information

MDPI

Article

Multidocument Arabic Text Summarization Based on Clustering and Word2Vec to Reduce Redundancy

Samer Abdulateef, Naseer Ahmed Khan, Bolin Chen and Xuequn Shang *

School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China; samirabdulateef@mail.nwpu.edu.cn (S.A.); naseerkhan@mail.nwpu.edu.cn (N.A.K.); blchen@nwpu.edu.cn (B.C.)

* Correspondence: shang@nwpu.edu.cn

Received: 25 December 2019; Accepted: 19 January 2020; Published: 23 January 2020

check for updates

Abstract: Arabic is one of the most semantically and syntactically complex languages in the world. A key challenging issue in text mining is text summarization, so we propose an unsupervised score-based method which combines the vector space model, continuous bag of words (CBOW), clustering, and a statistically-based method. The problems with multidocument text summarization are the noisy data, redundancy, diminished readability, and sentence incoherency. In this study, we adopt a preprocessing strategy to solve the noise problem and use the word2vec model for two purposes, first, to map the words to fixed-length vectors and, second, to obtain the semantic relationship between each vector based on the dimensions. Similarly, we use a k-means algorithm for two purposes: (1) Selecting the distinctive documents and tokenizing these documents to sentences, and (2) using another iteration of the k-means algorithm to select the key sentences based on the similarity metric to overcome the redundancy problem and generate the initial summary. Lastly, we use weighted principal component analysis (W-PCA) to map the sentences' encoded weights based on a list of features. This selects the highest set of weights, which relates to important sentences for solving incoherency and readability problems. We adopted Recall-Oriented Understudy for Gisting Evaluation (ROUGE) as an evaluation measure to examine our proposed technique and compare it with state-of-the-art methods. Finally, an experiment on the Essex Arabic Summaries Corpus (EASC) using the ROUGE-1 and ROUGE-2 metrics showed promising results in comparison with existing methods.

Keywords: arabic text summarization; multidocument text summarization; text clustering; word2vec

1. Introduction

Automatic text summarization (ATS) is a technique designed to automatically extract salient information from related documents, which helps to produce a summarized document from a related set of documents [1]. Nowadays, the amount of text data is increasing rapidly in areas such as news, official documents, and medical reports, so there is a need to compress such data using machine learning techniques, and text summarization can assist in extracting the significant sentences from various related documents [2].

The main problems related to document summary are redundancy, noisy information, incoherency, and diminished readability [3]. We propose an unsupervised technique to deal with these problems that is based on combined multilevel features, such as important phrases, sentence similarity with titles, and sentence location.

The text clustering technique is used for eliminating redundancy, and the sentences are categorized into semantically correlated sentences. Text summarization is used for selecting the key sentences (rich significant information) from correlated documents. When selecting two sentences and making the comparison between them based on a similarity method, one of these sentences is rendered redundant based on the similarity threshold [1,4]. Important issues addressed by our technique relate to solving the redundancy and noise problems without eradicating significant sentences and ordering sentences after selection.

One method that deals with natural language processing (NLP) is ATS, which extracts the important sentences from related documents. Many researchers have focused on examining European languages and English at the Text Analysis Conference (TAC) and the Document Understanding Conference (DUC), however, there is a shortage of research on the Arabic language [5]. There are many types of methods that can classify text summarization, and Figure 1 shows the techniques of text summarization. Our research has examined multidocument text summarization based on extracting related and significant information from the Arabic language within a generic context [6].

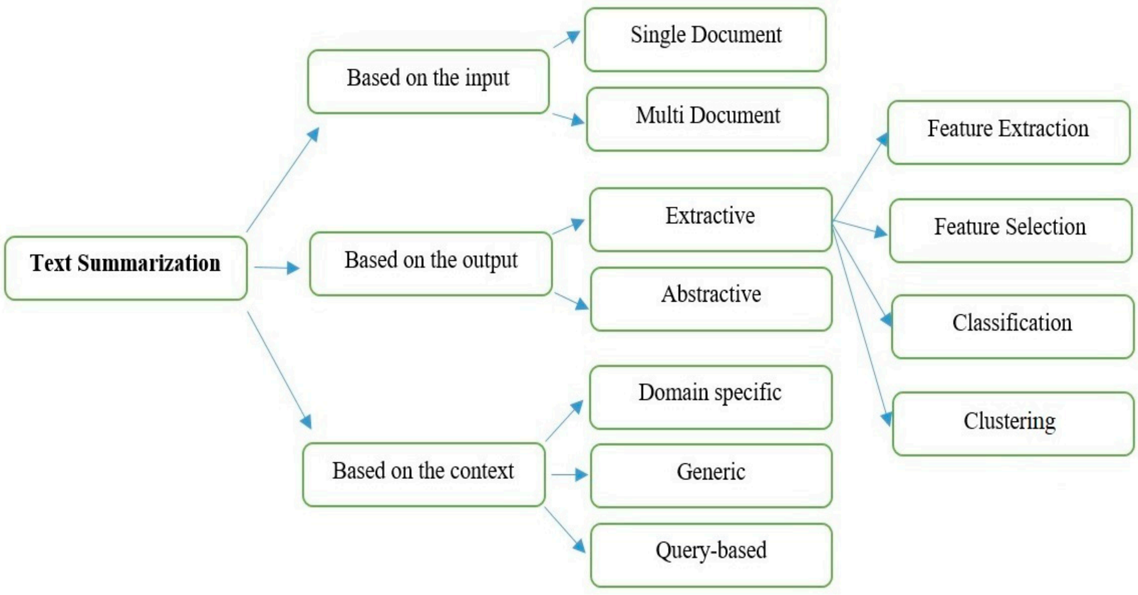


Figure 1. Text summarization methods.

We have devised a novel model, which is described as follows:

- We have adopted a continuous bag-of-words (CBOW) model to generate the semantics and relationships in order to retain the context information.
- We have applied a k-means algorithm for realizing text summarization with the Arabic language, which can aid in increasing the efficiency of the model.
- We have adopted a statistical model (weighted principal component analysis (W-PCA)) based on a list of features to solve the ranking and selection problems.

