Spanish Automatic Text Summarization: A Survey

Griselda Areli Matías-Mendoza, Yulia Ledeneva*, René Arnulfo García-Hernández

Autonomous University of the State of Mexico,
Mexico

{gris_9123, renearnulfo}@hotmail.com, yledeneva@yahoo.com

Abstract. The exponential growth of textual data has necessitated efficient summarization techniques. However, it is difficult for humans to summarize large text documents manually. As a result, automatic text summarization has emerged as a crucial and effective tool for helping to interpret and manage text information. Given the limited time available to read and fully comprehend each document before making decisions, there is a strong need for summarizing documents to convey a clear, representative idea of the original content. This has important practical applications in information retrieval, document classification, and knowledge extraction. Moreover, advanced summarization systems can effectively identify the core ideas of texts, significantly reducing the time users spend reading entire documents. While automatic text summarization has been extensively researched for languages, such as English, over the last 60 years, Spanish has received less attention. This paper addresses this gap by presenting key approaches, challenges, and methodologies in Spanish automatic text summarization. Through a comprehensive survey of relevant literature, we aim to provide a foundation for future research in this area. The presented survey is a compilation of important works in Spanish automatic text summarization and is intended to be a basis for research in the task. Also, we determine the main challenges for the task of Spanish automatic text summarization.

Keywords. Automatic text summarization, state-of-the-art methods for Spanish, abstractive text summarization, document summarization, news summarization, extractive summarization, natural language processing, corpus TER, corpus DUC, corpus TAC, evaluation of text summaries.

1 Introduction

Summaries are ubiquitous in our daily lives, from books and news articles to films, audio, scientific papers, and even social media platforms like Twitter. A summary can be defined as a condensed version of one or more texts that highlights key information while maintaining a length typically less than half of the original [1]. While traditionally applied to text, automatic summarization can also be used for other media, such as video and audio.

The explosion of information online has created a demand for tools that can quickly and efficiently extract key points from vast amounts of text. Automatic text summarization research has been ongoing since the 1950s, with Luhn's pioneering work in 1958 [2]. Over the decades, researchers have continually refined techniques to produce summaries that resemble those created by humans.

A summary can be generated through extractive, abstraction, and hybrid methods. Abstractive methods involve a complex process that requires significant computational resources and advanced linguistic techniques. Extractive methods create summaries by selecting and extracting the most important text elements, such as sentences, phrases, or paragraphs. The hybrid methods combine extractive and abstraction methods. The research community focuses more on extractive summaries, achieving more coherent and meaningful summaries [3].

The state-of-the-art methods for generating summaries also take into account the distribution of sentences and structure to identify and extract the most important ones [4-9]. These methods also use the text model to maintain the consistency of the summaries [10-13]. Another significant problem is the need for an equitable study of this task for different languages.

For example, before 2000, research on automatic text summarization primarily focused on

English because resources such as standard evaluation measures and corpora were available for this language. Despite this, other languages, like Spanish, have shown substantial growth. Spanish is now the world's second most spoken language and the third most used online, as noted in [14].

This creates an excellent prospect for advance study in Spanish automatic text summarization. This area has been the need for gold-standard summaries in Spanish. However, this is starting to improve, especially with the inclusion of the Spanish language in the corpora and tasks of the ACL 2013 MultiLing Workshop.

While other surveys and reviews cover general automatic text summarization, this one specifically examines Spanish-language summarization. It offers a comprehensive overview of existing research. Additionally, it covers the methods used in Spanish automatic text summarization, evaluates the outcomes, and presents relevant corpora, conferences, and workshops. The survey also addresses the most significant challenges in the area and completes with recommendations and suggestions for future research.

2 Natural Language Processing for Spanish Language

In 2023, over 599 million people spoke Spanish as their native language. Additionally, the number of potential Spanish users worldwide exceeds 585 million. Spanish ranks as the second most spoken native language globally, following Mandarin Chinese, and is also the second most spoken language overall when considering native speakers, those with limited proficiency, and Spanish learners.

Regarding institutional recognition, Spanish holds the third position as a working language within the United Nations and ranks fourth within the European Union. Spanish is the third most widely used language online, especially on platforms like Wikipedia, Facebook, and Twitter, where it holds second place in usage [14].

Spanish is said to come from the Romance languages, which do not derive from the Latin written in literature but from the Latin spoken in the streets and places [15]. While its roots trace back to the 3rd century A.C., its distinct development occurred centuries later.

Spanish is spoken in almost all the Iberian Peninsula, in the southwest of the U.S.A., throughout Mexico, and in Central and South America (except for Brazil and Guayana). In addition, it is the language of a minority group of speakers in the Philippines. This vast geographical spread brings, consequently, a significant range of dialectal variants.

However, despite being a language spoken in such distant areas, there is a certain uniformity in the cultured level of the language that allows people on either side of the Atlantic to understand each other relatively quickly. The most significant differences are suprasegmental, that is, the varied intonation, apparently the result of the different linguistic substrates in Spanish-speaking countries.

The Spanish language is composed of 26 letters of the Latin alphabet. Like Spanish, languages such as English (universal language), Portuguese, German, French, Swedish, and others use the Latin alphabet, so it is not difficult to become familiar with its symbology since it is not as complex as in languages such as Arabic or Russian. Currently, the universal language of world communication is English, so most of the research in the different areas of Natural Language Processing (PLN) has been carried out in this language, especially automatic text summarization.

One of the problems between languages is that specific characteristics depend on each language and simplify or make the relationship between groups of words more complete. However, English and Spanish use the same alphabet and have a basic order in the composition of their sentences: subject + verb + complement; this does not mean that this order is always fulfilled.

English has a stricter order, which must be conserved. However, the Spanish had more freedom, for example (see Table 1). The freedom of the Spanish language to create sentences complicates the automatic abstractive text summarization task. However, automatic extractive text summarization is a task very similar to that performed in English due to the use of the same alphabet and the coincidence between the composition of the sentences.

