hurricane arrived. All Jamaica-bound flights were canceled at Miami International Airport. Flights from the Cayman Islands, reportedly next in the path of the hurricane, arrived in Miami packed with travelers cutting short their vacations. "People were running around in the main lobby of our hotel (on Grand Cayman Island) like chickens with their heads cut off," said one man. A National Weather Service report said the hurricane was moving west at 17 mph with maximum sustained winds of 115 mph. It said Jamaica would receive up to 10 inches of rain that would cause flash floods and mud slides. "Right now it's actually moving over Jamaica," said Bob Sheets, director of the National Hurricane Center in Miami. "It looks like the eye is going to move lengthwise across that island, and they're going to bear the full brunt of this powerful hurricane," he said. Gilbert reached Jamaica after skirting southern Puerto Rico, Haiti and the Dominican Republic. Hurricane warnings were issued Monday for the south coast of Cuba east of Camaguey, the Cayman Islands, and Haiti, while warnings were discontinued for the Dominican Republic. High winds and heavy rain preceding the storm drenched Kingston overnight, toppling trees, causing local flooding and littering streets with branches. Most of Jamaica's 2.3 million people stayed home, boarding up windows in preparation for the hurricane. The popular north coast resort area, on the other side of the mountains, was expected to receive heavy rain but not as much damage from the hurricane as the south coast, where officials urged residents to seek higher ground. Havana Radio, meanwhile, reported Monday that 25,000 people were evacuated from coastal areas in Guantanamo Province on the nation's southeastern coast as Gilbert's winds and rain began to brush the island. In Washington, the Navy reported its bases at Guantanamo Bay, Cuba, and Roosevelt Roads, Puerto Rico, had taken various precautionary steps but appeared to be safe from

the brunt of the hurricane. Lt. Ken Ross, a spokesman, said the Navy station at Guantanamo reported that as of 2:30 p.m. EDT, the brunt of the storm appeared to be passing southeastern Cuba. "They have reported maximum winds of 25 knots and gusts up to 50 knots," said Ross. "But there are no reports of injuries or damage." The spokesman said earlier in the day, Guantanamo had moved to "Condition Two," meaning electrical power usage was cut back to only essential uses and "all non-essential personnel sent to their barracks." The storm also skirted Puerto Rico without causing any damage to military facilities, Ross said. Sheets said Gilbert was expected next to sweep over the Cayman Islands, on its westward track, and in two to three days veer northwest into the southern Gulf of Mexico. Residents of the neighboring Caymans, a British dependency to the northwest, were urged to "rush all preparatory actions." The National Weather Service warned that the Caymans could expect high waters and large waves "which may undermine buildings along the beaches." "All interests in the Western Caribbean should continue to monitor the progress of this dangerous hurricane," the service advised. Forecaster Hal Gerrish on Sunday described Gilbert "certainly one of the larger systems we've seen in the Caribbean for a long time."

Summary 1

The full force of Hurricane Gilbert slammed into Kingston, Jamaica, at noon on Monday. Torrential rain and 115 mph winds severely damaged the city and its airport. No casualties or injuries have been reported. The storm is expected to move lengthwise across the island, dropping as much as 10 inches of rain. Next landfall is expected to be the Cayman Islands, where tourists are attempting to evacuate and residents were making preparations. The United States Navy reported its bases on Puerto Rico and at Guantanamo Bay, Cuba, were not damaged by Gilbert as it moved through the central Caribbean…

Summary 2

Hurricane, "he said evacuated from coastal areas in Guantanamo Province on the nation's residents to seek higher ground Havana Radio, meanwhile, reported Monday that 25,000 people were Jamaica would receive up to 10 inches of rain that would cause reported that as of 2:30 p m EDT, the brunt of the storm appeared an hour before the second blow of the hurricane arrived southeastern coast as Gilbert's winds and rain began to brush the For half an hour, the hurricane lashed the city, tearing

overnight, toppling trees, causing local flooding and littering Cuba, and Roosevelt Roads, Puerto Rico, had taken…

Summary 3

Hurricane Gilbert slammed into Kingston, Jamaica on Monday, with torrential rains and 115mph winds that ripped roofs off buildings, uprooted trees, downed power lines and did heavy damage to the airport and parked aircraft. No fatalities in this city of 750,000 people have been reported. Jamaica's popular north coast is not expected to receive as much damage as the south coast, where officials urged residents to seek higher ground. The storm skirted Puerto Rico and is now tracking toward the Cayman Islands. The U.S. Navy reports that its Cuban bases at Guantanamo Bay and Roosevelt Island appear to be relatively …

Summary 4

Hurricane Gilbert slammed into Kingston on Monday with torrential rains and 115 mph winds that ripped roofs off homes and buildings, uprooted trees and downed power lines. No serious injuries were immediately reported in the city of 750,000 people, which was hit by the full force of the hurricane around noon. For half an hour, the hurricane lashed the city, tearing branches from trees, blowing down fences and whipping paper through the air. The National Weather Service reported heavy damage to Kingston's airport and aircraft parked on its fields. The first shock let up as the eye of the storm…

Summary 5

Hurricane Gilbert slammed into Kingston on Monday with torrential rains and 115 mph winds that ripped roofs off homes and buildings, uprooted trees and downed power lines. No serious injuries were immediately reported in the city of 750,000 people, which was hit by the full force of the hurricane around noon. The National Weather Service reported heavy damage to Kingston's airport and aircraft parked on its fields. The first shock let up as the eye of the storm moved across the city. Skies brightened, the winds died down and people waited for an hour before the second blow of the…

Summary 6

Hurricane Gilbert slammed into Kingston on Monday with torrential rains and 115 mph winds that ripped roofs off homes and buildings, uprooted trees and downed power lines. A National Weather Service report said the hurricane was moving west at 17 mph with maximum sustained winds of 115 mph. Hurricane warnings were issued Monday for the

> south coast of Cuba east of Camaguey, the Cayman Islands, and Haiti, while warnings were discontinued for the Dominican Republic High winds and heavy rain preceding the storm drenched Kingston overnight, toppling trees, causing local flooding and littering streets with branches. The popular north coast…

Out of the summaries for this trial, two are the gold standards, two were made by the heuristics and the other two, automatically by a machine; their correspondence is as follows:

- Summary 1 — Human 1 (gold standard)
- Summary 2 — *Baseline:random* (heuristic)
- Summary 3 — Matias (2016) (machine)
- Summary 4 — Human 2 (gold standard)
- Summary 5 — *Baseline:first* (heuristic)
- Summary 6 — Copernic (machine)

Appendix C

Example of the summary in Portuguese

Appendix C presents an example of a news item in Portuguese and two summaries: one generated by a commercial tool and the other by a novel scientific method.

No Turing Test has been carried out in Portuguese yet. However, in this section, an example of an item from TeMário is included. Additionally, two summaries are presented; one generated by a commercial tool and the other by a novel scientific method.

CIDADE É CANTADA EM MAIS DE 1.800 MÚSICAS

Aos 440 anos São Paulo já foi cantada em pelo menos 1.800 músicas. Essa história de citar a cidade começou em 1750, quando dois compositores (Calixto e Anchieta Arzão) decidiram fazer "Missa à São Paulo", partitura recuperada e gravada pela primeira vez em 1970 com regência de Júlio Medaglia. Depois disso, parece que não foi mais possível conter homenagens e desilusões musicais dos compositores pela que é hoje a terceira maior metrópole do mundo. Até a primeira frase do Hino Nacional menciona São Paulo: "Ouviram do Ipiranga..." Esses dados foram pesquisados por um paraibano de João Pessoa, radicado em São Paulo desde 75, que há cinco anos levanta a história musical da cidade em sebos e livrarias. O escritor e jornalista Assis Angelo, 41, está agora preparando o que ele chama de a primeira enciclopédia musical sobre São Paulo. Os primeiros 300 verbetes já estão escritos. Angelo acredita que o material resultante da sua pesquisa é suficiente para 900 verbetes e umas 600 páginas de livro, à espera de patrocinadores. E como seu trabalho é enciclopédico, vale dizer que a cidade já foi cantada de "A" a "Z", passando por "X" e "Y", por intérpretes e compositores de todos os Estados brasileiros. Por exemplo, com "Z", "Zona Leste Total" (de Luiz Carlos, 1991); com "X", "Xamego Paulista" (de Arlindo Bettio e Nhozinho, 1987); com "Y", "Yayá do Peruche" (de L. Correa e Rodolfo Vila) e com "A", entre outras, "A Baixada do Glicério" (de Enerdino Ortiz, João Marques e Manoel Lourenço). Se quantidade significar amor, Adoniram Barbosa foi o mais apaixonado dos cantores. Adoniram lidera o ranking com nada menos que 22 músicas sobre São Paulo. Tom Zé e a dupla Tonico e Tinoco também têm lá sua quedinha pela cidade. Cada um gravou 11 músicas. Quem pensa que Caetano Veloso parou na "arroz de festa" Sampa quando quis falar da cidade, mostra ou que não entende nada desta cidade ou que não sabe nada de Caetano. O doce bárbaro fala de São Paulo em outras cinco músicas, menos que Itamar Assumpção (dez músicas) e mais que Alvarenga e Ranchinho, que gravaram quatro composições. O ano em que mais se cantou São Paulo foi o do 4.º Centenário, 1954. Foram gravadas 72 músicas, com versões até japonês

enaltecendo a cidade. Nem Hebe Camargo deixou de gravar a sua. "A Hebe vai ficar uma arara, mas a música que ela gravou, "Paulicéia em Festa", era um horror, muito ruim", diz o pesquisador Angelo. Já a composição "IV Centenário", de Mário Zan e J.M. Alves, fez tanto sucesso na época, que vendeu, numa São Paulo quase provinciana, mais de cinco milhões de discos. Só para registro, Mário Zan é o autor do bolero que só estourou e ficou bastante conhecido depois de gravado em espanhol. A estrofe é esta: "Dizem que os homens/ não devem chorar/ por uma mulher/ que não soube amar..." Se o Campeonato Paulista algum dia for definido pela quantidade de música composta para os times, só vai dar Corinthians. Das 140 músicas compostas para os clubes paulistas, 85 foram dedicadas ao time do Parque São Jorge.

