ISSN 2007-9737

Evolutionary Automatic Text Summarization using Cluster Validation Indexes

Néstor Hernández Castañeda[1], René Arnulfo García Hernández[1],
Yulia Ledeneva[1], Ángel Hernández Castañeda[1,2]

[1] Autonomous University of the State of Mexico,
Mexico

[2] Cátedras CONACYT,
Mexico

nestor.hecada@gmail.com, renearnulfo@hotmail.com

Abstract. The main problem for generating an extractive automatic text summary (EATS) is to detect the key themes of a text. For this task, unsupervised approaches cluster the sentences of the original text to find the key sentences that take part in an automatic summary. The quality of an automatic summary is evaluated using similarity metrics with human-made summaries. However, the relationship between the quality of the human-made summaries and the internal quality of the clustering is unclear. First, this paper proposes a comparison of the correlation of the quality of a human-made summary to the internal quality of the clustering validation index for finding the best correlation with a clustering validation index. Second, in this paper, an evolutionary method based on the best above internal clustering validation index for an automatic text summarization task is proposed. Our proposed unsupervised method for EATS has the advantage of not requiring information regarding the specific classes or themes of a text, and is therefore domain- and language-independent. The high results obtained by our method, using the most-competitive standard collection for EATS, prove that our method maintains a high correlation with human-made summaries, meeting the specific features of the groups, for example, compaction, separation, distribution, and density.

Keywords. Automatic text summarization, cluster validation indexes, evolutionary method, extractive summaries.

1 Introduction

Large amounts of digital information are being generated and shared on the Internet every day. As a result, search engines are used to find ordinary information about a specific topic. However, people must read the full documents before deciding if the information is proper. Considering that a search returns thousands of documents, reading all of them would be impossible.

One way to solve this problem is using a system capable of obtaining the key ideas of the documents to return a volume-reduced document [10]. Thus, the new document should contain only the relevant information, avoiding the inclusion of secondary ideas. This allows the user to have a general idea about a topic by reading an automatically generated summary [18].

On the one hand, an abstractive summary is generated based on a more in-depth analysis in which the output summary can include new sentences that are not contained in the original text. On the other hand, an extractive summary uses text units of the original text to generate the summary, for example, words, sentence segments, or sentences. Therefore, an extractive approach provides summaries with the information available in the original texts. In both cases, the main problem in generating an automatic text summary is how to detect the key themes of a text.

doi: 10.13053/CyS-24-2-3392

For an extractive automatic text summarization (EATS) [24],[28] unsupervised approaches are based on clustering the sentences in the original text to find the key sentences that are part of an automatic summary. It is worth mentioning that the quality of an automatic summary is evaluated using similarity metrics with human-made summaries. According to different studies on EATS, it has been suggested that the use of clustering algorithms helps in detecting the key themes of a document. However, the relationship between the quality of the human-made summaries and the internal quality of the clustering remains unclear.

In this study, we first compare the correlation of the quality of human-made summaries with the internal quality of the clustering validation index to find the best correlation with a clustering validation index. To measure the impact of the cluster validation indexes on the quality of the summaries, three baselines are evaluated [17, 21], namely, top-line, first-line, and random-line summaries.

A top-line summary consists of the best-quality summary according to the summaries written by a human. A first-line summary consists of selecting the first sentences of the document with at least n-words. The first-line heuristic occurs because in some types of domains, such as in the news domain, the main content is located at the beginning of the document. A first-line summary is a hard-to-beat baseline for EATS systems. A random-line summary consists of selecting random sentences with at least n-words. We infer that the poor-quality summaries (random-line summaries) maintain some correlation with poor-quality clusters, and vice-versa.

In this study, we show the correlation between the quality of the baselines of human-made summaries and the quality of clustering with solid internal validation indexes, namely, Dunn, Davies Bouldin, and Silhouette. Our research findings demonstrate that these indexes are correlated with a high-quality summary generation.

In addition, we propose an evolutionary EATS method that applies as a fitness function the best internal clustering validation index described above.

Our proposed unsupervised EATS method has the advantage of not requiring information about specific classes or themes of a text; therefore, it is domain- and language-independent. The high results obtained by our method, based on the most-competitive standard collection for EATS, prove that our method maintains a good correlation with human-made summaries that meet the specific features of the groups, for example, compaction, separation, distribution, and density.

Clusters generated by the proposed approach indicate the summaries, where the objects in a group may be highlighted ideas from a particular text. In our search for related studies, no information regarding the correlation between the quality of the clustering and the quality of the human summaries was available; different validation indexes were applied to validate the generated clusters. To achieve this, we compared the results reported using validation indexes to those obtained through an external quality measure, namely, Rouge. The Rouge measure automatically compares a summary generated with one or more reference summaries. For this reason, the most-competitive EATS DUC02 dataset was used because it contains summaries written by human users. This allows a more faithful comparison of the system performance with a human reference.

2 Related Work

Multiple strategies for automatically generating summaries and processing large numbers of documents in an efficient manner have been developed. Thus, the general process of an EATS task is the identification of relevant information from a text to build a new summarized document.

