IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 May 2023, accepted 14 May 2023, date of publication 23 May 2023, date of current version 1 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3279101

RESEARCH ARTICLE

Toward the Automatic Generation of an Objective Function for Extractive Text Summarization

ÁNGEL HERNÁNDEZ-CASTAÑEDA[1,2], RENÉ ARNULFO GARCÍA-HERNÁNDEZ[2], AND YULIA LEDENEVA[2]

[1]Cátedras Conacyt, Mexico City 03940, Mexico
[2]Autonomous University of Mexico State, Toluca 50000, Mexico

Corresponding authors: Ángel Hernández-Castañeda (angelhc2305@gmail.com), René Arnulfo García-Hernández (renearnulfo@hotmail.com), and Yulia Ledeneva (yledeneva@yahoo.com)

This work was supported by the Consejo Nacional de Ciencia y Tecnología (CONACyT) under the Cátedras program.

ABSTRACT A fitness function is a type of objective function that quantifies the optimality of a solution; the correct formulation of this function is relevant, in evolutionary-based ATS systems, because it must indicate the quality of the summaries. Several unsupervised evolutionary methods for the automatic text summarization (ATS) task proposed in current standards require authors to manually construct an objective function that guides the algorithms to create good-quality summaries. In this sense, it is necessary to test each fitness function created to measure its performance; however, this process is time consuming and only a few functions are analyzed. This study proposes the automatic generation of heuristic functions, through genetic programming (GP), to be applied in the ATS task. Therefore, our proposed method for ATS provides an automatically generated fitness function for cluster-based unsupervised approaches. The results of this study, using two standard collections, demonstrate to automatically obtain an orientation function that leads to good quality abstracts.

INDEX TERMS Automatic text summarization, clustering, genetic programming, genetic algorithms, heuristic functions.

I. INTRODUCTION

Recently, the internet has held a large amount of textual information of different kinds, such as academic, disclosure and general knowledge documents. A search on the internet can lead to the selection of a subset of documents that are appropriate for the user objective (e.g., conducting research or writing an essay). Each document needs to be analyzed to understand the main purpose of the writer; in this process, the main ideas are classified from those that are secondary. Thus, the selected key ideas form the condensed version of the original document that should preserve the central, relevant or vital information. As a result, the summary can provide a general idea of a complex document (e.g., a book, scientific paper, etc.) allowing the reader learn the main points on it.

In the natural language processing (NLP) area, a specific area of artificial intelligence, different approaches have been proposed to automatically build summaries simulating human ability. Automatic text summarization (ATS) is a task that automatically produces summaries to identify key ideas of a source document [1]. In this sense, the ATS task is addressed by two approaches: abstractive [2] and extractive [3], [4] summarization.

On the one hand, abstractive summarization methods generate new text that cannot be contained in the original document. To do this, the internal semantic representation of the source documents is commonly learned to create a language model. The obtained model could create new sections paraphrasing the content of documents to generate the condensed document. The abstractive method may produce more strongly condensed documents but could lose the main meaning of the original document.

On the other hand, extractive summarization methods make use of the content in the source documents to generate the condensed version (summary). To that end, extractive systems may use different basic units such as words, sentences or paragraphs; most of the state of the art methods [5], [6] use sentences as a basic unit because it

The associate editor coordinating the review of this manuscript and approving it for publication was Yu-Da Lin.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

could express a complete idea of an author. Because of the complexity of the abstractive approach, most summarization systems are extractive.

In turn, extractive-based summaries can be addressed by supervised and unsupervised methods [7]. The supervised approach commonly needs a prelabeled corpus in which the key ideas are highlighted to train a supervised algorithm; then, it is possible to recognize key ideas in a new document. The unsupervised approach, for example, identifies key sentences based on a word count.

The unsupervised extractive ATS system has the advantage of not needing a training step, and in addition, it may be more appropriate for real cases where a labeled dataset is not always available. In this sense, clustering-based schemes are widely used by grouping sentences to discover general topics and then selecting the main idea of each group.

In addition, it is common to combine evolutionary and clustering algorithms for unsupervised approaches to address the combinatorial problem in group formation [8]. In turn, evolutionary methods need a guidance function to find the best clustering configurations, but finding optimal clusters is not guaranteed to generate a good-quality summary [9]. The guidance or fitness function is a means to know the quality of solutions generated by evolutionary approaches and is generally established based on the author's intuition.

In this study, the main objective is the automatic search for a fitness function that correlates with the quality of summaries by combining genetic programming (GP) and a genetic algorithm (GA). Integration of these algorithms has been used in previous studies using the GP to find the hidden relationships between features to build general structures and then using GA to identify relevance between them [10]. On the one hand, we propose a GP system that considers the internal validation indices to automatically build functions. On the other hand, the GA creates summaries using each function created by the GP as its fitness function. Finally, the relationship between the quality of the summaries and the clustering is established using the Rouge measure.

The general contributions of this study are the following: a) a system for the automatic generation of aptitude functions for the ATS task; b) performance analysis of the GP and GA integration for the ATS task; c) the generation of fitness functions correlated with human behavior for the generation of summaries; d) performance analysis of various text representation methods; e) comparison of our proposed method with attention-based methods.

The rest of the paper is organized as follows. Section II describes different approaches that use a fitness function as a guide to evolutionary methods. Section III details the basic concept applied in this study. Section IV describes the framework of the proposed approach for the ATS task. The performance of the proposed method is addressed in Section V. Finally, conclusions are drawn in Section VI.

II. RELATED WORK

The ATS task attempts to create a condensed version of a document synthesizing or abbreviating ideas of a more complex document while preserving the most relevant information. In the literature, various kinds of summarization problems have been raised, and in turn, various methods to solve those problems have been proposed.

For example, with regard to the source documents, an ATS system could automatically create an abstract by condensing a single document or multiple documents [11], [12], [13]; this makes the task more difficult because it increases redundancy in selected ideas and the amount of information to analyze.

In addition, on the one hand, summaries can be extractives, which are generated using only the information included in the original document, such as words, sentences or paragraphs; on the other hand, summaries can be abstract [14], [15], which could convey new information not included in the original document, by commonly combining a paraphrasing process with a language model.

The ATS systems can also be classified according to the approach used to select the relevant information as supervised and unsupervised. Supervised systems need a training process and, in turn, a prelabeled dataset highlighting the relevant information. Unlike unsupervised systems, unsupervised systems are more practical in the analysis of multidomain documents because they are not limited by training information.