Table 1. Example of the composition of sentences in Spanish [16]

Example	Structure
Juan vino a mi casa	Subject + Verb + Complement
A mi casa vino Juan	Complement + Verb + Subject
Vino Juan a mi casa	Verb + Subject + Complement
A mi casa Juan vino	Complement + Subject + Verb
Juan a mi casa vino	Subject + Complement + Verb
Vino a mi casa Juan	Verb + Complement + Subject

3 History of Text Summarization: Corpus and Evaluation

Automatic text summarization has been the research subject for 60 years, beginning in the 1950s with Luhn's pioneering work in 1958 [17]. Luhn was the first to apply automatic extractive text summarization using text similarity. Later, in 1969, Edmundson introduced features such as word frequency, sentence position, title, and pragmatic words, which are still relevant and utilized today [18].

The advance of automatic text summarization in the following years was stopped, and only some investigations were carried out, such as those of Rush et al.'s work in 1971-1975 [19, 20] and Gerald Francis DeJong's studies in 1982 [21]. In 1993, research took off again with work by Spärck-Jones [22] and 1995 Julia Kupiec et al. [23]. This research helped to revive an interest in studying automatic text summarization.

Among the works that followed are [24-27]. Until 2000, most research in automatic text summarization focused exclusively on the English language. It was conducted without a standard corpus or evaluation measures, making comparison across studies difficult.

For example, the research in [17] used 50 journalistic articles, [18] utilized 200 articles, [28] analyzed 30 documents, [23] examined 188 scientific documents, and [29] worked with 30 documents. In 2001, the Document Understanding Conferences (DUC) were established to promote progress in summarization for English and provide a large-scale platform for researchers. DUC consisted of seven conferences: DUC01 through DUC07. Each conference included several tasks, with a corresponding gold standard corpus developed for each task.

Building on the foundation laid by the DUC conferences, the Text Analysis Conference (TAC) emerged in 2008 as a significant player in automatic text summarization. TAC's workshops were designed to elevate system evaluation, focusing on multi-document summaries for end-users. The TAC corpus, which concentrated on summaries produced between 2008 and 2014, is a testament to TAC's commitment to advancing the field. Table 3 provides an overview of the TAC corpora, further highlighting its role in the field.

In 2011, the MultiLing task was introduced to evaluate language-independent summarization algorithms across different languages. MultiLing corpora were produced in 2011, 2013, 2015, and 2017 for multilingual automatic text summarization. While MultiLing includes multiple languages, the original texts are primarily in English and translated into various languages, so there is no native corpus for each language [30-32]. Table 2 presents the standard datasets for text summarization.

Despite existing research on Spanish, a standardized or specialized corpus is essential for developing effective automatic text summarization systems. Many researchers have adapted corpora from information extraction tasks or created their own for Spanish automatic text summarization [58- 64].

This inconsistency hinders direct comparisons and makes it difficult to assess the progress in this field. To address this issue, recent efforts have focused on developing a standardized Spanish corpus. The CNN corpus was created in 2019, with the Spanish version based on news articles sourced from the CNN Mexico website. These articles address various general-interest topics and are written in standard language.

The corpus features summaries written by the original authors in English, emphasizing the key points of the CNN texts. It also includes the original text, story highlights, and additional metadata such as author names, titles, subject classifications, and publication dates, all retrieved from the Spanish

Table 2. Overview of existing corpora for summarization

Corpus	Lang.	Domain	Single-Doc.	Multi-Doc	Size
DUC 2001 [33]	English	News	Yes	Yes	60 x 10
DUC 2002 [34]	English	News	Yes	Yes	60 x 10
DUC 2003 [35]	English	News	Yes	Yes	60 x 10, 30 x 25
DUC 2004 [36]	English/Arabic	News	Yes	Yes	100x10
DUC 2005 [37]	English	News		Yes	50 x 32
DUC 2006 [38]	English	News		Yes	50 x 25
DUC 2007 [39]	English	News		Yes	25 x 10
TAC 2008 [40]	English	News		Yes	48 x 20
TAC 2009 [41]	English	News		Yes	44 x 20
TAC 2010 [42]	English	News		Yes	46 x 20
TAC 2011 [43]	English	News		Yes	44 x 20
ICSI [44]	English	Meetings	Yes		57
AMI [45]	English	Meetings	Yes		137
Opinosis [46]	English	Reviews	Yes	Yes	51 x 100
Gigaword [47]	English	News	Yes		4,111,240
Gigaword 5 [48]	English	News	Yes		9,876,086
LCSTS [49]	Chinese	blogs	Yes		2,400,591
CNN/Daily Mail [50]	English	News	Yes		312,084
MSR Abstractive [51]	English	misc	Yes		6,000
arXiv [52]	English	science	Yes		194,000
PubMed [52]	English	science	Yes		278,000
EASC [53]	Arabic	News/Wikipedia	Yes		153
SummBank [54]	Chinese/English	News	Yes	Yes	40 x 10
CAST [55]	English	News	Yes		147
CNN-corpus [56]	English	News	Yes		3,000
TeMário [57]	Portugues	News	Yes		100

version of the CNN website. The development of the Spanish CNN corpus followed the methodology proposed by Lins et al. in 2019.

In 2020, the TER standard corpus for Automatic Text Summarization in Spanish was created. TER is a corpus of Mexican Spanish-language news from the "Crónica" newspaper.

The construction of the corpus is divided into two stages: the first for the selection, cleaning, and tagging of news, and the second for the selection of experts, construction, and tagging of summaries [66].

