ISSN 2007-9737

Ground Truth Spanish Automatic Extractive Text Summarization Bounds

Griselda Areli Matias Mendoza[1], Yulia Ledeneva[1,2], René Arnulfo García Hernández[1], Mikhail Alexandrov[2,3], Ángel Hernández Castañeda[1]

[1] Universidad Autónoma del Estado de México,
Unidad Académica Profesional Tianguistenco,
Mexico

[2] Presidential Academy of National Economy and Public Administration,
Department of system analysis and informatics,
Russia

[3] Autonomous University of Barcelona,
Spain

{gris_9123, renearnulfo}@hotmail.com, yledeneva@yahoo.com,
malexandrov@mail.ru, angelhc2305@gmail.com

Abstract. The textual information has accelerated growth in the most spoken languages by native Internet users, such as Chinese, Spanish, English, Arabic, Hindi, Portuguese, Bengali, Russian, among others. It is necessary to innovate the methods of Automatic Text Summarization (ATS) that can extract essential information without reading the entire text. The most competent methods are Extractive ATS (EATS) that extract essential parts of the document (sentences, phrases, or paragraphs) to compose a summary. During the last 60 years of research of EATS, the creation of standard corpus with human-generated summaries and evaluation methods which are highly correlated with human judgments help to increase the number of new state-of-the-art methods. However, these methods are mainly supported for the English language, leaving aside other equally important languages such as Spanish, which is the second most spoken language by natives and the third most used on the Internet. A standard corpus for Spanish EATS (SAETS) is created to evaluate the state-of-the-art methods and systems for the Spanish language. The main contribution consists of a proposal for configuration and evaluation of 5 state-of-the-art methods, five systems and four heuristics using three evaluation methods (ROUGE, ROUGE-C, and Jensen-Shannon divergence). It is the first time that Jensen-Shannon divergence is used to evaluate AETS. In this paper the ground truth bounds for the Spanish language are presented, which are the heuristics *baseline:first*, *baseline:random*, *topline* and *concordance*. In addition, the ranking of 30 evaluation tests of the state-of-the-art methods and systems is calculated that forms a benchmark for SAETS.

Keywords. Spanish automatic text summarization, ROUGE, ROUGE-C, Jensen Shannon divergence, corpus TER.

1 Introduction

The information has become a necessary resource whose growth is increasing in different languages spoken in the world. Among the most spoken languages according to their number of native speakers are Chinese, Spanish, English, Arabic, Hindi, Portuguese, Bengali, and Russian, among others [1]. To have access to the information that is generated day by day, it is suggested to use the methods of Automatic Text Summarization (ATS). ATS aims to extract the most relevant information of a document [2].

The most of the state-of-the-art methods have been based on Automatic Extractive Text Summarization (AETS) because of its easy implementation and competent results. The methods of AETS extract essential parts of a text

(sentences, key phrases, or paragraphs) considered important for the original version; therefore, do not require complex, sophisticated methods.

AETS has 60 years of research; its study started in the ’50s with Luhn’s work in 1958 [3]. Luhn was the first to perform AETS.

Subsequently, the investigation of the AETS has continued with the research of [4-14] and others. Research of AETS up to the year 2000 focused on the English language because the resources (corpus and standard evaluation measures) were available for this language. However, other most spoken languages have an accelerated growth according to [1], e.g., the Spanish language is the second most spoken language in the world and the third most used on the internet.

The problem is that there is not standard corpus with human-generated summaries and evaluation methods, which are highly correlated with human judgments; therefore, there are not state-of-the-art methods for Spanish AETS (SAETS).

Consequently, to be able to update the SAETS, it is necessary to know how the study of this task has progressed for the English language over the 60 years of research.

Up to 2000, all research was focused on the English language and was carried out without having a standard corpus or evaluation measure, so a comparison could not be made. In 2001, the Document Understanding Conferences (DUC) was created with the objective of further progress in summarization in the English language and enable researchers to participate at a large scale.

Several DUC corpus was created over the years 2001 - 2007. DUC01 and DUC02 focus on the automatic text summarization for single and multiple documents; DUC03 to DUC07 for multiple documents with different tasks.

As a continuation of the DUC conferences in 2008, the conference TAC (Text Analysis Conference) is organized by a series of evaluation workshops created to improve systems evaluation. Corpus TAC focused on summaries created over the years: 2008, 2009, 2010, 2011, and 2014, being its main area of study the summaries for multiple documents focused on end-user.

In 2011, the MultiLing task was created to evaluate language-independent summarization algorithms on a for different languages. Several MultiLing corpora were created in 2011, 2013, 2015 and 2017 for the multilanguage automatic text summarization. The MultiLing task already works with different languages; the original texts are collected in English and translated into different languages, so there is no real corpus for each language.

Due to the number of papers published in Google Scholar, it is possible to obtain an approximation of the number of researches that resort to the standard corpus DUC (250 papers), TAC (100 papers) and MultiLing (30 papers). However, despite the efforts made to create a standard corpus of AETS, the most commonly used corpus to test methods and systems has been DUC02, and it is still currently used. [15-19]. DUC02 was built with specific features (news domain, labeling, model summaries, specific length, the measure of *baseline:first* heuristic) that make it robust and usable.

Another essential factor for AETS is the assessment method. Initially, evaluation methods for AETS were manually processed, that is, were evaluated by humans. However, these manual processes were costly and time-consuming. Subsequently, automatic evaluation methods were developed to reduce the costs presented by manual methods.

The evaluation methods of summaries are classified into two categories: intrinsic and extrinsic [20]. For intrinsic methods, it has a reference text, usually a summary created by a human (gold standard). However, other text or the same original document can also be used [21]. The methods of the extrinsic evaluation determine the effect of the summary on other tasks (e.g., relevance evaluation) [22].

Currently, the most used intrinsic evaluation method is ROUGE. The evaluation method ROUGE compares the summary to be evaluated (candidate summary) with the summary created by the human (model summary or reference summary) [23]. Because ROUGE uses as a reference to the summary created by the human, the evaluation is made concerning the criteria that the human used to generate the summary.

To make a more objective evaluation of the AETS, other intrinsic methods are proposed: ROUGE-C and Jensen Shannon divergence (JS).

These two evaluators, unlike ROUGE, use the original document as the reference text instead of the summary made by the human, which allows them to evaluate the performance of the methods concerning the entire content of the document.

