Journal of King Saud University – Computer and Information Sciences xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

An efficient single document Arabic text summarization using a combination of statistical and semantic features

Aziz Qaroush *, Ibrahim Abu Farha, Wasel Ghanem, Mahdi Washaha, Eman Maali

[a] Department of Electrical and Computer Engineering, Birzeit University, Palestine

ARTICLE INFO

Article history:
Received 3 October 2018
Revised 5 March 2019
Accepted 16 March 2019
Available online xxxx

Keywords:
Arabic language
Single document summarization
Machine learning
Score-based
Statistical
Semantic
NLP

ABSTRACT

The exponential growth of online textual data triggered the crucial need for an effective and powerful tool that automatically provides the desired content in a summarized form while preserving core information. In this paper, we propose an automatic, generic, and extractive Arabic single document summarizing method aiming at producing a sufficiently informative summary. The proposed extractive method evaluates each sentence based on a combination of statistical and semantic features in which a novel formulation is used taking into account sentence importance, coverage and diversity. Further, two summarizing techniques including score-based and supervised machine learning were employed to produce the summary and then assist leveraging the designed features. We demonstrate the effectiveness of the proposed method through a set of experiments under EASC corpus using ROUGE measure. Compared to some existing related work, the experimental evaluation shows the strength of the proposed method in terms of precision, recall, and F-score performance metrics.

1. Introduction

An automatic mechanism to summarize a text is available now in response to the dramatic increase online textual information via different resources including social websites, news agencies, etc. Currently, news agencies are publishing online news massively on daily basis. Admitting the fact that people these days have a busy life, they find it troublesome to read redundant texts. It is natural that humans tend to save their time and effort to access the most important/relevant and salient information in a document. For example, the authors in Modaresi et al. (2017) investigated the (commercial) benefits of the summarization systems in handling news articles. Their results indicated that incorporating even simple summarization systems (e.g query-based extractive approach) can dramatically save the processing time of the employees without significantly reducing the quality of their work.

For these reasons, automatic text summarization that started in 2001 has quickly grown into a major research area in the fields of Natural Language Processing (NLP) as illustrated by interests of Text Analysis Conference (TAC) and Document Understanding Conference (DUC) series. Text summarization proved to be beneficial in different domains such as medicine, legal proceedings, news circulation, and web pages (Hua et al., 2017). Hu and Liu proposed a system to summarize Amazon clients reviews (Hu and Liu, 2004). Meanwhile, Ya-Han Hu et al. proposed a summarization system for hotel reviews automatically (Hua et al., 2017). Tseng et al. employed a single-text summarization system that produces patent summaries(Tseng et al., 2007). Furthermore, Kallimani offered a score-based statistical method for summarizing news articles (Kallimani et al., 2012).

A summary can be defined as "a text which is produced from one or more texts and conveys core information in the original texts; typically, it is no longer than half of the original text(s) and usually less than that (Radev et al., 2002). There are often several related parameters, features, and properties that determine different types or categories of text summarization. The main parameters used in classifying text summarization are the number of source or input documents (span), the number of languages in the document, details of a summary (summary length), targeted audience, and summary formation (Hovy and Lin, 1998; Radev et al., 2011; Al-Saleh and Menail, 2016; Lagrini et al., 2017).

* Corresponding author.
E-mail addresses: aqaroush@birzeit.edu (A. Qaroush), iabufarha@birzeit.edu (I. Abu Farha), ghanem@birzeit.edu (W. Ghanem), mahdi.washaha@gmail.com (M. Washaha), emaali@birzeit.edu (E. Maali).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2019.03.010
1319-1578/

Please cite this article as: A. Qaroush, I. Abu Farha, W. Ghanem et al., An efficient single document Arabic text summarization using a combination of statistical and semantic features, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.03.010

For instance, the span parameter differentiates between a single-document summary where the summary is generated from only one document or a multi-document summary where the summary is spawned from a group of related documents. Furthermore, based on the number of languages, summarization systems might be monolingual, if they summarize documents written in one language only, or multi-lingual if they can summarize documents written using at least two different languages. With respect to details of summary parameters, the summary could be an indicative one when the most important idea of the document is preserved in a way that helps the user to get the main idea of the text; in addition, the summary could be informative when it is intended to cover all important topics or details cutting the word count. Based on audience parameter, it could be a generic summary where all information/topics are equally important or it could be a query-based (topic-based) summary where it relies on an initially submitted user query to summarize the available related documents. Finally, summary formation mechanisms yield either an extractive or an abstractive summary.

To illustrate, summaries that are extractive in nature combine the significant fragments (important sentences) from the text based on some extracted features (statistical or/and linguistic) without any modification on the selected text. It is evident that extractive methods are easier to construct but their summaries are less readable, and have less coverage and coherence, in addition to the higher probability of redundancy occurrence. On the other hand, abstractive summarization process strives to interpret and paraphrase the text based on the information extracted from the document or corpus using linguistic features or methods so as to generate novel coherent and grammatically correct sentences. Despite the fact that the summaries that are generated using linguistic methods look more human-like and produce more condensed summaries, these techniques are much harder to implement compared to extractive techniques; hence, researchers are motivated to focus on extractive summarization approaches.

Researches on forming Arabic text summaries have not been done sufficiently when compared to the research accomplished in English or other languages. This is due to some issues and challenges that slow down the progress in Arabic Natural Language Processing. These challenges were inherited from the complexity of the Arabic language and the lack of automated Arabic NLP tools. These complications can be briefed in Al-Saleh and Menail (2016): (i) Arabic is a highly derivational and inflectional language, which makes morphological analysis such as lemmatization and stemming a very complex task, (ii) Arabic lacks capitalization leading to a great challenge in the process of Named Entity Recognition (NER) system, (iii) the absence of Diacritics called "Tashkeel" that is integral in Arabic texts increases the complexity of inferring s' meaning, (iv) Arabic is considered highly ambiguous in comparison to other languages, (v) and the lack of Arabic corpora besides essential automated Arabic NLP tools such as lexicons, semantic role labelers, and named entity recognition complicate the process more.

Several methods were presented in the recent research literature for Arabic single-document extractive text summarization. However, these methods focus on one or some text summarization objectives including content coverage, diversity between sentences, readability, and compression ratio. In addition, the previous studies didn't provide sufficient analysis and formulations regarding the features used by summarization methods. Moreover, the presented studies for Arabic single document summarization are below the desired level of performance compared to other languages. In this paper, we present a generic, extractive, single document summarization method aiming at maximizing content coverage and diversity between sentences within the summary. The proposed method evaluates each sentence based on a combination of the most informative statistical and semantic features using a novel formulation to achieve both contradictory semantics objectives namely coverage and diversity. In addition, two summarization techniques including score-based and supervised machine learning are used to test the strength of these features. The effectiveness of the proposed method can be demonstrated through intensive experiments conducted on Essex Arabic Summaries Corpus (EASC) (EL-Haj et al., 2010). Hence, the main contributions of this paper are as follows: First: the proposed method is domain independent that does not need any domain-specific knowledge or features. Second: it presents a novel formulation of the most informative statistical and semantic features to produce an information-rich summary. Third: the study investigates the performance of the proposed combination of features using two summarization techniques i.e. score-based and machine learning techniques. Finally: compared to the state-of-the-art methods, the experimental evaluation shows the efficiency of the proposed work in terms of precision, recall, and F-measure.