Shvoong (commercial tool)

CIDADE É CANTADA EM MAIS DE 1.800 MÚSICAS VICTOR AGOSTINHO Aos 440 anos São Paulo já foi cantada em pelo menos 1.800 músicas. Essa história de citar a cidade começou em 1750, quando dois compositores (Calixto e Anchieta Arzão) decidiram fazer "Missa à São Paulo", partitura recuperada e gravada pela primeira vez em 1970 com regência de Júlio Medaglia. Até a primeira frase do Hino Nacional menciona São Paulo: "Ouviram do Ipiranga..." Adoniram lidera o ranking com nada menos que 22 músicas sobre São Paulo. O doce bárbaro fala de São Paulo em outras cinco músicas, menos que Itamar Assumpção (dez músicas) e mais que Alvarenga e Ranchinho, que...

AG-Multi (novel scientific method)

Aos 440 anos São Paulo já foi cantada em pelo menos 1.800 músicas. Depois disso, parece que não foi mais possível conter homenagens e desilusões musicais dos compositores pela que é hoje a terceira maior metrópole do mundo. Esses dados foram pesquisados por um paraibano de João Pessoa, radicado em São Paulo desde 75, que há cinco anos levanta a história musical da cidade em sebos e livrarias. Os primeiros 300 verbetes já estão escritos. Angelo acredita que o material resultante da sua pesquisa é suficiente para 900 verbetes e umas 600 páginas de livro, à espera de patrocinadores...

Appendix D

Example of a summary in Russian

Appendix D presents an example of a news item in Russian and two summaries; one generated by a commercial tool and another by a novel scientific method.

No Turing Test has been carried out in Russian yet. However, this section presents an example of a news item from *corpus TEXTRUSS*. Additionally, two summaries are presented: one made by a commercial tool and the other by a novel scientific method.

Как не стать жертвой автоподставы
Как узнать автоподставщика на дороге и что делать при встрече с автомошенниками

Фотография: Shutterstock

17.08.2015, 20:58 | Алина Распопова

Автоподставщики сами бросаются под колеса, а после вымогают у водителей деньги. Они просят компенсировать вред здоровью или поломку дорогих часов. Другие мошенники подставляются на своих старых иномарках под водителей-новичков и выманивают по пять тысяч евро за раз. Им удается убеждать водителей, что случай не страховой, и угрожают расправой. Как не попасться на уловки автоподставщиков, «Газете.Ru» рассказали эксперты ГУ МВД России по Москве. Чтобы добраться до кошельков наивных водителей, автоподставщики используют как новые, так и классические способы обмана. Так, некоторое время назад сотрудники Московского уголовного розыска задержали 43-летнего мужчину, который ловко изобразил, как его якобы сбил проезжающий мимо автомобиль. На самом деле он специально бросался под машины, а его сообщник наносил удар по автомобилю жертвы для имитации звука удара. Далее «пешеход» демонстрировал водителю сломанные дорогостоящие часы или планшетный компьютер. Для решения проблемы он требовал выплатить компенсацию. Для усиления психологического давления на водителя аферист представлялся адвокатом и показывал поддельное удостоверение. Чтобы дополнительно надавить на свою жертву, он звонил по громкой связи своему подельнику, выступающему в роли инспектора ГИБДД. Его сообщник уверенным голосом заявлял, что такой проступок влечет за собой лишение водительских прав. Обманутые люди отдавали «потерпевшему» крупные суммы, порой доходящие до миллиона рублей. По такой проверенной схеме действовали еще несколько злоумышленников, которые также попались в руки полиции. Еще одна организованная группа из трех человек обманом выманивала деньги у водителей, убеждая их, что они повредили их дорогую машину, а случай нестраховой. Доверчивые автомобилисты выкладывали до €5 тыс., только бы избавить себя от неприятностей. По просьбе «Газеты.Ru» специалисты в ГУ МВД России по Москве рассказали о том, как вычислить

автоподставщиков. Используя свой опыт, они объяснили, как работают автомошенники и как правильно себя вести при встрече с ними. Как выбирают жертву «Подставлялы — хорошие психологи, и чаще всего их жертвами становятся неопытные автолюбители, — рассказали «Газете.Ru» в ГУ МВД России по Москве. — В первую очередь это «чайники» и любители болтать за рулем по мобильному телефону. Еще один тип легкой добычи — начинающий водитель со знаком «У» на стекле. В числе потенциальных клиентов — те, кто водит агрессивно и постоянно перестраивается из ряда в ряд». Со слов самих преступников, в качестве жертвы они выбирают только мужчин. Женщины начинают звонить мужьям и друзьям, после чего приезжают люди и начинаются ненужные разборки. Вне зоны риска обладатели новых дорогих иномарок. Редко подставляются под машины, в которых едут несколько человек — свидетели мошенникам ни к чему. Для работы автоподставщики используют подержанные машины, но известных престижных марок. Это могут быть старенькие Mercedes, BMW, Audi, Volvo. На деле цена таких средств передвижения не выше $10 тыс. Неповрежденные машины практически никогда не подставляются.

Одиночные разводки

«Самый грубый способ одиночной подставы — обогнать жертву, подрезать и резко оттормозиться, подставив под удар корму, — отметили в ГУ МВД России по Москве. — Если предполагаемый «спонсор» не успел затормозить — мошенники будут стараться повесить на него всю вину. Ведь всегда виноват тот, кто сзади». Еще один распространенный вариант работы в паре: «охотник» заходит по правому борту жертвы сзади. Этот водитель ведет себя так, как будто и не собирается приближаться, и ждет перестроения «спонсора» в правый ряд. Следуя ПДД, «жертва» заблаговременно показывает правый «поворотник». В ответ на это «подстава» всем своим поведением дает понять «спонсору», что пропускает его. Но как только «жертва» подает вправо, «подстава» резко ускоряется и подставляет свой левый борт под удар. Для удобства инсценировки такого ДТП «подстава» выбирает темное время суток, машину темных цветов и едет только с «габаритами». Попасть на удочку может и водитель, который пытается выехать из левого или среднего ряда, в том числе и на круговом движении. В этот момент автомобилиста подсекают справа. Как правило, не до, а сразу же после удара следует возмущенное «бибиканье». Это тоже часть спектакля, рассчитанная на возможных свидетелей. Мошенники также любят подставлять правый борт под тех, кто бодро мчится по правому ряду, не уступая дорогу соседям слева от себя при объезде припаркованных машин.

Работа в паре

Есть и классические примеры парной работы. Потенциальная жертва должна двигаться по крайнему левому ряду. Ей на хвост плотно садится спешащий водитель и начинает сигналить дальним светом. Логика большинства нормальных водителей — уступить. Не ожидая подвоха, жертва начинает перестраиваться правее. В этот момент ее цепляет машина, которая до того спокойно двигалась чуть поодаль. Ранее водитель мог не обращать на нее внимания, либо она просто находилась в мертвой зоне. Согнавшая жертву с полосы машина якобы уезжает, жертва остается один на один с «невинно пострадавшим». Еще пример. На относительно свободной дороге, двигаясь по крайнему левому ряду с хорошей скоростью, потенциальная жертва «развода» догоняет вяло ползущего впереди якобы «чайника», который упорно отказывается уступить дорогу. Жертва делает рывок вправо, а там его уже поджидает перехватчик. Или такой способ: потенциальная «жертва» едет в среднем ряду. Справа подкатывает «подстава», слева – «сгоняющий» – дорогая машина, возможность контакта с которой инстинктивно отметается всяким нормальным водителем. Так они и едут втроем – параллельно и рядом. Вдруг, «сгоняющий» делает резкий поворот руля в сторону жертвы. Та, чтобы избежать контакта, тоже уходит вправо. Даже если

«жертва» контролирует свой правый борт, то «подстава» может неожиданно подвинуться к «жертве», оставаясь при этом в пределах своей полосы движения. В результате «жертва» въезжает правым бортом в левый борт «подставы». По ПДД в ДТП виновата «жертва». «Сгоняющий» уезжает, его задние номера не читаются.

Как понять, что вас разводят?

После инцидента подставщики немедленно начинают убеждать якобы виноватого водителя, что он должен заплатить добровольно и прямо на месте. Нередко злоумышленники устанавливают жесткие временные ограничения. Например, «деньги нужны через час, через 40 минут сервис закроется, у тебя есть два часа собрать деньги». Некоторые мошенники практикуют хорошо зарекомендовавший себя метод «злого-доброго». «Злой» требует максимальную сумму, а «добрый» соглашается на существенно меньшую, мотивируя тем, что, так уж и быть, ремонтироваться они будут у знакомых.

Что делать?

«Законный путь подстащикам изначально невыгоден. Они не получат желаемую сумму, и у них нет желания светиться в правоохранительных органах и страховых компаниях, — предупреждают в ГУ МВД России по Москве. — Договориться с ними по-хорошему невозможно. За один только поцарапанный бампер с вас потребуют баснословную сумму. А уж если вы поедете с

«подставлялами» в их сервис, будьте готовы, что мастер объявит о поврежденных лонжеронах и выставит внушительный счет. Поэтому закройтесь в машине и не выходите до приезда ГИБДД. Позвоните домой, друзьям, обрисуйте ситуацию, в которую попали, опишите марки, номера, приметы автомашин «подставлял» и их самих. Фотографируйте «подставлял» и номера их машины. В большинстве случаев уже одно это заставит их спешно залезть в свой автомобиль и уехать. При этом обязательно продержитесь до приезда сотрудников ГИБДД».