Clustering-based unsupervised approaches [16], [17] typically use an evolutionary algorithm to optimize groups of sentences to generate good-quality abstracts by selecting the best candidate sentences from each group. Thus, the most relevant part of evolutionary algorithms is the definition of a fitness function that evaluates each solution. For example, Alguliyev et al. [8] use a differential evolution algorithm to maximize the fitness function that measures the summary quality. This function is focused on the relevance and diversity of the information contained in summaries.

Another evolutionary approach was proposed by Rautray and Balabantaray [18], where a Cuckoo search is performed to address the problem of multidocument summarization. The authors consider different aspects to build the fitness function, such as coverage, nonredundancy, cohesion and readability.

Similar to the works above, Sanchez-Gomez et al. [19] proposed a multiobjective artificial bee colony to automatically generate good-quality summaries. The objective function proposed by the authors addresses coverage, where the main topics in the source document should be considered, and redundancy reduction, where similar sentences existing in the source document should not be repeated in the generated summary.

As mentioned above, authors who propose clustering-based ATS systems manually configure their fitness functions considering different aspects of texts, such as sentence relevance, topic diversity, non-redundancy between sentences, and readability, among others. This information is obtained by processing texts at different levels (for example, words, sentences, or paragraphs) and then formulating an objective function that, according to the author's intuition, correlates with the quality of the summaries.

Therefore, the information considered to formulate the fitness function, which is commonly a linear function, must result in a single value that determines the quality of the solution.

Accordingly, unsupervised evolutionary approaches for generating summaries do not require a prelabeled corpus to learn; however, they need a guidance function (fitness function) to build good-quality summaries. Fitness functions can be maximized or minimized and are usually set manually based on intuition. Manual configuration limits the exploration of new functions that could improve results; therefore, in this work, we propose to create guidance functions automatically using genetic programming.

III. METHODOLOGY

The proposed method (detailed in Section IV) aims to create fitness functions for evolutionary cluster-based methods for automatic text summarization. In this process, on the one hand, texts or documents to be summarized should be represented as numeric vectors by means of different mapping methods, as described in Section III-A; on the other hand, internal quality measures, detailed in Section III-B, are provided as operands to the genetic programming with the purpose of identifying its correlations with the quality of summaries.

A. MAPPING METHODS

In this study, four methods to represent documents as numerical vectors are proposed to explore the relevance of lexical and semantic information for the identification of relevant sentences in the ATS task. These methods are detailed below.

1) FEATURES BASED ON TERM FREQUENCY–INVERSE DOCUMENT FREQUENCY

The term frequency - inverse document frequency ($tf - idf$) [20] is widely used in natural language processing and information retrieval. In the ATS task, $tf - idf$ (equation 1) provides information about how relevant a word within a document is in relation to the collection. Specifically, tf (equation 2) is the ratio between the number of times that a word appears in a document and the total number of words in that document. Instead, IDF (equation 3) shows how relevant a word is relative to the collection of documents by computing the ratio between the number of documents in the collection and the number of documents in the collection that contains the word. Thus, tf shows how relevant a word is in a specific document, while idf assigns more importance to those unique words to a small percentage of documents than to those words that are very common (e.g., the, a, and).

$$tf - idf = (t, d, D) = tf(t, d) * idf(t, D) \tag{1}$$

$$tf(t, d) = \frac{f_d(t)}{\max_{w \in d} f_d(w)} \tag{2}$$

$$idf(t, D) = ln(\frac{|D|}{|d \in D : t \in d|}) \tag{3}$$

where:

$f_d(t)$ = frequency of term t in document d
D = collection of documents

To create the vector representation of a document, the vocabulary V of the collection of documents is listed, that is, a list of each different word w in the collection (namely, types). Then, given a document d, the $tf - idf$ value is calculated for each $w_i \in V$ and set in position i of the representative vector of d. Therefore, each w has some relevance relative to each d in the collection, and in turn, $|V|$-dimentional vectors are created.

2) ONE-HOT ENCODING

One-hot encoding (OHE) [21] is one of the simpler methods for representing text as a numeric vector; however, it has proven to provide relevant information to NLP models. The process to create vectors is similar to $tf - idf$; however, in the OHE method, the values of each position of the representative vector are binary. Therefore, for each $w_i \in V$, position i of the representative vector is set to 1 if $w \in d$ and is set to 0 otherwise. As a result, the resultant OHE vectors consist of $|V|$ dimensions.

3) LATENT DIRICHLET ALLOCATION

Latent Dirichlet allocation (LDA) [22] is a topic modeling algorithm that, in NLP, commonly attempts to generate documents based on a priori sampled distributions of documents over topics and, in turn, distributions of words over topics. Therefore, LDA sees a document as a mixture of topics and sees a topic as a mixture of words. To obtain the correct distribution given a collection of documents, this topic modeling algorithm attempts to maximize the following formula:

$$p(W, Z, \theta, \varphi; \alpha, \beta) = \prod_{j=1}^{M} p(\theta_j; \alpha) \prod_{i=1}^{K} p(\varphi_i; \beta) \times \prod_{t=1}^{N} p(Z_{j,t}|\theta_j) p(W_{j,t}|\varphi z_{j,t})$$

where the first two factors are related to the Dirichlet distribution of topics over terms and the distribution of documents over topics, respectively, while the last two factors represent the probability of a topic appearing given a document and the probability of a word appearing given a topic.

In this study, we use the latent topic distribution obtained by the LDA model to represent documents in terms of the themes that make them up.

The main advantage of LDA is that it allows obtaining the latent structure of a document; that is, we can obtain a distribution of topics in a vector representation used as input to the clustering phase (Section IV-B). Therefore, this representation provides a topic-based grouping that the fitness function evaluates to determine the quality of the summaries.

4) Doc2Vec

Commonly, text representation ignores the relationship in a sequence of words, such as many bag of words methods that ignore the word order in phrases. Doc2Vec [23], [24] is a method that learns vector representations of words and, in turn, of sentences and documents. To that end, the doc2vec algorithm considers the context of words by computing the probability that a certain word is in the context of other words.