The rest of the paper is structured as follows: Section two gives insights into the state-of-the-art Arabic extractive text summarization techniques and some relevant systems. Section three presents the problem definition along with its formulation. The design of the proposed method is described in section four. The data set, evaluation measures, tools, experimental setup and a series of conducted experiments are described in section five. Finally, section six concludes the work by providing perspectives.

2. Related works

Several techniques are proposed in the literature for single document text summarization. These techniques are categorized under a set of approaches including semantic-based, statistical-based, machine learning-based, cluster-based, graph-based, discourse-based summarization, and an optimization-based approach taking into consideration the large overlapping between these techniques (El-Haj, 2012; Lagrini et al., 2017; Al-Saleh and Menai, 2018; Qassem et al., 2017).

2.1. Semantic-based summarization

XThe semantic analysis is greatly concerned with the meaning of the words as well as the connections/relations between words, phrases, and sentences to construct the intended concepts of the text. Several semantic analysis techniques can be applied to summarize texts including lexical chains and natural language processing methods such as latent analysis (Barzilay and Elhadad, 2015; Ozsoy et al., 2011). I. Imam et al. have utilized users desired query keywords or topics to generate the summary of the original text (Imam et al., 2013). In addition to the statistical techniques, the method applies linguistic analysis such as part of speech tagging. The user is asked to enter a query which determines the desired field that the user is interested in. This query is expanded using the Arabic WordNet. Then, the user is asked to finalize the expanded form by removing irrelevant terms. The scoring of the sentences depends on the existing words in the original and expanded queries. The sentences with the highest scores are extracted to form the summary. AL-Khawaldeh and Samawi have applied both lexical cohesion and text entailment-based segmentation as scoring measures to prevent redundant and less important sentences from being generated in the summary (AL-Khawaldeh and Samawi, 2015).

The lexical cohesion is responsible for evaluating the importance of the certain sentence contribution to the summary; hence, poor sentences will be removed by dividing the text into tokens and using the lexical chains between tokens that have semantic relations. Then, possibly redundant important sentences are collapsed into

Please cite this article as: A. Qaroush, I. Abu Farha, W. Ghanem et al., An efficient single document Arabic text summarization using a combination of statistical and semantic features, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.03.010

one in the text entailment stage using directional cosine similarity and specified threshold values. T. Shishtawy et al. have also accomplished a combinational method of statistical and linguistic analysis (El-Shishtawy and El-Ghannam, 2012). They have used the key-phrases as attributes to evaluate the importance of the sentences within texts because key-phrases represent the most important concepts of the text. They have built their work on existing Arabic Key-phrase Extractor (AKE) with some modifications such as adding new sets of syntax rules. Indicative key-phrases are extracted from the input/processed text at a lemma level; lemma refers to the set of all word forms that have the same meaning. Then the extraction performed at levels of one, two, or three consecutive words. Following that, these phrases go through a filtering process according to syntactic rules. After that, some statistical features are extracted. The score of each sentence within the text is determined depending on the extracted key-phrases. The output summary is actually formed by extracting the top-ranked sentences within the specified summary length or percentage.

The use of these methods in automatic text summarization has contributed to the widely promoted quality by generating more coherent, less redundant and more informative summaries. However, it is a challenging task since it has difficulties in using high-quality semantic analysis tools and linguistic resources (WordNet, Lexical Chain, etc.) as they require memory for saving the semantic information like WordNet and processor capacity because of additional linguistic and semantic knowledge and complex linguistic processing (Khan, 2014).

2.2. Statistical-based summarization

Statistical approaches widely used in summarizing texts. The concept of relevance score which depends on the extraction of a set of features is the decisive factor that reflects the importance of a sentence regardless of its meaning. In Al-Hashemi (2010), the sentence selection depends on key-phrase extraction. The extracted key-phrases is based on some features like Term Frequency (TF), inverse document frequency (IDF), font types and their existence in the document title. The extracted key-phrases are then assessed to their ability to reflect sentence importance. Gholamerezazadeh, Fattah et al., Abuobieda et al., C. Nobata et al., Rajesh et al., Gupta et al., and Rafael et al. have used other features to score the sentence including indicator phrases, uppercase words, sentence length, similarity with the title, and sentence position in the document (Gholamrezazadeh et al., 2009; Fattah and Ren, 2009; Abuobieda et al., 2012; Nobata et al., 2009; Prasad and Kulkarni, 2010; Gupta and Pendluri, 2011; Ferreira et al., 2012; Abdelkrime et al., 2015; Litvak et al., 2016). In Abdelkrime et al. (2015) and Litvak et al. (2016) a weighted linear combination of statistical features is used for sentence ranking. In addition, they obtained the optimal weights using a genetic algorithm (GA). Using statistical features alone might not provide good results, because they don't take into consideration the meaning of the words and the relations between them as well as the relations between the sentences themselves. Furthermore, another expected problem is redundancy in the selected sentences. Bearing this in mind, this approach might yield better results if it is combined with other approaches. For example, T. El-Shishtawy et al. have built a Key-phrase Based Arabic Summarizer. The system uses a combination of semantic features and some statistical features to identify the key-phrases (El-Shishtawy and El-Ghannam, 2012; El-shishtawy et al., 2012). These features are phrase relative frequency (PRF), word relative frequency (WRF), sentence location, phrase location, sentence length, and phrase length. In addition, many different systems use the statistical features in order to enhance their results. Schlesinger et al. employed statistical features to enhance the selection or elimination of sentences prior to the summarization process (Schlesinger et al., 2008). Statistical-based approaches are easy to implement and can be used to enhance the selection of important sentences or for the elimination of redundancy. However, it fails to understand the text since it sometimes only depends on statistical measures.