2. Related Work

Various methods have been mentioned in the literature regarding multidocument summarization, which are classified as machine learning methods, semantic methods, cluster methods, statistical methods, graph methods, optimization methods, and discourse method summarization.

2.1. Machine Learning Method

This approach, which has been applied in binary classification with multidocument summarization and has shown promising results, requires labeled data to train the model. In other words, the performance is affected by selecting the significant set of features, and the representation of these features plays an important role in this approach's performance.

With this method, sentences are transferred into vectors. These vectors are calculated for different levels of features, for example, tokening the documents to paragraphs, sentences, or words, and frequency is calculated to extract the relationship between them. Belkebir et al. [7] examined the Support Vector Machine (SVM) and AdaBoost algorithms for Arabic document summarization based on machine learning. This method decides which sentences will be selected for the final summary. The first step of this method is to apply two classifier techniques, namely, SVM and AdaBoost. The second step is predicting the sentence for the final summary by the AdaBoost and SVM classifiers. The performance of machine learning is affected by the selected classifier method, the set of features selected, and the set of features represented, which play a significant role in the performance of this method.

Moscato et al. [8] proposed a system that is related to online social networks which focuses on online medical health records. Their system applied cluster-computing in order to provide medical suggestions via both web-app and chat-bot modules based on deep learning to train the chat-bot so that interaction between software and users could be utilized.

2.2. Semantic Method

This method focuses on extracting the relationships between the words based on the semantics of the words. These kinds of techniques can be used for text summarization problems also, however, the drawback of these methods is that they need a specific tool to achieve a high-quality summary, such as in linguistic resources and semantic analysis (Lexical, WordNet). This approach requires intensive memory to store semantic relationships and also requires high-performance processors for complex linguistic processing, semantic knowledge, and additional linguistics. The words' meaning can be presented using the semantic relationship between terms and sentences for providing useful information from the text corpus [9].

2.3. Cluster-Based Method

The main goal of these methods is to classify the objects into sub-classes using the similarity calculated between the word vectors, then, a clustering technique is used to cluster sentences using the similarity measure obtained from the vectors. This technique generates an initial summary and solves redundancy problems by categorizing the similarity of sentences into the same class [10].

Alguliyev et al. [1] suggested the approach of combining the optimization and clustering techniques, based on the generated extractive summaries. Their approach is based on two properties, namely, sentence length and summary coverage. The first stage of their research was to use a k-means algorithm for grouping sentences into clusters based on the different sub-topics that select topics in a document. The second stage was to select representative sentences from each cluster for generating a primary summary, considering the high level of diversity and content coverage. The last stage was to use the final summary as the ideal document summary, then apply the harmonic mean to an objective function to diversify and provide coverage for selecting the sentences for the final summary. Their work was applied to the DUC2001 and DUC2002 datasets and it achieved a ROUGE value of 0.490.

Clustering and word2vec have been applied for extracting the keywords of Arabic language [11]. The main idea of this paper is grouping the similar keywords based on the semantic similarity of words.

The clustering based on cosine similarity was applied based on word vectors for grouping the semantically similar keywords and grouping these words based on the synonyms, words, and common stems clustered to the group, then, using the unigram, bigram, and trigrams, with weights for the final selection. An evaluation has been carried out in terms of the F-measure, recall, and precision using The Universal Declaration of Human Rights corpus, and this method achieved a F-measure of 0.63.

Flora et al. [12] proposed a system for detecting and managing disasters and emergencies using a real life dataset, based on an online social media network, for the purpose of generating a sensor application for an emergency alert system. Their research used a clustering event discovery method with online social media network analysis and a bio-inspired impact analysis technique. The experiments were conducted using the real Twitter dataset.

2.4. Statistical Method

Summarizing text can also be done using various statistical methods. Sentence selection may be based on the set of features which depend on inverse document frequency (IDF), term frequency (TF), similarity with the document title, binary term occurrences, and term occurrences. This method is simple to execute and also can be used to eliminate redundancy. In other words, there are various approaches that are based on statistical methods that can obtain a set of features in order to improve the final results, like the approach used in [13], which improved results when used in combination with statistical and other methods. This method was based on combining sentence location and semantic score, and it examined a single document of Arabic language (Essex Arabic Summaries Corpus (EASC) Corpus), achieving a F-measure of 0.57.

For solving the redundancy problem, we applied statistical-based methods to improve the selection of significant sentences. In addition, the statistical methods enhance the results when combined with other methods, also known as a hybrid approach [6].

2.5. Graph Method

In these methods, text data are shown as a graph, where the nodes of the graph represent the sentences, while the edges among the nodes represent similarity relationships among sentences. The approach presented by [14] applied multidocument sensitive ranking. Their method highlights the effect of the set of documents as global information for evaluating local sentences based on the document-to-document relations and document-to-sentence relations. This method is based on a graph model that assigns relationships between different sets of documents and sentences, where it then examines the sentence evaluations based on the sets of entire document relationships. Experiments with this method on the DUC2004 and DUC2007 datasets have achieved good accuracy.

Wan et al. [15] applied a graph-based method to relate sentences to documents, where these relationships were calculated by a graph-based ranking algorithm. The graph document model was combined to identify the document impact by determining sentence to document relationships and document importance for ranking the sentences. The evaluation results using the DUC2002 and DUC2001 datasets achieved good accuracy based on the proposed approach.

Flora et al. [16] proposed a multimedia summary technique based on an online social media network for generating multimedia stories. Their paper focused on the sharing and management of multimedia data on a social media website, and it focused on the influence of analysis methodologies and graph-based models for discovering the most significant data that was related to one hot topic. They modified their Artificial Bee Colony (ABC) algorithm for ranking, selection, and the semantic correlation between two objects (mixture of texts and pictures). The experiments were conducted using the YFCC100M dataset, and the ROUGE-2 and ROUGE-SU4 metrics were used to test the evaluation.