Specifically, given a training set of words $W = w_1, w_2, \ldots, wT$, the goal is to maximize the probability of w_t appearing, such as n words appearing before (Equation 4). Thus, the prediction of w_t can be performed by the softmax multiclass classifier (Equation 5).

$$\sum_{t=k}^{T-k} log\ p(w_t \mid w_{t-k}, \ldots, w_{t+k}), \tag{4}$$

$$p(w_t \mid w_t - k, \ldots, w_{t+k}) = \frac{e^{y_{w_t}}}{\sum_{e^{y_i}}} \tag{5}$$

where each y_i is given by $y = b + Uh(w_{t-k}, \ldots, w_{t+k}; W)$, h is constructed by the concatenation of vectors in the word matrix W and U, b are the softmax parameters.

As a result of the training process and the inclusion of the paragraph context information, vectors of fixed dimensions are generated. Furthermore, these vectors involve semantic relations where sentences or paragraphs with similar meaning are closer in the vector space.

B. INTERNAL QUALITY MEASURES

In pattern recognition, clustering is an unsupervised classification where an algorithm attempts to organize objects or patterns into k-groups. The main goal of this task is that objects in the same group should be as compact as possible, while objects of different groups should be as different as possible.

The cluster validation indices are measures to evaluate the quality of a clustering given two main characteristics: the internal homogeneity and the external separability. The former evaluates how compact a group is, while the latter evaluates how far apart one group is from another.

In the works of Liu et al. [25] and Rendón et al. [26], different cluster validation indices are tested through different synthetic datasets. Both studies conclude that the Dunn, Davies Bouldin and Silhouette indices are highlighted from other indices on the proposed synthetic datasets. Hernández-Castañeda et al. [9], based on the results of Liu and Rendón et al., search for the correlation between Dunn, Davies Bouldin and Silhouette indices and the quality of summaries. The authors propose three forms to generate groups called baselines: 1) top-line, where summaries written by humans are used as reference; 2) first-line, where key ideas are those n first sentences of the documents; and 3) random-line, where key ideas were selected randomly from the documents. Research results show that the Silhouette index has more correlation with the quality of summaries because it shows high performance when relevant information is selected by humans (top-line), while it shows low performance when relevant information is selected randomly (random-line).

In this study, in view of the above, we propose to use the three indices, defined below, that show the best performance in the clustering tasks.

The **Dunn index** [27] measures the relation between the maximal distance in the same group and the minimum distance between groups of the partition. That is, for each cluster, the pairwise distance between each of the objects in the cluster and the objects of the remainder of the clusters is computed. Then, the minimum pairwise distance (min-separation) is obtained. Next, for each cluster, the distance between all objects of the same group is calculated; the maximum distance (max-diameter) is selected. Formally, the Dunn index is defined as follows:

$$Dunn = \frac{min_{1 \le i < j \le c} f(c_i, c_j)}{max_{1 \le k \le c(d(X_k))}}$$

where $f(c_i, c_j)$ defines the intercluster separation and $d(X_k)$ stands for the intracluster compactness. Thus, the Dunn index should be maximized.

The **Davies Bouldin index** [28] computes, for each cluster, the average distance between the objects and its centroid to measure the compactness of the clusters. In addition, to identify the cluster separation, the distance between centroids is computed. This index is defined as follows:

$$DB = \frac{1}{c} \sum_{i=1, i \ne j}^{c} Max\{\frac{\delta_i + \delta_j}{d(c_i, c_j)}\}$$

where c is the number of clusters, δ_i defines the average distance between each object in Cluster i and its centroid (δ_j follows the same process), and $d(c_i, c_j)$ defines the distance between the centroids of the clusters. Small values of the index stand for compact clusters whose centroids are well separated from each other. Thus, the partition that minimizes the Davies Bouldin index is considered optimal.

The **Silhouette coefficient** [29] measures how close each centroid in the cluster is to each other object in the neighboring clusters. Thus, for each object i, compute the average proximity a_i between i and all other objects in the cluster to which i belongs. Then, for the remaining Clusters c, calculate the average proximity $f(i, c)$ to all objects in c. The smallest value of $f(i, c)$ is defined as $b_i = min_c f(i, c)$. The coefficient is defined as follows:

$$s(i) = \frac{b_i - a_i}{max\{a_i, b_i\}}$$

where $SC = \frac{1}{c} \sum_{i=1}^{c} s(i)$ computes the coefficient for the complete partition.

IV. PROPOSED METHOD

The proposed approach of this study (figure 1) is performed in two general stages; in the first stage, a genetic programming (GP) algorithm generates aptitude functions, and in the second stage, a genetic algorithm (GA) evolves clusters of sentences to produce summaries based on the aptitude

function built by the GP (see Algorithm 1). These two steps are detailed as follows.

First, the GP creates a set of functions, considering the selected internal validation indices (Davies Bouldin, Dunn and Silhouette) as operands. Then, each expression built by the GP is taken as a fitness function in the evolutionary clustering approach where the summaries are created. Therefore, the fitness function should evaluate the clustering that makes up the good-quality summaries as detailed below.

In the clustering representation, each document is divided into sentences. To encode a document as an individual of the GA (genotype), it is represented by a binary vector where each gen represents a sentence. The active genes from this codification are considered the centroid sentences in the clustering; therefore, a vector with n active genes represents a clustering of n groups of sentences. Then, to measure the quality of solutions, the clusters are evaluated using the fitness function generated by the GP. Finally, the active genes of the best solution indicate which sentences will take part in the summary. This process is repeated through each document in the collection.

To ensure that the summaries created are of good quality, the Rouge measure is used. Rouge [30] is a measure to automatically determine the quality of a summary by comparing it to ideal summaries written by humans. This measure has different versions that count the number of overlapping units such as n-grams (Rouge-1 and Rouge-2), word sequences (Rouge-L), and word pairs (Rouge- SU) between the computer-generated summary to be evaluated and the ideal summaries.

To establish the correlation between the quality of the summaries and the grouping, our proposed method pursues two objectives: to maximize the fitness function created by the GP (which measures the quality of the clusters) and Rouge measure (which measures the quality of the summaries). Therefore, we attempt to establish the correlation between the Rouge measure and the function created by the GP. Specifically, the GP algorithm invokes the GA algorithm, the latter creates the summaries and returns their best fitness value. Finally, the GP calculates the Rouge measure on the generated summaries and seeks to optimize both the best fitness value obtained by the GA and the value obtained by Rouge (i.e., the GP has a multiobjective fitness function).