2.3. Machine learning-based summarization

In supervised machine learning based approach, extractive text summarization process is modeled as a binary classification problem. It relies on a set of statistical features to train a binary classifier over a set of training documents along with their human extractive summaries. Each sentence in the document is represented as a vector of features that are extracted from different levels; token, sentence, paragraph, and document. The common features between these levels depend highly on term frequency, the position of the sentence in the paragraph or document, the similarity with the title, sentence length, etc. In this approach, the probability of a sentence to belong to the summary class is depicted by the score of the sentence itself. Fattah and Ren employed 10 features to train various machine learning methods including Support Vector Machine (SVM), Neural network, and Gaussian mixture models over a manually created corpus of 50 English documents and 100 Arabic documents (Fattah and Ren, 2009). Then, the trained classifier model was used to rank the sentences based on their score (the probability of a sentence to be in the summary class) to generate the final summary. In this regard, Boudabous et al. have trained a binary SVM classifier using 15 features over a manually created corpus of 500 Arabic newspaper articles on different topics (Boudabous and Belguith, 2010). Belkebir and Guessoum have proposed an extractive machine learning-based summarizer based on two stages using a set of statistical features extracted from each sentence (Belkebir, 2015). The first stage includes training of two classifiers AdaBoost and SVM. Then, in the second stage, AdaBoost enhances the SVM classifier to predict whether a sentence is a summary sentence or not. The authors collected their own corpus which is composed of 20 Arabic news articles along with their manually generated summaries. Machine learning methods have been shown to be very effective and successful in single and multi-document summarization. However, they need a set of training documents (labeled data) to train the classifier. In addition, their performance are affected by the chosen classifier, features, and features representation which play an important role in the performance of this approach.

2.4. Cluster-based summarization

The Clustering process aims at grouping objects into classes drawing on the similarities. While summarizing texts, the objects are the sentences, the classes are the clusters that the sentences belong to. In this approach, the formation of the summary is performed by selecting a sentence or more from each cluster based on the closeness to their cluster centroid (Froud et al., 2013; El-Gedawy, 2014; Fejer and Omar, 2014). Although clustering techniques are used to decrease the data redundancy by categorizing similar data, its generated summary may not be meaningful enough since the selected sentences are mainly ranked depending on the closeness to the cluster centroid; these sentences are computed through distance measures without paying any attention to the meaning of the text in the sentence or centroid.

2.5. Graph-based summarization

In this approach, the document is illustrated in a graph like the model. In this model, the nodes of the graph represent the sentences, while the links/edges between the connected nodes

Please cite this article as: A. Qaroush, I. Abu Farha, W. Ghanem et al., An efficient single document Arabic text summarization using a combination of statistical and semantic features, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.03.010

represent the similarity relation between sentences. Therefore, a sentence is considered important if it is strongly connected to many other sentences (Al-Taani et al., 2014; Erkan and Dragomir, 2004). LexRank (Erkan and Radev, 2004; Thomas et al., 2015) and TextRank (Mihalcea and Tarau, 2004) are two well-known graph-based ranking systems that are used in this approach. The use of graph-based methods has a positive contribution in multi-document research communities since it has the ability to capture distinct topics from unconnected sub-graphs. However, the construction of sub-graphs, depending on statistical similarity measures without paying any attention to the meaning of the text, risks the production of less-informative summary (Lagrini et al., 2017).

2.6. Discourse-based summarization

Discourse structure is essential in determining the content or the information conveyed by text. In this structure, instead of treating the text as a continuity of words and sentences, texts are represented or organized in a way where discourse-units are related to each other to ensure both discourse coherence and cohesion. Building successful discourse structures mainly depend on the availability of robust discourse parsers which rely on four factors including the type of discourse theory, the data structure used for representing structure (tree, or graph), the nature and the hierarchy of the relations (semantic, intentional or lexically grounded) and finally the language (Lagrini et al., 2017). There are several existing discourse theories that are used to represent or generate the discourse structure of text including the Rhetorical Structure Theory (RST) (Elghazaly and Ibrahim, 2012; Azmi and Al-Thanyyan, 2012), and Segmented Discourse Representation Theory (SDRT) (Keskes, 2015). Within a discourse, texts organized in a way such that discourse units are related to each other so as to achieve both coherence and cohesion. However, building automatic parsers for discourse information has proven to be a hard task and computationally expensive. In addition, discourse structure is only as useful for content selection as simpler text structure built using lexical similarity (Louis et al., 2010).

2.7. An optimization-based summarization

Text summarization considered by many researchers as a single/Multi- objective optimization problem, where a set of objectives considered to produce a high-quality summary including coverage, redundancy (diversity), coherence, and balance. Coverage means that summary should contain all important aspects appearing in the documents, while diversity aims to reduce the similar sentences in the output summary. On the other hand, coherence aims to generate a coherent text flow. Moreover, balance means that summary should have the same relative importance of different aspects of the original documents. However, searching for the optimal summary given these objectives is an NP-hard problem. Therefore, several methods had been used to approximate the solution including population-based methods (Alguliev et al., 2013; John et al., 2017), swarm intelligence (Alguliev and Aliguliyev, 2013; Alguliev et al., 2011), artificial bee colony (Sanchez-Gomez et al., 2017), ant colony (Mosa et al., 2017), and cuckoo search (Rautray and Balabantaray, 2018). Optimization-based approaches produce promising results; however, it needs more formulations besides being time-consuming.

To sum up, several approaches were presented in the literature for Arabic text summarization. Some of them such as cluster-based, graph-based, and optimization-based are more suitable for multi-document summarization. In addition, they distinct from each other in terms of their main goal such as identifying relevant sentences, reducing redundancy, or maximizing coverage and diversity. Researches on Arabic single-document extractive text summarization concentrate on one or more of these goals. However, they did not provide sufficient analysis and formulation regarding the features used by the summarization methods to provide rich information summary. Unlike these studies, our work focuses on deeply analyze and formulate these features taking into account properties of the Arabic text. In addition, we provide a combination of statistical and semantic features to identify the most relevant sentences to achieve both contradictory semantics objectives namely coverage and diversity.

3. Problem definition and formulation

The problem is defined and formulated as follows: Given an input Arabic single document D_{in} represented as a set of sentences $D_{in} = \{S_1, S_2, \cdots, S_n\}$ ordered based on their location in D_{in} where S_i corresponds to the i^{th} sentence in the document and n is the total number of sentences that comprise it. In addition, each sentence S_i in D_{in} represented as set of tokens (e.g. words) $S_i = \{t_1, t_2, \cdots, t_m\}$, where t_k is the k^{th} token in sentence S_i and m is the total number of tokens in the sentence S_i. Therefore, an automatic extractive text summarization system is a reductive/selective transformation of a single input text document D_{in} into an output document D_{out}, consisting of single or multiple target statements $D_{out} = \{S_1, S_2, \cdots, S_k\}$. This transformation process tries to achieve three main objectives: (i) the target statements (selected statements) must contain a significant portion of the information that exists in the original document, i.e. the main information, (ii) minimizing text redundancy while maximizing diversity and coherence in the summary, and (iii) the output document D_{out} has a size, i.e. number of statements, no longer than half of the input document (Radev et al., 2002). In order to achieve these objectives, a set of the most important statistical and semantic features $F = \{f_1, f_2, \ldots, f_i\}$ are employed to evaluate each sentence S_i to reflect its importance. Finally, the summary S_i is generated by combining the highest scored sentences based on the predefined summary ratio while considering text coherence.

4. Proposed work

The proposed extractive text summarization method consists of three main stages named: text preprocessing, features extraction, sentence evaluation, and selection stage. In the preprocessing stage, the document is prepared and represented in a structured/unified way to facilitate working on coming stages. In the second stage, a set of statistical and semantic features computed for each sentence to reflect its importance and used in sentence evaluation and selection stage where two different methods are used to assess the selected features and their formulation including score-based and supervised machine learning.