2.6. Optimization Method

The researchers consider document summarization issues as a multiobjective optimization problem. The main objective is to produce a high-quality text summary that has characteristics such as diminished redundancy, coverage, and coherence of the generated summary. Diminished redundancy means reducing sentence similarity in the final summary and removing repeating information, while the coverage aims for selecting all significant characteristics in the original documents so that important concepts in the documents are not missed. Coherence means generating the flow of a sentences that are semantically correct and continuous. Based on these objectives, the search for the ideal summary in NLP is a challenging task. Al-Radaideh et al. [17] proposed a single-document text summarization method with Arabic language, combining genetic algorithms, statistical features, and domain knowledge for selecting the final summary, based on the EASC corpus. ROUGE was used as an evaluation framework, and the method achieved a F-measure of 0.605. Al-Abdallah et al. [18] suggested the use of a particle swarm optimization algorithm for a single-document text summarization method with

Arabic language, and their approach examined the EASC corpus, based on the combined features like sentence length, term frequency-inverse document frequency (TF-IDF), title similarity, and term frequency. ROUGE was applied as an evaluation framework, and the method achieved a F-measure of 0.553.

2.7. Discourse Method

The structure of the discourse text is necessary for defining the information or context transferred by the text. This method based on text is organized or represented as discourse-units, where each unit is related to other units to ensure cohesion and coherence of the discourse. There are four factors to build structures for successful discourse, namely, language, text structure (used to represent structure (graph or tree)), relationships (lexically grounded, intentional, or semantic), and the discourse theory type [19].

Many methods have been listed in the related work for Arabic language text summarization. Based on our discussion in the previous sections, there are some methods that are more suitable for multidocument Arabic text summarization, such as the optimization method, graph method, and cluster method. The main goals for these approaches are to maximize diversity, coverage, diminished redundancy, and select the most significant sentences. Unlike other studies, our suggested method focuses on examining text summarization with Arabic language based on the EASC corpus in terms of multidocument (MD) summarization, based on combined semantic and statistical features with clustering as an unsupervised technique.

3. The Challenge of the Redundancy

The problem for Arabic multidocument summarization is formulated as follows:

Given a set of documents, MD, that is, MD = (D1, D2, ..., Dn) where Di indicates the i^{th} document in MD, *n* represents the fold document in all text, then, we tokenize the original document D to a list of sentences, that is, D = (S1, . . . , Sn), where Si represents the i^{th} sentence in D, and *n* represents the total number of sentences in each document. The goal of the final summary is to select the set of sentences from MD covering various related topics from related documents, [20].

Natural language processing of the Arabic language is challenging and has the following key properties. In Table 1, various forms of words based on one root (دارس) have been shown.

Table 1. Words with different sub-parts based on one root.

Word	Prefixes	Infixes	Suffixes	Meaning
دارسون	-	أ	و+ ن	Scholars
مدرسات	م	-	أ + ت	Teachers
المدارس	م + ل + ا	أ	-	Schools

- The Arabic language is diacritical and derivative, making morphology analysis a hard task.
- Arabic words are often imprecise, since the system is based on a tri-literal root.
- Broken plurals, where a broken plural in linguistics is an irregular plural form of a noun or adjective initiate in the Semitic Arabic languages.
- Characters can be written in different ways based on the location of the character in a word.

The main idea for this paper is to reduce the redundancy issue by focusing on extracting key sentences. These sentences should contain the main idea for correlated documents for making a comparison between two sentences. One of these sentences is considered redundant if the similarity between these two sentences is high (based on a chosen threshold of similarity), therefore, only one of the associated sentences is selected thereafter [21,22].

4. Proposed System

For this study, a novel text extractive summary method is proposed for Arabic language. There are six main stages for this approach, namely, data collection, text preprocessing, selecting the discriminative documents for generating the initial summary, sentence tokenization, sentence weight mapping, and selecting sentences based on the best weight as the final summary. The final step is to evaluate our suggested approach using the ROUGE metric and compare the results with state-of-the-art methods from the literature. Figure 2 demonstrations the main steps for the suggested approach.

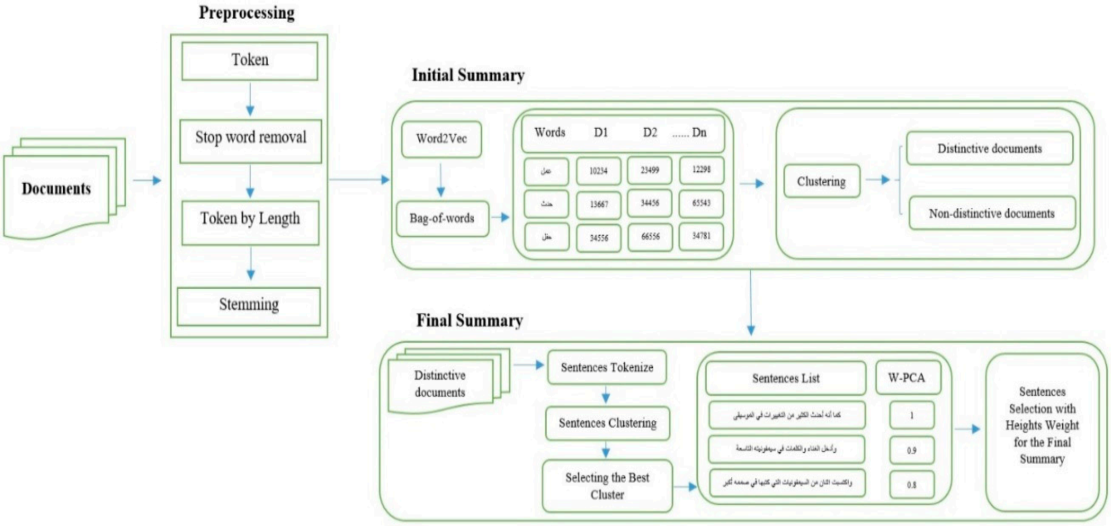


Figure 2. Steps of the proposed method.