Tool4noobs (commercial tools)

Как не попасться на уловки автоподставщиков, «Газете.Ru» рассказали эксперты ГУ МВД России по Москве. Мошенники также любят подставлять правый борт под тех, кто бодро мчится по правому ряду, не уступая дорогу соседям слева от себя при объезде припаркованных машин. Как выбирают жертву «Подставлялы — хорошие психологи, и чаще всего их жертвами становятся неопытные автолюбители, — рассказали «Газете. Ru» в ГУ МВД России по Москве. Одиночные разводки «Самый грубый способ одиночной подставы — обогнать жертву, подрезать и резко оттормозиться, подставив под удар корму, — отметили в ГУ МВД России по Москве. Они не получат желаемую сумму, и у них нет желания светиться в правоохранительных органах и…

AG-Multi (novel scientific method)

Они просят компенсировать вред здоровью или поломку дорогих часов. Как не попасться на уловки автоподставщиков, «Газете.Ру» рассказали эксперты ГУ МВД России по Москве. Цхтобы добраться до кошельков наивных водителей, автоподставщики используют как новые, так и классические способы обмана. На самом деле он специально бросался под машины, а его сообщник наносил удар по автомобилю жертвы для имитации звука удара. Еще одна организованная группа из трех человек обманом выманивала деньги у водителей, убеждая их, что они повредили их дорогую машину, а случай нестраховой. Используя свой опыт, они объяснили, как работают автомошенники и как правильно себя вести при встрече с ними. В первую очередь это…

Appendix E

Stop words in English

Appendix E contains the list of stop words for AGTS in English.

A, ABLE, ABOUT, ABOVE, ACCORDING, ACCORDINGLY, ACROSS, ACTUALLY, AFTER, AFTERWARDS, AGAIN, AGAINST, AIN'T, ALL, ALLOW, ALLOWS, ALMOST, ALONE, ALONG, ALREADY, ALSO, ALTHOUGH, ALWAYS, AM, AMONG, AMONGST, AN, AND, ANOTHER, ANY, ANYBODY, ANYHOW, ANYONE, ANYTHING, ANYWAY, ANYWAYS, ANYWHERE, APART, APPEAR, APPRECIATE, APPROPRIATE, ARE, AREN'T, AROUND, AS, ASIDE, ASK, ASKING, ASSOCIATED, AT, AVAILABLE, AWAY, AWFULLY, B, BE, BECAME, BECAUSE, BECOME, BECOMES, BECOMING, BEEN, BEFORE, BEFOREHAND, BEHIND, BEING, BELIEVE, BELOW ,BESIDE, BESIDES, BEST, BETTER, BETWEEN, BEYOND, BOTHBRIEF, BUT, BY, C, C'MON, C'S, CAME, CAN, CAN'T, CANNOT, CANT, CAUSE, CAUSES, CERTAIN, CERTAINLY, CHANGES, CLEARLY, CO, COM, COME, COMES, CONCERNING, CONSEQUENTLY, CONSIDER, CONSIDERING, CONTAIN, CONTAINING, CONTAINS, CORRESPONDING, COULD, COULDN'T, COURSE, CURRENTLY, D, DEFINITELY, DESCRIBED, DESPITE, DID, DIDN'T, DIFFERENT, DO, DOES, DOESN'T, DOING, DON'T, DONE, DOWN, DOWNWARDS, DURING, E, EACH, EDU, EG, EIGHT, EITHER, ELSE, ELSEWHERE, ENOUGH, ENTIRELY, ESPECIALLY, ET, ETC, EVEN, EVER, EVERY, EVERYBODY, EVERYONE, EVERYTHING, EVERYWHERE, EX, EXACTLY, EXAMPLE, EXCEPT, F, FAR, FEW, FIFTH, FIRST, FIVE, FOLLOWED, FOLLOWING, FOLLOWS, FOR, FORMER, FORMERLY, FORTH, FOUR, FROM, FURTHER, FURTHERMORE, G, GET, GETS, GETTING, GIVEN, GIVES, GO, GOES, GOING, GONE, GOT, GOTTEN, GREETINGS, H, HAD, HADN'T, HAPPENS, HARDLY, HAS, HASN'T, HAVE, HAVEN'T, HAVING, HE, HE'S, HELLO, HELP, HENCE, HER, HERE, HERE'S, HEREAFTER, HEREBY, HEREIN, HEREUPON, HERS, HERSELF, HI, HIM, HIMSELF, HIS, HITHER, HOPEFULLY, HOW, HOWBEIT, HOWEVER, I, I'D, I'LL, I'M, I'VE, IE, IF, IGNORED, IMMEDIATE, IN, INASMUCH, INC, INC., INDEED, INDICATE, INDICATED, INDICATES, INNER, INSOFAR, INSTEAD, INTO, INWARD, IS, ISN'T, IT, IT'D, IT'LL, IT'S, ITS, ITSELF, J, JUST, K, KEEP, KEEPS, KEPT, KNOW, KNOWS, KNOWN, L, LAST, LATELY, LATER, LATTER, LATTERLY, LEAST, LESS, LEST, LET, LET'S, LIKE, LIKED, LIKELY, LITTLE, LOOK, LOOKING, LOOKS, LTD, M,

MAINLY, MANY, MAY, MAYBE, ME, MEAN, MEANWHILE, MERELY, MIGHT, MORE, MOREOVER, MOST, MOSTLY, MUCH, MUST, MY, MYSELF, N, NAME, NAMELY, ND, NEAR, NEARLY, NECESSARY, NEED, NEEDS, NEITHER, NEVER, NEVERTHELESS, NEW, NEXT, NINE, NO, NOBODY, NON, NONE, NOONE, NOR, NORMALLY, NOT, NOTHING, NOVEL, NOW, NOWHERE, O, OBVIOUSLY, OF, OFF, OFTEN, OH, OK, OKAY, OLD, ON, ONCE, ONE, ONES, ONLY, ONTO, OR, OTHER, OTHERS, OTHERWISE, OUGHT, OUR, OURS, OURSELVES, OUT, OUTSIDE, OVER, OVERALL, OWN, P, PARTICULAR, PARTICULARLY, PER, PERHAPS, PLACED, PLEASE, PLUS, POSSIBLE, PRESUMABLY, PROBABLY, PROVIDES, Q, QUE, QUITE, QV, R, RATHER, RD, RE, REALLY, REASONABLY, REGARDING, REGARDLESS, REGARDS, RELATIVELY, RESPECTIVELY, RIGHT, S, SAID, SAME, SAW, SAY, SAYING, SAYS, SECOND, SECONDLY, SEE, SEEING, SEEM, SEEMED, SEEMING, SEEMS, SEEN, SELF,SELVES, SENSIBLE, SENT, SERIOUS, SERIOUSLY, SEVEN, SEVERAL, SHALL, SHE, SHOULD, SHOULDN'T, SINCE, SIX, SO, SOME, SOMEBODY, SOMEHOW, SOMEONE, SOMETHING, SOMETIME, SOMETIMES, SOMEWHAT, SOMEWHERE, SOON, SORRY, SPECIFIED, SPECIFY, SPECIFYING, STILL, SUB, SUCH, SUP, SURE, T, T'S, TAKE, TAKEN, TELL, TENDS, TH, THAN, THANK, THANKS, THANX, THAT, THAT'S, THAT'S, THE, THEIR, THEIRS, THEM, THEMSELVES, THEN, THENCE, THERE, THERE'S, THEREAFTER, THEREBY, THEREFORE, THEREIN, THERES, THEREUPON, THESE, THEY, THEY'D, THEY'LL, THEY'RE, THEY'VE, THINK, THIRD, THIS, THOROUGH, THOROUGHLY, THOSE, THOUGH, THREE, THROUGH, THROUGHOUT, THRU, THUS, TO, TOGETHER, TOO, TOOK, TOWARD, TOWARDS, TRIED, TRIES, TRULY, TRY, TRYING, TWICE, TWO, U, UN, UNDER, UNFORTUNATELY, UNLESS, UNLIKELY, UNTIL, UNTO, UP, UPON, US, USE, USED, USEFUL, USES, USING, USUALLY, UUCP, V, VALUE, VARIOUS, VERY, VIA, VIZ, VS, W, WANT, WANTS, WAS, WASN'T, WAY, WE, WE'D, WE'LL, WE'RE, WE'VE, WELCOME, WELL, WENT, WERE, WEREN'T, WHAT, WHAT'S, WHATEVER, WHEN, WHENCE, WHENEVER, WHERE, WHERE'S, WHEREAFTER, WHEREAS, WHEREBY, WHEREIN, WHEREUPON, WHEREVER, WHETHER, WHICH, WHILE, WHITHER, WHO, WHO'S,

WHOEVER, WHOLE, WHOM, WHOSE, WHY, WILL, WILLING, WISH, WITH, WITHIN, WITHOUT, WON'T, WONDER, WOULD, WOULDN'T, X, Y, YES, YET, YOU, YOU'D, YOU'LL, YOU'RE, YOU'VE, YOUR, YOURS, YOURSELF, YOURSELVES, Z, ZERO.

Appendix F

Stop words in Spanish

Appendix F contains the list of stop words for AGTS in Spanish.

UN, UNA, UNAS, UNOS, UNO, SOBRE, TODO, TAMBIÉN, TRAS, OTRO, ALGÚN, ALGUNO, ALGUNA, ALGUNOS, ALGUNAS, SER, ES, SOY, ERES, SOMOS, SOIS, ESTOY, ESTA, ESTAMOS, ESTAIS, ESTAN, COMO, EN, PARA, ATRÁS, PORQUE, POR QUÉ, ESTADO, ESTABA, ANTE, ANTES, SIENDO, AMBOS, PERO, POR, PODER, PUEDE, PUEDO, PODEMOS, PODEIS, PUEDEN, FUI, FUE, FUIMOS, FUERON, HACER, HAGO, HACE, HACEMOS, HACEIS, HACEN, CADA, FIN, INCLUSO, PRIMERO, DESDE, CONSEGUIR, CONSIGO, CONSIGUE, CONSIGUES, CONSEGUIMOS, CONSIGUEN, IR, VOY, VA, VAMOS, VAIS, VAN, VAYA, GUENO, HA, TENER, TENGO, TIENE, TENEMOS, TENEIS, TIENEN, EL, LA, LO, LAS, LOS, SU, AQUÍ, MIO, TUYO, ELLOS, ELLAS, NOS, NOSOTROS, VOSOTROS, VOSOTRAS, SI, DENTRO, SOLO, SOLAMENTE, SABER, SABES, SABE, SABEMOS, SABEIS, SABEN, ULTIMO, LARGO, BASTANTE, HACES, MUCHOS, AQUELLOS, AQUELLAS, SUS, ENTONCES, TIEMPO, VERDAD, VERDADERO, VERDADERA, CIERTO, CIERTOS, CIERTA, CIERTAS, INTENTAR, INTENTO, INTENTA, INTENTAS, INTENTAMOS, INTENTAIS, INTENTAN, DOS, BAJO, ARRIBA, ENCIMA, USAR, USO, USAS, USA, USAMOS, USAIS, USAN, EMPLEAR, EMPLEO, EMPLEAS, EMPLEAN, AMPLEAMOS, EMPLEAIS, VALOR, MUY, ERA, ERAS, ERAMOS, ERAN, MODO, BIEN, CUAL, CUANDO, DONDE, MIENTRAS, QUIEN, CON, ENTRE, SIN, TRABAJO, TRABAJAR, TRABAJAS, TRABAJA, TRABAJAMOS, TRABAJAIS, TRABAJAN, PODRIA, PODRIAS, PODRIAMOS, PODRIAN, PODRIAIS, YO, AQUEL.