4.1. Text preprocessing

This stage is the initial stage in almost all summary methods. Its main purpose is to prepare the input text document for processing in other stages. It mainly transforms the input document into a unified representation. The proposed text summary system includes the following preprocessing sequenced operations: tokenization, letters normalization, stop-words removal, and stemming, as shown in Fig. 1 (Abdelkrime et al., 2015; Litvak et al., 2016; Thomas et al., 2015).

Tokenization

Text preprocessing starts with the tokenization process which split the input documents into their units with different levels to

Please cite this article as: A. Qaroush, I. Abu Farha, W. Ghanem et al., An efficient single document Arabic text summarization using a combination of statistical and semantic features, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.03.010

facilitate accessing all parts of the input document. These units are paragraphs, sentences, tokens, numbers, or any other appropriate unit (Attia, 2007). To illustrate, the proposed tokenization is a morphological decomposition based on punctuation which starts with finding the paragraphs that the document consists of, where the newline character (|n) is the paragraph delimiter. After that paragraphs are split into a set of sentences based on full stop (.), question mark (?), and exclamation mark(!) as delimiters. Finally, these sentences are divided into tokens based on delimiters like white space, semicolons, commas, and quotes. We employed AraNLP tool with little modification to handle the above sequence of operations (Althobaiti et al., 2014).

Normalization

In the Arabic language, some Arabic letters might appear in different forms, while other characters are used instead of others because their shapes are similar. Moreover, writers use diacritics in their texts. These create a set of variations for the same term; and thus affect the computation of some features such as Term Frequency (TF). Therefore, a normalization process is required to unify the different forms of the same letter to avoid such variations. The proposed normalization step employs AraNLP tool to do the following tasks (Althobaiti et al., 2014): (i) removing non-Arabic letters such as special symbols and punctuations, (ii) removing diacritics, (iii) replacing آ ، إ ، أ with ا, ي with ى, and ة with ه (Ayedh et al., 2016). and (iv) removing tatweel (stretching character).

Stop-word removal

Stop words (i.e. pronouns, prepositions, conjunctions, etc.) are insignificant words that frequently appear in the documents to form sentences (Kanan et al., 2004). Since these words are not informative (do not add information), they can be eliminated from sentences without affecting the core content of the sentence. Indeed, this step is crucial since some of the calculations are based on the words' frequencies in the sentence/document. Thus, by removing stop words, these calculations become more relevant and accurate. There are several stop-list methods that are used to remove stop-words from the text including, General Stop-list, Corpus-Based Stop-list, and Combined Stop-list. The proposed method depends on general stop-list using AraNLP tool, which performed better than the other two methods (El-Khair, 2006; Althobaiti et al., 2014).

Stemming

Arabic is a highly inflectional and derivational language, which means that Arabic words can have many different forms but share the same abstract meaning of action. This has, evidently, affected many natural languages processing methods such as building bag-of-word model and text similarity calculation. Therefore, Stemming is the process of removing some or all affixes (e.g. prefixes, infixes, postfixes, and suffixes) from a word. In other words, stemming transforms the different forms/derivatives of a word to a single unified form (e.g. root or stem) from which all the derivatives are generated. In Arabic, there are two common stemming approaches; Morphological root-based stemming and light stemming (Mustafa et al., 2017). The work presented in Alami et al. (2016) compares between these approaches regarding text summarization using two well-known Arabic stemmers including Khoja root stemmer[1]). Their experiments showed that, in Arabic text summarization, root stemming is preferred to light stemming. Based on those finding, we adapted a Khoja root stemmer to handle the stemming operation as a preprocessing task for the proposed work. Fig. 2 shows the output of the proposed preprocessing methods on a sample input text.

[1] http://zeus.cs.pacificu.edu/shereen/research.htm and Larkeys light stemmer (Larkey et al., 2007

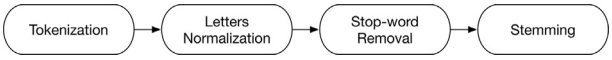

Fig. 1. Sequence of the preprocessing steps.

4.2. Feature extraction and formulation

An extractive-based text summary that involves selecting sentences of high relevance or importance is based on employing a set of features to generate coherent summaries that state the main idea of the given document. Therefore, selecting and designing these features will greatly affect the quality of the generated summaries. A large number features are proposed for automatic extractive text summarization by various researchers (Ferreira et al., 2012; Meena and Gopalani, 2014; Prasad and Kulkarni, 2010; Kiyoumarsi, 2014; Neto et al., 2002; Al-Saleh and Menail, 2016; Mendozaab et al., 2014; Prasad et al., 2012). These features are classified into four levels including word-based level, sentence-based level, paragraph-based level, and graph-based features. Since the quality of the generated extractive summary is highly affected by the selected features along with their design, our target in this paper is to redesign the most important or prominent features that identify the most important sentences in addition to maximizing content coverage and diversity between sentences within the summary. Using statistical features alone may not provide rich information summary, because they don't take into consideration the meaning and may cause some redundancy in the generated summary. On the other hand, relying on semantic features alone will not capture very important statistics like TF-ISF. Therefore, to handle these shortcomings, a combination of these types were used El-Shishtawy and El-Ghannam (2012) and El-shishtawy et al. (2012). Table 1 summarizes the selected features along with their level, category, and contribution in the quality of the generated summary in terms of sentence importance, coverage, and diversity. The selection and formulation are based on some hypothesis, our observations/analysis, set of experiments, and some previous studies (Ferreira et al., 2012; Meena and Gopalani, 2014; Meena and Gopalani, 2016; Meena et al., 2015). The importance of explaining these features along with their design stems from the fact that there are some differences in the formulation of both summarization methods that are used to evaluate the performance of the selected features.

Key-phrases feature

Key-phrases is a short list of important and topical keywords that provide a condensed summary of the main topic in the document (Turney, 2000). They might be a single word or a composite of multiple words. Many applications in information retrieval, including text summarization, employ key-phrase extraction (Hasan and Ng, 2014; Najadat et al., 2016;7(2).). The possibility of having a core idea in a sentence is conditioned by containing a key-phrase/s. Indeed, this would increase its importance with respect to other sentences (El-Shishtawy and El-Ghannam, 2012; Sarkar, 2014). The score of the key-phrase feature depends on many factors, including the frequency of the candidate phrase, number of words in each phrase, frequency of the most recurring single word in a candidate phrase, location of the phrase within the document, location of the candidate phrase within its sentence, relative phrase length to its containing sentence, and assessment of the phrase sentence verb content (El-shishtawy et al., 2012). In this work, we used Kp-Miner (El-Beltagy and Rafea, 2009) tool to

Please cite this article as: A. Qaroush, I. Abu Farha, W. Ghanem et al., An efficient single document Arabic text summarization using a combination of statistical and semantic features, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.03.010