To build the GP functions, 10% of the DUC02 dataset is used; thus, the rest of DUC02 and the CNN/Daily mail dataset are used to test the efficiency of each generated function.

Algorithm 1 Proposed System Pseudocode

Input: Source documents
Output: Summaries
Initialization:
Randomly create the initial population $P(0)$
LOOP Process
for $t = 1$ to *NumberOfGenerations* **do**
  $P'(t) = \emptyset$
  Evaluate each function in $P(t)$ with the GA
  Copy the best individual from $P(t)$ to $P'(t)$
  while $P'(t)$ is not filled **do**
    if insertion probability $P_i < rand[0-1]$ **then**
      Select an individual i based on roulette
      Insert i in $P'(t)$
    end if
    if crossover probability $P_c < rand[0-1]$ **then**
      Select two individuals (i, j) based on tournament
      $i = crossover(i, j)$
      Insert i into $P'(t)$
    end if
    if mutation probability $P_m < rand[0-1]$ **then**
      Randomly select an individual i
      $i = mutate(i)$
      Insert i into $P'(t)$
    end if
  end while
end for
Select the best individual *gpBest* from $p'(t)$
Use *gpBest* as GA fitness function and create summaries
return Summaries

A. POSSIBLE VARIATIONS AND IMPROVEMENTS OF THE CURRENT FRAMEWORK

The model proposed in this work would be able to analyze more mapping methods that provide a vector representation of texts at different linguistic levels. For example, Large Language Models (LLM) can be used as a mapping method (Section III-A) for a possible improvement of the proposed framework. In accordance with the above, our method can benefit from the advantages of LLMs, and consequently also acquire the disadvantages. For example, LLMs like BERT or GPT are often resource intensive to train and generate text representations.

In this sense, on the one hand, the model presented in this study is an evolutionary-based algorithm; and according to Neumann [31] and Nopiah et al. [32], the time complexity of this type of algorithms (in general cases) is $O(n)$ or $O(n \log n)$. Figure 2 shows the execution time of our proposed summary system with respect to the percentage of data used; our summary system generates a model on average over three hours (when all data are considered) and the resulting plot shows linear complexity. On the other hand, the attention-based systems have a time complexity of $O(n^2)$ [33] (per attention layer) which implies a major time of processing in exchange for a more accurate language model. However, despite the large amount of data to train these models, traditional word embeddings methods are still better at some tasks [34].

It is worth noting that we consider some improvements to the current framework as future work, such as the use of LLMs as mapping methods. An important advantage of LLMs over other context-based methods, such as Doc2Vec, is that the former can evaluate the context of a word bidirectionally. This feature produces more accurate representations of the text. Also, unlike recurrent neural networks (RNNs),

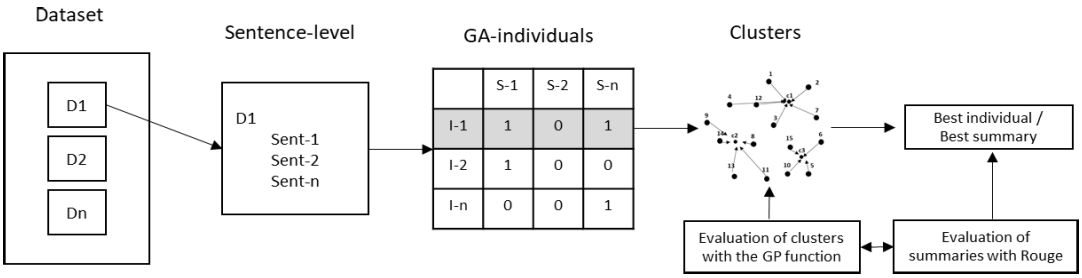

FIGURE 1. Block diagram of the proposed approach.

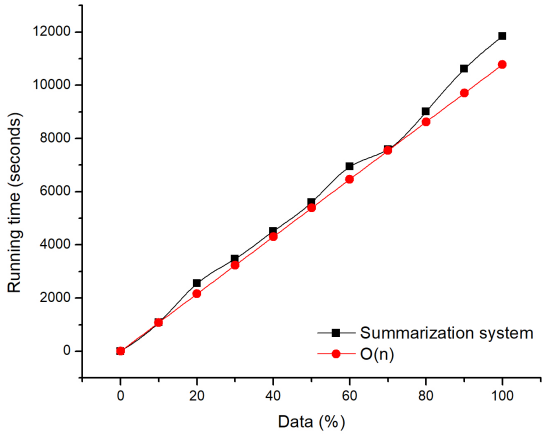

FIGURE 2. Execution time of the proposed system.

LLMs have attentional mechanisms that allow parallel data processing [33]. Therefore, in our proposal, LMMs can provide more precise semantic information of the sentences in a document and thus improve the process of selecting key information.

Another possible improvement is the expansion of the set of terminals of the GP. That is, the objective functions generated by the GP, in the current configuration, only consider internal information of the clusters (external separability and internal homogeneity), but they can be improved by adding external information such as: title similarity, redundancy and length of sentences, coverage, etc.

Despite the improvements that can be made to our method, as can be seen in Table 1, the results of this work outperform the established baselines [35] by the basic systems for the analyzed datasets. SumaRuNNer [36] and SUMO [37] are RNN and transformer-based systems, respectively; and Lead-3 is the selection of the first three sentences of the documents. It is worth noting that some attention-based systems baselines, such as SUMO, show similar performance with our proposed model.

B. CLUSTERING REPRESENTATION

The partitional clustering strategy is used in this study to organize the sentences of the documents. To that end, documents are divided into sentences, and each sentence is converted to a numeric vector using feature generation methods (see Section III-A).

TABLE 1. Comparison of the proposed system with baseline systems.

Method	Rouge-1	Rouge-2	Rouge-L
Duc02			
This work	48.6	23.2	-
Lead-3	43.6	21.2	40.2
DUC 2002 Best	48.0	22.8	-
CNN-Daily Mail			
This work	41.6	17.9	36.8
SummaRuNNer	39.6	16.2	35.3
SUMO	41.0	18.4	37.2

In the next step, we built a Euclidean distance matrix M for each document D^n in the collection, where n represents the number of sentences in D. That is, the Euclidean distance is obtained between each sentence s_i and $s_j \in D^n$. As a result, M is a bidimentional matrix of n x n.