In addition, a Corpus, composed of documents from various languages, has been generated, such as Multilingual Summarization Corpus (MLSUM).MLSUM is the first extensive dataset of its kind, featuring over 1.5 million article-summary pairs across five languages: Turkish, Spanish, Russian German, and French. Sourced from online newspapers, this valuable resource is a

Table 3. Overview of existing corpora for summarization in Spanish

Corpus	Lang.	Domain	Single-Document	Multi-Document	Size
ABC	Spanish	News	Yes		109
Medical articles	Spanish	Science	Yes		20
Desastres	Spanish	News		Yes	300
CNN-Corpus Spanish	Spanish	News	Yes		1117
TER	Spanish	News	Yes		240
MLSUM	Spanish	News	Yes		290,645
DACSA	Spanish	News	Yes		2,120,649
Bernoldi	Spanish	News	Yes		93,913

cornerstone for advancing multilingual summarization research.

For the Spanish language, the newspaper El País was used in that article [67]. Segarra et al.'s research describes the construction of a corpus of Catalan and Spanish newspapers, the Dataset for Automatic Summary of Catalan and Spanish period Articles (DACSA).

It is a large-scale, high-quality corpus that can be used to train summary models for Catalan and Spanish [68]. In [69], a corpus is built from the website of the Spanish newspaper "20 Minutos", which has a history of news that is freely accessible and downloadable. This corpus's main objective is to generate abstract summaries of news in Spanish automatically. Table 3 provides a brief description of the corpora for summary in Spanish.

Standard construction data (corpus) and various evaluation methods are necessary to assess automatically generated summaries. These evaluation methods are divided into intrinsic and extrinsic categories [70]. Intrinsic methods directly analyze the automatically produced summary, evaluating grammatical correctness, cohesion, and coherence to determine its quality.

These methods typically compare automatically generated summaries with expert-created ones to evaluate coverage. On the other hand, extrinsic evaluation methods assess the summary in the context of the task for which it was created, aiming to measure its impact on the performance of related tasks. These tasks may include, for example, relevance evaluation [71].

The most widely used evaluation method in automatic text summarization is ROUGE (Recall-Oriented Understudy for Gisting Evaluation), introduced by Lin and Hovy [72], [73]. ROUGE compares system-generated summaries with human-created (gold standard) summaries using n-gram statistics. ROUGE offers several automatic evaluation metrics for this purpose:

— ROUGE-N (n-grams co-occurrence).

This metric measures the recall or coverage of n-grams between a candidate summary and a set of reference summaries. It is calculated using the following formula (Formula 1):

$$ROUGE-N = \frac{\sum_{set\{PeerSummary\}}\sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{PeerSummary\}}\sum_{gram_n \in S} Count(gram_n)}, \quad (1)$$

where n is the length of the n-gram and $Count_{match}(gram_n)$ is the maximal number of n-grams that co-occur in the candidate summary and in the set of reference summaries.

ROUGE-N evaluates the quality of candidate summaries by quantifying the overlap of n-grams between the candidate and reference summaries. The score ranges from 0 to 1, where 0 signifies no overlap between the candidate and reference texts, while 1 indicates a complete overlap. ROUGE-N helps determine how well a system captures key content and linguistic details.

This metric, which evaluates the occurrence of noncontiguous bigrams, is a crucial component in automatic text summarization. Noncontiguous bigrams are any two words that appear in the same order within a sentence, regardless of the number of intervening words. The co-occurrence of noncontiguous bigrams provides a statistical

Table 4. Summary of survey

Name	Language
A Survey for Multi-Document Summarization [77]	English
A Survey on Automatic Text Summarization [78]	English
A Comprehensive Survey on Text Summarization Systems [79]	English
A Survey of Text Summarization Extractive Techniques [80]	English
Query-Based Summarization: A survey [81]	English
A Survey of Text Summarization Techniques [82]	English
A Survey of Unstructured Text Summarization Techniques [83]	English
A Survey on Automatic Text Summarization [84]	English
Automatic Arabic text summarization: a survey [85]	Arabic
Recent automatic text summarization techniques: a survey [86]	English
Automatic Arabic Summarization: A survey of methodologies and systems [87]	Arabic
Text Summarization Techniques: A Brief Survey [88]	English

measure of how well the candidate summary captures the noncontiguous bigrams from the reference summaries. Lin [72] demonstrated that this measure can effectively assess the quality of automatically generated summaries, achieving a 95% correlation with human judgments.

Since the introduction of standard corpora, automatic text summarization has gained importance, leading to over 400 studies focusing on the English language.

Few studies have focused on researching automatic text summarization for the Spanish language. In 2001, Acero et al. [58] presented the automatic generation of personalized summaries using their corpus, built from news articles from the ABC newspaper. Villatoro [61] used a similar corpus to extract and adapt information for automatic multi-document summarization in Spanish [74]. Other studies related to Spanish automatic summarization include [58-59], [61-62], [64], and [75-76].

However, despite these efforts, progress remains unclear because researchers have used either custom or adapted corpora, which prevents consistent comparisons between different methods. While a standard corpus exists, many state-of-the-art techniques have not yet been tested to evaluate their performance. In recent years, there has been growing interest in compiling research on automatic text summarization across various languages. Table 4 provides a list of different surveys conducted in this field. However, we still need an overview of the study of automatic text summarization for the Spanish language.

4 Spanish Automatic Text Summarization Approaches

Several generic automatic text summarization algorithms have been developed, each with advantages and disadvantages and different classifications depending on the technique or the input type. This section presents a survey of the literature on Spanish automatic text summarization. Due to the few Spanish automatic text summarization investigations, each state-of-the-art method that works with Spanish is described.

− Automatic Generation of Personalized Summaries [58]. This work is a practical application within Hermes, a personalized news dispatcher that handles information in English and Spanish. This system effectively utilizes three heuristics to select phrases to realize the summary.

1 Sentence position heuristic. It consists of giving a higher score to the first five sentences of a text.

2 Keyword heuristic. It consists of extracting the M most significant words from each

text and then checking how many of these keywords are found in each phrase. In this way, the highest number of phrases with the highest number of keywords is assigned.

3 Personalization heuristic. It consists of promoting phrases most relevant for a user model to personalize the summary.

The corpus consists of 109 news obtained in the electronic edition of the newspaper ABC.