In Maña López [16], depending on the linguistic level, the techniques of automatic text summarization are classified as extractive and abstractive. On the one hand, extractive techniques are based on a superficial analysis of the text when considering only a syntactic level where the output summary uses text units from the original text, for example, words, sentence segments, or sentences. The sentence is considered the unit that represents an idea with a complete meaning of the author.

In most of the standard datasets, an EATS task is at the sentence level.

On the other hand, abstractive techniques consider a deeper analysis; for instance, they include a semantic analysis in which the output summary may include new sentences not contained within the original text. In this sense, abstractive summaries have the risk of reformulating sentences with an altered interpretation different from that of the original author.

Most studies have focused on extractive summaries by considering key sentences and their position in the text [1], by measuring words frequencies [6], or by assigning importance levels to the sentences [31], among others.

At the lexical-level, n-grams are frequently used to generate text models. For instance, in Ledeneva [13], sequences of n-grams are extracted from the text by using a model of maximal frequent sequences. However, Bando et al. [5] use n-grams to build paragraphs with the most representative terms in the document.

The extracted features from the documents are evaluated through supervised and unsupervised methods to create models that allow detecting the main components of the key ideas.

Supervised approaches have been widely explored [33, 4] to generate extractive and abstractive summaries. In Neto et al.'s approach [19], each sentence from a document is labeled as ”positive” if the sentence belongs to a summary, whereas the remainder of sentences are labeled as ”negative.” The authors then generate a variety of features by applying statistics-and linguistics-oriented procedures, for instance, the sentence position, sentence length, similarity to the title, and proximities among the centroids and sentences. The sentences are classified using two algorithms: a C4.5 decision-tree and naïve Bayes.

Similar to Neto et al.'s method [19], Fattah and Ren [9] proposed a trainable summarizer by extracting a variety of features. However, the authors consider the relevance of the features by assigning a weight to them. This assignation is given by a genetic algorithm [21] and a regression model [29].

These models are exploited to obtain an appropriate set of weights by processing 50 manually summarized English documents. The results have reported a precision of up to 44.94.

The main problem of supervised approaches lies in the fact that a set of labeled data is needed. In addition, the domain of the training samples is commonly insufficiently general to process new multi-domain samples.

Recently, unsupervised machine learning approaches have been exploited using clustering algorithms [11] to group sentences based on the structure and frequency of the words. The most representative sentences of the formed groups are used to generate a summary.

In clustering approaches, to obtain high-quality summaries, groups of sentences need to be evaluated. There are two validation methods for evaluating the quality of the partitions: internal measures and external measures. The former does not consider any external information about the dataset classes; the latter requires the class labels to be applied. Different authors have compared internal and external quality measures of clustering validation. Experiments were conducted to prove which of these approaches can evaluate the optimal number of groups from a dataset. Different quality measures are tested based on the building of groups of clustering algorithms. Studies have proven that internal measures perform better than external measures.

Most methods focused on an unsupervised approach have used external quality measures to validate the model performance (e.g., F-measure); however, cluster validation indexes (internal quality measures) have been little explored in EATS tasks.

Soto et al. [18] developed an automatic summarization system by using unsupervised learning. The authors used three text models to build numeric vectors: bag-of-words, n-grams, and maximal frequent sequences. In turn, these methods are mapped to numeric vectors by applying different methods of weighing terms, namely, Boolean plus standard term frequency (tf) and inverse document frequency (idf) plus tf-idf. The resulting vectors are grouped using a K-means algorithm and the final clusters are evaluated through an external measure (f-score).

Research findings have shown that the maximal frequent sequences provide relevant information to the model to improve the results.

In general, unsupervised approaches for EATS are those that use clustering techniques to group units of documents at different levels, for instance, words, sentences, or paragraphs. The goal is to separate the key ideas from those that are secondary. However, it is important to know which clustering algorithms perform better before beginning to produce a summary. For this purpose, as detailed in Section 2.1, previous studies have analyzed both validation measures and clustering algorithms to provide information regarding the possible best combination.

2.1 Cluster Validation Indexes

As described below in Section 4.1, we follow with the clustering techniques used in the process of an automatic summary generation. It is therefore necessary to choose a measure to validate the quality of the clustering.

Different internal cluster validation indexes have been described in the literature. Because each index has its own advantages and disadvantages over different datasets, we decided to select them according to their properties and performance on different synthetic datasets.

The goal of clustering is to build groups where the objects of the same group are similar but the objects between groups are as different as possible. Therefore, internal measures are used to evaluate two aspects of the clusters, namely, the compactness and separation. The compactness measures how homogeneous the objects are in the same group. The separation measures how well the separated groups are from other groups.

To determine a good-quality index of the clustering, there are certain properties that each index meets at a higher or lower degree. Liu et al. [15] explores the use of five validation properties: monotonicity, noise, density, subclusters, and skewed distributions. Synthetic datasets allow determining the performance of each property for different indexes.

In a similar manner, Rendón et al. [20] evaluated the internal quality indexes on 12 synthetic datasets. Although the property to be measured is not labeled, each dataset is built to measure the clustering index performance in different scenarios, that is, in a distinct organization of objects.