For the English language, ROUGE-C and JS evaluation methods have not been used to assess state-of-the-art methods.

Since the creation of the standard corpus DUC and the creation of automatic evaluation methods, it has been possible to find out the progress for the AETS, also, different heuristics have been calculated, among them are a *baseline: random*, *baseline: first*, *topline* and *concordance*. The heuristics have served as a reference for the evaluation of the AETS.

Baseline: random baseline consists of randomly choosing the sentences that will constitute the summary [24]. So, when a method or system generates a summary, it is expected to be better quality than just random. *Baseline: first* consists of taking the first n sentences to make up the summary [25]. For state-of-the-art methods and systems, the goal is to overcome this heuristic. Mainly, for the news, it turns out to be very high, since this type of texts contains the most important information at the beginning of the document.

Topline consists of obtaining the best combination of sentences of every possible combination. This allows ascertaining what the maximum result that can be reached is when evaluating the summaries generated with a standard corpus [26-27].

Concordance consists of obtaining the correspondence or conformity that exists between summaries made by humans [28], for which it is only an informative heuristic and not a reference for the evaluation of method and system.

The heuristics serve as a reference to know the performance of state-of-the-art methods and systems. For the Spanish language, these heuristics have not been calculated due to the lack of resources.

As mentioned, most of the research on AETS is done for the English language. However, the methods performed and tested in the English language are not exclusive to this language.

Many of the state-of-the-art methods mention being language-independent [29-32] and some others, despite not saying they are independent of language, work within structures (extractive) that allow them to work with different languages [17,33-35]. The best methods that have performed are those based on graphs [36] and those based on genetic algorithms [17, 33-35].

In addition to state-of-the-art methods, systems for AETS are also currently available. AETS systems are methods available to the public and their use in some cases requires a payment.

For the Spanish language, few efforts have been made in the research of AETS. In 2001, Acero et al. [37], presents the automatic work generation of personalized summaries using their proper corpus, built with news from newspaper ABC.

Villatoro et al. [38] use the corpus created for the task of extracting information and adapts to apply it to the automatic multi-document text summarization for the Spanish language [39]. There are also other investigations on the SAETS as: [20, 37-38, 40-43].

However, despite the research carried out for SAETS, the current progress is not known because proper or adapted corpora have been used, which does not allow a comparison between the methods and second to the lack of standard corpus.

Currently, it is known which the best state-of-the-art methods and systems are for English Automatic Extractive Text Summarization (EAETS). Then, if they are tested in a standard corpus in Spanish and their performance is measured with different evaluation methods, the research in Spanish can be updated 60 years after the beginning of the task of AETS.

In addition, one can calculate the heuristics that are considered reference for comparison for the methods and systems of AETS.

This paper presents an update of SAETS to motivate research in the Spanish language. The results obtained from the evaluations with ROUGE, ROUGE-C, and JS are presented for state-of-the-art methods and systems of AETS with a standard corpus in Spanish. Also, the results obtained for the heuristics are presented (*baseline:random*, *baseline:first*, *topline* and *concordance*).

2 State of the Art Summarization Methods

In the task of AETS, state-of-the-art methods are using different techniques such as the use of genetic algorithms, neural networks, use of graphs, among others. In this article, two of the most used techniques for the task of AETS are taken up, to know how they work and test them.

2.1 Use of Graphs

The work of [34] has been one of the most referenced and resumed for new research, so in this article tests its operation for the task of SAETS.

2.1.1 TextRank

This method consists of a graph-based weighting algorithm. According to Rada Mihalcea [36], it constructs a graph to represent the text so that the nodes are words interconnected by arcs with significant relationships. For the task of extracting sentences, the objective is to qualify whole and classic sentences from higher to lesser importance.

Therefore, an arc is added to the graph for each sentence in the text. To establish the connections between sentences, a relation of similarity is defined, where the relationship between two sentences can be seen as a process of "recommendation". A sentence that indicates a certain concept in the text of a reader as a "recommendation" to refer to other sentences in the text that refer to the same concepts and, therefore, a link can be established between two sentences that share common content.

2.2 Use of Genetic Algorithms

Genetic algorithm's techniques have worked very well for the AETS and the state-of-the-art methods based on this type of technique have obtained acceptable results (surpassing the heuristic *baseline:first* for the English language). In this article, some of the state-of-the-art methods based on genetic algorithms are tested with a standard corpus in Spanish.

2.2.1 GA-Bag of Words

The method proposed by [35], uses a genetic algorithm based on the bag-of-words text model. The used fitness function takes two main features, which are mentioned below:

– The first sentences are more important. It is considered that the first sentences of a text as candidates to be part of the summary. For a text with n sentences, if the sentence i was selected for the summary (it is, the chromosome$|C_i| = 1$) then its relevance is defined as: $t(i - x) + x$, where $x = 1 + (n - 1)/2$ and t is the slope for discovering. To normalize the sentence position measure (δ), it is calculated the relevance of the first k sentences, where k is the number of selected sentences:

$$\delta = \frac{\sum_{|C_i|=1}^{n} t(i-x)+x}{\sum_{j=1}^{k} t(j-x)+x}, \; x = 1 + \frac{(n-1)}{2} \; . \qquad (1)$$

– It is evaluated that the summary has different ideas, it is not repetitive, but at the same time, it has important words using the measures of precision-recall. For generating a summary (S), the maximum-words threshold (m) of a summary is considered. Consequently, the number of recovery units always is limited by the maximum-word threshold. Therefore, the golden summary must have, for one side, the most relevant words of the original text (T) and, for the other side, must have expressivity, it means, it must not be redundant. The relevance of a word w is represented by the appearing frequency of the word in the original text (frequency(w, T)), and the expressivity is represented if only are considered the different words that the summary can have ({word ∈ S}).

In this sense, the best summary would contain the most frequent words concerning the original text, and each word must be different. To have a normalized measure the sum of the frequencies of the different words in the summary is divided by the sum of the frequencies of the most frequent words concerning the original text:

$$\beta = \frac{\sum_{p=\{word \in S\}}^{m} frequency(p,T)}{\sum_{q=\{word \in T\}}^{m} frequency(q,T)}. \quad (2)$$

Therefore, the fitness function was calculated as: $fitness = \beta \times \delta$.