– Towards a Linguistic Model of Automatic Summary of Medical Articles in Span-ish [60]. It focuses on the specialized Spanish automatic text summarization, specifically in medicine. The corpus he uses consists of 20 medical articles in Spanish that are part of the Technical Corpus of the Institut Universitari de Lingüística Aplicada (IULA) of the Fabra University of Barcelona. The method that is used consists of four stages.

1 Selection of work corpus. The selected corpus is divided into two subcorpus, reference and contrast.

2 Analysis of the texts of the reference subcorpus. The text structure of the medical article, its representative lexical units, and its discursive, syntactic, and communicative structure are analyzed.

3 Development of the model.

– Definition of the summary model.

– Development of linguistic rules.

– Manual validation of the operation of the rules.

– Implementation of the rules.

– Application of the rules on texts of the contrast subcorpus.

4 Evaluation of the model.

– Approach to the Automatic Summary as a tool to help legal translation in the field of tourism law [59]. This research is done for documents in Spanish in the tourism law field. However, it does not present any method for automatic text summarization since it only applies to the Copernic Summarizer tool to generate the summaries that later serve to translate.

– The Platform for Language Independent Summarization [64] introduces a summarization platform that operates independently of language. It supports tasks such as corpus acquisition, language classification, translation, and text summarization across 25 different languages. When the input text is in English, it is processed by an automatic extractive summarization module. This module selects the most important sentences from the original text using well-established sentence scoring methods, known for their high efficiency in extractive summarization. For texts in other languages, the platform employs language-independent summarization algorithms, and various translation tools are used to convert the sentences into English. Since automatic translation may cause some semantic loss, utilizing multiple translation tools can help mitigate these issues. The resulting translated versions are then fed into the extractive summarization module, where each version generates scores for the sentences in relation to the original text. The Sentence Scoring and Selection Module evaluates the chosen sentence sets and produces a final summary by selecting the corresponding sentences from the original text.

The corpus used in this platform is CNN-Spanish, with the current version containing 400 texts classified into eight categories: sports, entertainment, world, national, opinion, technology, travel, and health news.

– Automatic Summarization of Multiple Documents [61]. Villatoro's work utilizes a classifier and supervised learning tools. The core concept is that an inductive process automatically builds a classifier by analyzing the characteristics of a set of previously summarized documents. The learning algorithm receives pairs of (documents and summaries), turning the task of generating summaries into a supervised learning process. The Disaster dataset was used for experimentation with Spanish-language corpora [89]. Although the corpus was originally designed for classification, it was adapted for automatic text summarization. The

Disaster dataset consists of 300 news articles collected from Mexican newspapers. Each sentence was labeled with two tags: Relevant and Non-Relevant. To minimize subjectivity in the labeling process, experts were instructed to label a sentence as "Relevant" only if it contained at least one factual detail about the event, such as the date, location, the number of affected people or homes, economic damages, or the scale or magnitude of the disaster.

– Automatic Generation of Summaries [90]. A method based on supervised learning techniques is proposed, specifically in classification. The corpus he uses is composed of more than 8000 documents containing nine years of rectoral resolutions of the Catholic University of Salta. The method uses a labeling process to determine whether sentences are relevant. In addition, each sentence must have a label that indicates whether it belongs to the summary. They used the We-ka software tool for the experiments, which included a vast collection of classification techniques. Among the classifiers this method uses are ADTree, ID3, C4.5 with pruning, C4.5 without pruning, Decision Table, Ripper, and Naïve Bayes. The construction of decision trees obtains summaries of adequate quality, which serve as indicative summaries for the user of a semantic search engine in the proposed corpus in this research.

– A New Cross-Lingua Automatic Summarization Approach Based on Textual Energy [91]. This method introduces a cross-language summarization system that incorporates textual energy and translation time measurement, improving the reliability of the final news summaries. The automatic summarization technique, which uses textual energy, is inspired by statistical physics and combines a Vector Space Model (VSM) with neural networks. The ENERTEX method [92] treats words in the text as units that interact and are influenced by the field generated by each unit. As a result, each word is assigned a score based on its textual energy. Additionally, this approach factors in the translation time of each sentence. A textual energy matrix is generated, aiding in the summary creation process. The system's performance was evaluated using the FRESA framework, which compared the automatically generated summaries with baseline summaries for varying percentages of the original texts.

– PuertoTex: A Data Mining Software Based on Ontologies for Automatic Summarization in the Port and Coastal Engineering Domain [93]. This research focuses on developing and evaluating an ontology-based software designed to automatically generate summaries in the field of Ports and Coastal Engineering. The tool's development incorporates techniques from discourse analysis and cognitive methods to create rules for processing texts. It also involves constructing an ontology to support labeling processes, utilizing the capabilities of the Resource Description Framework and Extensible Markup Language. A set of agents was created to act on the ontology, defining its essential elements. The resulting product is the PuertoTex software, which generates ontology-based automatic summaries. This method was tested in both English and Spanish. Three evaluation approaches were employed: usability evaluation, information retrieval evaluation, and an assessment of the automatically generated summary.

– Automatic Sentence Compression: a Study towards the Generation of Summaries in Spanish [76]. This research explores sentence compression techniques for Spanish summarization. A linear model that predicts the removal of intra-sentence segments based on a set of text-based features were proposed. The model was trained on a large dataset of over 60,000 sentences, considering the entire context and the generated summary. Through statistical analysis, the most significant features for predicting segment deletion with 75% accuracy were identified. Then, two algorithms are proposed for generating summaries with compressed sentences after summaries are evaluated with a test similar to the Turing Test.

– Automatic Generation of Summaries with Support in Ontologies Applied to the Biomedical Domain [94]. This research

proposes an architecture for generating informative summaries of a single document in a specific domain: biomedicine. A method of extracting sentences is presented, based on the theory of complex networks, which maps the text to the concepts of the UMLS ontology and represents the document and the sentences as graphs. The selection of sentences is based on the degree of connection of their concepts in the graph of the document, using a grouping algorithm based on connectivity. A system that implements the proposed method is developed, and the empirical results of applying different heuristics to select the summary sentences are shown.