Both works [15, 20] highlighted the Davies Bouldin, Silhouette, and Dunn indexes, and for this reason, we tested these indexes in this study. Each is briefly described in the following.

The **Dunn index** [8] measures the relation between the maximal distance in the same group and the minimum distance between groups of a partition. That is, for each cluster, it computes the pairwise distance between each object in the cluster and the objects of the remaining clusters. The minimum pairwise distance (min-separation) is then obtained. Next, for each cluster the distance between all objects of the same group is calculated, and the maximum distance (max-diameter) is selected. Formally, Dunn index is defined as follows:

$$Dunn = \frac{min_{1\leq i<j\leq c} d(c_i, c_j)}{max_{1\leq k\leq c(\delta_k)}}, \tag{1}$$

where $d(c_i, c_j)$ defines the inter-cluster separation and $d(X_k)$indicates the intra-cluster compactness. Thus, the Dunn index should be maximized.

The **Davies Bouldin index** [7] computes for each cluster the average distance between the objects and its centroid to measure compactness of the clusters. In addition, to identify the cluster separation, the distance between centroids is computed. This index is defined as follows:

$$DB = \frac{1}{c} \sum_{i=1, i\neq j}^{c} Max\{\frac{\delta_i + \delta_j}{d(c_i, c_j)}\}, \tag{2}$$

where c is the number of clusters, δ_i defines the average distance between each object in the 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 in the index indicate compact clusters, the centroids of which are well-separated from each other. Thus, the partition that minimizes the Davies Bouldin index is considered optimal.

The **Silhouette coefficient** [22] measures how close each centroid in the cluster is to each other objects in the neighboring clusters. Thus, for each object i, the average proximity a_i is computed between i and all other objects in the cluster to which i belongs. Then, for the remaining clusters c, the average proximity $f(i,c)$ is calculated for all objects in c. The smallest value of $f(i,c)$ is defined as $b_i = min_c f(i,c)$, and the coefficient is defined as follows:

$$s(i) = \frac{b_i - a_i}{max\{a_i, b_i\}}, \tag{3}$$

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

2.2 Proximity Measures

Clusters are commonly defined as grouped objects that are similar to each other, whereas the objects of the different clusters are not. Thus, determining the closeness of the objects is an extremely important process for providing high-quality clusters. Different measures have been proposed to calculate the proximity between objects in a partition [12]. In this study, the three proximity measures commonly used in text and summarization are as follows.

The **cosine similarity** measures the similarity between two patterns with the cosine of the angle of its feature vectors. If two vectors consist of the same terms, the cosine value is 1; otherwise, the cosine value may decrease to -1. The cosine similarity is defined as follows:

$$CS = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}, \tag{4}$$

where A_i and B_i are attributes of vectors A and B, respectively.

The **Euclidean distance** is a standard metric that indicates the ordinary distance between two points. This measure is widely used in clustering problems. As a true metric, it meets the following properties:

— Symmetry, $D(x_i, x_j) = D(x_j, xi)$,

— Positivity, $D(x_i, x_j) \geq 0 forall x_i, x_j$,

— Triangle inequality, $D(x_i, x_j) \leq D(x_i, x_k) + D(x_k, x_j) \forall x_i, x_j$ and x_k,

— Reflexivity, $D(x_i, x_j) = 0$, if $x_i = xj$.

The Euclidean distance tends to form hyper spherical clusters. Furthermore, it is invariant to translations and rotations. The distance between two points is described as follows:

$$d_E(P,Q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}, \tag{5}$$

where P and Q are two points of an n-dimensional space.

The **Normalized Google Distance** (NGD) computes the semantic distance between two concepts by measuring the logarithm of the maximum frequency with regard to the k-term among n-sentences. This measure is described as follows:

$$NGD(x,y) = \frac{max\{logf(x), logf(y)\} - logf(x,y)}{logN - min\{logf(x), logf(y)\}}, \tag{6}$$

where $f(x)$ indicates the number of documents containing the term x, and $f(x,y)$ indicates the number of documents containing both the x and y terms.

3 Validation Indexes for Automatic Text Summarization

As mentioned above (see Section 2), numerous studies have used external quality measures to validate the performance of automatic summarization systems. However, the main disadvantage of an external measure is that previous knowledge of the specific classes is needed. By contrast, validation indexes do not require prior information.

Following our proposed approach, sentence clustering allows detecting the key ideas in the documents; in turn, the validation indexes are focused on evaluating the generated clusters. That is, these validation indexes evaluate the homogeneity and separability of the groups. However, the correlation between these indexes and the production of high-quality summaries is not clear. Furthermore, it is necessary to prove that such indexes can provide summaries as well as a human can.

In view of the above, to measure the impact of validation indexes on the quality of a summary, three baselines were evaluated, namely, top-line, first-line, and random-line. These baselines were generated based on the DUC02 dataset (detailed in Section 4.2), and are defined below.

The top-line baseline consists of summaries written by humans, and thus we apply the reference summaries of the dataset because they were written by humans. The first-line baseline consists of summaries generated by selecting the first sentences of the documents. Finally, the random-line baseline consists of summaries generated through a random selection of sentences. We can infer that poor-quality summaries (random-line) maintain a specific correlation with poor-quality clusters, and vice-versa.