Thus, given a set of objects $\Omega = X_i \in R^d; i = 1, \ldots, N$, a partional clustering has the goal of organizing the objects in K Clusters $K = C_1, C_2, \ldots, C_K$, while a criterion function is maximized or minimized.

Finally, groups are generated by selecting the objects closer to each centroid (group representative) and following the next rules:

$$1)\ C_i \neq \emptyset, i = 1, \ldots, K$$
$$2)\ \bigcup_{i=1}^{K} Ci = \Omega$$
$$3)\ C_i \bigcap C_j = \emptyset, i, j = 1, \ldots, K\ y\ i \neq j$$

In addition, the proposed method uses a genetic algorithm to optimize the clustering process that becomes a combinatorial problem.

C. GENETIC ALGORITHM CODIFICATION

The genetic algorithm (GA) is an optimization method based on the theory of natural selection, where the survival and reproduction of individuals depends on their genetic characteristics. GA randomly creates a population that evolves by g generations with the aim of improving the individuals (solutions) while applying crossover and mutation operators.

Each individual or chromosome in the population is a possible solution for some specific problem (phenotype), commonly represented by a binary vector (genotype). In this study, each individual is the representation of a document. Each position of the vector, namely, gen, represents a

sentence of the document, and its value {1, 0} indicates if the sentence is taken as the centroid. In turn, centroids represent the key sentences that make up the summary.

The operators of the GA used in this study are the crossover in two points and the standard mutation, and the selection method is roulette. On the one hand, the crossover and mutation rates are set to 0.7 and 0.1, respectively; on the other hand, the population evolves over 50 generations.

The fitness function is a heuristic function that assigns a value to each individual and indicates the quality of the solutions. Instead of other works where this function is set manually, we propose to generate it automatically with genetic programming.

D. THE SEARCH OF A FITNESS FUNCTION WITH GENETIC PROGRAMMING

Genetic programming (GP) is a technique that evolves computer programs. Similar to genetic algorithms (GA), GP has a series of operators to evolve the population, such as crossover and mutation. Instead, the representation of individuals in GP is through a tree structure. This structure allows us to represent a mathematical expression where each nonterminal node has an operator function and every terminal node has an operand.

In this study, a GP algorithm is performed to build an objective function to guide the search for good-quality summaries. To that end, the internal validation indices were selected to be part of the operands; that is, the Davies Bouldin, Dunn and Silhouette index could be located at the terminal nodes of the tree. It is worth noting that these indices have proven to be correlated with the quality of abstracts [9]. Additionally, a random constant in the range of [0,1] could also be added to the terminal nodes to provide a weighting of terms. On the other hand, the basic operators could be located in the nonterminal nodes.

The Rouge measure is defined as the fitness function of the GP because it is widely used to assess the quality of summaries. Thus, the better the quality of the summaries determined by Rouge, the better the ability of the GP-generated function to detect good-quality summaries.

The advantage of the generated function is that it evaluates the summaries considering only the internal information, that is, it evaluates the clustering representation in which the key ideas are selected based on the configuration of each group.

In the GP system, we define the following parameters: population of 100 individuals, 60% crossover rate and 10% mutation rate, and the population evolved over 500 generations.

E. DATASETS

To validate the proposed approach of this study, two standard collections are analyzed: DUC02 and CNN/Daily Mail. Table 2 details the basic statistics of the source documents and abstracts.

The DUC02 dataset was selected because every news item was written by two expert humans; this fact works as a reference point between automatic and manual summaries. In addition, the evolutionary process can generate objective functions by learning the process of generating human abstracts.

In addition, the widely used CNN/Daily Mail dataset [38] was selected to measure the performance of our proposal relative to other current standards addressing both supervised and unsupervised methods.

TABLE 2. Dataset statistics.

Corpus	Source docs.		Summaries	
	docs.	Avg. length	docs.	Avg. length
DUC02	567	649	1,112	114
CNN/Daily Mail	11,490	778	11,490	58

TABLE 3. Results of DUC02.

Metric	Rouge-1	Rouge-2	Rouge-SU
TF–IDF	0.46911	0.21346	0.23240
OHE	0.46791	0.21118	0.23081
LDA	0.46212	0.20757	0.22697
Doc2Vec	0.45352	0.20075	0.22048
LDA+TF–IDF	0.47465	0.22000	0.23829
Doc2Vec+TF–IDF	0.47208	0.21577	0.23511
LDA+OHE	0.47466	0.22018	0.23744
Doc2Vec+OHE	0.47646	0.22032	0.23891
LDA+Doc2Vec	0.47511	0.22091	0.23945
LDA+Doc2Vec+TF–IDF	**0.48628**	**0.23288**	**0.24900**

TABLE 4. Results of CNN.

Metric	Rouge-1	Rouge-2	Rouge-SU
TF–IDF	0.36379	0.13839	0.15066
OHE	0.35915	0.13551	0.14866
LDA	0.36014	0.13552	0.14847
Doc2Vec	0.35944	0.13538	0.14830
LDA+TF–IDF	0.40587	0.17185	0.17715
Doc2Vec+TF–IDF	0.40587	0.17182	0.17712
LDA+OHE	0.35923	0.13487	0.14813
Doc2Vec+OHE	0.35811	0.13444	0.14763
LDA+Doc2Vec	0.40555	0.17154	0.17694
LDA+Doc2Vec+TF–IDF	**0.41643**	**0.17930**	**0.18466**

V. RESULTS AND DISCUSSION

In this study, internal validation indices are used to provide information on the quality of the clusters and, in turn, the quality of the summaries. The inference is that each index may have some degree of relevance in the ATS task. In this sense, the main goal is to create an objective function that only considers the internal information of the documents to create good-quality summaries. To that end, a GP algorithm is performed to select and bind the correct components and weights (operands) and the correct operators to automatically generate a fitness function. As a result, the decision to add some operand or operator to the objective function is made automatically. It should be noted that in previous works, the fitness functions were adjusted manually according to the author's intuition [39], [40], [41]; this makes it very time-consuming for the authors to analyze various functions.