More information about our proposed model is given as follows: Firstly, we apply text cleaning to solve the noisy problem. Then, a CBOW model is used to capture the semantic relationship between the terms (word tokens). Then, we apply a K-means algorithm with a cosine similarity measure as the threshold to select the distinctive documents from each category based on the distance metric. These documents will also be used in generating the final summary. The final step is to tokenize these documents to lists of sentences and apply another iteration of the K-means algorithm to select distinctive sentences, where W-PCA is then applied using three kind of features, as discussed, in order to assign the weights for each sentence. These weights help us to select and order the sentences for the final summary.

5. Experimental Results

5.1. Data Collection

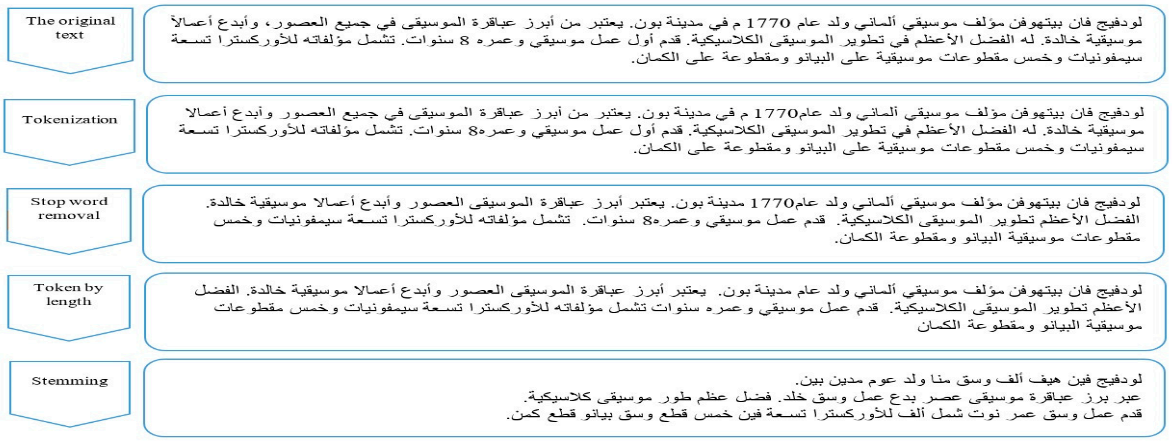
We used an Arabic language corpus called the Essex Arabic Summaries Corpus (EASC), generated by Mechanical Turk [23]. Table 2 shows the details of the corpus.

Table 2. Essex Arabic Summaries Corpus (EASC) statistics.

The Name of the Corpus	Essex Arabic Summaries Corpus (EASC)
Number of Documents	153
Number of Sentences	1652
Number of Words	29,045
Number of Distinct Words	12,785
Number of Gold-Standard Summaries	10 (one for each category)

5.2. Text Preprocessing

Text preprocessing is the main stage for the text simplification method, and the steps are divided into five sub-steps. We started with text tokenization. In this step, we tokenized the text to a set of sentences. Then, we applied filtering to remove the stop word list, deleting all unnecessary words like (في, على) (in, on). After that, we applied a tokenization by length technique and selected 3 to 25 characters in each term. This helps us keep words with three letters. The fourth step was to use the stemming to transfer all the selected words, for example التمويلwill be مول, changing from funding to fund. The last step is to represent the significant words as vectors use the word2vec model [24], based on the continuous bag of words (CBOW) model. This approach was trained by utilizing Wikipedia and Google News, with 100 as the dimension size of each word. The final parameters for the CBOW were a window size of 5, layer size of 100, and a minimum vocab frequency of 2. This approach was used to predict the occurrence probability $p = (r_t | r_{t-c}, r_{(t-c)-1}, \ldots\ldots r_{t-1}, r_{t+1}, r_{t+2}, \ldots\ldots, r_{t+c}$, of a root,$r_t$, given the context roots $r_{t-c}, r_{(t-c)-1}, \ldots\ldots r_{t-1}, r_{t+1}, r_{t+2}, \ldots\ldots, r_{t+c}$. In this approach, preselected window size is represented by c and r_t is the root vector in a featured word. We adopted this approach for word vector training [25]. At the end of this stage, we resolved the noise problem. Figure 3 shows the summary for all the above steps. Figure 4 shows the vectors of the root based on the CBOW, an important observation from the picture is that the word موسيقى, which means music, is an important word in the "Music and Art" category. The output of this step was used to solve the noise problem [26].



The original text

لودفيج فان بيتهوفن مؤلف موسيقي ألماني ولد عام 1770 م في مدينة بون. يعتبر من أبرز عباقرة الموسيقى في جميع العصور، وأبدع أعمالاً
موسيقية خالدة. له الفضل الأعظم في تطوير الموسيقى الكلاسيكية. قدم أول عمل موسيقي وعمره 8 سنوات. تشمل مؤلفاته للأوركسترا تسعة
سيمفونيات وخمس مقطوعات موسيقية على البيانو ومقطوعة على الكمان.

Tokenization

لودفيج فان بيتهوفن مؤلف موسيقي ألماني ولد عام1770 م في مدينة بون. يعتبر من أبرز عباقرة الموسيقى في جميع العصور وأبدع أعمالا
موسيقية خالدة. له الفضل الأعظم في تطوير الموسيقى الكلاسيكية. قدم أول عمل موسيقي وعمره8 سنوات. تشمل مؤلفاته للأوركسترا تسعة
سيمفونيات وخمس مقطوعات موسيقية على البيانو ومقطوعة على الكمان.

Stop word removal

لودفيج فان بيتهوفن مؤلف موسيقي ألماني ولد عام1770 مدينة بون. يعتبر أبرز عباقرة الموسيقى العصور وأبدع أعمالا موسيقية خالدة.
الفضل الأعظم تطوير الموسيقى الكلاسيكية. قدم عمل موسيقي وعمره8 سنوات. تشمل مؤلفاته للأوركسترا تسعة سيمفونيات وخمس
مقطوعات موسيقية البيانو ومقطوعة الكمان.