2.2.2 GA-Multilanguage

The *GA-Multilanguage* method proposed by [44] has been applied to the ATS for different languages. The method is based on a genetic algorithm that uses n-grams with $n = 1, 2,3,4\ and\ 5$ as a text model.

For the fitness function, two of the most used features are considered on the-state-of-the-art [35], which are: term frequency (see Eq. 2) and sentences position. For feature sentence position Eq. 3 is used, calculated using the work of [18-18] sentence position using symbolic regression is calculated:

$$PS = \frac{(-28.7 - N)}{-.57.4}, \quad (3)$$

where N is the number of sentences in a text. Therefore, the fitness function is calculated as:

$$fitness = frequency\ of\ the\ terms \times sentence\ position. \quad (4)$$

2.2.3 GA-4feature

In [17], a method to optimize the combination of the four features is presented: similarity with the title (δ), the position of sentences (β) based on [35] (see Eq. 1), length of the sentence (γ) and coverture (α) based on [35] (see Eq. 2), based on a genetic algorithm for each step. For the similarity with the title obtains a weighting of the sentence according to the similarity with the document title, as it contains it relevant words that can be taken as unsupervised keywords.

Some similarity measures have been proposed, to mention some: Cosine, Euclidean, Dice, Jaccard, recently Soft Cosine [45], and other measures. However, these measures usually depend on the term selecting and weighting steps. Specifically, [33] uses the classical cosine similarity as term weighting and 1-grams (words) as term selection, described in the Eq. 5:

$$RT_S = \sum_{\forall S_i \in Summary} \frac{sim_{\cos(S_i,t)}}{0}, \quad (5)$$

where $sim_{\cos(S_i,t)}$ is the cosine similarity of sentence S_i with the title t, O is the number of sentences in the summary, RT_S is the average of the similarity in the summary S with the title, $max_{\forall Summary}\, RT$ is the average of the maximum values obtained from the similarities of all sentences in the document with the title (that is the average top more significant O similarities of all sentences with the title), and RTF_S is the similarity factor of the sentences of the summary S with the title, and is calculated in Eq. 6:

$$RTF_S = \frac{RT_S}{max_{\forall Summary}\, RT} = \delta\,. \quad (6)$$

For the length of the sentence, the Eq. 10 is used [31]. The fitness function used is presented in Eq. 7:

$$fitness = w_1\alpha + w_2\beta + w_3\gamma + w_4\delta\,. \quad (7)$$

2.2.4 MA-SingleDocSum

The method *MA-SingleDocSum* proposed by Mendoza [33] is based on a memetic algorithm, focused on the generation of summaries for a single document. In addition to using genetic operators for the generation of summaries, local search is used.

The features that are considered for the fitness function are sentencing position, the relation of sentences with title, sentence length, cohesion, and convergence (known as thematic of the text).

For sentence position, the scheme proposed by [46], was used, where a standard calculation is applied from the position based on Eq.8:

$$P = \sum_{\forall S_i \in Summary} \sqrt{\frac{1}{q_i}}, \quad (8)$$

where q_i indicates the position of the sentence S_i in the document and P is the result of the calculation for all sentences of the summary.

Calculation of the relation of the sentences with the title begins with the representation through the

vector space model, and the cosine similarity measure [47] is used, as shown in Eq. 9:

$$RT_S = \sum_{\forall S_i \in Summary} \frac{sim_{cos}(S_i, t)}{O}, \qquad RTF_S = \frac{RT}{max_{\forall Summary} RT}, \tag{9}$$

where $sim_{cos}(S_i, t)$ is the cosine similarity of sentence S_i with title t, O is the number of sentences in the summary, RT_S is the average of the similarity of the sentences in the summary S with the title, $max_{\forall Summary} RT$ is the average of the maximum values obtained from the similarities of all sentences in the document with the title (i.e., the average top greater O similarities of all sentences with the title), and RTF_S is the similarity factor of the sentences of the summary S with the title. RTF is close to one (1) when sentences, in summary, are closely related to the document title, and RTF is close to zero (0) when sentences, in summary, are very different to the document title.

For sentence length, it is considered that a sentence that is not too short will obtain a good grade in this characteristic. Based on this premise, Eq. (10) shows the calculation of length for the sentences of a summary (L):

$$\mathrm{L} = \sum_{\forall S_i \in Summary} \frac{1 - e^{\frac{-TL(S_i) - \mu(1)}{std(1)}}}{1 + e^{\frac{-TL(S_i) - \mu(1)}{std(1)}}}, \tag{10}$$

where $TL(S_i)$ is the length of sentence S_i (measured in characters), μ(l) is the average length of the sentence of the summary, and std(l) is the standard deviation of the lengths of the sentences of the summary.

For the calculation of cohesion, the cosine similarity measure of one sentence to another is used, see Eq. 11:

$$CoH = \frac{\log(C_S * 9 + 1)}{\log(M * 9 + 1)}, \qquad C_S = \frac{\sum_{\forall S_i S_j \in Summary} sim_{cos}(S_i S_j)}{N_S} \; N_S = \frac{(O) * (O - 1)}{2}, \qquad M = maxSim_{cos}(i, j), \;\; i, j \leq N, \tag{11}$$

where CoH corresponds to the cohesion of a summary, C_S is the average similarity of all sentences in the summary S, $sim_{cos}(S_i S_j)$ is the cosine similarity between sentences S_i and S_j, N_S is the number of nonzero similarity relationships in the summary, O is the number of sentences in the summary, M corresponds to the maximum similarity of the sentences in the document and N is the number of sentences in the document.

This way, CoH tends to 0 when the summary sentences are very different between them, while that CoH tends to 1 when these sentences are too similar between them.

Coverage is defined as the similarity between the sentences that produce a summary and the full document. Therefore, for each of the sentences, the document is consequently represented through the vector space model and is weighted by calculating its relative frequency according to Eq. 12:

$$Cov = \sum_{\forall S_i \in Summary} \sum_{\forall S_i \in Summary \; j > i} [sim_{cos}(D, S_i) + sim_{cos}(D, S_i)], \tag{12}$$

where D is the vector of weights of the terms in the document, and S_i and S_i are the vectors of weights of the terms in the sentences i and j, respectively, belonging to the summary.