The three cluster validation indexes Dunn, Davies Boldin, and Silhouette were selected because they have achieved good results (see Section 2). Each index offers an interpretation of the quality of the clustering, that is, Dunn is a maximization index, and thus the higher its result, the better the clustering is. By contrast, Davies Bouldin is a minimization index, and Silhouette considers a range of real values between 1 and -1, where the values closest to 1 represent a better clustering.

Our study results shown in Figure 1 indicate that the Davies Bouldin (a) and Silhouette (b) indexes obtain better results in the top-line baseline, which indicates that the summaries generated by humans provide better index results, unlike the summaries generated randomly or those generated by selecting the first sentences. By contrast, the Dunn index (c) does not show the same behavior because different baselines have reached the best results.

According to [15],[20], dedicated measuring the efficiency of the validation indexes, the Davies Bouldin and Silhouette indexes performed better than the Dunn index. This fact is reaffirmed based on the degree of correlation achieved by the three indexes (Dunn's index showing the least correlation). Therefore, given that the Silhouette index showed the best correlation with the quality of the summaries, we selected this index for application in the following experiments.

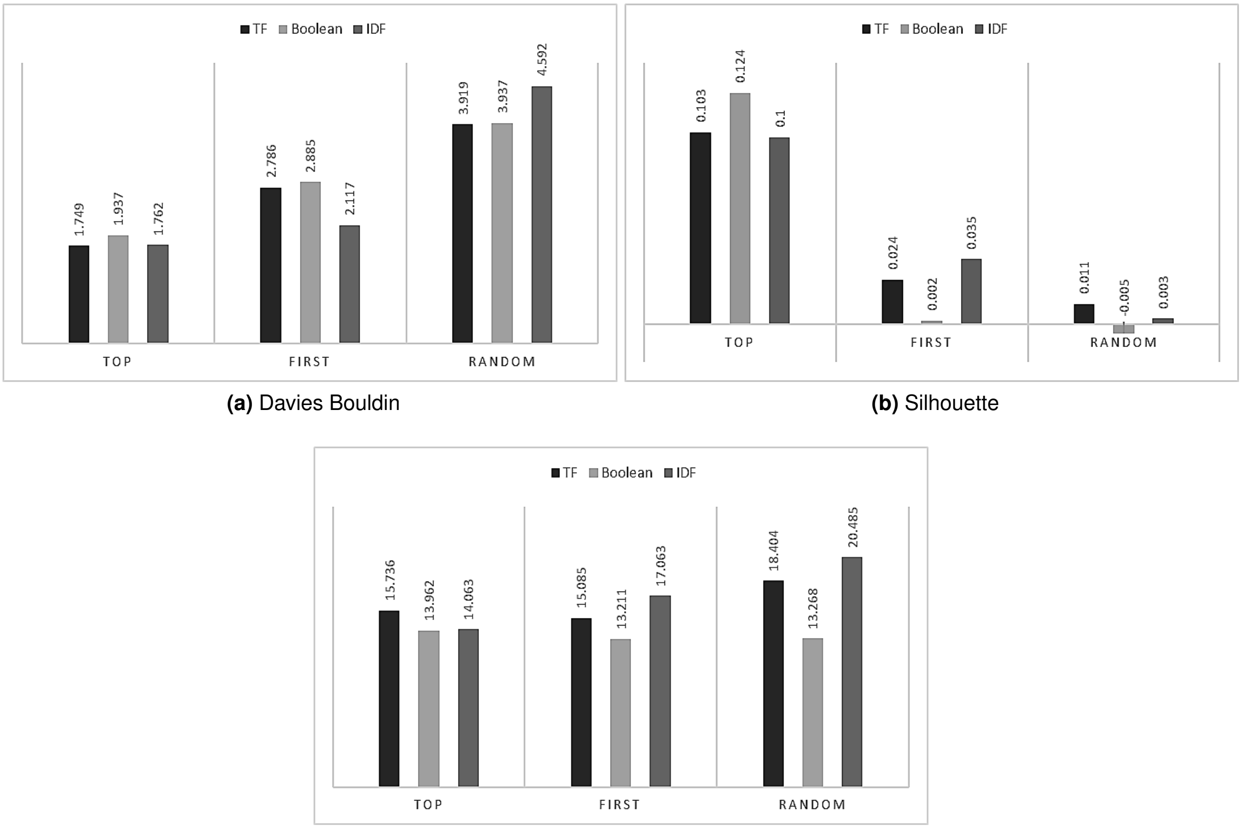
4 An Evolutionary Algorithm for Automatic Text Summarization

Given that the formation of groups of sentences turns into a combinatorial problem, the proposed approach attempts to generate clusters based on the genetic algorithm (GA). That is, each sentence of the text is considered as a centroid of each group. These centroids are represented as 1 among the individuals of the GA, which allows for the search of a variety of solutions. Each solution was evaluated based on the validation index, and the centroids of the best solutions were the sentences of the final summary. The main advantage of this configuration is that the automatic text summarization system is domain- and language-independent.