Appendix G

Stop words in Portuguese

Appendix G contains the list of stop words for AGTS in Portuguese.

DE, A, O, QUE, E, DO, DA, EM, UM, PARA, COM, NÃ£O, UMA, OS, NO, SE, NA, POR, MAIS, AS, DOS, COMO, MAS, AO, ELE, DAS, Ã, SEU, SUA, OU, QUANDO, MUITO, NOS, JÃ¡, EU, TAMBÃ©M, SÃ³, PELO, PELA, ATÃ©, ISSO, ELA, ENTRE, DEPOIS, SEM, MESMO, AOS, SEUS, QUEM, NAS, ME, ESSE, ELES, VOCÃª, ESSA, NUM, NEM, SUAS, MEU, Ã S, MINHA, NUMA, PELOS, ELAS, QUAL, NÃ³S, LHE, DELES, ESSAS, ESSES, PELAS, ESTE, DELE, TU, TE, VOCÃªS, VOS, LHES, MEUS, MINHAS, TEU, TUA, TEUS, TUAS, NOSSO, NOSSA, NOSSOS, NOSSAS, DELA, DELAS, ESTA, ESTES, ESTAS, AQUELE, AQUELA, AQUELES, AQUELAS, ISTO, AQUILO, ESTOU, ESTÃ¡, ESTAMOS, ESTÃ£O, ESTIVE, ESTEVE, ESTIVEMOS, ESTIVERAM, ESTAVA, ESTÃ¡VAMOS, ESTAVAM, ESTIVERA, ESTIVÃ©RAMOS, ESTEJA, ESTEJAMOS, ESTEJAM, ESTIVESSE, ESTIVÃ©SSEMOS, ESTIVESSEM, ESTIVER, ESTIVERMOS, ESTIVEREM, HEI, HÃ¡, HAVEMOS, HÃ£O, HOUVE, HOUVEMOS, HOUVERAM, HOUVERA, HOUVÃ©RAMOS, HAJA, HAJAMOS, HAJAM, HOUVESSE, HOUVÃ©SSEMOS, HOUVESSEM, HOUVER, HOUVERMOS, HOUVEREM, HOUVEREI, HOUVERÃ¡, HOUVEREMOS, HOUVERÃ£O, HOUVERIA, HOUVERÃAMOS, HOUVERIAM, SOU, SOMOS, SÃ£O, ERA, Ã©RAMOS, ERAM, FUI, FOI, FOMOS, FORAM, FORA, FÃ´RAMOS, SEJA, SEJAMOS, SEJAM, FOSSE, FÃ´SSEMOS, FOSSEM, FOR, FORMOS, FOREM, SEREI, SERÃ¡, SEREMOS, SERÃ£O, SERIA, SERÃAMOS, SERIAM, TENHO, TEM, TEMOS, TÃ©M, TINHA, TÃNHAMOS, TINHAM, TIVE, TEVE, TIVEMOS, TIVERAM, TIVERA, TIVÃ©RAMOS, TENHA, TENHAMOS, TENHAM, TIVESSE, TIVÃ©SSEMOS, TIVESSEM, TIVER, TIVERMOS, TIVEREM, TEREI, TERÃ¡, TEREMOS, TERÃ£O, TERIA, TERÃAMOS, TERIAM.

Appendix H

Documents in *corpus* TER

Appendix H presents the documents contained in *corpus* TER, created for AGTS tasks in Spanish. Each stage of its construction is described, and finally the structure of the *corpus*.

H.1 Introduction

This document presents the construction of a *corpus* composed of texts in Spanish for summaries with a view to serving as a support in Spanish natural language processing, mainly in Automatic Summary Generation. The *corpus* was created in the context of Red Temática en Tecnologías del Lenguaje [Thematic Network for Language Technology], Red TTL.

It comprises items of news in Mexican Spanish and two summaries made by two humans. The main objective of the document is to serve as a *corpus* to assess the commercial tools and novel scientific methods on the generation of extractive summaries. However, it may be utilized for various ends such as the linguistic analysis of texts, either for summaries or only text analyses, information retrieval systems or topic detection.

At present, some novel scientific methods work with the generation of extractive summaries, such as: Ledeneva (2008), Ledeneva, (2008a), García (2008), Montiel, (2009), García, (2013), Mendoza, (2014), Meena, (2015), Bhargava, (2016), among others. However, all of them work in English only. There are others that are independent from language and can be used in a number of collections, namely: Mihalcea, (2005), Patel, (2007), Last, (2010), Saggion, (2011). In spite of testing with more than one collection in English, Portuguese, Chinese, among others, but they leave one of the most important aside, Spanish.

According to Cervantes (2013), experts predict that by 2050 there will be more than 530 million Spanish speakers, of which 100 million will be living in the United States. This reveals a vast field work for NLP in Spanish, so it is important to have a *corpus* in this regard, which besides is in Mexican Spanish to understand better the behavior of various methods and tools to produce extractive summaries in our language.

H.2 *Corpus* of texts in Spanish for summaries

H.2.1 General characteristics

The *corpus* created in Mexican Spanish is exclusive for the generation of extractive summaries from single documents. It is presented in digital format and is composed of news items.

H.2.1.1 Compilation protocol

Search and access to the information

The *corpus* was created out of electronic items of news available on the Internet. The items were retrieved from the official website of "Crónica" newspaper ("La Crónica de Hoy | La noticia hecha diario," 2014). A total of 20 items were selected in each of the following categories: academia, wellbeing, city, culture, sports, entertainment, states, world, national, business, opinion and society; totaling 240 news items. Those selected corresponded to April 2014. One of the most important considerations to select the items was that they had various lengths, all of them longer than one hundred words.

Preprocessing

The news items were downloaded from the Web in an .html format, so the following cleaning and normalization process was carried out.

Figure H.1 Stages of preprocessing

- Locating their important parts. In addition to the text itself, the news items available on the Internet can contain other information such as advertisements, photographs, link to other pages, etc. Owing to this, it was necessary to detect the part that offer relevant and necessary information to build the *corpus*.

The segments chosen were the code of the items, which is a single number that identifies it and is part of the file name; the items' title; its category; publication date; and, the text proper.

- Cleaning. The cleaning process consists in suppressing all .html tags, text, images, links, images, etc., leaving only the item's title, its category, publication date and the text. The cleaning was carried out using a program in Java, resorting to regular expressions so that it would be automatically applied to all texts.
- Normalizing. Once clean, the texts were tagged to more easily identify all the parts of the item. **Table H.1** displays the tags used and an example of it is provided.

Table H.1 Description of tags for the full texts

Tags	Description
<DOC></DOC>	Indicates the beginning and end of a document
<DOCNO> </DOCNO>	Indicates the document's name
<FILEID></FILEID>	Indicates the document's single number
<HEAD></HEAD>	Indicates the document's title
<CATEGORY> </ CATEGORY >	Indicates the document's category
<DATE></DATE>	Indicates the document's issuing date
<TEXT></TEXT>	Indicates the text to summarize
<s><\s>	Indicates the beginning and end of a sentence

Example of a full text tagged

<DOC>
<DOCNO>
<s docid="09ED020414_825542" num="1" wdcount="1"> 825542 </s>
</DOCNO>

<FILEID>
<s docid="09ED020414_825542" num="2"
wdcount="1">09ED020414_825542</s>
</FILEID>
<HEAD>
<s docid="09ED020414_825542" num="3" wdcount="10"> Atención digna a grupos vulnerables distingue a Toluca, afirma alcaldesa </s>
</HEAD>
<CATEGORY>
<s docid="09ED020414_825542" num="4" wdcount="1"> Estados </s>
</CATEGORY>
<DATE>
<s docid="09ED020414_825542" num="5" wdcount="1"> 2014-04-02 </s>
</DATE>
<TEXT>
<s docid="09ED020414_825542" num="6" wdcount="61"> La alcaldesa de Toluca, Martha Hilda González Calderón, encabezó la entrega de auxiliares auditivos, sillas de ruedas y lentes, en beneficio de 240 personas del municipio, acompañada por la presidenta del sistema DIF Toluca, Diana Elisa González Calderón, y el representante del secretario de Salud de la entidad, César Nomar Gómez Monge, Pedro Montoya Moreno, coordinador de Salud de dicha secretaría</s>

<s docid="09ED020414_825542" num="7" wdcount="82"> En su mensaje, González Calderón refrendó su compromiso con las personas con discapacidad y dijo que la atención a grupos vulnerables distingue a Toluca como Municipio Educador, además de que hay la responsabilidad de brindar las herramientas necesarias a aquellos grupos con algún grado de vulnerabilidad para contribuir a mejorar su calidad de vida Reconoció el apoyo y coordinación de la Secretaría de Salud estatal que, dijo, se traduce en mejores condiciones para los toluqueños y las toluqueñas que más lo necesitan</s>

<s docid="09ED020414_825542" num="8" wdcount="40"> En presencia de miembros del Cabildo y de autoridades municipales, la presidenta del sistema DIF de Toluca, Diana Elisa González Calderón, indicó que en esta ocasión se entregaron 110 auxiliares auditivos, 100 juegos de lentes y 30 sillas de ruedas.</s>

</TEXT>
</DOC>

Storing

To name each of the files, the following considerations were taken: as they were 20 files per category, a consecutive number was assigned (1-20); after this, two letters for each category were taken: academia (AC); wellbeing (BI); city (CI); culture (CU); Sports (DE); entertainment (ES); states (ED); world (MU); national (NA); business (NE); opinion (OP); and, society (SO). Followed by the category abbreviation, the date was placed while an underscore separated the item's code. An example of a file name: 01AC010414_825278.txt. Finally, there were 12 folders with 12 files each; in total, there were 240 files.