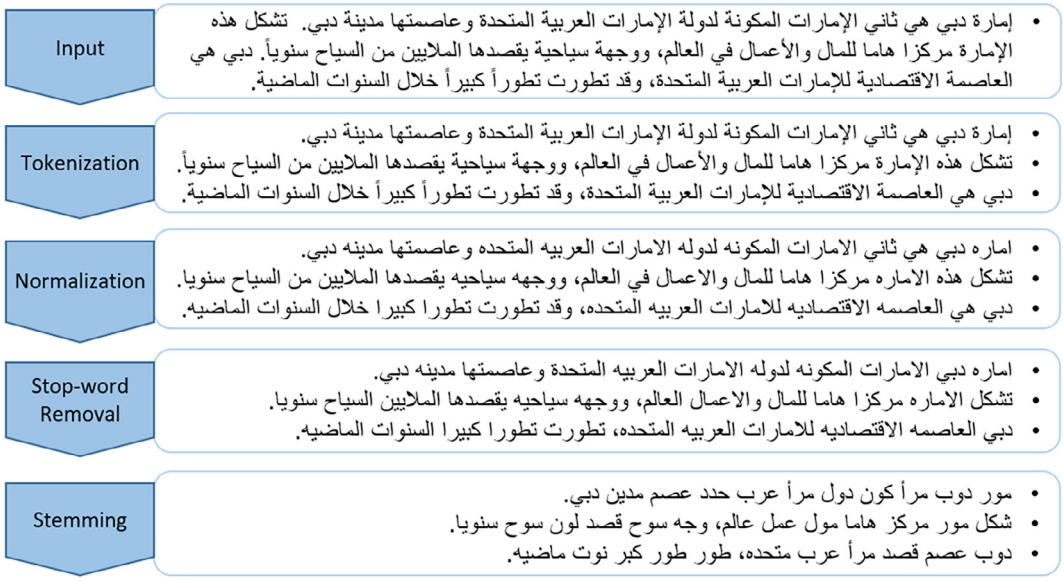

Fig. 2. Sample output of the text preprocessing methods proceeded in sequence.

Table 1
Description of the selected features along with their level, category, and contribution.

Feature Name	Brief Description	Level	Category	Contribution
Key-Phrases	A short list of important terms that provide a condensed summary of the main topics of a document.	Word-level	Statistical, Semantic	Coverage and diversity
Sentence location	Relating to the position of a sentence to the paragraph and document.	Paragraph-level	Statistical	Sentence relevance
Similarity with title	Similarity or overlapping between a given sentence and the document title.	Word-level	Statistical	Sentence relevance
Sentence centrality	The similarity or the overlapping between a sentence and other sentences in the document.	Graph-level	Statistical	Coverage and diversity
Sentence length	Counting the number of words in the sentence (can be used to classify sentence as too short or too long).	Sentence-level	Statistical	Sentence relevance and coverage
Cue words	Words in the sentence such as "therefore, finally and thus" can be a good indicators of significant content.	Word-level	Semantic	Sentence relevance and coverage
Positive key-words	Words that are used to emphasize or focus on special idea such as "have outstanding, and support for".	Word-level	Semantic	Sentence relevance
Sentence inclusion of numerical data	Existence of numerical data in the sentence.	Sentence-level	Statistical	Sentence relevance
Occurrence of Non-essential Information	Words that serve as an explanation words such as "for example"	Word-level	Semantic	Sentence relevance

extract key-phrases while the score of a key-phrase feature is computed based on three of the most important or prominent factors that can be defined as follows:

- *Key-phrase Frequency*: it indicates how many times the key-phrase appeared in the sentence and it is calculated by $KPF = \frac{\#SKP_i}{\#KP_d}$, where KPF is the key-phrases frequency, $\#SKP_i$ is the number of sentences that contain the key-phrase (KP_i), and $\#KP_d$ is the total number of key-phrases in the document.
- *Key-phrase Length*: it is the number of words that the key-phrase has. The length of the key-phrase plays a role in its importance and consequently the sentence importance. We found that long key-phrase is more important than shorter ones (El-shishtawy et al., 2012). The value of this feature is calculated as $\sqrt{KPL}$, where KPL represents the length of KP_i. The aim of using square root is to smoothly increase the score if the length is more than one term.
- *Proper Name*: the importance of a key-phrase increase if it is a proper name which is a noun corresponding to a particular person, place, or thing (Fattah and Ren, 2009; Nobata et al., 2009). In order to check if a key-phrase has a proper name, Stanford part of speech (POS) tagger is used (Adhvaryu and Balani, 2015).[2] Thus, if the key-phrase is a proper name then the value of this feature is set to 2. Otherwise, it is set to 1.

Using the aforementioned factors, key-phrases feature score is computed as following:

$$Key\ Phrases\ Score = \sum_{KP_i \in S_i} (KPF_i * \sqrt{KPL} * PNV) \quad (1)$$

where KPF_i is the key phrase frequency of KP_i, KPL is the length of KP_i, and PNV is the proper name value of KP_i. The above equation will give a higher score if the length of the key phrase is more than one or if it is a proper noun, and will give more score if both factors are found. For machine learning method, key-phrases concept formulated as a three features defined as: (i) key-phrase frequency which is calculated in same way of the score-based approach, (ii) key-phrase length which represented as a binary value indicating whether the sentence contains a key-phrase consisting of multiple words, and (iii) proper name key-phrase represented as a binary

[2] https://stanfordnlp.github.io/CoreNLP/index.html.

value acting as an indicator to whether the sentence contains a key-phrase of proper name type.

Sentence location feature

This feature has been firstly proposed by Baxendle (Baxendale, 1958) for sentence assessment where the importance of a sentence is dependent on its location in the paragraph/document regardless of the document domain/topic (Lin and Hovy, 1997; Gupta and Pendluri, 2011). In extractive text summarization, several formulations proposed regarding sentence location (Abuobieda et al., 2012; Gupta and Pendluri, 2011; Baxendale, 1958; Radev et al., 2004; Barrera and Verma, 2012; Bossard et al., 2008; Prasad and Kulkarni, 2010). These formulations are modeled based on one or more of the following four hypothesis: (i) the first paragraph and the last paragraph are important since they provide a summary about the whole document, (ii) in each paragraph, the first sentence and the last sentence are very important and strong candidates to be included in the summary, (iii) the first sentence in the first paragraph is the most important sentence and this assumption is mostly considered while processing news data being treated as the baseline summary (Saggion and Poibeau, 2013; EL-Haj et al., 2010), and (iv) sentences that are away from the beginning of the document are less important. Therefore, based on these observations, we formulated the sentence location score using the following rules:

Sentence Location Score

$$= \begin{cases} 3 & \text{for the first sentence in the first paragraph} \\ 2 & \text{for the first sentence in the last paragraph} \\ 1 & \text{for the first sentence in any paragraph} \\ \frac{1}{\sqrt{S_{in}}} & \text{for the first/last pargraph} \\ \frac{1}{\sqrt{S_{in}+P_{in}^2}} & \text{for any paragraph excluding the first and last ones} \end{cases} \tag{2}$$

where S_in is the index of the sentence S_i in the paragraph P_i, and P_{in} is the index of the paragraph P_i in the document.