The weights found for the objective function are: α = 0.35, β = 0.35, γ = 0.29, δ = 0.005, ρ = 0.005; which correspond to the features of Position (P), Relationship to the title (RT), Length (L), Cohesion (CoH) and Coverage (Cov), respectively.

To assess the quality of a summary represented by a representation of a solution X_k, an objective function is required, which will be maximized according to Eq. 13:

$$Max(f(X_k)) = \alpha P(X_k) + \beta RT(X_k) + \gamma L(X_k) + \delta CoH(X_k) + \rho Cob(X_k). \tag{13}$$

The fitness function was calculated as:

$$\alpha + \beta + \gamma + \delta + \rho = 1. \tag{14}$$

3 Summarization System

In this paper, we test to describe the commercial tools are tested to compare to the state-of-the-art methods to have a complete update of the SEATS.

3.1 Open Text Summarization

The Open Text Summarizer[1,2] (OTS) is an open source tool for summarizing texts. The OTS reads a text and decides which sentences are important and which are not. OTS will create a short summary or will highlight the main ideas in the text. OTS is both a library and a command line tool. Word processors such as AbiWord[3] and KWord[4] can link to the library and summarize documents while the command line tool can summarize text on the console. The program shows the summarized text as plain text or HTML.

3.2 Text Compactor

Text Compactor[5] is a free online summarization tool was created to help struggling readers a lot of information. The web app calculates the frequency of each word in the passage. Then, a score is calculated for each sentence based on the frequency count associated with the words it contains. The most important sentence is deemed to be the sentence with the highest frequency count.

3.3 Copernic Summarizer

The system was developed exclusively for ATS. According to [48], Copernic Summarizer uses the following methods:

– A common statistical model (S-Model) can be applied to a multi-language, to a certain degree, to approximate the topic-specific vocabulary. It includes Bayesian estimates and rule systems derived from an analysis of thousands of documents.
– Knowledge intensive processes (K-Process) consider how human beings summarize texts. Considering the following steps: language detection, sentences limit, extraction of concepts, segmentation of documents, and sentence selection.

3.4 Microsoft Office Word (MOW)

This tool has the option of ATS only in the versions Microsoft Office Word 2003 and Microsoft Office Word 2007. The summary created by Microsoft Office Word is the result of a keyword analysis; the selection of each keyword is done by assigning a score to each word. The tool offers several ways to view summaries. The most frequent words in the document will be higher scores that are considered important. The sentences that contain these words will be included in the summary.

3.5 Summarizing

Summarizing[6] is an online tool for EATS articles. The stages used are based on detecting the main ideas of the text, obtaining a description of the ideas, which reflects the author's writing style, to rewrite finally the text in summary. The Summarizing tool has the following parameters to generate summaries of 100, 150, 200, and 300 words.

4 Evaluation

In this section, three evaluation methods used in the AETS task are presented. ROUGE is the most evaluation method used in the evaluation of summaries that uses one or several gold standard summaries (summary made by the human) to perform its evaluation. While ROUGE-C and JS divergence are focused on the evaluation of the summaries concerning the original document, however, although they have different evaluation approaches, state-of-the-art methods that evaluate with ROUGE, ROUGE-C, and JS divergence must

[1] https://github.com/neopunisher/Open-Text-Summarizer
[2] https://www.splitbrain.org/services/ots
[3] AbiWord is a free word processing program. https://www.abisource.com/
[4] KWord, is a word processing program
[5] https://www.textcompactor.com/
[6] https://www.summarizing.biz/best-summarizing strategies/article-summarizer-online/

use the standard corpus to be compared with other methods.

4.1 ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) was proposed by Lin y Hovy [49-50]. ROUGE compares the summaries generated by a system to the human-generated (gold standard) summaries. For comparison, it uses n-gram statistics.

ROUGE includes the following automatic assessment measures.

– ROUGE-N (n-grams co-occurrence). It expresses the coverage or recall of n-grams between a candidate summary and a set of reference summaries. It is calculated as follows:

$$ROUGE-N = \frac{\sum_{set\{PeerSummary\}} \sum_{gram_n \epsilon S} Count_{match}(gram_n)}{\sum_{S \epsilon \{PeerSummary\}} \sum_{gram_n \epsilon S} Count(gram_n)}, \quad (15)$$

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 the set of reference summaries.

– ROUGE-S (noncontiguous bigram co-occurrence): a noncontiguous bigram is any pair of words in the order of the sentence, which allows for an arbitrary number of spaces. The co-occurrence of noncontiguous bigrams statistically measures the coverage of noncontiguous bigrams between the candidate summary and the set of reference summaries. Lin [49] shows that this sort of measure can be applied to assess the quality of automatically generated summaries, as 95% correlation between human judgments is managed.

4.2 ROUGE-C

ROUGE-C is presented as a tool to evaluate summaries without the reference summary made by the human [51]. The ROUGE-C method alternatively by replacing the reference summaries with source document as well as query-focused information (if any), therefore it enables a fully manual-independent way of evaluating multi-document summarization.

In ROUGE-C, for a summary of a document, they were defined as those used by ROUGE. For example, ROUGE-C-N, it is defined as shown in the Eq. (16):

$$ROUGE-C-N = \frac{\sum_{set\{PeerSummary\}} \sum_{gram_n \epsilon S} Count_{match}(gram_n)}{\sum_{S \epsilon \{SourceDocument\}} \sum_{gram_n \epsilon S} Count(gram_n)}, \quad (16)$$

where n stands the length of $n-gram$, $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a peer summary and the source document. ROUGE-C-N is the proportion of the overlapping grams in total n-grams of the source document. ROUGE-C-N is a precision-related measure the denominator of the equation is occurring on the Test side.

4.3 Jensen-Shannon Divergence (JS)

Jensen-Shannon divergence [52] is a method that evaluates the content of a summary that does not require models made by humans (gold standard). It assumes that the distribution of the words in the source document and the generated summary should be similar to each other.

The Jensen-Shannon divergence is a measure that compares two probability distributions of words: the text of the original document, (P), and the evaluated summary text, (Q). Low divergence from the input document(s) by the produced summary is taken as a signal of a good summary. Given two probability distributions over words: (P and Q), Jensen-Shannon divergence is defined as:

$$D_{JS}(P||Q) = \frac{1}{2}\sum_w P_w \, log_2 \frac{2P_w}{P_w + Q_w} + Q_w log_2 \frac{2Q_w}{P_w + Q_w}, \quad (17)$$

The measure can be applied to the distribution of units in system summaries P and reference summaries Q. The value obtained may be used as a score for the system summary.