– Evaluation of Summaries in Spanish with Latent Semantic Analysis: A Possible Implementation [63]. This research seeks to identify an effective method for evaluating summaries using Latent Semantic Analysis (LSA). Secondary school students from Valparaíso, Chile, wrote the summaries. To achieve this goal, the scores assigned by three teachers to 244 summaries of primarily expository texts and 129 summaries of mostly narrative texts were compared with the scores produced by three computational methods based on LSA. The methods include:

1 Comparison of summaries with the source text.

2 Comparison of summaries with a summary developed by the consensus of a group of linguists.

3 Comparison of summaries with three summaries constructed by three language teachers.

– Text Summarization of Spanish Documents [95]. This research aimed to develop an extraction-based automatic text summarization algorithm. The proposed method involves constructing a directed weighted graph from the original text. A ranking algorithm is then applied to identify the most important sentences based on the weighted graph, ensuring that these critical sentences are included in the summary. The project's primary objective was to summarize 642 news articles computationally while ensuring no essential information was omitted from the summaries.

– Ground Truth Spanish Automatic Extractive Text Summarization Bounds [66]. This research introduces the TER standard corpus, designed to evaluate state-of-the-art methods and systems for automatic summarization in the Spanish language. The essential contribution lies in proposing the configuration and evaluation of five state-of-the-art methods, five systems, and four heuristics using three evaluation metrics: ROUGE, ROUGE-C, and Jensen-Shannon divergence. Notably, this study marks the first use of Jensen-Shannon divergence to assess automatic summarization in Spanish. In Matias (2020), ground truth bounds for Spanish were presented, including the heuristic baselines of first, random, topline, and concordance. Additionally, a ranking of 30 evaluation tests for state-of-the-art methods and systems was established, creating a benchmark for automatic summarization in Spanish.

– Evaluating Extractive Automatic Text Summarization Techniques in Spanish [96]. This study assesses both traditional and innovative extractive text summarization techniques in Spanish. The Corpus-TER [66], a dataset compiled from Mexican-Spanish news websites, was used for this evaluation. The primary objectives of the research are:

1. Select and develop specific summarization methods,

2. Choose a suitable corpus for testing these methods,

3. Design a concise and reusable interface and

4. Evaluate the summarization techniques.

The evaluation process utilizes the ROUGE and BLUE tools to assess performance.

– Generación Automática de Resúmenes Abstractivos de Noticias en Español [69]. In this work, we propose and evaluate a BERT-based processing pipeline for generating abstractive summaries of Spanish news. Specifically, it uses the BERTSUM framework on BETO [98], a model pre-trained exclusively in Spanish. On this basis, the model parameters are adjusted with a corpus of Spanish news. The work evaluates its results using the ROUGE metric and compares them

with some results obtained in English with the CNN/Daily Mail corpus.

- esT5s: A Spanish Model for Text Summarization [99]. The paper is about building a deep learning model for the task of Spanish text summarization based on the T5 (Text-to-Text Transfer Transformer) architecture. Such models have made significant progress in natural language processing, especially in English, but Spanish and other languages require specific models, the training of which is often computationally expensive. The work described in the paper addresses building a Spanish text summarization model from a large multilingual model, in this case, the mT5 model, which includes 101 languages. The authors managed to create a specialized model for Spanish called esT5, which is more efficient in terms of training time and computational power required. This model can be trained in less than an hour using a single GPU and produces summaries of comparable quality to larger models, significantly faster at inference.

- XL-Sum: Introduces a large-scale multilingual dataset designed for automatic abstractive summarization. This dataset includes over one million article-abstract pairs in 44 languages, including Spanish. The dataset was collected from BBC news articles using an automated process that extracts professional summaries written by human authors. It is highlighted that the dataset includes summaries in Spanish, which is significant due to the scarcity of high-quality public datasets in this language for abstractive summarization tasks.

5 Discussion

In the previous sections, several research studies on automatic text summarization were addressed, first general and later focused on the Spanish language. The main objective was to present a general overview of the task to understand the Span-ish automatic text summarization problem. While there are more than 400 studies for the English language and various studies on automatic text summarization, less than 24 research are available for the Spanish language.

The investigations in Spanish for automatic text summarization cannot be compared because each works with different corpora and various objectives. Even though the Spanish automatic text summarization research is approximately 20 years old, there has yet to be much progress; this is likely because Spanish did not hold significant global importance or was not extensively utilized. However, due to the growth of native and foreign speakers, and above all, on the Internet, automatic text summarization in Spanish has become essential.

In recent years, state-of-the-art methods began to present language independence [61], [100-104]; however, they have been tested in other languages, such as English, Arabic, and Portuguese, but not in Spanish. This is mainly due to the need for a standard corpus.

The nature of the Spanish language is very similar to that of English. English is the most studied language in automatic text summarization, so state-of-the-art methods of automatically generating summaries, mainly extractive and multilanguage, are created and tested in English. However, applying these methods to the Spanish language would be possible due to the language's nature.

There is no investigation into automatic abstractive text summarization for the Spanish language. Moreover, most of the investigations carried out are for extractive summaries of a single document; only one of those presented is for multiple documents. Therefore, this represents a great research opportunity for Spanish automatic text summarization.

The evaluation methods proposed for the English language [72] can be used since most of them are based on the correlation between the words of the automatically generated summary and the gold standard (made by the human).

6 Conclusion

This paper provides a comprehensive overview of the existing literature on Spanish automatic text summarization. We explore a range of methods used for both summary generation and evaluation, highlighting the relatively recent and understudied nature of this research area.

To advance Spanish automatic text summarization, future studies should consider adapting state-of-the-art methods from English and exploring related research in the field of natural language processing. A significant challenge in Spanish summarization is the lack of high-quality gold-standard summaries. Addressing this issue through the creation of a standardized corpus would enable researchers to test existing extractive summarization methods and fine-tune their parameters for Spanish.