These values are chosen with respect to the importance of other used features and their formulation. For machine learning method, we modeled sentence location using five features where each of them is represented as follows: (i) first sentence in the first paragraph, the first sentence in the last paragraph, and first sentence in any paragraph excluding first and last paragraphs will be represented as a binary value that indicates if the condition is valid or not, and (ii) any sentence in any paragraph excluding first and last paragraphs, and any sentence in the first or last paragraph will be calculated using Eq. (2).

Similarity with title feature

This feature has been firstly proposed by Edmundson (Edmundson, 1969) and defined as the similarity or the overlap between a given sentence and the document title (Fattah and Ren, 2009; Abuobieda et al., 2012; Nobata et al., 2009). The importance of this feature comes from the idea that if a sentence consists of words appearing in the title, then it might be an important sentence. Moreover, if a sentence shares a key-phrase with the title, this will significantly increase its score. Therefore, the title-similarity score for a sentence is computed using the following equation:

$$Title\_Similarity\ Score = Similarity(Title, S_i) * \sqrt{1 + (KPT \cap KPS_i)} \tag{3}$$

where S_i is the current Sentence, KPT is the list of Key-Phrases that appear in document's title, KPS_i is the list of Key-Phrases extracted from Sentence S_i, $(KPT \cap KPS_i)$ is the number of common key-phrases between S_i and T. The aim of using square root is to smoothly increase the score if the intersection is realized in more than one. Finally, $Sim(Title, S_i)$ is the degree of similarity between S_i and the document's title computed by a cosine similarity measure, which is a well-known text similarity method (Gomaa and Fahmy, 2013; Qazvinian et al., 2008; Shareghi and Hassanabadi, 2008). To compute the similarity, the sentence and the title are represented using the bag-of-words model. In this model, each sentence S_i is represented as an N-dimensional vector $Si = \{w_{i1}, w_{i2}, \cdots, w_{ik}, \cdots, w_{in}\}$, where w_{ik} is the weight of term t_k that exists in the sentence S_i, and n is the number of all possible unique words in the target document. Therefore, based on this representation, the cosine similarity can be computed as follows:

$$Cosine\ Similarity(S_i, T) = \frac{\sum_{w \in S,T} tf_{w,S} * tf_{w,T} (isf_w)^2}{\sqrt{\sum_{S_i \in S} (tf_{S_iS} * isf_{S_i})^2} \sqrt{\sum_{T_i \in T} (tf_{T_iT} * isf_{T_i})^2}} \tag{4}$$

where tf_{w,S_i} is the frequency of word w in sentence S, which is defined as $tf_{w,S_i} = 1 + \log(tf_{w,S_i})$, and isf_w is the Inverse Sentence Frequency which is a special version of Inverse Document Frequency (IDF) that measures how much information a term provides. Thus, the term is considered important if it is dense in the given sentence and rare in the entire document (Doko et al., 2013). Inverse sentence frequency is defined as $isf_w = \log \frac{N}{1+|S_i \in S : w \in S_i|}$, where N is the number of sentences in the document and $|S_i \in S : w \in S_i|$ is the number of sentences where the word w appears (Patil et al., 2011). For machine learning method, title feature is formulated depending on two features: the first one is computed using cosine similarity measure, as defined in Eq. (4) and the second one is represented as a binary value indicating the possibility for the sentence to share key-phrases with the title or not.

Sentence centrality feature

This feature is defined as the similarity or the overlap between a sentence and other sentences in the document. Thus, a sentence might be central in the document, and many sentences might explain it. Thus, a sentence is given a high score when its words occur in a greater number of other sentences in the document. Employing centrality feature will eliminate the problem of sentence redundancy and thus increases diversity (Abuobieda et al., 2012; Qazvinian et al., 2008; Shareghi and Hassanabadi, 2008; Prasad and Kulkarni, 2010; Mendozaab et al., 2014). The adopted method in computing centrality score starts by computing the similarity matrix using cosine similarity measure similar to Eq. (4) where each item in the matrix represents the similarity between the corresponding sentences pair (Al-Gaphari et al., 2013; Erkan and Radev, 2004; ChoSeoung and Kim, 2015). Since we are interested in significant similarities, we can eliminate some low similarity values in the similarity matrix by defining a threshold value (i.e. 0.1). After eliminating low similarities, the centrality score feature is computed and normalized using Eq. (5), where *similarity degree of* S_i represents the number of sentences that are similar to Si with a similarity value above a threshold. Also, it is calculated the same way as in Eq. (5) for the machine learning method.

$$Centrality\ Score_{Si} = \frac{Similarity\ degree\ of\ S_i}{Maximum\ similarity\ degree\ in\ the\ documnt} \tag{5}$$

Sentence length feature

The length of a sentence might affect its importance. Thus, too long or too short sentences might be excluded from the summary (Kupiec et al., 1995; Fattah and Ren, 2009; Neto et al., 2002; Gupta

Please cite this article as: A. Qaroush, I. Abu Farha, W. Ghanem et al., An efficient single document Arabic text summarization using a combination of statistical and semantic features, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.03.010

et al., 2012). Indeed, a too long sentence will increase its information content; however, this is not always the case since there is a constraint. Mainly, the limit is exceeded when a sentence becomes an over-detailed one i.e. it might be an explanation for another one. Also, short sentences tend to include less information compared to other sentences and thus they are less important. We employ the Interquartile Range (IQR) statistical method to identify outlier sentences based on their length in order to penalize sentences that are too short or too long. Therefore, very short and very long sentences are given a score equal to 0. For other sentences, their scores are calculated and normalized using Eq. (6). For machine learning method, two features are derived. The first one is represented as a binary value that indicates if the current sentence is very short/long based on the proposed IQR method. The second one computes the normalized sentence length using Eq. (6).

$$Length\ Score = \frac{\#\ words\ in\ the\ sentence}{\#\ words\ in\ the\ longest\ sentence} \quad (6)$$

Cue-phrases feature

The existence of some phrases in the sentence such as الاهم، بالتحديد كملخص ، كنتيجة، الافضل (in English: as a summary, as a result, the most important, precisely) can be a good indicator of the importance of the content (Edmundson, 1969; Fattah and Ren, 2009; Gupta and Pendluri, 2011; Prasad and Kulkarni, 2010; Prasad et al., 2012; Lakshmi et al., 2015; Barzilay and Elhadad, 2015). Thus, sentences that contain these phrases are given a higher score compared to other sentences. The score of this feature is computed and normalized using Eq. (7). Also, it is calculated the same way as in Eq. (7) for machine learning method.

$$Cue\_Phrase\ Score = \frac{\#\ Cue\ phrases\ in\ the\ sentence}{\#\ Cue\_phrases\ in\ the\ document} \quad (7)$$