JS divergence formula is given in Eq. 17 is implemented here with the following specification (see Eq. 18) for the probability distribution of words w:

Table 1. Description of tags used in the text

Tags	Description
<DOC></DOC>	Tag indicating the start and end of the document
<DOCNO> </DOCNO>	Tag indicating the name of the document
<FILEID></FILEID>	Tag indicating a unique number of the document
<TITLE></TITLE>	Tag indicating the title of the document
<CATEGORY></ CATEGORY >	Tag indicating the category to which the document belongs
<DATE></DATE>	Tag indicating the date of issue of the document
<TEXT></TEXT>	Tag indicating what is the text of the news
<s><\s>	Tag indicating the beginning and end of a sentence

Table 2. Description of tags used in the summary

Tags	Description
<SUM></SUM>	Tag indicating the beginning and end of the summary made by the human
CATEGORY	Tag indicating the category to which the news belongs
TYPE	Tag indicating the type of summary, in this case it is per document
SIZE	Tag Indicating the minimum number of words that the summary should have
DOCREF	Tag that shows the name of the base document for the generation of the extractive summary
SELECTOR	Tag with the unique key of the human that created the summary
SUMMARIZER	Tag that indicate which of the two generated abstracts is. A (first) and B (second).

Table 3. Features of the full texts of the corpus TER

Newspaper	Category	Documents	Words	Average of words	Sentences	Average sentences
Crónica	Academy	20	10966	548	382	19
	Wellness	20	11801	590	405	20
	City	20	7568	378	219	11
	Culture	20	8631	432	297	15
	Sports	20	9519	476	363	18
	Entertainment	20	8869	443	311	16
	States	20	7471	374	185	9
	World	20	7108	355	247	12
	National	20	7533	377	186	9
	Business	20	7523	376	229	11
	Opinion	20	12716	636	443	22
	Society	20	6507	325	228	11
	Total	240	106212		3495	
	Average			442		15

$$P_w = \frac{C_w^T}{N},$$

$$Q_w \begin{cases} \frac{C_w^S}{N_S} & \text{if } w \in S, \\ \frac{C_w^T + \delta}{N + \delta * B} & otherwise. \end{cases} \quad (18)$$

where P is the probability distribution of words w in text T and Q is the probability distribution of words w in summary S; N is the number of words in text and summary $N = N_T + N_S$, $B = 1.5|V|$ where V is the size of the vocabulary of the documents, C_w^T is the number of words in the text and C_w^T is the number of words in the summary. For smoothing the summary's probabilities, we have used δ = 0.005 [53].

It uses the versions smoothed (JS-SMT) and unsmoothed (JS-WSMT) versions of the divergence as features.

5 Experiment and Results

This section shows the experiments carried out on the best methods and systems for AETS, tested in a standard corpus in Spanish. First, the corpus used is described; second, the results of the different heuristics, state-of-the-art methods, and systems are presented using the evaluation methods ROUGE, ROUGE-C, and JS. In addition to showing the results obtained by the *concordance* between the summaries made by humans. Third, the ranking matrix is calculated for the methods and systems for the SAETS.

5.1 Corpus

The standard corpus used for the experimentation is called "*Textos en Español para Resúmenes*" (TER[7]). TER is a corpus composed by Mexican Spanish language news obtained from the newspaper "Crónica[8]".

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

In the first stage, 20 news items were randomly selected from the following categories: Academy, Wellness, City, Culture, Sports, Entertainment, States, World, National, Business, Opinion, and Society, giving a total of 240 news. The texts were cleaned of tags and images by extracting only the title, the category, the date and the main text of the news. Subsequently, a normalization of the texts was carried out, through the tagging of the texts.

The tagging of the text helps mainly to know where a sentence starts and ends. In this way, its use is facilitated, and it is guaranteed that the methods that use it will use the same separation of sentences. The tags used are shown in Table 1.

In the second phase for the creation of human-made summaries (gold standard), a group of humans of Mexican nationality and minimal university education was selected.

The human was given the text separated by sentences with the number of words corresponding to each of them so that they only read the text and select the sentences they considered important. Of prayers chosen, he was asked to create a more extensive summary of 100 words. Then for each document, two humans made an extractive summary of more than 100 words. The summaries were also tagged for their best use. Next, the tags used for the summaries are described.

Then there is a corpus of 240 news in the Spanish language of Mexico with two summaries made by humans for each news item. It is worth mentioning that the corpus was built considering the main features of DUC02. Table 3 presents a summary of how the TER corpus is constituted.

5.2 Concordance

The results of the *concordance* heuristic for the corpus TER are shown for three evaluation methods: ROUGE (see Table 4), ROUGE-C (see Table 5), and JS (see Table 6).

For ROUGE, the heuristic *concordance* shows a level of agreement between the experts of 66%. It shows that there are the two experts chose more than half of the sentences.

For ROUGE-C and JS, the *concordance* heuristic was applied to evaluate the first summary of human 1 with respect to the source text. Later the summary of human 2 was evaluated with respect to the source text. It is to fulfill the main

[7] https://github.com/gmatiasm/Corpus-TER

[8] http://www.cronica.com.mx/noticias.php

Table 4. Results of ROUGE *concordance* with TER

Measure	F-measure
ROUGE-1	0.6665
ROUGE-2	0.5432
ROUGE-SU4	0.5552

Table 5. Results of ROUGE-C *concordance* with TER

Measure	Precision
ROUGE-C-1	0.3709
ROUGE-C-2	0.3602
ROUGE-C-L	0.3684
ROUGE-C-SU4	0.3441

Table 6. Results of JS *concordance* with TER

Measure	Correlation
JS-SMT	0.7866
JS-WSMT	0.7720

features of ROUGE-C and JS to evaluate with respect to the original document. Finally, the average between the two summaries of the experts was established.

The results using JS show a higher *concordance* between the summaries of humans, while in ROUGE-C, the agreement is lower.

5.3 Experimental Results

We present the results of the heuristics, state-of-the-art methods, and systems evaluated with ROUGE (see Table 7), ROUGE-C (see Table 8), and JS (see Table 9).