Subsequently, the parameters of the methods for the Spanish language can be adjusted. There is a large field of research in generating automatic abstractive text summarization.

Finally, the development of automatic abstractive summarization systems for Spanish remains a promising area for future research.

References

1. **Radev, D. R., Hovy, E., McKeown, K. (2002).** Introduction to the special issue on summarization. Computational linguistics, Vol. 28, No. 4, pp. 399–408.

2. **Luhn, H. P. (1958).** The automatic creation of literature abstracts. IBM Journal of research and development, Vol. 2, No. 2, pp. 159–165. DOI: 10.1147/rd.22.0159.

3. **El-Kassas, W. S., Salama, C. R., Rafea, A. A., Mohamed, H. K. (2021).** Automatic text summarization: A comprehensive survey. Expert Systems with Applications, Vol. 165, p. 113679. DOI: 10.1016/j.eswa.2020.113679.

4. **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, Vol. 41, No. 9, pp. 4158–4169. DOI: 10.1016/j.eswa.2013.12.042.

5. **Nandhini, K., Balasundaram, S. R. (2014).** Extracting easy to understand summary using differential evolution algorithm. Swarm and Evolutionary Computation, Vol. 16, pp. 19–27. DOI: 10.1016/j.swevo.2013.12.004.

6. **Qazvinian, V., Hassanabadi, L. S., Halavati, R. (2008).** Summarizing text with a genetic algorithm-based sentence extraction. International Journal of Knowledge Management Studies, Vol. 2, No. 4, pp. 426–444. DOI: 0.1504/IJKMS.2008.01975.

7. **Mateo, P. L., González, J. C., Villena, J., Martínez, J. L. (2003).** Un sistema para resumen automático de textos en castellano. Procesamiento del lenguaje natural, Vol. 31, pp. 29–36.

8. **Babar, S., Patil, P. D. (2015).** Improving performance of text summarization. Procedia Computer Science, Vol. 46, pp. 354–363. DOI: 10.1016/j.procs.2015.02.031.

9. **Kiyoumarsi, F. (2015).** Evaluation of automatic text summarizations based on human summaries. Procedia-Social and Behavioral Sciences, Vol. 192, pp. 83–91. DOI: 10.1016/j.sbspro.2015.06.013.

10. **Sidorov, G. (2019).** Syntactic n-grams in computational linguistics. Cham, Switzerland: Springer International Publishing. DOI: 10.1007/978-3-030-14771-6.

11. **Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., Chanona-Hernández, L. (2014).** Syntactic n-grams as machine learning features for natural language processing. Expert Systems with Applications, Vol. 41, No. 3, pp. 853–860. DOI: 10.1016/j.eswa.2013.08.015.

12. **Ledeneva, Y., García-Hernández, R. A. (2017).** Automatic generation of text summaries. Challenges, Proposals and Experiments, Universidad Autónoma del Estado de México, Toluca.

13. **Matias, G., Ledeneva, Y., García Hernández, R. A. (2020).** Detección de ideas principales y composición de resúmenes en inglés, español, portugués y ruso. 60 años de investigación (Alfaomega - UAEMex.). Alfaomega Grupo Editor, SA de CV.

14. **Vítores, D. F., Cervantes, I. (2023).** El español: una lengua viva. Informe 2023. El español en el mundo 2023: Anuario del Instituto Cervantes, pp. 23–142.

15. **Huidobro, J. M. (2016).** Origen y evolución del castellano. Acta, pp. 85–91.

16. **Haro, S. N. G., Gelbukh, A. (2007).** Investigaciones en análisis sintáctico para el

español. Instituto Politécnico Nacional, Dirección de Publicaciones.

18. **Edmundson, H. P. (1969).** New methods in automatic extracting. Journal of the ACM (JACM), Vol. 16, No. 2, pp. 264–285. DOI: 10.1145/321510.32151.

19. **Rush, J. E., Salvador, R., Zamora, A. (1971).** Automatic abstracting and indexing. II. Production of indicative abstracts by application of contextual inference and syntactic coherence criteria. Journal of the American Society for Information Science, Vol. 22, No. 4, pp. 260–274. DOI: 10.1002/asi.4630220405.

20. **Pollock, J. J., Zamora, A. (1975).** Automatic abstracting research at chemical abstracts service. Journal of Chemical Information and Computer Sciences, Vol. 15, No. 4, pp. 226– 232.

21. **DeJong, G. (1982).** An overview of the FRUMP system. Strategies for natural language processing, Vol. 113, pp. 149–176.

22. **Jones, K. S. (1993).** What might be in a summary? Information retrieval, Vol. 93, pp. 9– 26.

23. **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 pp. 68–73.

24. **Endres-Niggemeyer, B. (1998).** Summarizing Information (Springer).

25. **Mani, I., Maybury, M. T. (1999).** Advances in automatic text summarization MIT Press. Vol. 293.

26. **Moens, M. F. (2000).** Automatic indexing and abstracting of document texts. Boston: Kluwer Academic Publishers.

27. **Mani, I. (2001).** Automatic summarization. John Benjamins. Retrieved from.

28. **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, pp. 77–85.

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

30. **Giannakopoulos, G., El-Haj, M., Favre, B., Litvak, M., Steinberger, J., Varma, V. (2011).** TAC 2011 MultiLing pilot overview.

31. **Elhadad, M., Miranda-Jiménez, S., Steinberger, J., Giannakopoulos, G. (2013).** Multi-document multilingual summarization corpus preparation, part 2: Czech, hebrew and spanish. Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization, pp. 13–19.

32. **Giannakopoulos, G., Kubina, J., Conroy, J., Steinberger, J., Favre, B., Kabadjov, M., Poesio, M. (2015).** Multiling 2015: multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 270–274.

33. **Over, P., Yen, J. (2001).** Introduction to DUC-2001: An intrinsic evaluation of generic news text summarization systems. Proceedings of DUC 2001 Document Understanding Conference, Vol. 49.