Token by length

لودفيج فان بيتهوفن مؤلف موسيقي ألماني ولد عام مدينة بون. يعتبر أبرز عباقرة الموسيقى العصور وأبدع أعمالا موسيقية خالدة. الفضل
الأعظم تطوير الموسيقى الكلاسيكية. قدم عمل موسيقي وعمره سنوات تشمل مؤلفاته للأوركسترا تسعة سيمفونيات وخمس مقطوعات
موسيقية البيانو ومقطوعة الكمان

Stemming

لودفيج فين هيف ألف وسق منا ولد عوم مدين بين.
عبر برز عباقرة موسيقى عصر بدع عمل وسق خلد. فضل عظم طور موسيقى كلاسيكية.
قدم عمل وسق عمر نوت شمل ألف للأوركسترا تسعة فين خمس قطع وسق بيانو قطع كمن.

Figure 3. Illustrated all the steps of text cleaning.

موسيقى
حيا
عوم
عمل
خامس
حفل

Figure 4. Illustrated root vectors.

5.3. Initial Summary Generation

We used the clustering technique (k-means algorithm) to generate the primary summary from related documents [27]. This stage is divided into two sections, the first one is classifying the documents into distinctive and non-distinctive sets, meaning clustering similar documents into a group (distinctive text). The second stage is tokenizing these documents into sentences and clustering sentences into distinctive and non-distinctive sentences based on the cosine similarity measure, and, after that, selecting the list of sentences and grouping them in a distinctive cluster [1]. The main idea of this algorithm is to select sentences randomly and using sentences as the central point for each cluster. Then, the sentences are distributed iteratively into the nearest clusters and each cluster centroid is recalculated until there no longer is any change in the centroids. This algorithm depends on some parameters, namely, k (number of clusters) and the measure type. For this study, we used cosine similarity as a numerical measure to calculate the similarity of vectors, as it is a common similarity metric (sentence-to-sentence). On the other hand, the vector-to-vector similarity metric is based on the angle between vectors [28].

The main issue here is to select the best cluster that can further be used in text summarization, as proposed in this paper, also k-means does not have any prior knowledge of the number of clusters used. To tackle this issue, we suggest adopting a cluster ordering technique, however, this comes with the additional problem of determining how to select the cluster based on the size of the cluster. Therefore, to this end, we suggest the use of an automated approach for ordering the cluster, where cluster ordering relies on the performance of cluster distance. The main parameters used in this algorithm were a K-value of 2, maximum run count of 10, a cosine similarity measure type, and a max optimization of 100 steps. Table 3 shows the clustering result. The output of this step is the generated initial summary based on the similarity metric to solve the redundancy problem.

Table 3. The clustering results.

Name of Category	Cluster Distance Performance	Best Cluster	Distinctive Documents
Music and Art	Cluster-0: 0.694 Cluster-1: 0.678	Cluster-1	1, 2, 3, 4, and 10
Education	Cluster-0: 0.639 Cluster-1: 0.534	Cluster-1	5, 6, and 7
Tourisms	Cluster-0: 0.660 Cluster-1: 0.700	Cluster-0	2, 3, 4, 5, 7, and 9
Environment	Cluster-0: 0.770 Cluster-1: 0.484	Cluster-1	1, 5, and 8
Health	Cluster-0: 0.583 Cluster-1: 0.722	Cluster-0	2, 3, 4, and 10
Finance	Cluster-0: 0.675 Cluster-1: 0.694	Cluster-0	1, 2, 3, 7, and 9
Politics	Cluster-0: 0.724 Cluster-1: 0.612	Cluster-1	3, 6, 7, and 8
Science and Technology	Cluster-0: 0.709 Cluster-1: 0.706	Cluster-1	5, 7, 8, 9, and 10
Religion	Cluster-0: 0.533 Cluster-1: 0.687	Cluster-0	1, 3, and 5
Sport	Cluster-0: 0.537 Cluster-1: 0.716	Cluster-0	1, 2, and 10

5.4. Final Summary Generation

We used the weighted principal component analysis for mapping the list of distinctive sentences using the weights. W-PCA generates the feature weights of the list of sentences by using a component created by the principal component analysis, based on a list of features like phrase frequency, and is calculated using Equation (1) [29]. Sentence similarity with the topic title is calculated by cosine similarity [30], and the location of a sentence is assigned based on the first sentence in the first paragraph, the first sentence in the last paragraph, or the first sentence in any paragraph, and is calculated using Equation (2) [31]. For the length of sentence, we adopted the statistical interquartile range (IQR) to avoid very long or very short sentences based on a threshold score between 0.2 and

0.9, because short sentences may not contain key ideas of the document topic and long sentences may contain irrelevant information. The selection was carried out based on Equation (3) [30].

Most of the recent studies have used PCA for solving the dimensional reduction problem [32], so this paper also used PCA for overcoming readability and coherency problems. We applied W-PCA for mapping the distinctive sentences list to the weights that ranged from 0 to 1. Then, we used the selection by weight method and assigned the parameter P with a value of 0.3, where this thereby means that we selected the top 30% from the list of sentences as the final summary. Figure 5 shows the samples of the list of sentences when mapping by weight.

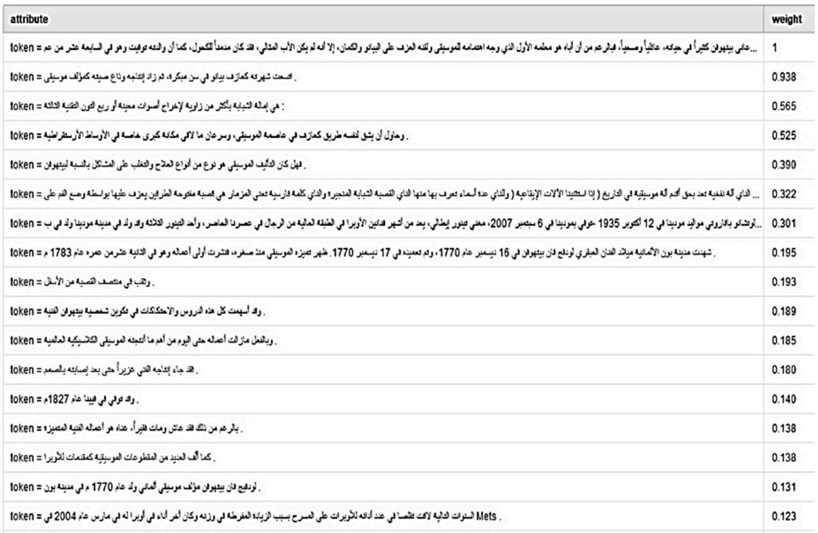
$$IPF = \frac{SIP_i}{IF_d} \tag{1}$$

where IPF refers to important phrase frequency, SIP_I refers to sentence number that contains important phrases, and IF_d refers to the total number of the important frequency in documents.