Although there are several techniques used to automatically generate a summary, they need prior knowledge regarding the language, characteristics, or domain of the documents (supervised approach). Some approaches are even carried out using unsupervised methods; however, they apply external measures that require a priori information. This information is commonly unavailable in a real problem.

In this study, an automatic summarization is approached using a clustering technique (as detailed in Section 4.1) based on the GA (see Section 4.2). In addition, the Silhouette index is applied as a fitness function in the GA to evaluate a summary (see Section 3). The proposed approach is briefly described as follows (see Figure 2). First, each document is separated into sentences that are fed into the GA.

Next, the binary individuals of the GA stand for the sentences of a certain document where the algorithm provides the best tentative solutions of the clusters. To provide the best solutions,



(a) Davies Bouldin

(b) Silhouette

(c) Dunn

Fig. 1. Index comparison between baselines

the clusters are evaluated using the Silhouette index validation. Finally, the centroids of the best solutions are chosen as part of the summary. This process is repeated through each document in the collection.

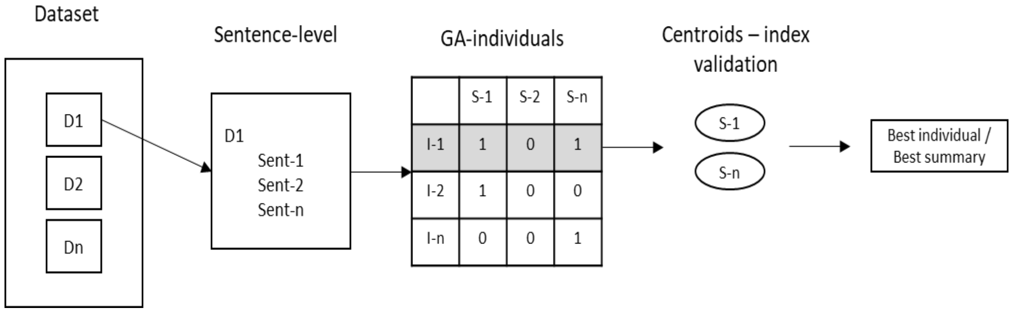
The proposed approach has the advantages of being language- and domain-independent because it does not require any a priori information.

4.1 Partitional Clustering Representation

Following the human behavior where summaries are generated by choosing the most important sentences in a document, we attempted to capture the key sentences by considering that they are surrounded by other secondary ideas such as a centroid surrounded by attracted patterns.

To achieve a distance measure between documents, it is necessary to create a proximity matrix that consists of distances between all objects or patterns. In the framework of this research, the objects are the sentences in a certain document. Thus, for N sentences, we define an $N \times N$ symmetric matrix where the intersection of i and j represents the proximity measure between the i^{th} and j^{th} sentences. Thus, the generation of this matrix requires two steps, namely, choosing methods for mapping the sentences to numeric vectors, and choosing the proximity measures.

Three common methods have been used for mapping texts to numeric vectors [26]: term frequency (tf), a Boolean representation, and the inverse document frequency (idf), which are briefly define below. A list of all words $W_1, W_2, ..., W_n$



	S-1	S-2	S-n
I-1	1	0	1
I-2	1	0	0
I-n	0	0	1

Fig. 2. Framework of the proposed approach

in the document is formed to obtain a Boolean representation. Next, we analyze each document, searching for whether W_n exists in the current text; if so, feature n (F_n) is set to 1; otherwise, it is set to zero. The tf representation is carried out similar to a Boolean representation, although the word frequency is considered.

Unlike the term frequency and Boolean representation, which consider only the terms in the processed document, the inverse document frequency considers the frequency of the terms in all documents in the collection. Therefore, this representation suggests that a term that is very frequent in several documents is less relevant.

Three proximity measures were implemented to compute the closeness between vectors: normalized Google, cosine, and the Euclidean distance. These measures are described in Section 2.2.

To generate groups of similar objects, the basics of partitional clustering algorithms are used; however, determining the number of groups to be generated to find the best solution turns into a combinatorial problem. That is, the partitional algorithms may organize a set of documents into K clusters; therefore, given a set of documents $x_i \in \mathcal{R}^d, i = 1, ..., N$, it is possible to enumerate all possibilities to find the best solution.

However, this brute force approach is infeasible because it turns into an extremely expensive computational problem [32], as suggested through the following formula:

$$P(N, K) = \frac{1}{K!} \sum_{m=1}^{K} (-1)^{K-m} C_K^m m^N, \qquad (7)$$

where N is the number of attributes (dimensions) of the patterns, K is the number of clusters in the partition, and C_k stands for a particular cluster.

Following the formula, the possible solutions to grouping 30 documents into three clusters is 2×10^{14}. Therefore, we opted for the use of a heuristic to provide the best approximate solutions.

4.2 Generating Partitions using Genetic Algorithm

A GA representation is proposed to find the best combination of sentences to provide high-quality summaries. Therefore, the individuals are configured as follows: The number of chromosomes in each individual is equal to the number of sentences in the document to be summarized. In turn, the individual codification is binary, and thus, each chromosome may be set to 1 or 0, where 1 indicates that the sentence is a centroid and 0 indicates that it is not.