As detailed above (Section IV-D), each function generated by the GP is sent to the genetic algorithm that creates

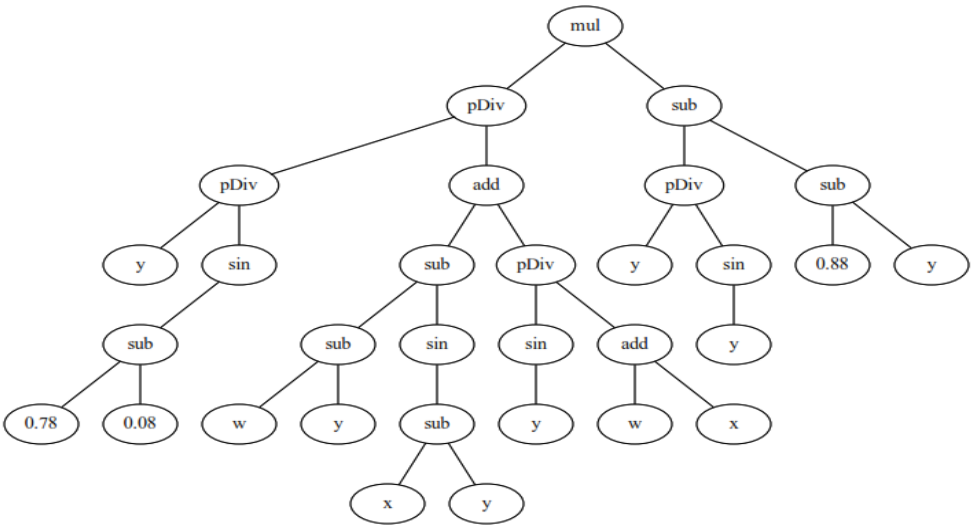

FIGURE 3. An automatically created function to find good-quality summaries where x, y and w represents Dunn, Davies Bouldin and Silohuette index, respectively.

TABLE 5. Comparison of the results of the proposed approach with those of other approaches for the DUC02 dataset. The number in brackets represents a ranking among the proposed systems.

Approach	Rouge-1	Rouge-2	Rouge-SU	Average
This work	**0.48628(1)**	**0.23288(1)**	**0.24900**	**0.35958**
FEOM [42]	0.46575(6)	0.12490(4)	-	0.29532
GA approach [43]	0.48270(4)	-	-	0.24135
UnifiedRank [44]	0.48478(2)	0.21462(3)	-	0.34970
SFR [41]	0.48423(3)	0.22471(2)		0.35447
COSUM [8]	0.46694(5)	0.12368(5)	-	0.29531
NetSum [45]	0.44963(7)	0.11167(6)	-	0.28065
CRF [46]	0.44006(8)	0.10924(7)	-	0.27465

TABLE 6. Comparison of the proposed approach with respect to other approaches on CNN/Daily mail dataset.

Approach	ROUGE-1	ROUGE-2	ROUGE-L	Average
This work	41.6	17.9	36.8	32.1
Narayan et al. [47]	40.3	17.7	36.6	31.5
See et al. [48]	39.5	17.2	36.3	31.0
Zheng et al. [49]	**54.7**	**30.4**	**50.8**	**45.3**

and selects the best summary. Once all the summaries are generated, the GP fitness function calculates the quality of the collection using Rouge and selects the best generated function.

Figure 3 shows an automatically generated function for the text summary task where the variables x, y and w represent the Dunn, Davies Bouldin and Silhouette index, respectively. Specifically, this figure shows the graphical representation of a binary tree generated by the GP; this tree is made up of binary or unary operators (parent nodes), and operands (leaf nodes). Therefore, the GP is in charge of adjusting the tree structure by applying genetic operators (i.e., crossover and mutation) to find new solutions. In this sense, the GP is capable of omitting operators and operands if they are not relevant for the solution; however, all indices were added to the best solutions found.

It should be noted that the fitness function construction process is carried out considering only 10% of the DUC02 collection. The resultant function is then tested on the remaining DUC02 documents and the CNN/Daily mail dataset.

In Tables 3 and 4, the results on DUC02 are shown. Various feature generation methods, which obtain information from texts at the lexical and semantic levels, are compared and combined to achieve the best performance. As seen, the combination of LDA, Doc2Vec and TF-IDF obtained the best result for both corpora. This suggests that topic information, context-based semantics, and word relevance is the combination that provides the best clustering representation for selecting key sentences. Therefore, the generation of good quality summaries, within the framework of our proposal, depends on: 1) the topic: the topics that make up the sentence; 2) the semantic context: how similar are the sentences in a semantic space; and 3) the relevance of sentences to a document in a collection.

Tables 5 and 6 show a comparison between the approach proposed in this study and other studies that proposed supervised and unsupervised methods. Our proposed approach achieves the best performance on the DUC02 collection for the Rouge-1, Rouge-2 and Rouge-SU measures. It is worth noting that our system outperforms the results with respect to studies that focus on clustering [42] or evolutionary [43] methods, where the objective function is created based on the author's intuition. Additionally, this study shows competitive performance for the CNN/Daily mail collection compared to supervised methods based on neural networks.

Table 7 shows a couple of examples of automatically created summaries from the CNN/Daily mail dataset and its respective human-made sums built from the same document. As seen, the summary made by humans is more compact since this feature is proper of the abstractive summaries; however,

TABLE 7. Example of human-made and automatically obtained summaries.

Human-made summary	Automatically generated summary
Anderson Silva met with Brazilian taekwondo officials on Wednesday. Silva is currently suspended by UFC after failing drug tests. However, the former UFC champion will fight for Olympics taekwondo spot.	The announcement was made on Wednesday after a meeting with Brazilian taekwondo officials. The former UFC champion said he is 'trying to give back to the sport' in which he began his career. Silva is a taekwondo ambassador and a black belt in the sport. The 40-year-old Brazilian posted a photo of himself via his Twitter page practicing taekwondo last week.
Stephen Ward in contention for Burnley after overcoming ankle injury. Matt Taylor close to return having been out since August. Tottenham Hotspur without goalkeeper Hugo Lloris through knee injury. Danny Rose and Roberto Soldado also fitness concerns for Spurs.	Here is all the information you need for Burnley's home clash with Tottenham... Midfielder Matt Taylor, who has not featured since August, is also nearing a return after resuming training following his recovery from achilles surgery. The Clarets have no fresh injury concerns following the international break but Dean Marney and Kevin Long (both cruciate ligament) remain on the long-term absentee list. Provisional squad: Heaton, Gilks, Mee, Duff, Shackell, Keane, Reid, Ward, Trippier, Barnes, Wallace, Arfield, Jones, Boyd, Kightly, Vokes, Jutkiewicz, Ings, Sordell. Tottenham goalkeeper Hugo Lloris will miss Sunday's Barclays Premier League match at Burnley because of a knee injury.

most ideas of this summary can be inferred by the summary obtained automatically.