Strong words feature

Some words such as وثّق ، أكّد (in English: trust, he stressed) are used to emphasize or focus on a core idea in the sentence (Fattah and Ren, 2009). Thus, the existence of these words in a sentence must increase its score. The score of this feature is computed and normalized using Eq. (8). The machine learning method is defined as a binary value that indicates whether the sentence contains strong words or not.

$$Strong\_Word\ Score = \frac{\#\ strong\ words\ in\ the\ sentence}{\#\ Strong\ words\ in\ the\ document} \quad (8)$$

Existence of numerical data

The existence of numerical data rather than enumerations or bullets such as numbers, dates, and time can affect the importance of a sentence (Fattah and Ren, 2009; Meena and Gopalani, 2014). Indeed, they might point to some important stats of the core idea or some result in statistical form, and this might increase the importance of the sentence. The normalized score of this feature is computed using Eq. (9). Also, it is calculated the same way as in Eq. (9) for the machine learning method.

$$Numbers\ Score = \frac{\#\ occurrences\ of\ numbers\ in\ the\ sentence}{\#\ occurrences\ of\ numbers\ in\ the\ document} \quad (9)$$

Occurrence of non-essential information

Some phrases like بالاضافة، على سبيل المثال (in English: in addition, for example) are speech markers serving as explanation words. Such phrases weaken the sentence because they imply that the coming sentence is an extra information with respect to the core idea (Neto et al., 2002; Gupta and Lehal, 2010). Therefore, the score of sentences that have these phrases must be decreased. To set the score of the explanation phrases, the following rules should be considered: (i) the score of explanation phrase set to −2, if the first word in the sentence is an explanation word, (ii) the score of explanation words other than the first word is calculated and normalized using Eq. (10). For machine learning method, the following two features are derived. The first one is represented as a binary value to indicate that the first word in the sentence is an explanation word or not. The second feature is the score of other explanation words and can be calculated by Eq. (10).

$$Weak\ Words\ Score = \frac{\#\ weak\ words\ in\ the\ sentence}{\#\ words\ in\ the\ sentence} \quad (10)$$

4.3. Sentence extraction and summary generation

In extractive text summarization, important text segments (e.g. sentences) of the original document usually are identified based on a set of important features extracted from different levels (e.g. tokens, sentence, paragraph, and document). In this paper, two summarization methods have been used to evaluate the effectiveness of the proposed features including greedy score-based method and a supervised learning method. These methods are widely used in extractive single document summarization and their results are significantly affected by the chosen features along with their design and formulation. In the score-based method, important sentences are extracted based on the total scores that are assigned to them. According to Meena and Gopalani (2016), which compares the performance of different sentence-based voting methods (e.g., BordaFuse, CombMNZ, expCombANZ, etc.), we adopt a weighted linear sum of normalized features score to evaluate each sentence in the document as defined in Eq. (11):

$$Sentence\ Score = \sum_{i=1} W_i * S_i \quad (11)$$

where S_i and W_i represent the weight and the score of $feature_i$ defined previously. W_i set to one because we take into consideration the importance and contribution of each feature during the formulation stage. For example, all features except key-phrase feature, title similarity, and sentence location have a value between 0 and 1. The other features are more important and have a value as following: sentence location have values either 3 or 2 or 1 or less than 1, key-phrase and title similarity have a value greater than or equal 0. These values reflect the contribution of the feature in the total score and for this we assign the weigh to be 1. After computing the total score, sentences are ranked in a descending order based on their total scores. Fig. 4.a shows the order of the sentences along with their total score of the input document shown in Fig. 3 after passing this stage. After that, top-ranked sentences will be selected to be included in the output summary based on the required summary ratio. Indeed, the sentences that have the highest scores will be representing the most important content of the document (e.g document main idea), and thus will be selected to be included in the final summary. Finally, the extracted sentences will be reorder based on their original position on the document to preserve text coherency in the generated summary. Fig. 5.a shows the golden standard summary of the input document shown in Figs. 3 and 5. b shows the generated summary by the score-based method. Algorithm 1 as shown in Fig. 6 summarizes the procedure of score-based method.

In the machine learning approach, the extractive summarization process is modeled as a binary classification problem as shown in Fig. 7. In this model, after text preprocessing, each sentence is represented by a feature vector of size 20 based on the features described and formulated in the previous section. Then, a binary (Yes/No) classifier is trained based on a set of training documents

Please cite this article as: A. Qaroush, I. Abu Farha, W. Ghanem et al., An efficient single document Arabic text summarization using a combination of statistical and semantic features, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.03.010

إمارة دبي هي ثاني الإمارات المكونة لدولة الإمارات العربية المتحدة وعاصمتها مدينة دبي. تشكل هذه الإمارة مركزا هاما للمال والأعمال في العالم، ووجهة سياحية يقصدها الملايين من السياح سنوياً. دبي هي العاصمة الاقتصادية للإمارات العربية المتحدة، وقد تطورت تطوراً كبيراً خلال السنوات الماضية. الاقتصاد الحر والنشط في الإمارة وعدم وجود نظام ضريبي لعب دورا كبيرا في جذب المستثمرين من جميع أنحاء العالم.
وتقع إمارة دبي بين إمارتي أبو ظبي والشارقة. وأهل إمارة دبي ينحدرون من قبائل عربية متنوعة، على رأسها قبيلة آل بو فلاسه التي تنحدر منها أسرة آل مكتوم الحاكمة. وتقطنها قبائل بني كعب وآل بو فلاح وآل بو مهير والسودان والشوامس والبلوش والمناصير والرميثات والشحوح وغيرهم.
وبها عوائل كثيرة من أصول أفريقية وفارسية. ودين أهالي دبي هو الإسلام على نهج أهل السنة والجماعة، والمذهب الرسمي في دبي هو المذهب المالكي.
آل مكتوم هم حكام دبي. وهم من آل بو فلاسه من بني ياس. حاكمها الأن هو الشيخ محمد بن راشد آل مكتوم. وهو أيضا نائب لرئيس الدولة ورئيس لمجلس الوزراء في الحكومة الاتحادية. ونائبيه في الحكم هما: شقيقه الشيخ حمدان بن راشد ال مكتوم وزير المالية والصناعة والشيخ مكتوم بن محمد بن راشد آل مكتوم. بينما يتولى منصب ولاية العهد بالإمارة الشيخ حمدان بن محمد بن راشد آل مكتوم رئيس المجلس التنفيذي للإمارة.
يرأس المجلس التنفيذي لحكومة دبي الشيخ حمدان بن محمد بن راشد آل مكتوم. ويجمع هذا المجلس في عضويته جميع مدراء الدوائر في حكومة دبي حيث يعقدون اجتماعاتهم الدورية لتسيير شؤون الإمارة.