The initial population is generated by assigning a random value to each gene. That is, given the individual $P = \{g_1, g_2, ..., g_n\}$, where n is the total number of sentences in the document, each $g_i = Random[0, 1]$. As the sole constraint, the generated summaries should consist of approximately 100 words, and thus the results are comparable with those of the current benchmark studies.

Therefore, it is possible to add sentences to an individual, i.e., a summary, until a maximum of 100 words is reached.

The activated genes ($g_n = 1$) act as attractors to the closer sentences. Thus, an individual formed of n-centroids will form n-clusters. Finally, the centroids of the groups are considered the main topics of the document, whereas the sentences attracted by the centroid are considered ideas that are close to the main topic.

The principle of evolution suggests that the recombination of good solutions tends to provide outperforming solutions. However, their diversity is also important. Thus, the parents' selection process is applied using a roulette operator that provides a high likelihood that the best solutions will be selected; however, it does not completely discriminate poor solutions.

Because frequently used methods are unsuitable for a summarization process, to generate an offspring, a recombination operator is proposed. Therefore, random genes in the parent individuals are selected to be a part of the new individual, taking into account that only genes with a value of 1 are considered. The minimum number of words forming the summary is verified each time a gene is selected to be part of the son chromosome.

According to the evolution scheme, there is a low probability that a mutation will occur; however, mutations play an important role in the diversification of the solutions. The standard mutation operator inverts the binary value of a selected gene. However, in this study, we propose applying this operator in the first instance to genes with a value of 1, and then to those with a value of 0. The purpose of this is to control the number of words in the summaries; as in the recombination process, the summary length is revised after each mutation is applied.

We use the DUC02 dataset, which consists of 567 news articles written in English, to validate our approach. Every news article was written by two expert humans, which allows comparing the summaries generates by the system with those made for humans.

As proved in Section 3 based on an empirical method, the Silhouette and Davies Bouldin indexes may be utilized by generating high-quality summaries. However, the index selected based on its aptitude function was Silhouetted owing to the fact that it was shown to have support for different grouping properties (as described in Section 2.1). Thus, as a result of evaluating individuals, the value obtained will be within a range of 1 to -1, where the values nearest to 1 will indicate promising individuals, whereas the values nearest to -1 will indicate that the sentences were improperly assigned.

5 Results and Discussion

In Section 3, it was proven that it is possible to use a solid validation index as an aptitude function to generate high-quality summaries. That is, a high correlation exists between a Silhouette index and a Rouge measure.

In this section, the proposed approach for an automatic text summarization is evaluated. The experiments shown below were validated using the Silhouette index as a fitness function.

In addition, the results are also shown using the Rouge measure, which provides a comparison with the validation index. This measure was chosen owing to the fact that it is widely used to evaluate automatic summarization tasks, allowing performance comparisons of our results with those of a previous study.

The Rouge measure [14] was proposed to automatically evaluate the similarity between summaries. This measure is able to compare an automatically generated summary (hypothesis) with multiple references (e.g. summaries generated by humans). In addition, it provides different methods to measure a similarity; for example, based on unigrams (Rouge-1), bigrams (Rouge-2), and a skip-bigram (Rouge-SU), among others.

The results of an evaluation with Rouge-1, Rouge-2, and Rouge-SU on the DUC02 dataset are shown in Tables 1, 2, and 3, respectively. Because the corpus consists of 567 documents, each was evaluated using Rouge and the average precision, recall, and F-measure were calculated. In addition, the best and worst results (of the 567 summarized documents) and the coefficient of variation of the F-measure also were obtained.

Owing to the fact that a part of the proposed approach includes a clustering task, we decided to evaluate different proximity measures, namely, the cosine, Euclidean, and normalized Google distances.

Table 1. Automatic text summarization results using different metrics (Cosine, Euclidean, and Normalized Google Distance) as proximity measures and Silhouette index. The results are shown based on the precision, recall, and F-measure by applying Rouge-1

Metric	Avg-F-Rouge	Avg-P-Rouge	Avg-R-Rouge	min-F	max-F	Coefficient of variation
Cosine	0.478	0.476	0.480	0.214	0.788	0.182
Euclidean	0.455	0.454	0.457	0.170	0.763	0.203
Cosine-Euclidean	0.470	0.470	0.472	0.197	0.799	0.186
NGD	0.476	0.474	0.478	0.222	0.795	0.182
NGD-Cosine-Euclidean	**0.481**	**0.478**	**0.483**	0.213	0.788	0.182

Table 2. Automatic text summarization results using different metrics (cosine, Euclidean, and normalized Google distances) as proximity measures and a Silhouette index. The results are shown based on the precision, recall, and F-measure by applying Rouge-2