$$SL(S_d) = \frac{m-1}{m} \tag{2}$$

where SL refers to sentence location, S_d refers to the d^{th} sentence in the document, and m is the maximum number of a sentence in document d.

$$Sentence\ Score = \frac{words\ in\ sentense}{words\ in\ the\ longest\ sentence} \tag{3}$$



attribute	weight
token = ...عاني بيتهوفن كثيراً في حياته، عاطفياً وصحياً، فبالرغم من أن أباه هو معلمه الأول الذي وجه اهتمامه للموسيقى ولقنه العزف على البيانو والكمان، إلا أنه لم يكن الأب المثالي، فقد كان مدمناً للكحول، كما أن والدته توفيت وهو في السابعة عشر من عم	1
token = اصبحت شهرته كعازف بيانو في سن مبكره، ثم زاد إنتاجه وذاع صيته كمؤلف موسيقي .	0.938
token = هي إمالة الشبابة بأكثر من زاوية لإخراج أصوات معينة أو ربع التون التقنية الثالثة :	0.565
token = وحاول أن يشق لنفسه طريق كعازف في عاصمة الموسيقى، وسرعان ما لاقى مكانة كبرى خاصة في الأوساط الأرستقراطية .	0.525
token = فهل كان التأليف الموسيقي هو نوع من أنواع العلاج والتغلب على المشاكل بالنسبة لبيتهوفن .	0.390
token = ...الناي آلة نفخية تعد بحق أقدم آلة موسيقية في التاريخ (إذا استثنينا الآلات الإيقاعية) وللناي عدة أسماء تعرف بها منها الناي القصبة الشبابة المنجيرة والناي كلمة فارسية تعني المزمار هي قصبة مفتوحة الطرفين يعزف عليها بواسطة وضع الفم على	0.322
token = ...لوتشانو بافاروتي مواليد مودينا في 12 أكتوبر 1935 -توفي بمودينا في 6 سبتمبر 2007، مغني تينور إيطالي، يعد من أشهر فناني الأوبرا في الطبقة العالية من الرجال في عصرنا الحاضر، وأحد التينور الثلاثة ولد ولد في مدينة مودينا ولد في ب	0.301
token = شهدت مدينة بون الألمانية ميلاد الفنان العبقري لودفيغ فان بيتهوفن في 16 ديسمبر عام 1770، وتم تعميده في 17 ديسمبر 1770. ظهر تميزه الموسيقي منذ صغره، فنشرت أولى أعماله وهو في الثانية عشر من عمره عام 1783 م .	0.195
token = وثقب في منتصف القصبة من الأسفل .	0.193
token = وقد أسهمت كل هذه الدروس والاحتكاكات في تكوين شخصية بيتهوفن الفنية .	0.189
token = وبالفعل مازالت أعماله حتى اليوم من أهم ما أنتجته الموسيقى الكلاسيكية العالمية .	0.185
token = لقد جاء إنتاجه الفني غزيراً حتى بعد إصابته بالصمم .	0.180
token = وقد توفي في فيينا عام 1827م .	0.140
token = بالرغم من ذلك فقد عاش ومات فقيراً، عداه هو أعماله الفنية المتميزة .	0.138
token = كما ألف العديد من المقطوعات الموسيقية كمقدمات للأوبرا .	0.138
token = لودفيغ فان بيتهوفن مؤلف موسيقي ألماني ولد عام 1770 م في مدينة بون .	0.131
token = السنوات التالية لاقت نقصا في عدد أدائه للأوبرات على المسرح بسبب الزيادة المفرطة في وزنه وكان آخر أداء في أوبرا له في مارس عام 2004 في Mets .	0.123

Figure 5. Ordering the list of sentences by weights.

6. Evaluation and Comparison

After creating the final text summary, a process of evaluation is required to examine the quality of the suggested approach. There are two types of evaluation, the first one is called manual evaluation, which means we need to submit the final result to humans to decide the quality of the method, but this is time-consuming and costly. The second one is called automatic evaluation, and it is very fast and also cheaper than the manual method. To this end, we used Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [1]. There are various kinds of ROUGE measures, like ROUGE-N, ROUGE-W, ROUGE-L, and ROUGE-S. We used ROUGE-N (ROUGE-1, ROUGE-2) in terms of the recall, precision, and F-score, and this helps us examine the overlap between the machine summary and the references summary by counting the similarity units between each of them as unigram or bigrams, as formulated in Equation (7). Table 4 shows the final results, which are based on the gold-standard human summary.

Table 5 shows the comparison results, which are based on the similarity of the dataset or/and the similarity of techniques.

$$\text{Recall} \ = \frac{gram_{ref} \ \cap \ grams_{gen}}{grams_{ref}} \tag{4}$$

$$\text{Precision} = \frac{gram_{ref} \ \cap \ grams_{gen}}{grams_{gen}} \tag{5}$$

$$\text{F} - 1 \ score = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \tag{6}$$

where the reference summary grams are represented by $grams_{ref}$ and the system summary grams are represented by $grams_{gen}$.

$$\text{ROUGE} - \text{N} = \frac{\sum_{S \in References \quad summaries} \sum_{N-gram \in S} Count_{match \ (N-gram)}}{\sum_{S \in References \quad summaries} \sum_{N-gram \in S} Count \ (N - gram)} \tag{7}$$

where N is the total-size of the n-gram, count match Ngram is the highest number of n-grams found in both the human and system summaries, and count over Ngram is the total number of n-grams that are in the human summary.