34. **Over, P., Liggett, W. (2002).** Introduction to DUC-2002: An intrinsic evaluation of generic news text summarization system. ACL 2002, Workshop on Text Summarization.

35. **Over, P., Yen, J. (2003).** Introduction to DUC-2003: An intrinsic evaluation of generic news text summarization systems.

36. **Over, P., Yen, J. (2004).** Introduction to DUC-2004: An intrinsic evaluation of generic news text summarization systems.

37. **Dang, H. T. (2005).** Overview of DUC 2005. Proceedings of the document understanding conference, Vol. 2005, pp. 1–12.

40. **Dang, H. T., Owczarzak, K. (2008).** Overview of the tac 2008 opinion question answering and summarization tasks. Proceedings of the First Text Analysis Conference, Vol. 2.

41. **Dang, H. T., Owczarzak, K. (2009).** Overview of the TAC 2009 summarization track. Proceedings of the Text Analysis Conference. pp. 1–25.

42. **Owczarzak, K., Dang, H. T. (2010).** Overview of the tac 2010 summarization track. Proceedings of the Third Text Analysis Conference, Gaithersburg, Maryland, USA, National Institute of Standards and Technology.

43. **Owczarzak, K., Dang, H. T. (2011).** Overview of the tac 2011 summarization track: Guide task and aesop task. Proceedings of the Second Text Analysis Conference (TAC2011). Gaithersburg, Maryland, USA.

44. **Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Stolcke, A. (2003).** The ICSI meeting corpus. 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, p. I–I.

45. **McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Karaiskos, V. (2005).** The AMI meeting corpus. Proceedings of the 5th international conference on methods and techniques in behavioral research, Vol. 88, p. 100.

46. **Ganesan, K., Zhai, C., Han, J. (2010).** Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions. Proceedings of the 23rd international conference on computational linguistics, pp. 340–348.

47. **David, G., Cieri, C. (2003).** English gigaword. Linguistic Data Consortium.

48. **Parker, R., Graff, D., Kong, J., Chen, K., Maeda, K. (2011).** English gigaword fifth edition. Technical Report, Linguistic Data Consortium, Philadelphia.

49. **Hu, B., Chen, Q., Zhu, F. (2016).** LCSTS: A large scale chinese short text summarization dataset. Computation and Languaje, arXiv preprint arXiv:1506.05865.

50. **Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, P. (2015).** Teaching machines to read and comprehend. Advances in neural information processing systems, Vol. 28, pp. 1693–1701.

51. **Toutanova, K., Brockett, C., Tran, K. M., Amershi, S. (2016).** A dataset and evaluation metrics for abstractive compression of sentences and short paragraphs.

52. **Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., Goharian, N. (2018).** A discourse-aware attention model for abstractive summarization of long documents. arXiv preprint arXiv:1804.05685.

53. **El-Haj, M., Kruschwitz, U., Fox, C. (2010).** Using mechanical turk to create a corpus of arabic summaries.

54. **Radev, D. (2003).** Summbank 1.0. web download. Linguistic Data Consortium, Philadelphia.

55. **Hasler, L., Orasan, C., Mitkov, R. (2003).** Building better corpora for summarization. Proceedings of corpus linguistics. pp. 309– 319.

56. **Lins, R. D., Oliveira, H., Cabral, L., Batista, J., Tenorio, B., Ferreira, R., Simske, S. J. (2019).** The CNN-corpus: A large textual corpus for single-document extractive summarization. Proceedings of the ACM Symposium on Document Engineering. pp. 1- 10.

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

58. **Acero, I., Alcojor, M., Díaz-Esteban, A., Gómez-Hidalgo, J. M., Maña-López, M. J. (2001).** Generación automática de resúmenes personalizados. Procesamiento del lenguaje natural, No 27, pp. 281–290.

59. **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 dialogo, Actas del IV Congreso El español, lengua de traducción, Madrid.

60. **Da-Cunha, I. (2008).** Hacia un modelo lingüístico de resumen automático de artículos médicos en español. Proyecto de investigación, Universidad Pompeu Fabra, Instituto Universitario de Lingüística Aplicada, Doctorado en Ciencias del Lenguaje y Lingüística Aplicada, http://www.tesisenxarxa.net.

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

62. **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. Universidad Complutense de Madrid, Madrid.

63. **Venegas, R. (2011).** Evaluación de resúmenes en español con análisis semántico latente: Una implementación posible. Revista signos, Vol. 44, No. 75, pp. 85–102.

64. **Cabral, L. 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, pp. 203–206.

65. **Lins, R. D., Oliveira, H., Cabral, L., Batista, J., Tenorio, B., Salcedo, D. A., Simske, S. J. (2019).** The CNN-corpus in spanish: a large corpus for extractive text summarization in the spanish language. Proceedings of the ACM Sym-posium on Document Engineering, pp. 1– 4.

66. **Matias, G. A., Ledeneva, Y., García-Hernández, R. A., Alexandrov, M., Hernández-Castañeda, Á. (2020).** Ground Truth Spanish Automatic Extractive Text Summarization Bounds. Computación y Sistemas, Vol. 24, No. 3, pp. 1241–1256. DOI: 10.13053/CyS-24-3-3484.

67. **Scialom, T., Dray, P. A., Lamprier, S., Piwowarski, B., Staiano, J. (2020).** MLSUM: The multilingual summarization corpus. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing EMNLP, pp. 8051–8067.

68. **Segarra-Soriano, E., Ahuir, V., Hurtado, L. F., González, J. (2022).** DACSA: A large-scale dataset for automatic summarization of catalan and spanish newspaper articles. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the As-sociation for Computational Linguistics: Human Language Technologies, pp. 5931– 5943.

69. **Bernoldi, R., Tolosa, G. H. (2022).** Generación automática de resúmenes abstractivos de noticias en español. Presented at the Simposio Argentino de Ciencia de Datos y Grades Datos.