Metric	Avg-F-Rouge	Avg-P-Rouge	Avg-R-Rouge	min-F	max-F	Coefficient of variation
Cosine	0.226	0.225	0.227	0.020	0.587	0.438
Euclidean	0.200	0.199	0.201	0.011	0.600	0.512
Cosine-Euclidean	0.217	0.216	0.218	0.015	0.657	0.464
NGD	0.223	0.222	0.224	0.019	0.637	0.448
NGD-Cosine-Euclidean	**0.227**	**0.226**	**0.228**	0.020	0.574	0.443

Table 3. Automatic text summarization results using different metrics (cosine, Euclidean, and normalized Google distances) as proximity measures and a Silhouette index. The results are shown based on the precision, recall, and F-measure by applying Rouge-SU

Metric	Avg-F-Rouge	Avg-P-Rouge	Avg-R-Rouge	min-F	max-F	Coefficient of variation
Cosine	**0.243**	**0.242**	0.244	0.055	0.593	0.358
Euclidean	0.221	0.221	0.222	0.043	0.601	0.404
Cosine-Euclidean	0.235	0.234	0.236	0.048	0.630	0.373
NGD	0.240	0.239	0.241	0.056	0.640	0.365
NGD-Cosine-Euclidean	**0.243**	**0.242**	**0.245**	0.054	0.593	0.360

As can be seen in Table 1, the combination of these three measures improves the results of the F-measure based on Rouge-1 when compared with each measure applied separately.

This means that the combination of distances can provide a special proximity matrix that can better represent the homogeneity and separability between groups.

An increase in performance also applies to Rouge-2 (see Table 2), whereas Rouge-SU shared the best results with the cosine measure (see Table 3).

To demonstrate the performance of the proposed approach for different summaries, a coefficient of variation was calculated.

This coefficient indicates whether the system performs well using documents in the dataset.

Table 4. Comparison of results with other approaches

Approach	Rouge-1	Rouge-2
This work	**0.48151(3)**	**0.22781(1)**
FEOM [25]	0.46575(5)	0.12490(4)
UnifiedRank [30]	0.48478(1)	0.21462(3)
SFR [28]	0.48423(2)	0.22471(2)
DE [2]	0.46694(4)	0.12368(5)
NetSum [27]	0.44963(6)	0.11167(6)
CRF [23]	0.44006(7)	0.10924(7)

Table 5. Global evaluation of our proposed approach with respect to other state-of-the-art methods

Method	Partial ranking							Global ranking
	1	2	3	4	5	6	7	
This work	1	0	1	0	0	0	0	1.714
UnifiedRank	1	0	1	0	0	0	0	1.714
SFR	0	2	0	0	0	0	0	1.714
FEOM	0	0	0	1	1	0	0	1.000
DE	0	0	0	1	1	0	0	1.000
NetSum	0	0	0	0	0	2	0	0.571
CRF	0	0	0	0	0	0	2	0.285

That is, if the results are highly different, then the coefficient of variation should increase. The confidence range suggests that values within the ranges of 0-0.10, 0.11-0.20, and up to 0.20, are considered as good, acceptable, and unreliable variations, respectively.

For all compared measures shown in Table 1, Rouge-1 reported a variation coefficient of less than 0.20; therefore, this indicates that the results are acceptable because the system shows an acceptable variation in the Rouge score.

However, the variation coefficient applied to Rouge-2 and Rouge-SU significantly increased. Thus, for these measures, the system obtained variable results for each generated summary. The reason for this variation is the search for a co-occurrence because Rouge-2 looks for bigrams and Rouge-SU looks for trigrams. Therefore, it is more probable in a summary that the co-occurrence of unigrams will be matched than bigrams or trigrams.

6 Conclusions

In this work, different proximity measures were evaluated to provide a vector space representation for the clustering process. As shown in Section 5, the combination of the Google, cosine, and Euclidean distances shows the best results in most cases.

This indicates that the combination of proven measures increases the model performance in creating summaries of the DUC2002 dataset.

In the view of the above, this research introduced a new approach for an automatic text summarization that applies Silhouette to find high-quality clusters. The research findings empirically show that, for the first instance, the Silhouette and Davies Boldin indexes provide a measure indicating the high quality of the summaries. That is, the results of the external measurement by Rouge are correlated with those obtained by the applied index.

The use of a validation index allows the model to find high-quality summaries without the need for prior information. This measure was then used in a genetic algorithm as an aptitude function when searching for key sentences in the documents.

As can be seen in Table 4, the proposed framework shows competitive outcomes compared to those methods that depend on the domain and language. A comparison between the proposed approach and the other methods showed competitive results for Rouge-1, whereas the proposed approach outperformed the results of Rouge-2.