According to the results presented in Table 4 for ROUGE, all methods and systems overcome the *baseline:random* heuristic. However, as regards *baseline: first,* only one method overcomes it. The *baseline:first* heuristic for the TER corpus is very high due to how the news items were written (the most important things are written at the beginning), as well as how the humans selected the sentences to produce the model summary. For the methods and systems that evaluate using the model summaries as a reference, they aim to overcome the heuristic *baseline:first*. The maximum result that can be reached when evaluating the summaries generated with a standard corpus TER is shown in the first row of results in Table 7.

The results of the evaluations of methods and systems of SAETS with ROUGE-C and JS are very similar with respect to the position of the methods and systems in the ranking. For ROUGE-C and JS, the *baseline:first* heuristic does not have much relevance because the evaluation reference is the complete document. According to the results presented in Table 8 for ROUGE-C (R-C) and Table 9 for JS, all methods and systems outperform the *baseline:random* heuristic and only one system does not overcome the *baseline:first* heuristic.

Despite the differences between the presented evaluation methods, it is observed that the state-of-the-art methods keep their order about their results.

5.3.1 The Ranking Results of the State-of-The-Art Methods and System

The main objective of the paper is to update the methods and systems for the SAETS. However, based on the results obtained by the evaluation methods, is generated a rank matrix to compare the position that the methods and systems have up to now.

Three evaluation methods were used (ROUGE, ROUGE-C, and JS), and each use a different way of calculating their output results (see Table 4-6), it is not possible to determine which of the methods or systems are the best. Therefore, a unification of the methods and systems is proposed considering the position that each method and system occupies according to its evaluation measure. Table 10 shows the position of each method and system with respect to the results obtained by each measure. The resulting ranking matrix was calculated as proposed in [54] as follows (see Eq. 19):

$$Ran = \sum_{r=1}^{n} \frac{(n - r + 1)R_r}{n}, \tag{19}$$

where n is the number of methods and systems involved for the comparison, and R_r refers to the number of times that the method or system affects the r - th position.

Table 7. Results of ROUGE for methods, systems and heuristics regarding TER

Method \ System	ROUGE - 1	ROUGE – 2	ROUGE - SU4
Topline	0.8344	0.7664	0.7649
GA-Multilanguage	**0.7274**	**0.6289**	**0.6378**
Baseline:first	0.7626	0.6229	0.6326
GA-4feature	0.7131	0.6072	0.6180
GA-Bag of words	0.6989	0.5852	0.5972
MA-SingleDocSum	0.6883	0.5706	0.5842
OTS	0.6761	0.5562	0.5698
Text Compactor	0.6749	0.5537	0.5678
TextRank	0.6606	0.5390	0.5532
Copernic	0.6187	0.4711	0.4898
MOW2007	0.6178	0.4691	0.4854
MOW 2003	0.6160	0.4649	0.4819
Summarizing	0.5775	0.4098	0.4290
Baseline:random	0.4969	0.2933	0.3201

Table 8. Results of ROUGE-C for methods, systems and heuristics regarding TER

Method \ System	ROUGE - C1	ROUGE - C2	ROUGE - CL	ROUGE – CSU4
GA-4feature	**0.5041**	**0.4968**	**0.5041**	**0.4864**
MA-SingleDocSum	**0.5044**	**0.4945**	**0.5044**	**0.4803**
TextRank	0.4402	0.4290	0.4402	0.4128
GA-Multilanguage	0.3915	0.3867	0.3915	0.3793
MOW2007	0.3688	0.3567	0.3654	0.3395
MOW 2003	0.3559	0.3438	0.3527	0.3266
GA-Bag of words	0.3477	0.3411	0.3477	0.3309
OTS	0.3509	0.3413	0.3490	0.3272
Text Compactor	0.3406	0.3315	0.3389	0.3177
Summarizing	0.2754	0.2636	0.2726	0.2466
Baseline:first	0.2791	0.2756	0.2764	0.2699
Copernic	0.2971	0.2852	0.2952	0.2733
Baseline:random	0.2538	0.2322	0.2475	0.2147

According to the results shown in Table 10, the best state-of-the-art method is *GA-4feature*, and the lowest result has the Summarizing syste6 Conclusions and Future Work. Automatic extractive text summarization has been under research for 60 years.

However, the progress made in the Spanish Automatic Extractive Text Summarization was not known until the present paper. In this paper, we tested a standard corpus in Spanish with the best state-of-the-art methods and systems of AETS.

The evaluation was carried out with ROUGE, ROUGE-C, and JS evaluation measures.

The results obtained with ROUGE show that the state-of-the-art methods and systems have a challenge to overcome because the *baseline:first*

Table 9. Results of JS for methods, systems and heuristics regarding TER

Method \ System	JS-SMT	JS-WSMT
GA-4feature	**0.8524**	**0.8436**
MA-SingleDocSum	0.8452	0.8362
TextRank	0.8223	0.8120
GA-Multilanguage	0.7950	0.7812
MOW2007	0.7920	0.7773
MOW 2003	0.7858	0.7702
GA-Bag of words	0.7796	0.7634
OTS	0.7745	0.7592
Text Compactor	0.7690	0.7526
Summarizing	0.7343	0.7124
Baseline:first	0.7321	0.7107
Copernic	0.7250	0.7061
Baseline:random	0.7105	0.6884

Table 10. The ranking of the state-of-the-art methods and systems

Method/system	R(r)											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**R**
GA-4feature	4	5	0	0	0	0	0	0	0	0	0	8.5
MA-SingleDocSum	2	4	0	3	0	0	0	0	0	0	0	7.8
GA-Multilanguage	3	0	0	6	0	0	0	0	0	0	0	7.3
TextRank	0	0	6	0	0	0	3	0	0	0	0	6.2
GA-Bag of words	0	0	3	0	0	1	4	1	0	0	0	5.1
MOW 2007	0	0	0	0	6	0	0	0	3	0	0	4.6
OTS	0	0	0	0	3	0	2	4	0	0	0	4.2
MOW 2003	0	0	0	0	0	5	0	1	0	3	0	3.6
Text Compactor	0	0	0	0	0	3	0	0	6	0	0	3.2
Copernic	0	0	0	0	0	0	0	3	0	4	2	2.0
Summarizing	0	0	0	0	0	0	0	0	0	2	7	1.0

heuristic is very high and only one method has managed to overcome it.