Table 4. Recall-Oriented Understudy for Gisting Evaluation (ROUGE)-1 and ROUGE-2 results.

Name of Category	ROUGE-1			ROUGE-2		
	Recall	**Precision**	**F-score**	**Recall**	**Precision**	**F-Score**
Music and Art	0.672	0.501	0.574	0.70	0.427	0.517
Education	0.603	0.579	0.590	0.580	0.351	0.431
Tourisms	0.504	0.304	0.379	0.548	0.360	0.405
Environment	0.532	0.307	0.389	0.464	0.308	0.355
Finance	0.605	0.589	0.467	0.558	0.339	0.418
Health	0.695	0.60	0.644	0.621	0.498	0.552
Politics	0.548	0.407	0.467	0.447	0.319	0.345
Religion	0.608	0.6	0.603	0.587	0.534	0.559
Science and Technology	0.607	0.564	0.584	0.615	0.421	0.501
Sport	0.364	0.313	0.336	0.462	0.418	0.332

Table 5. Comparison based on F-score results.

Author (s) Name & Year	Arabic Corpus	Methods	F-score Results
[18] Al-Abdallah 2017	EASC	Optimization Algorithm (single document)	0.553
[13] Al-Abdallah 2019	EASC	Firefly Algorithm (single document)	0.57
[17] Al-Radaideh 2018	EASC	Genetic Algorithms (single document)	0.605
[11] Suleiman 2019	The Universal Declaration of Human Rights	Word2Vec and Clustering (single document)	0.63
Our approach	EASC	Word2Vec, Clustering, and Statistical-based methods (multidocument)	0.644

7. Discussion

Due to large size of the text documents that is increasing day by day, an automatic summarization system is essential so that only useful and meaningful information from the large set of documents is stored and extracted. The main idea for this system is determining how to extract key sentences. We examined 153 related documents of Arabic language. We proposed an unsupervised technique to overcome the problems with Arabic natural language processing (ANLP), as it is one of the complex languages in the world. We used preprocessing to simplify the text to overcome the text noisy problem and word2vec for extracting the semantic relation between the lists of words. In the second step, we applied a k-means algorithm to solve the redundancy problem and generate the initial summary. In the final step, we applied W-PCA to solve readability and coherency problems.

The computational complexity of producing the threshold for each cluster depends on the distance metric which will be used to select the best cluster. Take for instance if there are 10 documents in each category, the computational complexity, a function of the total number of documents in all the categories, is a factor of 100, but, if we produce the gist of each category by extracting few documents from each of them by applying K-means algorithm, the complexity of model as well as the density of the resultant summary is reduced to a greater extent when compared with the primary summary, and also a significant decrease in the computational complexity in the selection of the final summary.

To improve the efficiency, our suggested method for generating the multidocument final text summary in terms of computational times, we observed that it mainly depends on average number of documents in each category and the length of the final summary. We used the top 30% selected sentences from all the sentence list related to a category. To check the computational efficiency, a small experiment was devised using the "Art and Music" category, finding that if there were 1, 2, and 3 number of documents, then it took an average time of 40 s, 80 s, 120 s, respectively, and a similar efficiency of data was observed in all other categories.

The ROUGE evaluation measure was adopted for evaluating the final system summary with a reference summary, we have also compared our method with the state-of-the-art methods. We believe that we have provided the NLP community working on multidocument summarization based on Arabic language with a new tool that will be valuable for future research in this specific domain.

8. Conclusions

In this study, we used an unsupervised technique based on multidocument Arabic text summarization and have focused on text summarization problems such as noisy information, redundancy elimination, and sentence ordering. We investigated word2vec, clustering, and W-PCA in terms of important phrase frequency in each sentence, sentence similarity with the topic title, and the location of sentences as list of features for multidocument text summarization. This paper had four objectives, namely, solving the noisy information problem, reducing redundancy, sentence selection, and sentence ordering. ROUGE has been used as an evaluation measure, based on the proposed approach on the EASC corpus, and the method has achieved an F-score of 0.644. The final results show that the suggested method outperforms the state-of-the-art methods. Finally, a combination of word2vec, clustering, and statistical methods are more suitable technique for Arabic multidocument text summarization.

Author Contributions: X.S. conceived the idea. S.A. proposed the model, architecture and devised implementing strategy for the proposed model. N.A.K. helped in editing manuscript and the implementation of the methodology. B.C. checked the overall progress of the methodology, results, and suggested edits. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under grant number 61772426.

Acknowledgments: We are very grateful to the Chinese Scholarship Council (CSC) for providing us financial and moral support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alguliyev, R.M.; Isazade, N.R.; Abdi, A.; Idris, N.J.E.S. COSUM: Text summarization based on clustering and optimization. *Wiley Online Libr.* **2019**, *36*, e12340. [CrossRef]