VI. CONCLUSION

This study proposes the automatic generation of an objective function for the unsupervised text summary task. A combination of a genetic algorithm and genetic programming was performed to build a maximization function that maintains a close correlation with the quality of the summaries (i.e., the higher the value of the objective function, the better the quality of summaries generated).

According to the results shown in this work (Section V), the combination of lexical and semantic information (LDA+Doc2Vec+TF-IDF) achieves the best results in detecting key ideas to form a summary. This combination of features includes information about the relevance of words (TF-IDF), the topics involved (LDA) and the contexts around a window of words (Doc2Vec).

The resulting objective function for the extractive ATS task considers only the internal information since it is formed from internal validation indices; that is, the created function only considers the quality of clustering. This fact allows the function to be applied without external information such as the true labels of each document.

The Rouge measure was used to correlate the fitness function created by the GP with the quality of summaries; this correlation allowed this study to automatically create objective functions and yield competitive results for the DUC02 and CNN/Daily mail datasets.

ACKNOWLEDGMENT

The authors would like to thank the Mexican Government (Cátedras CONACYT, SNI, Universidad Autónoma del Estado de México) for its support.

REFERENCES

[1] A. Elsaid, A. Mohammed, L. F. Ibrahim, and M. M. Sakre, "A comprehensive review of Arabic text summarization," *IEEE Access*, vol. 10, pp. 38012–38030, 2022.

[2] A. A. Syed, F. L. Gaol, and T. Matsuo, "A survey of the state-of-the-art models in neural abstractive text summarization," *IEEE Access*, vol. 9, pp. 13248–13265, 2021.

[3] J. N. Madhuri and R. G. Kumar, "Extractive text summarization using sentence ranking," in *Proc. Int. Conf. Data Sci. Commun. (IconDSC)*, Mar. 2019, pp. 1–3.

[4] R. Elbarougy, G. Behery, and A. El Khatib, "Extractive Arabic text summarization using modified PageRank algorithm," *Egyptian Informat. J.*, vol. 21, no. 2, pp. 73–81, Jul. 2020.

[5] S. G. Jindal and A. Kaur, "Automatic keyword and sentence-based text summarization for software bug reports," *IEEE Access*, vol. 8, pp. 65352–65370, 2020.

[6] H. Gupta and M. Patel, "Method of text summarization using LSA and sentence based topic modelling with BERT," in *Proc. Int. Conf. Artif. Intell. Smart Syst. (ICAIS)*, Mar. 2021, pp. 511–517.

[7] S. Ghodratnama, A. Beheshti, M. Zakershahrak, and F. Sobhanmanesh, "Extractive document summarization based on dynamic feature space mapping," *IEEE Access*, vol. 8, pp. 139084–139095, 2020.

[8] R. M. Alguliyev, R. M. Aliguliyev, N. R. Isazade, A. Abdi, and N. Idris, "COSUM: Text summarization based on clustering and optimization," *Exp. Syst.*, vol. 36, no. 1, Feb. 2019, Art. no. e12340.

[9] N. H. Castañeda, R. A. G. Hernández, Y. Ledeneva, and Á. H. Castañeda, "Evolutionary automatic text summarization using cluster validation indexes," *Computación Y Sistemas*, vol. 24, no. 2, pp. 583–595, Jun. 2020.

[10] M. G. Smith and L. Bull, "Genetic programming with a genetic algorithm for feature construction and selection," *Genetic Program. Evolvable Mach.*, vol. 6, no. 3, pp. 265–281, Sep. 2005.

[11] R. Alqaisi, W. Ghanem, and A. Qaroush, "Extractive multi-document Arabic text summarization using evolutionary multi-objective optimization with K-medoid clustering," *IEEE Access*, vol. 8, pp. 228206–228224, 2020.

[12] T. Uçkan and A. Karci, "Extractive multi-document text summarization based on graph independent sets," *Egyptian Informat. J.*, vol. 21, no. 3, pp. 145–157, Sep. 2020.

[13] L. Dong, M. N. Satpute, W. Wu, and D. Du, "Two-phase multidocument summarization through content-attention-based subtopic detection," *IEEE Trans. Computat. Social Syst.*, vol. 8, no. 6, pp. 1379–1392, Dec. 2021.

[14] A. Sahu and S. G. Sanjeevi, "Better fine-tuning with extracted important sentences for abstractive summarization," in *Proc. Int. Conf. Commun., Control Inf. Sci. (ICCISc)*, Jun. 2021, pp. 11328–11339.

[15] M. T. Nayeem, T. A. Fuad, and Y. Chali, "Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 1191–1204.

[16] N. Alami, M. Meknassi, N. En-nahnahi, Y. El Adlouni, and O. Ammor, "Unsupervised neural networks for automatic Arabic text summarization using document clustering and topic modeling," *Exp. Syst. Appl.*, vol. 172, Jun. 2021, Art. no. 114652.

[17] M. M. Haider, Md. A. Hossin, H. R. Mahi, and H. Arif, "Automatic text summarization using Gensim Word2Vec and K-means clustering algorithm," in *Proc. IEEE Region Symp. (TENSYMP)*, Jun. 2020, pp. 283–286.

[18] R. Rautray and R. C. Balabantaray, "An evolutionary framework for multi document summarization using cuckoo search approach: MDSCSA," *Appl. Comput. Informat.*, vol. 14, no. 2, pp. 134–144, Jul. 2018.