H.2.2 Summary construction

Once the *corpus* of news items in Spanish was built, two summaries for each file were made by human beings.

Selection of humans. The considerations to select a human were Mexican nationality, university graduated and were given the following indications.

H.2.2.1 Construction of summaries

The people were given a news item divided into sentences with the number of words for each of them. They were asked to read the full item and select the sentences they considered important. Of those selected, they were asked to make a summary longer than one hundred words. In Appendixes, there is an example of the instructions given to the humans and the list with the names of each of them.

Table H.2 describes the tags in the summaries made by humans and presents an example of tagging.

Table H.2 Description of tags for the summaries

Tags	Description
<SUM></SUM>	Indicates the beginning and end of the human-made summary
CATEGORY	Indicates the item's category
TYPE	Indicates the type of summary, in this case on a document basis
SIZE	Indicates the least number of words the summary must have
DOCREF	Shows the name of the source document to produce the extractive summary.
SELECTOR	Indicates the initials of the human who makes the summary
Summarizer	Indicates which summary is A - the fist- and B -the second-

Example of a tagged summary

<SUM
CATEGORY="ESTADOS"
TYPE="PERDOC"
SIZE="100"
DOCREF="09ED020414_825542"
SELECTOR="EX"
Summarizer="B">

La alcaldesa de Toluca, Martha Hilda González Calderón, encabezó la entrega de auxiliares auditivos, sillas de ruedas y lentes, en beneficio de 240 personas del municipio, acompañada por la presidenta del sistema DIF Toluca, Diana Elisa González Calderón, y el representante del secretario de Salud de la entidad, César Nomar Gómez Monge, Pedro Montoya Moreno, coordinador de Salud de dicha secretaría. En su mensaje, González Calderón refrendó su compromiso con las personas con discapacidad y dijo que la atención a grupos vulnerables distingue a Toluca como Municipio Educador, además de que hay la responsabilidad de brindar las herramientas necesarias a aquellos grupos con algún grado de vulnerabilidad para contribuir a mejorar su calidad de vida Reconoció el apoyo y coordinación de la Secretaría de Salud estatal que, dijo, se traduce en mejores condiciones para los toluqueños y las toluqueñas que más lo necesitan.

</SUM>

H.2.2.2 Collection of human-made summaries

Once humans produced the extractive summary, each was assigned a key to name the summary files, as follows: example of file name: SUM_01AC010414_825278_LX.sum.

As noticed, in order to identify that the file belongs to the model summaries, the SUM label was added to everyone; after this, the name of the original item and finally, the key assigned to the human was added. These files' extension is .sum.

Finally, there were twelve files with forty files each, totaling 4810 model summary files.

H.2.3 *CORPUS* DESCRIPTION

As mentioned, the *corpus* comprises 240 news items from a number of categories. The collection is presented tagged, in which each part of the text is described. It is important to mention that the *corpus* was divided into sentences, which are also tagged to simplify the text analysis.

Following, **Table H.3** shows the categories of the *corpus*, the number of documents in it and the number of sentences.

The summaries generated by humans are longer than one hundred words. However, to assess the methods and tools to generate the summaries, assessment is carried out at one hundred words.

H.2.4 *CORPUS* ORGANIZATION

As noticed in **figure H.2**, the *corpus* contains two folders. Resúmenes, in which the summaries made by two humans for each of the original documents are located. Finally, as displayed in **figure H.3**, Textos Originales folder holds two folders Textos por archivos and Textos por categoría, in which the original full texts tagged are located.

Table H.3 Characteristics of the *corpus*' full texts

	Category	No. of texts	Number of words	Word average	No. of sentences	Sentence average
Crónica newspaper	Academia	20	10966	548,3	382	19,1
	Wellbeing	20	11801	590,05	405	20,25
	City	20	7568	378,4	219	10,95
	Culture	20	8631	431,55	297	14,85
	Sports	20	9519	475,95	363	18,15
	Entertainment	20	8869	443,45	311	15,55
	States	20	7471	373,55	185	9,25
	World	20	7108	355,4	247	12,35
	National	20	7533	376,65	186	9,3
	Business	20	7523	376,15	229	11,45
	Opinion	20	12716	635,8	443	22,15
	Society	20	6507	325,35	228	11,4
	Total	240	106212		3495	
	Media			442,55		14,5625

The *corpus* is organized as follows

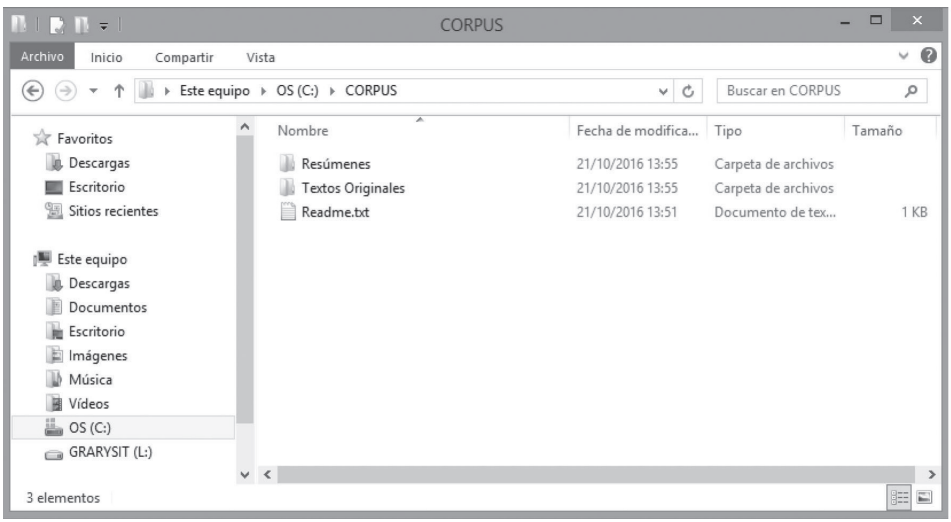


Figure H.2 Directory of the *corpus*

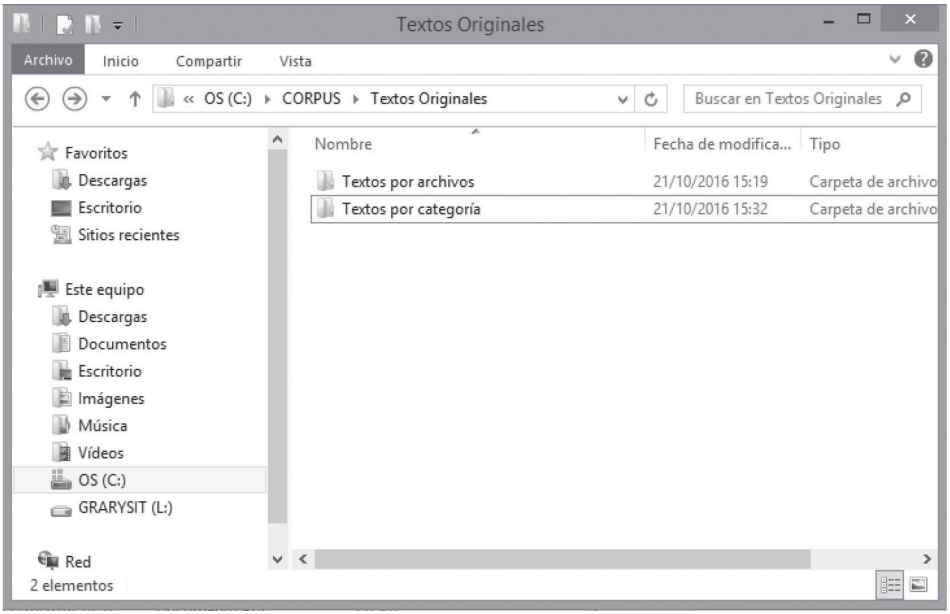


Figure H.3 Directory of the original texts

The only difference between these two files is that in Textos por categoría, there are 12 folders, one for each category of the *corpus* in which there are 20 files of the category. While, in Textos por archivos, there are 240 files. Figure H.4 shows these folders' content.

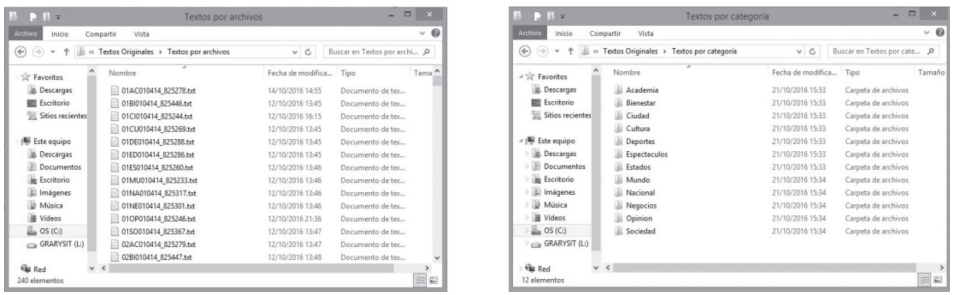


Figure H.4 Directory of Textos por archivos and Textos por categoría

H.3 Final considerations

As previously mentioned, the *corpus* proposed in this work is built to be used mainly in the study of extractive summaries in Spanish. The *corpus* presents this tagging so that only the full texts and the summaries in various folders are shown. However, tagging allows suppressing the parts that are not considered for the analysis of this collection; for example, if it is considered to work only with the text, only that contained between tags <TEXT></TEXT> is considered. One of the important contributions of this work is that text is separated into sentences, which means a standardization for future uses.