For ROUGE-C and JS all the state-of-the-art methods and three of four proven systems overcome the *baseline:first* heuristic.

All state-of-the-art methods and systems overcome *baseline:random* heuristic.

For English, the following methods: *MA-SingleDocSum*, *GA-Multilanguage*, *GA-Bag of words*, and *GA-4feature* outperform the heuristic *baseline:first*.

However, for Spanish, only AG-Multilanguage exceeds it for ROUGE measures (ROUGE-1, ROUGE-2, and ROUGE-SU4). There is no evidence of evaluations of state-of-the-art methods in English regarding ROUGE-C and JS. Therefore, the results show that the conclusions obtained for English are not supported for Spanish.

The degree of progress for Spanish was ascertained using the ranking of the state-of-the-art methods and systems for the AETS shown in Table 10.

Based on the results shown in this paper, the opportunity to generate new research is opened using the TER corpus to try state-of-the-art methods as [55-62], among others Also, the methods and systems tested in this paper could be adjusted their parameters to obtain better results in SAETS.

Acknowledgments

The authors are grateful to the SEP-SES for their support.

References

1. **Fernández, D.V. (2017)**. *El español una lengua viva.* Informe 2017, Instituto Cervantes.
2. **Gambhir, M. & Gupta, V. (2017).** Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, Vol. 47, No. 1, pp. 1–66.
3. **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.
4. **Edmundson, H.P. (1969).** New methods in automatic extracting. *Journal of the ACM (JACM),* Vol. 16, No. 2, pp. 264–285. DOI:10.1145/ 321510.321519.
5. **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.
6. **Paice, C.D. (1990).** Constructing literature abstracts by computer: techniques and prospects. *Information Processing & Management*, Vol. 26*,* No. 1, pp. 171– 186.
7. **Jing, H., Barzilay, R., McKeown, K., & Elhadad, M. (1998).** Summarization evaluation methods: Experiments and analysis. *AAAI Symposium on Intelligent Summarization*, pp. 51–59.
8. **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. *EACL 97, Workshop Intelligent Scalable Text Summarization,* pp. 25–31.
9. **Barzilay, R. & Elhadad, M. (1999).** Using lexical chains for text summarization. *Advances in automatic text summarization*, pp. 111–121.
10. **Benbrahim, M. & Ahmad, K. (1995).** Text summarisation: The role of lexical cohesion analysis. *The New Review of Document & Text Management*, Vol. 1, pp. 321–335.
11. **Carbonell, J. & Goldstein, J. (1998).** The use of MMR, diversity-based reranking for reordering documents and producing summaries. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 335–336.
12. **Marcu, D. (1997).** From discourse structures to text summaries. *Intelligent Scalable Text Summarization*, pp. 82–88.
13. **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,* pp. 74–82.
14. **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.
15. **Verma, R. & Lee, D. (2017).** Extractive summarization: limits, compression, generalized model and heuristics. *Computación y Sistemas,* Vol. 21, No. 4, 2017, pp. 787–798. DOI: 10.13053/CyS-21-4-2855
16. **John, A., Premjith, P.S., & Wilscy, M. (2017).** Extractive multi-document summarization using population-based multicriteria optimization. *Expert Systems with Applications*, Vol. 86, pp. 385–397. DOI: 10.1016/j.eswa.2017.05.075.
17. **Vazquez-Vazquez, E., García-Hernández, R.A., & Ledeneva, Y. (2018).** Sentence features relevance for extractive text summarization using genetic algorithms. *Journal of Intelligent & Fuzzy Systems. Applications in Engineering and Technology*, Vol. 35, No. 1, pp. 353–365. DOI:10.3233/JIFS-169594.
18. **Vazquez-Vazquez, E., Ledeneva, Y., & García-Hernández, R.A. (2019).** Learning relevant models using symbolic regression for automatic text summarization. *Computación y Sistemas,* Vol. 23, No. 1, pp. 127–141. DOI: 10.13053/CyS-23-1-2921.
19. **Hernández-Castañeda, N., García-Hernández, R. A., Ledeneva, Y., & Hernández-Castañeda, Á. (2020).** Evolutionary automatic text summarization using cluster validation indexes. *Computación y Sistemas,* Vol. 2, No. 2, pp. 583–595. DOI: 10.13053/CyS-24-2-3392.

20. **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.
21. **Steinberger, J. & Ježek, K. (2012).** Evaluation measures for text summarization. *Computing and Informatics*, Vol. 28, No. 2, pp. 251–275.
22. **Berker, M. (2011).** Using genetic algorithms with lexical chains for automatic text summarization. pp.1–54.
23. **Lin, C.Y. (2004).** Rouge: A package for automatic evaluation of summaries. *Association for Computational Linguistics,* pp. 74–81.
24. **Ledeneva, Y.N. (2008).** *Automatic Language-Independent Detection of Multiword Descriptions for Text Summarization.* Centro de Investigación en Computación, pp. 1–120.
25. **Document Understanding Conferences - Past Data (2001).** https://duc.nist.gov/data.html
26. **Rojas, J. (2018).** Calculating the significance of automatic extractive text summarization using a genetic algorithm. *Journal of Intelligent & Fuzzy Systems.* Vol, 35, No. 1, pp. 293–304. DOI: 10.3233/JIFS-169588.
27. **Rojas-Simón, J., Ledeneva, Y., & García-Hernández, R.A. (2018).** Calculating the upper bounds for multi-document summarization using genetic algorithms. *Computación y Sistemas*, Vol. 22, No. 1, pp. 11–26. DOI: 10.13053/CyS-22-1-2903.
28. **Mitrat, M., Singhal, A., & Buckleytt, C. (1997).** Automatic text summarization by paragraph extraction. *Intelligent Scalable Text Summarization*. pp. 39–49.
29. **Patel, A., Siddiqui, T., & Tiwary, U.S. (2007).** A language independent approach to multilingual text summarization. *Conference RIAO´07,* pp. 1–10.
30. **Mihalcea, R. & Tarau, P. (2005).** *A language independent algorithm for single and multiple document summarization*. pp. 19–24.
31. **Litvak, M., Lipman, H., Ben-Gur, A., Last, M., Kisilevich, S., & Keim, D. (2010).** Towards multi-lingual summarization: A comparative analysis of sentence extraction methods on English and Hebrew corpora. *Proceedings of the 4 th Workshop on Cross Lingual Information Access,* pp. 61–69.
32. **Saggion, H. (2011).** Using SUMMA for Language Independent Summarization at TAC 2011. *TAC*.
33. **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.
34. **Matias, G. (2016).** *Generación Automática de Resúmenes Independientes del Lenguaje.* Universidad Autónoma del Estado de México.
35. **García-Hernández, R.A. & Ledeneva, Y. (2013).** Single extractive text summarization based on a genetic algorithm. **Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F., Rodríguez, J.S., di Baja G.S. (eds)** *Pattern Recognition. MCPR´13. Lecture Notes in Computer Science,* Vol. 7914. DOI: 10.1007/978-3-642-38989-4_38.
36. **Mihalcea, R. (2004).** Graph-based ranking algorithms for sentence extraction, applied to text summarization. A*ssociation for Computational Linguistics.*
37. **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.
38. **Villatoro E. (2007).** *Generación automática de resúmenes de múltiples documentos*. Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla.
39. **Téllez, A., Montes, M., & Villaseñor-Pineda, L. (2009).** Using Machine learning for extracting information from natural disaster news reports. *Computación y Sistemas*, Vol. 9, No.1, pp. 33–44.
40. **Toledo-Báez, M.C. (2010).** *Aproximación al resumen automático como herramienta de ayuda a la traducción jurídica en el ámbito del Derecho turístico*. pp. 568–578.
41. **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 ACM symposium on Document Engineering*, pp. 203–206. DOI:10.1145/2644866.2644890.
42. **Molina, A. (2013).** Compresión automática de frases: un estudio hacia la generación de resúmenes en espanol. *Inteligencia Artificial*, Vol. 16, No. 51, pp. 41–62.
43. **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.
44. **Mendoza**, **G.A.M., Ledeneva, Y., & García-Hernández, R.A. (2019.).** Determining the importance of sentence position for automatic text summarization. *Journal of Intelligent & Fuzzy*