[19] J. M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Pérez, "Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach," *Knowl.-Based Syst.*, vol. 159, pp. 1–8, Nov. 2018.
[20] I. Arroyo-Fernández, C.-F. Méndez-Cruz, G. Sierra, J.-M. Torres-Moreno, and G. Sidorov, "Unsupervised sentence representations as word information series: Revisiting TF–IDF," *Comput. Speech Lang.*, vol. 56, pp. 107–129, Jul. 2019.
[21] M. K. Dahouda and I. Joe, "A deep-learned embedding technique for categorical features encoding," *IEEE Access*, vol. 9, pp. 114381–114391, 2021.
[22] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
[23] T. Mikolov, L. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2, Dec. 2013, pp. 3111–3119.
[24] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. 31st Int. Conf. Mach. Learn.*, vol. 32, Jan. 2014, pp. 1188–1196.
[25] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2010, pp. 911–916.
[26] E. Rendón, I. Abundez, A. Arizmendi, and E. M. Quiroz, "Internal versus external cluster validation indexes," *Int. J. Comput. Commun.*, vol. 5, no. 1, pp. 27–34, 2011.
[27] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *J. Cybern.*, vol. 4, no. 1, pp. 95–104, Jan. 1974.
[28] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.
[29] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.
[30] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81.
[31] F. Neumann, "Computational complexity analysis of multi-objective genetic programming," in *Proc. 14th Annu. Conf. Genetic Evol. Comput.*, Jul. 2012, pp. 799–806.
[32] Z. Nopiah, M. Khairir, S. Abdullah, M. Baharin, and A. Arifin, "Time complexity analysis of the genetic algorithm clustering method," in *Proc. 9th WSEAS Int. Conf. Signal Process., Robot. Autom.*, vol. 10, 2010, pp. 171–176.
[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[34] M. Pranjic, M. Robnik Sikonja, and S. Pollak, "An evaluation of BERT and Doc2Vec model on the IPTC subject codes prediction dataset," in *Proc. 24th Int. Multiconference*, D. Mladenic and M. Grobelnik, Eds. 2021, pp. 25–28.
[35] M. Gambhir and V. Gupta, "Deep learning-based extractive text summarization with word-level attention mechanism," *Multimedia Tools Appl.*, vol. 81, no. 15, pp. 20829–20852, Jun. 2022.
[36] R. Nallapati, F. Zhai, and B. Zhou, "SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, 2017, pp. 3075–3081.
[37] Y. Liu, I. Titov, and M. Lapata, "Single document summarization as tree induction," in *Proc. Conf. North*, 2019, pp. 1745–1755.
[38] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," in *Proc. Adv. Neural Inf. Process. Syst.* Cambridge, MA, USA: MIT Press, 2015, pp. 1693–1701.
[39] I. Tanfouri, G. Tlik, and F. Jarray, "An automatic Arabic text summarization system based on genetic algorithms," *Proc. Comput. Sci.*, vol. 189, pp. 195–202, Jan. 2021.
[40] Q. A. Al-Radaideh and D. Q. Bataineh, "A hybrid approach for Arabic text summarization using domain knowledge and genetic algorithms," *Cognit. Comput.*, vol. 10, no. 4, pp. 651–669, Aug. 2018.
[41] E. Vázquez, R. A. Garcia-Hernández, and Y. Ledeneva, "Sentence features relevance for extractive text summarization using genetic algorithms," *J. Intell. Fuzzy Syst.*, vol. 35, no. 1, pp. 353–365, Jul. 2018.
[42] W. Song, L. C. Choi, S. C. Park, and X. F. Ding, "Fuzzy evolutionary optimization modeling and its applications to unsupervised categorization and extractive summarization," *Exp. Syst. Appl.*, vol. 38, no. 8, pp. 9112–9121, Aug. 2011.
[43] R. A. Garcia-Hernández and Y. Ledeneva, "Single extractive text summarization based on a genetic algorithm," in *Proc. Mex. Conf. Pattern Recognit.* Cham, Switzerland: Springer, 2013, pp. 374–383.
[44] X. Wan, "Towards a unified approach to simultaneous single-document and multi-document summarizations," in *Proc. 23rd Int. Conf. Comput. Linguistics*, 2010, pp. 1137–1145.
[45] K. Svore, L. Vanderwende, and C. Burges, "Enhancing single-document summarization by combining ranknet and third-party sources," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn. (EMNLP-CoNLL)*, 2007, pp. 448–457.
[46] D. Shen, J.-T. Sun, H. Li, Q. Yang, and Z. Chen, "Document summarization using conditional random fields," in *Proc. IJCAI*, vol. 7. Burlington, MA, USA: Morgan Kaufmann, 2007, pp. 2862–2867.
[47] S. Narayan, S. B. Cohen, and M. Lapata, "Ranking sentences for extractive summarization with reinforcement learning," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2018, pp. 1747–1759.
[48] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 1073–1083.
[49] H. Zheng and M. Lapata, "Sentence centrality revisited for unsupervised summarization," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6236–6247.

ÁNGEL HERNÁNDEZ-CASTAÑEDA received the M.Sc. and Ph.D. degrees (Hons.) in computer science from the Centre for Computing Research (CIC), National Polytechnic Institute (IPN), in 2013 and 2017, respectively. He is currently a Research Professor with the Autonomous University of Mexico State and a member of the National System of Researchers (SNI) of Mexico. His research interests include natural language processing, data mining, and pattern recognition.

RENÉ ARNULFO GARCÍA-HERNÁNDEZ received the B.E. degree in computer systems engineering from the Toluca Institute of Technology, Mexico, in 2001, the M.S. degree in computer science from the National Centre of Research and Technology Development (Cenidet), Mexico, in 2003, and the Ph.D. degree in computer science from the National Institute of Astrophysics, Optics, and Electronics (INAOE), Mexico, in 2017. He is currently a Full Research Professor with the School of Software Engineering and the Postgraduate School, Autonomous University of the State of Mexico (UAEM). He has authored over 70 articles in top journals and international conferences, and three books. He is an adviser of 34 theses. He is recognized as a second-level national researcher, a higher level. His research interests include pattern recognition, evolutionary computation, text mining, and natural language processing. He is a member of the Mexican Association for the Natural Language Processing.

YULIA LEDENEVA received the B.Sc. and M.Sc. degrees in engineering from the Peoples' Friendship University of Russia, in 2002 and 2004, respectively, the M.Sc. degree in computer science from the National Institute for Astrophysics, Optics, and Electronics, Mexico, in 2006, and the Ph.D. degree in computer science from the Centre for Computing Research, National Polytechnic Institute (IPN), Mexico. She is currently a Research Professor with the Autonomous University of the State of Mexico and a member of the National System of Researchers (SNI) of Mexico. She is the author of more than 70 publications. Her research interests include computational linguistics, natural language processing, text mining, graph, and genetic algorithms. She received the Presea Lázaro Cárdenas from the hands of the President of Mexico, in 2009.