References (Appendix H)

Bhargava, 2016	Bhargava, R., Sharma, Y., & Sharma, G. (2016). ATSSI: Abstractive Text Summarization Using Sentiment Infusion. Procedia Computer Science, 89, 404-411.
Crónica. (s.f.)	© La Crónica Diaria S.A. de C.V. Obtenido de © La Crónica Diaria S.A. de C.V: http://www.cronica.com.mx/noticias.php
García, 2008	García R., Montiel, R., Ledeneva, Y., Rendón, e., Gelbukh, A. & Cruz, R. (2008). Text Summarization by Senten ce Extraction Using Unsupervised Learning. 7° Conferencia Internacional Mexicana de Inteligencia Artificial (MICAI08); Notas de la conferencia de Inteligencia Artificial, Springer-Verlag, Vol 5317, pp133-143.
García, 2013	García-Hernández, R. A., & Ledeneva, Y. (2013, June). Single extractive text summarization based on a genetic algorithm. In Mexican Conference on Pattern Recognition (pp. 374-383). Springer Berlin Heidelberg.
Last, 2010	Last, M. & Litvak, M. (2010). Language-independent Techniques for Automated Text Summarization. NATO Science for Peace and Security Series - D: Information and Communication Security. Vol. 27: Web Intelligence and Security, pp. 207-237.

Ledeneva, 2008	Ledevena, Y. N. (2008). Automatic Language-Independent Detection of Multiword Descriptions for Text Summarization. México, D.F.: Presentada en el Instituto Politécnico Nacional, para obtención del grado de Doctor.
Ledeneva, 2008[a]	Ledeneva, Y., Gelbukh, A. & García, R. (2008). Keeping Maximal Frequent Sequences Facilitates Extractive Summarization. Research in Computing Science, Vol. 34, pp.163-174.
Meena, 2015	Meena, Y. K., & Gopalani, D. (2015). Feature Priority Based Sentence Filtering Method for Extractive Automatic Text Summarization. Procedia Computer Science, 48, 728-734.
Mendoza, 2014	Mendoza, M., Bonilla, S., Noguera, C., Cobos, C., & León, E. (2014). Extractive single-document summarization based on genetic operators and guided local search. Expert Systems with Applications, 41(9), 4158-4169.
Mihalcea, 2005	Mihalcea, R. & Taran, P.. (2005). A Language Independent Algorithm for Single and Multiple Document Summarization. Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), Vol. 1, pp. 602-607.
Montiel, 2009	Montiel, R. (2009). Generación automática de resúmenes mediante aprendizaje no supervisado. Edo. de México: Presentada en el Instituto Tecnológico de Toluca, para obtención del Título de Ingeniero en Sistemas Computacionales.
Patel, 2007	Patel, A., Siddiqui, T & Tiwary, U. (2007). A language independent approach to multilingual text summarization. Conference RIA2007, Pittsburgh PA, U.S.A., 123-132.
Saggion, 2011	Saggion, H., Szasz, S., & Grupo, T. A. L. N. (2011). A Bilingual Summary *Corpus* for Information Extraction and other Natural Language Processing Applications. on Iberian Cross-Language Natural Language Processings Tasks (ICL 2011), 28.

Appendix I

Documents in *corpus* TeMário

Appendix I presents the documents contained in *corpus* TeMário. A translation of the document, originally in Portuguese, is made[TN] as it presents such *corpus* (Pardo and Rino, 2003); this with a view to accounting for the details of TeMário.

[TN]The present English version was translated from the authors' translation from Portuguese to Spanish.

Universidad de San Paulo - USP
Universidad Federal de San Carlos - UFSCar
Universidad Estadual Paulista - UNESP

TeMário: A *corpus* for automatic text summaries

Thiago Alexandre Salgueiro Pardo
Lucia Helena Machado Rino

NILC-TR-03-09
October 2003
Series of Reports from the Interinstitutional Center for Computational Lnguistics,
NILC - ICMC-USP, ZIP 668, 13560-970 San Carlos, SP, Brazil

Summary

This appendix describes TeMário, which is a *corpus* oriented to automatic text summaries. Developed for a number of purposes such as linguistic analysis, production of automatic summaries and their later assessment, TeMário is basically composed of news items and their summaries are in Portuguese. This were made by an expert and writer for text publication in Portuguese. This *corpus* is mainly utilized for specific research of automatic summarization methods in the context of project EXPLOSA.[33]

Index

[33]Developed with support from BY FAPESP (PROC. NRO. 01/08849-8).

I.1 Introduction

This appendix describes TeMário (acronym for '***TE**xtos com su**MÁRIO**s*'), which is a *corpus* produced with a view to obtaining automatic summaries in the context of project EXPLOSA[34] (exploration of several automatic summarization methods).

This *corpus* comprises news items and their corresponding manual summaries, produced by an expert writer for their publication in Portuguese.[35] In addition to serving for various purposes such as linguistic analysis and construction and production of automatic summaries and assess such systems, which will deal with other related tasks, whose current interest areas entail information retrieval and topic detection.

In EXPLOSA there are various systems that may be benefitted from this formation and evaluation *corpus*, for example GistSumm (Pardo *et. al*., 2003a), NeuralSumm (Pardo *et al*. 2003b), DMSumm[36] (Pardo, 2002), SuPor (Módolo, 2003) and UNLSumm (Martins, 2002). In addition to these systems, whose generic information can be found at NILC (http:// www.nilc.icmc. usp.br/), other activities may be developed with TeMário. For example, studies on the way an expert recognizes the relevant information in a text to compose the summaries, or the identification of parameters that indicate the criteria to summarize the modeling of computing systems. Details on these tasks and their relationship with automatic summaries were originally in Pardo and Rino (2003) and Martins *et al*. (2001).

[34]http://www.dc.ufscar.br/lucia~/projects/EXPLOSA.htm (FAPESP, Proc. Emisión. 01/08849-8).

[35]The manual summaries used in the present work made by a professional. In English, the name of professional summaries or human summarizers is also used by some authors.

[36]TAll these are available for download at http://www.nilc.icmc.usp.br/nilc/index.php/team?id=23#resource

Besides the tasks directly related to automatic summaries, at present other NILC projects may use TeMário, namely project LAZIO-WEB, which is a construction of resources for various research works, among them information retrieval as well as text or information tagging in Portuguese. In a broader context, the program will be part of Linguateca,[37] which is a large international resource repository that contains data and information on automatic processing in Portuguese.

I.2 TeMário

I.2.1 General characteristics

The name of TeMário to focus on the *corpus* was chosen owing to two reasons: to refer to the objects that compose it -summaries and texts- and the word tema [topic] in its name, whose recognition is essential in the task of summarizing.

To build TeMário, one hundred news items were collected, amounting 61.412 words. Sixty online texts from newspaper Folha de Sao Paulo (henceforward **FSP**), evenly distributed into: special, world, opinion; the other forty were published in the Jornal do Brasil newspaper (henceforward **JB**), also online and evenly distributed as well: international and politics. **Table I.1** summarizes these data, showing the number of words by section and the average words per section and their average in each text section. According to **table I.1**, the average words per section is 12.282, while the average per text are 613; this corresponds to texts that range from 1 to 2 pages and a half.

The news items were chosen to produce the *corpus* because they used a register aimed at a large readership, and so a Portuguese-language coverage in terms of vocabulary and in terms of grammar structures. This way, more supplements such as newspaper comments were automatically excluded from the section; FSP, for example, is aimed at more cultured readers.

[37] http://www.linguateca.pt/

Table I.1 Characteristics of the *corpus* of "source-texts"

Newspapers	Sections	Number of texts	Number of words	Word average/ text
Folha de São Paulo	Special	20	12340	617
	World	20	13739	686
	Opinion	20	10438	521
Jornal do Brasil	Internacional	20	12098	604
	Politics	20	12797	639
	Total	100	61412	
	General average		12282	613

This limitation has as a main objective goal to facilitate tasks related to automatic summaries; it is usual to resort to specialized labor force to produce assessment of automatic summaries. A more exaggerated style makes reading, comprehending and assessing difficult, which leads to wrong results for the focus of the task.

This relation is also noticed in the fact that at present news items are the most used in large-scale assessment of automatic summaries: international competitions of evaluation of automatic summaries such as text SUMMArization evaluation conference (SUMMAC) and DUC, which have used large-data-volume news items.

To build TeMário, once the news items were collected, the corresponding summaries were made, so the texts are called "source-texts".

I.2.2 Summary construction

The gathered texts are sent to the expert and writer to be published in Portuguese and to carry out two tasks; the creation of summaries corresponding to Task 1 (TeMário appendix), and for each source-text its main idea Task 2 (TeMário appendix). Hence, in task 1 the professor produced informative summaries; in task 2, a simple text reading was undertaken to learn the most important.

Relating both tasks in the identification of the main idea in a text is essential to produce good informative summaries; this is to say, to identify the phrases that refer to the main idea, these will lead the expert to compose the summaries, as they all must have (a significant part) main information from the "source-text" and even they can replace (the main condition of informative summaries). It is considered here the alternance between expert, reader and writer functions, the task commonly known as "rewriting the source-text in a condensed form" (Mani, 2001).

Besides the need of producing informative summaries, the automatic system generator of summaries had an additional restriction, the summary size had to be about 25-30% of the source-text. From the standpoint of the automatic summary, this is equivalent to establish "source-text" compression rates for the 70-75% interval, i.e., 70 or 75% of these texts' content has to be discarded to produce summaries.

The instructions the expert carried out in both tasks, the summary and the phrase marking are found in Appendix A.