Systems, Vol 39, No. 2, pp. 2421–2431. DOI: 10.3233/JIFS-179902.

45. **Sidorov, G., Gelbukh, A., Gómez-Adorno, H., & Pinto, D. (2014).** Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, Vol. 18, No. 3, pp. 491–504. DOI: 10.13053/CyS-18-3-2043.
46. **Bossard, A., Généreux, M., & Poibeau, T. (2008).** Description of the LIPN System at TAC: Summarizing Information and Opinions. *Presented at the TAC,* pp. 282–291.
47. **Qazvinian, V., Hassanabadi, L.S., & Halavati, R. (2008).** Summarising text with a genetic algorithm-based sentence extraction. *International Journal of Knowledge Management Studies*, Vol. 2, No. 4, pp. 426–444. DOI: 10.1504/IJKMS.2008.01975.
48. **Copernic Summarization-Technologies White Paper. (2003).** http://www.copernic.com/data/pdf/summarization-whitepaper- eng.pdf
49. **Lin, C.Y. & Hovy, E. (2003).** Automatic evaluation of summaries using n-gram co-occurrence statistics. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology,* Vol.1, pp. 71–78.
50. **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.*
51. **He, T., Chen, J., Ma, L., Gui, Z., Li, F., Shao, W., & Wang, Q. (2008).** ROUGE-C: A fully automated evaluation method for multi-document summarization. *IEEE International Conference,* pp. 269–274. DOI:10.1109/GRC.2008.4664680.
52. **Louis, A. & Nenkova, A. (2009).** Automatically evaluating content selection in summarization without human models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* Vol. 1, pp. 306–314.
53. **Torres-Moreno, J.M., Saggion, H., da-Cunha, I., San Juan, E., & Velázquez-Morales, P. (2010).** Summary evaluation with and without references. *Polibits*, Vol. 42, pp. 13–20.
54. **Aliguliyev, R.M. (2009).** Performance evaluation of density-based clustering methods. *Information Sciences*, Vol. 179, No. 20, pp. 3583–3602. DOI: 0.1016/j.ins.2009.06.012.
55. **Wan, X. (2010).** Towards a unified approach to simultaneous single-document and multi-document summarizations. *Proceedings of the 23rd international Conference on Computational Linguistics, Association for Computational Linguistics,* pp. 1137–1145.
56. **Aliguliyev, R.M. (2009).** A new sentence similarity measure and sentence based extractive technique for automatic text summarization. *Expert Systems with Applications*, Vol. 36, No. 4, pp. 7764–7772. DOI:10.1016/j.eswa.2008.11.022
57. **Song, W., Choi, L.C., Park, S.C., & Ding, X.F. (2011).** Fuzzy evolutionary optimization modeling and its applications to unsupervised categorization and extractive summarization. *Expert Systems with Applications*, Vol. 38, No. 8, pp. 9112–9121. DOI: 10.1016/j.eswa.2010.12.102.
58. **Svore, K., Vanderwende, L., & Burges, C. (2007).** Enhancing single-document summarization by combining RankNet and third-party sources. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,* pp. 448–457.
59. **Shen, D., Sun, J.T., Li, H., Yang, Q., & Chen, Z. (2007).** Document summarization using conditional random fields. *IJCAI*, Vol. 7, pp. 2862–2867.
60. **Dunlavy, D.M., O'Leary, D.P., Conroy, J.M., & Schlesinger, J.D. (2007).** QCS: A system for querying, clustering and summarizing documents. *Information processing & management*, Vol. 43, No. 6, pp. 1588–1605. DOI:10.1016/j.ipm.2007.01.003.
61. **Yeh, J.Y., Ke, H.R., Yang, W.P., & Meng, I.H. (2005).** Text summarization using a trainable summarizer and latent semantic analysis. *Information Processing & Management*, Vol. 41, No. 1, pp. 75–95. DOI: 10.1016/j.ipm.2004.04.003.
62. **Fors-Isalguez, Y., Hermosillo-Valadez, J., & Montes-y-Gómez, M. (2018).** Query-oriented text summarization based on multiobjective evolutionary algorithms and word embeddings. *Journal of Intelligent & Fuzzy Systems*, Vol. 34, No.5, pp. 3235–3244. DOI:10.3233/JIFS-169506.

Article received on 05/06/2020; accepted on 22/07/2020.
Corresponding author is Yulia Ledeneva.