70. **Sparck-Jones, K., Galliers, J. R. (1995).** Evaluating natural language processing systems: An analysis and review Springer Science & Business Media, Vol. 1083.

71. **Berker, M., Güngör, T. (2012).** Using genetic algorithms with lexical chains for automatic text summarization. ICAART, Vol. 1, pp. 595–600.

72. **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, Vol. 1, pp. 71–78.

73. **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. p. 605.

74. **Téllez, A., Montes, M., Villaseñor-Pineda, L. (2009).** Using machine learning for extracting information from natural disaster news reports. Computación y Sistemas, Vol. 13, No. 1, pp. 33–44.

75. **Da-Cunha, I., Torres-Moreno, J. M., Velázquez-Morales, P., Vivaldi, J. (2009).** Un algoritmo lingüístico-estadístico para resumen automático de textos especializados. Linguamática, Vol. 1, No. 2, pp. 67–79.

76. **Molina, A. (2013).** Compresión automática de frases: un estudio hacia la generación de resúmenes en español. Inteligencia Artificial, Vol. 16, No. 51, pp. 41–62.

77. **Sekine, S., Nobata, C. (2003).** A survey for multi-document summarization. Proceedings of the HLT-NAACL 03 on Text summarization workshop, Vol. 5, pp. 65–72.

78. **Das, D., Martins, A. F. (2007).** A survey on automatic text summarization. Literature

Survey for the Language and Statistics II course at CMU, Vol. 4, pp. 192–195.

79. **Gholamrezazadeh, S., Salehi, M. A., Gholamzadeh, B. (2009).** A comprehensive survey on text summarization systems. Computer Science and its Applications, 2009. CSA'09. 2nd International Conference on pp. 1–6.

80. **Gupta, V., Lehal, G. S. (2010).** A survey of text summarization extractive techniques. Journal of emerging technologies in web intelligence, Vol. 2, No. 3, pp. 258–268.

81. **Damova, M., Koychev, I. (2010).** Query-based summarization: A survey.

82. **Nenkova, A., McKeown, K. (2012).** A survey of text summarization techniques. Mining text data, Springer, pp. 43–76.

83. **Elfayoumy, S., Thoppil, J. (2014).** A survey of unstructured text summarization techniques. International Journal of Advanced Computer Science and Applications, Vol. 5, No. 4.

84. **Saranyamol, C. S., Sindhu, L. (2014).** A survey on automatic text summarization. International Journal of Computer Science and Information Technologies, Vol. 5, No. 6, pp. 7889–7893.

85. **Al-Saleh, A. B., Menai, M. E. B. (2016).** Automatic arabic text summarization: a survey. Artificial Intelligence Review, Vol. 45, No. 2, pp. 203–234.

86. **Gambhir, M., Gupta, V. (2017).** Recent automatic text summarization techniques: a survey. Artificial Intelligence Review, Vol. 47, No. 1, pp. 1–66.

87. **Al-Qassem, L. M., Wang, D., Al-Mahmoud, Z., Barada, H., Al-Rubaie, A., Almoosa, N. I. (2017).** Automatic arabic summarization: a survey of methodologies and systems. Procedia Computer Science, Vol. 117, pp. 10–18.

88. **Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., Kochut, K. (2017).** Text summarization techniques: a brief survey. arXiv pre-print arXiv:1707.02268.

89. **Téllez-Valero, A., Montes-y-Gómez, M., Villaseñor-Pineda, L. (2009).** Using machine learning for extracting information from natural disaster news reports. Computación y sistemas, Vol. 13, No. 1, pp. 33–44.

90. **Cardoso, A. C., Abelleira, M. A. P. (2013).** Generación automática de resúmenes. 1er Congreso Nacional de Ingeniería Informática / Sistemas de Información, CoNaIISI.

91. **Careaga-Moya, J. A., Medina-Urrea, A., Torres-Moreno, J. M. (2012).** A new cross-lingua automatic summarization approach based on textual energy. Journées internationales d'Analyse statistique des Données Textuelles, pp. 247–255.

92. **Fernández, S., SanJuan, E., Torres-Moreno, J. M. (2007).** Textual energy of associative memories: Performant applications of Enertex algorithm in text summarization and topic segmentation. Mexican International Conference on Artificial Intelligence, Springer, pp. 861–871.

93. **Leiva-Mederos, A., Domínguez-Velasco, S., Senso, J. A. (2012).** PuertoTex: a data mining software based on ontologies for automatic summarization on port and coastal engineering domain. Transinformação, Vol. 24, No. 2, pp. 103–115.

94. **Plaza-Morales, L. (2008).** Generación automática de resúmenes con apoyo en ontologías aplicada al dominio biomédico.

95. **Umadevi, K. S., Chopra, R., Singh, N., Aruru, L., Kannan, R. J. (2018).** Text summarization of Spanish documents. 2018 International Conference on Advances in Computing, Communications and Informatics pp. 1793–1797.

96. **Caparrós-Laiz, C., García-Díaz, J. A., Valencia-García, R. (2021).** Evaluating extractive automatic text summarization techniques in spanish. Technologies and Innovation, pp. 79–92.

97. **Cañete, J., Chaperon, G., Fuentes, R., Ho, J. H., Kang, H., Pérez, J. (2023).** Spanish pre-trained BERT model and evaluation data. DOI: 10.48550/arXiv.2308.02976.

98. **Vogel-Fernandez, A., Calleja, P., Rico, M. (2022).** esT5s: A spanish model for text summarization. Towards a Knowledge-Aware AI (pp. 184–190).

99. Hasan, T., Bhattacharjee, A., Islam, M. S., Samin, K., Li, Y. F., Kang, Y. B., Shahriyar, R. (2021). XL-Sum: Large-scale multilingual abstractive summarization for 44 languages.

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

101. 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.

102. Last, M., Litvak, M. (2010). Language-independent techniques for auto-mated text summarization. pp. 207–237.

103. Saggion, H. (2011). Using SUMMA for language independent summarization at TAC 2011. Presented at the TAC.

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