I.2.3 *Corpus* complement

TeMário is composed of 100 "source-texts" and their summary manuals are a significant (though relatively small) set of data for various automatic-summary tasks such as the formation of automated systems and the personalization of manual text summaries of the same genre and domain. However, for assessing tasks, considering that manual summaries are the result of a process of rewriting the content of "source-texts" the writer considers most relevant, using those manual as "ideal" summaries with a view to understanding those automatically generated is not an easy task; it is hardly an explicit connection. Therefore, the reviews of manual and automatic summaries, by and large, are expensive and complex to apply. In order to minimize this problem, it is common to use extract of "ideas" instead of ideal summaries to compare with automatic results, especially in the production of an extractive summary.

In this case, the terminology indicates that the extracts of ideas and the extracts from this summary process itself come from the extraction methodology, the application of textual segments selected for the summarized text, whose main characteristic is that of reproducing part of the "source-text" literally. This way, it is possible to simply consider the similar patterns between the ideal extracts and their corresponding extracts to assess them in order to define whether they are good representatives of the main idea in the "source-texts" and if the extracts are ideal. Obviously, this stage may be performed automatically in most of the cases and the assessment above may be different, with considerable benefits in terms of cost and complexity. Owing to this, TeMário was complemented with extracts produced by a generator of extracts of ideas with of the manual summaries.

The generator of extracts of ideas identifies and extracts phrases of juxtaposed sentences from "source-texts" with the same content as the phrases of the corresponding manual summaries. To do so, it utilizes Salton's cosine following Rino and Pardo's (2003) methodology. It is important to state that the extracts of ideas cannot be actually ideal to make the most of the full and relevant ideas from the "source-text", as an expert would: a measurement of the cosine as it is purely based on the concurrence of words in the manual summary and the "source-text" may produce extracts with inappropriate phrases. However, these extracts may be considered ideal to be used as best as possible from the cost/benefit standpoint of automatic production.

Table I.2 relates the sizes of manual summaries and extracts of ideas. It is worth underscoring that the manual summaries' average number of words is significantly lower than the average of words in the extracts of ideas; such difference may come from the fact that the expert is capable of summarizing the desired content as best as possible to meet the summaries requirements by rewriting it. In the case of the extraction of ideas that meets these restrictions, it is not always trivial as it is previously fixed in the minimal unit to extract the "source-texts" in general; the phrases are fully extracted to compose the summaries. Owing to this, it is more common to have more extracts from manual summaries.

Table I.2 Characteristics of manual summaries and extracts of ideas

		Manual summaries		Extracts of ideas	
Newspapers	**Sections**	**Number of words**	**Word average / sections**	**Number of words**	**Word average /sections**
Folha de São Paulo	Special	4313	215	4450	222
	World	4234	211	4706	235
	Opinion	3373	168	3980	199
Jornal do Brasil	International	3734	186	5676	283
	Politics	3791	189	4451	222
	Total	19445		23263	
	General averages	3889	193	4652	232

I.2.4 Organization of TeMário

Considering a hierarchical environment in which the files may be stored by using Microsoft Windows, this program organizes in a single folder with two subfolders, respectively added, "source-texts" and "summaries", as displayed in **figure I.1**.

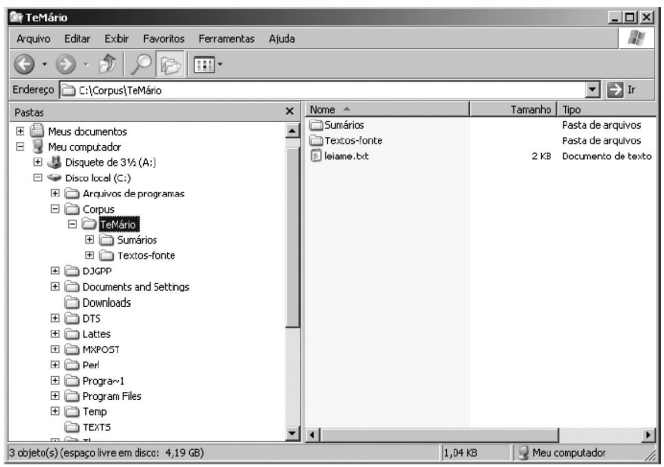


Figure I.1 TeMário's directory

In the folder "source-texts" there are three folders (**figure I.2**):

a) The first contains the original objects with their titles, organized by sources; this is to say, the texts group according to newspaper (FSP or JB, as shown in **figure I.3**) and section (special, world and opinion for texts from FSP; and international politics for JB); totaling 60 texts from FSP and 40 from JB.
b) The second contains all the "source-texts" with their titles with no origin discrimination.
c) The third contains the "source-texts" with no information regarding source or title.

The text files are unformatted and have a .txt extension as it enables automatic processing. Additional to their prefixes, all the file names include the year (NN), month (AA) and publication date (1-31).[38] The prefixes indicate the corresponding section of the newspapers as follows:

- "source-text" from the special section of **FSP are added the "ce" prefix**;
- "source-texts" from the global section of **FSP are added the "mu" prefix**;

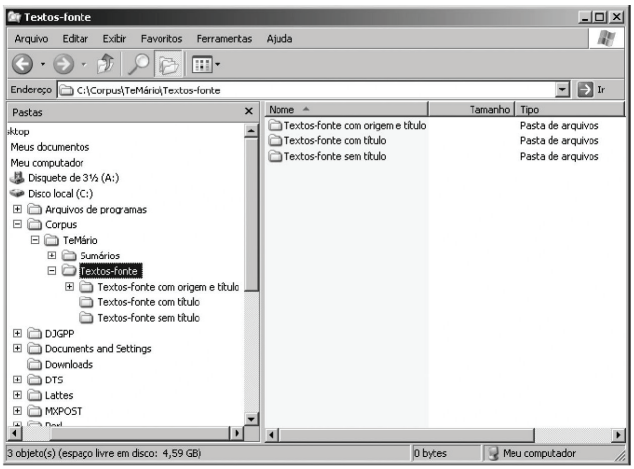


Figure I.2 Directory of "source-texts"

[38]NN for two numbers and AA for the initial letters of the corresponding month

- "source-texts" from the opinion **FSP section are added the "op" prefix**;
- "source-texts" from **JB International section are added the "in" prefix**;
- "source-texts" from **JB politics section are added the "po" prefix**.

This way, for example, the filename 'in96fe29-a.txt' indicates a text from JB international section, published on February 29th, 1996; the file 'mu94ag07- b.txt' indicates a text from FSP world, published on August 7th, 1994. Moreover, the "source-texts" are in untitled files identified **by "ST-"** before the aforementioned prefixes.

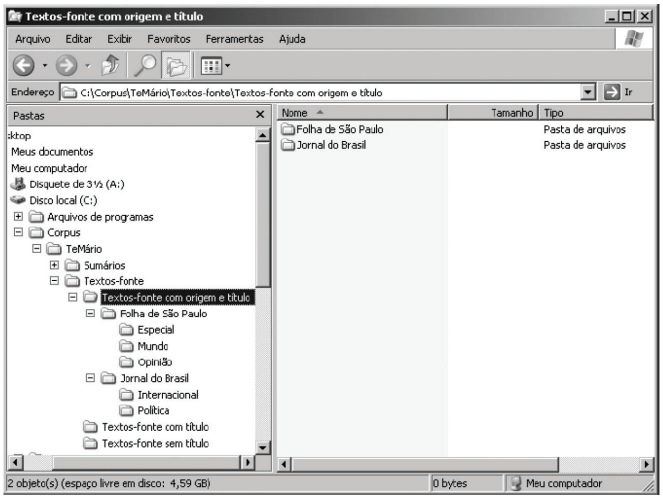


Figure I.3 Directory of the full "source-texts"

This subdivision of the "source-text" in various folders has the aim to help retrieve data by specific interests; for example, if the goal is to develop an assessment process, the "source-texts" with a title allow associating summaries or extracts automatically produced to the title in order to compare. In this case, it may be considered that the title is a legitimate representative of the source-text's main idea, which has been chosen by the writer with a view to verifying the automatic results to preserve. And for linguistic studies, as the verification of the particular characteristics of a genre and domain, an analyst may be appointed to retrieve specific books, those which already indicate a generic classification.

Moreover, in the very process of the summary, manual or automatic, it is convenient that "source-text" are visible without any title. Not in the case of a manual summary so that the writer is not influenced to write explicit information related to the title. In the case of automatic summaries, the reason is different; the summary of a text does not include the processing of its title.

There are also the folders in sumários (**figure I.4**): one with manual summaries (***sumários manuais***); other with marked manual summaries (***sumários manuais marcados***); finally, another with the ideal extracts produced by the generator of extracts of ideas, as previously described.

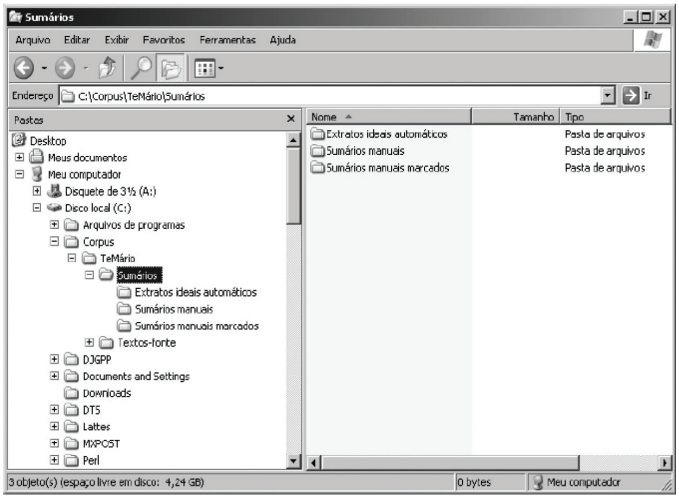


Figure I.4 Subfolders in "Resúmenes"

The unformatted files (.txt) with the summaries made by experts are in the folder of manual summaries (sumários manuais). Their names have the exact names of files of the corresponding "source-textslus the prefix "Suma" to indicate these are summaries instead of full texts. The folder of the marked manual summaries (manuais marcados) are the summaries, though with the phrases that indicate the professional summary and the main ideas of the corresponding "source-texts" in red (Task 2 according to the expert's request, see appendix of TeMário). These files are also called "source-texts", but with the "summ-" prefix (manually marked summaries). Their extension .doc is to preserve the

format, and so, the phrases remain unaltered. Consequently, it is recommended always preserving the data.