To show the final ranking between the results of the related studies and our proposal, we used Equation 8 proposed by Aliguliyev [3], where r_s indicates the number of times that the method appears in the s rank, and m indicates the number of methods included in the ranking. It can be seen in Table 5 that our proposed approach provides a competitive result based on this global ranking:

$$rank(method) = \sum_{s=1}^{m} \frac{(m-s+1)r_s}{m}. \qquad (8)$$

References

1. **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, n^o 27 (septiembre 2001); pp. 281-290*.
2. **Alguliyev, R. M., Aliguliyev, R. M., Isazade, N. R., Abdi, A., & Idris, N. (2019).** Cosum: Text summarization based on clustering and optimization. *Expert Systems*, Vol. 36, No. 1, pp. e12340.
3. **Aliguliyev, R. M. (2009).** Performance evaluation of density-based clustering methods. *Information Sciences*, Vol. 179, No. 20, pp. 3583–3602.
4. **Amini, M.-R. & Gallinari, P. (2002).** The use of unlabeled data to improve supervised learning for text summarization. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 105–112.
5. **Bando, L. L., Lopez, K. R., Vidal, M. T., Ayala, D. V., & Martinez, B. B. (2007).** Comparing four methods to select keywords that use n-grams to generate summaries. *Electronics, Robotics and Automotive Mechanics Conference (CERMA 2007)*, IEEE, pp. 724–728.
6. **Da Cunha, I., Fernández, S., Morales, P. V., Vivaldi, J., SanJuan, E., & Torres-Moreno, J. M. (2007).** A new hybrid summarizer based on vector space model, statistical physics and linguistics. *Mexican International Conference on Artificial Intelligence*, Springer, pp. 872–882.
7. **Davies, D. L. & Bouldin, D. W. (1979).** A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, , No. 2, pp. 224–227.
8. **Dunn, J. C. (1974).** Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, Vol. 4, No. 1, pp. 95–104.
9. **Fattah, M. A. & Ren, F. (2008).** Automatic text summarization. *World Academy of Science, Engineering and Technology*, Vol. 37, pp. 2008.
10. **Gambhir, M. & Gupta, V. (2017).** Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, Vol. 47, No. 1, pp. 1–66.
11. **García-Hernández, R. A., Montiel, R., Ledeneva, Y., Rendón, E., Gelbukh, A., & Cruz, R. (2008).** Text summarization by sentence extraction using unsupervised learning. *Mexican International Conference on Artificial Intelligence*, Springer, pp. 133–143.
12. **Huang, A. (2008).** Similarity measures for text document clustering. *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, pp. 49–56.
13. **Ledeneva, Y., García-Hernández, R. A., & Gelbukh, A. (2014).** Graph ranking on maximal frequent sequences for single extractive text summarization. *International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, pp. 466–480.
14. **Lin, C.-Y. (2004).** Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.
15. **Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010).** Understanding of internal clustering validation measures. *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, IEEE, pp. 911–916.
16. **Maña López, M. (2003).** *Generación automática de resúmenes de texto para el acceso a la información*. Ph.D. thesis, Tesis Doctoral, Universidad de Vigo). Recuperado de http://www. uhu. es . . . .
17. **Mendoza, V. N., Ledeneva, Y., & García-Hernández, R. A. (2019).** Abstractive multi-document text summarization using a genetic algorithm. *Mexican Conference on Pattern Recognition*, Springer, pp. 422–432.
18. **Montiel Soto, R., Ledeneva, Y., García-Hernández, R. A., & Cruz Reyes, R. (2009).** Comparación de tres modelos de texto para la generación automática de resúmenes. *Procesamiento del Lenguaje Natural*, , No. 43.
19. **Neto, J. L., Freitas, A. A., & Kaestner, C. A. (2002).** Automatic text summarization using a machine learning approach. *Brazilian Symposium on Artificial Intelligence*, Springer, pp. 205–215.
20. **Rendón, E., Abundez, I., Arizmendi, A., & Quiroz, E. M. (2011).** Internal versus external cluster validation indexes. *International Journal of computers and communications*, Vol. 5, No. 1, pp. 27–34.
21. **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.

22. **Rousseeuw, P. J. (1987).** Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, Vol. 20, pp. 53–65.

23. **Shen, D., Sun, J.-T., Li, H., Yang, Q., & Chen, Z. (2007).** Document summarization using conditional random fields. *IJCAI*, volume 7, pp. 2862–2867.

24. **Simón, J. R., Ledeneva, Y., & García-Hernández, R. A. (2018).** Calculating the significance of automatic extractive text summarization using a genetic algorithm. *Journal of Intelligent & Fuzzy Systems*, , No. Preprint, pp. 1–12.

25. **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.

26. **Steinbach, M., Karypis, G., Kumar, V., et al. (2000).** A comparison of document clustering techniques. *KDD workshop on text mining*, volume 400, Boston, pp. 525–526.

27. **Svore, K., Vanderwende, L., & Burges, C. (2007).** Enhancing single-document summarization by combining ranknet and third-party sources. *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*.

28. **Vázquez, E., Arnulfo García-Hernández, R., & Ledeneva, Y. (2018).** Sentence features relevance for extractive text summarization using genetic algorithms. *Journal of Intelligent & Fuzzy Systems*, , No. Preprint, pp. 1–13.

29. **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.

30. **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.

31. **Wang, F. L., Yang, C. C., & Shi, X. (2006).** Multi-document summarization for terrorism information extraction. *International Conference on Intelligence and Security Informatics*, Springer, pp. 602–608.

32. **Xu, R. & Wunsch, D. (2008).** *Clustering*, volume 10. John Wiley & Sons.

33. **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.

Article received on 30/10/2019; accepted on 07/03/2020. Corresponding author is Néstor Hernández Castañeda.