2. Sanchez-Gomez, J.M.; Vega-Rodríguez, M.A.; Pérez, C.J. Comparison of automatic methods for reducing the Pareto front to a single solution applied to multi-document text summarization. *Knowl. -Based Syst.* **2019**, *174*, 123–136. [CrossRef]
3. Verma, P.; Om, H. MCRMR: Maximum coverage and relevancy with minimal redundancy based multi-document summarization. *Expert Syst. Appl.* **2019**, *120*, 43–56. [CrossRef]
4. Patel, D.B.; Shah, S.; Chhinkaniwala, H.R. Fuzzy logic based multi Document Summarization with improved sentence scoring and redundancy removal technique. *Expert Syst. Appl.* **2019**. [CrossRef]
5. Mallick, C.; Das, A.K.; Dutta, M.; Das, A.K.; Sarkar, A. Graph-Based Text Summarization Using Modified TextRank. In *Soft Computing in Data Analytics*; Springer: Berlin, Germany, 2019; pp. 137–146.
6. Kanapala, A.; Pal, S.; Pamula, R. Text summarization from legal documents: A survey. *Artif. Intell. Rev.* **2019**, *51*, 371–402. [CrossRef]
7. Belkebir, R.; Guessoum, A. A supervised approach to arabic text summarization using adaboost. In *New Contributions in Information Systems and Technologies*; Springer: Berlin, Germany, 2015; pp. 227–236.
8. Amato, F.; Marrone, S.; Moscato, V.; Piantadosi, G.; Picariello, A.; Sansone, C. HOLMeS: eHealth in the Big Data and Deep Learning Era. *MDPI Inf.* **2019**, *10*, 34. [CrossRef]
9. Gerani, S.; Carenini, G.; Ng, R.T. Language. Modeling content and structure for abstractive review summarization. *Comput. Speech Lang.* **2019**, *53*, 302–331. [CrossRef]
10. Abualigah, L.; Bashabsheh, M.Q.; Alabool, H.; Shehab, M. Text Summarization: A Brief Review. In *Recent Advances in NLP: The Case of Arabic Language*; Springer: Berlin, Germany, 2020; pp. 1–15.
11. Suleiman, D.; Awajan, A.A.; Al Etaiwi, W. Arabic Text Keywords Extraction using Word2vec. In Proceedings of the 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS), Amman, Jordan, 9–11 October 2019; pp. 1–7.
12. Amato, F.; Moscato, V.; Picariello, A.; Sperli'ì, G. Extreme events management using multimedia social networks. *Future Gener. Comput. Syst.* **2019**, *94*, 444–452. [CrossRef]
13. Al-Abdallah, R.Z.; Al-Taani, A.T. Arabic Text Summarization using Firefly Algorithm. In Proceedings of the 2019 Amity International Conference on Artificial Intelligence (AICAI), Dubai, United Arab Emirates, 4–6 February 2019; pp. 61–65.
14. Wei, F.; Li, W.; Lu, Q.; He, Y. A document-sensitive graph model for multi-document summarization. *Knowl. Inf. Syst.* **2010**, *22*, 245–259. [CrossRef]
15. Wan, X.; Yang, J. Multi-document summarization using cluster-based link analysis. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, Singapore, 20–24 July 2008; pp. 299–306.
16. Amato, F.; Castiglione, A.; Mercorio, F.; Mezzanzanica, M.; Moscato, V.; Picariello, A.; Sperlì, G. Multimedia story creation on social networks. *Future Gener. Comput. Syst.* **2018**, *86*, 412–420. [CrossRef]
17. Al-Radaideh, Q.A.; Bataineh, D.Q. A hybrid approach for arabic text summarization using domain knowledge and genetic algorithms. *Cogn. Comput.* **2018**, *10*, 651–669. [CrossRef]
18. Al-Abdallah, R.Z.; Al-Taani, A.T. Arabic single-document text summarization using particle swarm optimization algorithm. *Procedia Comput. Sci.* **2017**, *117*, 30–37. [CrossRef]
19. Lagrini, S.; Redjimi, M.; Azizi, N. Automatic Arabic Text Summarization Approaches. *Int. J. Comput. Appl.* **2017**, 164. [CrossRef]
20. Bialy, A.A.; Gaheen, M.A.; ElEraky, R.; ElGamal, A.; Ewees, A.A. Single Arabic Document Summarization Using Natural Language Processing Technique. In *Recent Advances in NLP: The Case of Arabic Language*; Springer: Berlin, Germany, 2020; pp. 17–37.
21. Al Qassem, L.M.; Wang, D.; Al Mahmoud, Z.; Barada, H.; Al-Rubaie, A.; Almoosa, N.I. Automatic Arabic summarization: A survey of methodologies and systems. *Procedia Comput. Sci.* **2017**, *117*, 10–18. [CrossRef]
22. Badry, R.M.; Moawad, I.F. A Semantic Text Summarization Model for Arabic Topic-Oriented. In Proceedings of the International Conference on Advanced Machine Learning Technologies and Applications, Cairo, Egypt, 28–30 March 2019; pp. 518–528.
23. El-Haj, M.; Kruschwitz, U.; Fox, C. *Using Mechanical Turk to Create a Corpus of Arabic Summaries*; University of Essex: Essex, UK, 2010.
24. Alami, N.; Meknassi, M.; En-nahnahi, N. Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning. *Expert Syst. Appl.* **2019**, *123*, 195–211. [CrossRef]

25. Blagec, K.; Xu, H.; Agibetov, A.; Samwald, M. Neural sentence embedding models for semantic similarity estimation in the biomedical domain. *BMC Bioinform.* **2019**, *20*, 178. [CrossRef]
26. Elbarougy, R.; Behery, G.; El Khatib, A. Extractive Arabic Text Summarization Using Modified PageRank Algorithm. *Int. Conf. Adv. Mach. Learn. Technol. Appl.* **2019**. [CrossRef]
27. Deng, X.; Li, Y.; Weng, J.; Zhang, J. Feature selection for text classification: A review. *Multimed. Tools Appl.* **2019**, *78*, 3797–3816. [CrossRef]
28. Mosa, M.A.; Anwar, A.S.; Hamouda, A. A survey of multiple types of text summarization with their satellite contents based on swarm intelligence optimization algorithms. *Knowl. -Based Syst.* **2019**, *163*, 518–532. [CrossRef]
29. Adhvaryu, N.; Balani, P. Survey: Part-Of-Speech Tagging in NLP. In Proceedings of the International Journal of Research in Advent Technology (E-ISSN: 2321-9637) Special Issue 1st International Conference on Advent Trends in Engineering, Science and Technology "ICATEST 2015", Amravati, Maharashtra, India, 8 March 2015.
30. Abuobieda, A.; Salim, N.; Albaham, A.T.; Osman, A.H.; Kumar, Y.J. Text summarization features selection method using pseudo genetic-based model. In Proceedings of the 2012 International Conference on Information Retrieval & Knowledge Management, Kuala Lumpur, Malaysia, 13–15 March 2012; pp. 193–197.
31. Al-Saleh, A.B.; Menai, M.E.B. Automatic Arabic text summarization: A survey. *Artif. Intell. Rev.* **2016**, *45*, 203–234. [CrossRef]
32. Li, H. Multivariate time series clustering based on common principal component analysis. *Neurocomputing* **2019**, *349*, 239–247. [CrossRef]