In order to difference the manual summaries ("Sum" prefix) from manually marked summaries ("Summ"), the files in the folder "automatic extractys of ideas" have the "Ext-" prefix (these are also the files with no format, i.e., .txt files).

I.3 Final considerations

In this appendix TeMário is described; it is a *corpus* with 100 news items plus their corresponding manual summaries and the extract of ideas. The manual summaries were made by an expert and writer in Portuguese, while the extracts of ideas are automatically produced. Owing to the specific nature of the data repository, it also contains the original texts and now the main ideas in the "source-texts" are marked, which guided the decision on the summary made by the expert. The generation of such information does not represent a significant load for the human expert and preserving the "source-text" titles consists of a single representation, from summary titles and standpoint manual summaries of the automatic summaries that are a good data depot, both for comparative studies on automatic assessment of results and the exploration of other techniques for automatic summarizing. For example, the titles themselves may serve as a basis for the election of the corresponding text segments to compose a summary: a title may be considered an essential phrase such as GistSumm (Pardo *et al*., 2003a), in this case.

The potential use of TeMário may increase if a language such as XML (eXtensible Markup Language) is considered, according to LACIOWEB project.[39]

In this case, the files with a .doc extension (that indicate the ideas that guide the expert's decisions) can also be converted for XML notation with no meaning loss, as in this language the style tags are preserved.

[39]http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm

Bibliographic references (Appendix I)

Mani, I. (2001). Automatic Summarization. John Benjamins Publishing Co., Amsterdam.

Martins, C.B. (2002). U*NLSumm: Um Sumarizador Automático de Textos UNL*. Dissertação de Mestrado. Departamento de Computação. Universidade Federal de São Carlos. São Carlos - SP.

Martins, C.B.; Pardo, T.A.S.; Espina, A.P.; Rino, L.H.M. (2001). *Introdução à Sumarização Automática*. Relatório Técnico RT-DC 002/2001, Departamento de Computação, Universidade Federal de São Carlos.

Módolo, M. (2003). SuPor: *um Ambiente para a Exploração de Métodos Extrativos para a Sumarização Automática de Textos em Português*. Dissertação de Mestrado. Departamento de Computação, UFSCar. São Carlos - SP.

Pardo, T.A.S. (2002). *DMSumm: Um Gerador Automático de Sumários*. Dissertação de Mestrado. Departamento de Computação. Universidade Federal de São Carlos. São Carlos - SP.

Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2003a). GistSumm: ASummarization Tool Based on a New Extractive Method. In N.J. Mamede, J. Baptista, I. Trancoso, M.G.V. Nunes (eds.), *6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken* - PROPOR, pp. 210-218 (Lecture Notes in Artificial Intelligence 2721). Springer-Verlag, Germany.

Pardo, T.A.S; Rino, L.H.M.; Nunes, M.G.V. (2003b). NeuralSumm: Uma Abordagem Conexionista para a Sumarização Automática de Textos. *Anais do IV Encontro Nacional de Inteligência Artificial*. Campinas-SP.

Rino, L.H.M. e Pardo, T.A.S. (2003). A Sumarização Automática de Textos: Principais Características e Metodologias. *Anais do XXIII Congresso da Sociedade Brasileira de Computação, Vol. VIII: III Jornada de Minicursos de Inteligência Artificial (III MCIA)*, pp. 203-245. Campinas-SP.

Salton, G. (1989) *Automatic Text Processing. The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley.

Appendix J

Documents in *Corpus* TEXTRUSS

Appendix J presents the documents that compose *corpus* TEXTRUSS, created for AGTS tasks in Russian. Its structure and composition are described.

J.1 Creation of *corpus* TEXTRUSS

The *corpus* comprises news items each with a summary made by a human expert on Russian. The items were downloaded from the news portal gazeta.ru. The *corpus* considers a number of domains and has 11 categories as follows:

ПОЛИТИКА (POLITICS)
БИЗНЕС (BUSINESS)
ОБЩЕСТВО (FIRMS)
МНЕНИЯ (CRITIQUES)
КУЛЬТУРА (CULTURE)
НАУКА (SCIENCE)
ТЕХНОЛОГИИ (TECHNOLOGY)

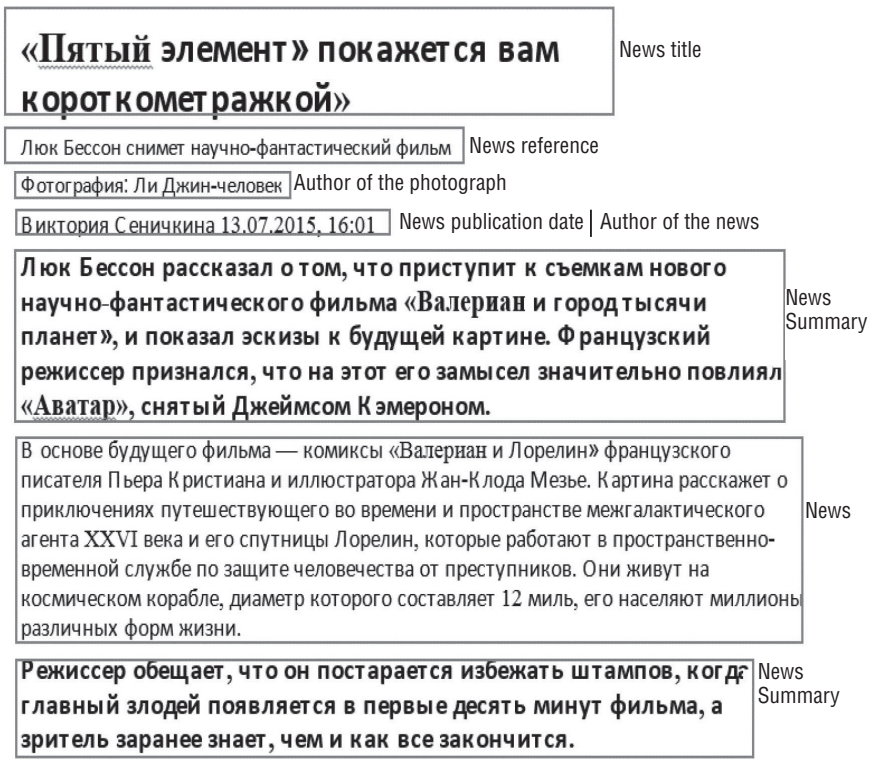


Figure J.1 Structure of article 10CU140815_7654545.TXT from *corpus TEXTRUSS*

НЕДВИЖИМОСТЬ (REAL STATE)
АВТО (AUTOMOBILES)
СТИЛЬ ЖИЗНИ (LIFESTYLE)
СПОРТ (SPORTS)

There are 22 items per category, totaling 242 news items.

The parts of each article's structure are the following (**figure J.1**):

To build TEXTRUSS, after downloading the items, each one was classified. The original texts are called source-texts, while their summaries are called summaries.

J.2 *CORPUS* ORGANIZATION

The *corpus* has 3 different formats (**figure J.2**):

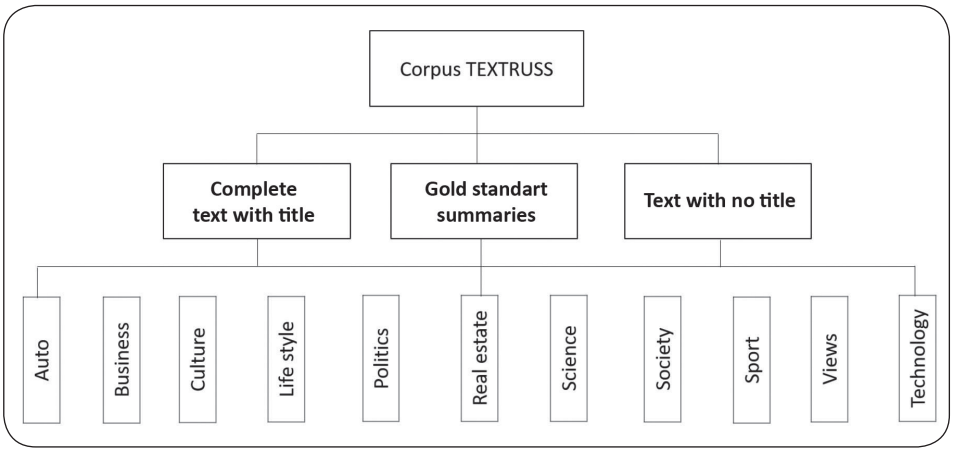


Figure J.2 Directory of *corpus* TEXTRUSS

Detection of main ideas and summary production
in English, Spanish, Portuguese and Russian.
60 years of research

by

Griselda Areli Matias Mendoza,
Yulia Ledeneva
and René Arnulfo García Hernández,

This book is a co-edition between the publisher Alfaomega Grupo
Editor and the Secretary of Research and Advanced Studies
of UAEM, through the Direction of Dissemination
and Promotion of Research and Advanced Studies.

Detection of main ideas and production of summaries in English, Spanish, Portuguese and Russian

60 years of research

Detection of main ideas and summary production in English, Spanish, Portuguese, and Russian: 60 years of research is a book that deals with the tasks of automatic generation of summaries from a qualitative and quantitative standpoint. Firstly, the results of Turing tests carried out on machines that produce summaries at present in the most spoken and written languages such as English, Spanish, Portuguese and Russian are presented with a view to finding out if a summary made by a machine has sufficient quality to confuse a human so that they fail to notice the summary was made by a machine. Later on, the integration and quantitative reports of the novel methods developed so far are presented, as well a comparison with the systems that generate automatic summaries is made.

The book is easy to read and accessible to anyone, in spite of using technical vocabulary in some sections, each and every one of the terms are comprehensively explained.

9 786076 333419

9 786075 388090

Computación
ÁREA

Nuevas tecnologías
SUBÁREA

www.alfaomega.com.mx

atencionalcliente@alfaomega.com.mx

Alfaomega Grupo Editor