The Emirate of Dubai is the second Emirate of the United Arab Emirates and Dubai its capital. This Emirate is an important financial and business center in the world and a tourist destination for millions of tourists annually. Dubai is the economic capital of the United Arab Emirates and has developed significantly over the past years. The free and active economy of this emirate and the absence of a tax system played a major role in attracting investors from all over the world.
Emirate of Dubai is located between Abu Dhabi and Sharjah. The people of the emirate of Dubai come from various Arab tribes, led by the tribe of the Al-Felisha, which descends from the ruling Al-Actium family. The tribes of Bani Ka'ab, Al Bo Falah, Al Bu Muheir, Sudan, Shawams, Baloch, Manasir, Rumaythat, Al Shahouh and others, inhabit it. With many families of African and Persian origin. The religion of the people of Dubai is Islam on the approach of the Sunnis and the community, and the official doctrine in Dubai is the Maliki school.
Al Actium is the rulers of Dubai. They are from Al Bu Falsa from Bani Yes. The ruler now is Sheikh Mohammed bin Rashid Al Actium. He is also Vice President and Prime Minister of the Federal Government. His two deputies in the government are his brother Sheikh Hamdan bin Rashid Al Actium, Minister of Finance and Industry and Sheikh Actium bin Mohammed bin Rashid Al Actium. While Sheikh Hamdan bin Mohammed bin Rashid Al Actium, the Chairman of the Executive Council of the Emirate, is the Crown Prince. The Dubai Executive Council is headed by Sheikh Hamdan bin Mohammed bin Rashid Al Actium. This council brings together all the directors of the departments in the Government of Dubai, where they hold regular meetings to manage the affairs of the Emirate.

Fig. 3. Sample input document along with its English translation.

5.41	إمارة دبي هي ثاني الإمارات المكونة لدولة الإمارات العربية المتحدة وعاصمتها مدينة دبي.
4.53	يرأس المجلس التنفيذي لحكومة دبي الشيخ حمدان بن محمد بن راشد آل مكتوم.
3.54	وأهل إمارة دبي ينحدرون من قبائل عربية متنوعة، على رأسها قبيلة آل بو فلاسه التي تنحدر منها أسرة آل مكتوم الحاكمة.
2.91	بينما يتولى منصب ولاية العهد بالإمارة الشيخ حمدان بن محمد بن راشد آل مكتوم رئيس المجلس التنفيذي للإمارة.
2.71	ويجمع هذا المجلس في عضويته جميع مدراء الدوائر في حكومة دبي حيث يعقدون اجتماعاتهم الدورية لتسيير شؤون الإمارة.
2.63	تشكل هذه الإمارة مركزا هاما للمال والأعمال في العالم، ووجهة سياحية يقصدها الملايين من السياح سنوياً.
2.27	الاقتصاد الحر والنشط في الإمارة وعدم وجود نظام ضريبي لعب دورا كبيرا في جذب المستثمرين من جميع أنحاء العالم.
2.26	وتقطنها قبائل بني كعب وآل بو فلاح وآل بو مهير والسودان والشوامس والبلوش والمناصير والرميثات والشحوح وغيرهم.
2.24	ونائبيه في الحكم هما: شقيقه الشيخ حمدان بن راشد آل مكتوم وزير المالية والصناعة والشيخ مكتوم بن محمد بن راشد آل مكتوم.
2.18	دبي هي العاصمة الاقتصادية للإمارات العربية المتحدة، وقد تطورت تطوراً كبيراً خلال السنوات الماضية.
2.04	ودين أهالي دبي هو الإسلام على نهج أهل السنة والجماعة، والمذهب الرسمي في دبي هو المذهب المالكي.
1.64	آل مكتوم هم حكام دبي.
1.46	وتقع إمارة دبي بين إمارتي أبو ظبي والشارقة.
1.34	وهم من آل بو فلاسه من بني ياس.
1.11	حاكمها الأن هو الشيخ محمد بن راشد آل مكتوم.
0.74	وهو أيضا نائب لرئيس الدولة ورئيس لمجلس الوزراء في الحكومة الاتحادية.
0.65	وبها عوائل كثيرة من أصول أفريقية وفارسية.

a. Order of sentences based on their score

No	إمارة دبي هي ثاني الإمارات المكونة لدولة الإمارات العربية المتحدة وعاصمتها مدينة دبي.
Yes	تشكل هذه الإمارة مركزا هاما للمال والأعمال في العالم، ووجهة سياحية يقصدها الملايين من السياح سنوياً.
No	دبي هي العاصمة الاقتصادية للإمارات العربية المتحدة، وقد تطورت تطوراً كبيراً خلال السنوات الماضية.
Yes	الاقتصاد الحر والنشط في الإمارة وعدم وجود نظام ضريبي لعب دورا كبيرا في جذب المستثمرين من جميع أنحاء العالم.
No	وتقع إمارة دبي بين إمارتي أبو ظبي والشارقة.
No	وأهل إمارة دبي ينحدرون من قبائل عربية متنوعة، على رأسها قبيلة آل بو فلاسه التي تنحدر منها أسرة آل مكتوم الحاكمة.
Yes	وتقطنها قبائل بني كعب وآل بو فلاح وآل بو مهير والسودان والشوامس والبلوش والمناصير والرميثات والشحوح وغيرهم.
Yes	وبها عوائل كثيرة من أصول أفريقية وفارسية.
Yes	ودين أهالي دبي هو الإسلام على نهج أهل السنة والجماعة، والمذهب الرسمي في دبي هو المذهب المالكي.
Yes	آل مكتوم هم حكام دبي.
Yes	وهم من آل بو فلاسه من بني ياس.
Yes	ونائبيه في الحكم هما: شقيقه الشيخ حمدان بن راشد آل مكتوم وزير المالية والصناعة والشيخ مكتوم بن محمد بن راشد آل مكتوم.
Yes	بينما يتولى منصب ولاية العهد بالإمارة الشيخ حمدان بن محمد بن راشد آل مكتوم رئيس المجلس التنفيذي للإمارة.
Yes	يرأس المجلس التنفيذي لحكومة دبي الشيخ حمدان بن محمد بن راشد آل مكتوم.
No	ويجمع هذا المجلس في عضويته جميع مدراء الدوائر في حكومة دبي حيث يعقدون اجتماعاتهم الدورية لتسيير شؤون الإمارة.

b. Predicted labels based on trained machine learning model

Fig. 4. An example of sentence reordering in score-based method and sentence prediction in machine-learning method.

Please cite this article as: A. Qaroush, I. Abu Farha, W. Ghanem et al., An efficient single document Arabic text summarization using a combination of statistical and semantic features, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.03.010

إمارة دبي هي ثاني الإمارات المكونة لدولة الإمارات العربية المتحدة وعاصمتها مدينة دبي. دبي هي العاصمة الاقتصادية للإمارات العربية المتحدة، وقد تطورت تطوراً كبيراً خلال السنوات الماضية. وتقع إمارة دبي بين إمارتي أبو ظبي والشارقة. حاكمها الآن هو الشيخ محمد بن راشد آل مكتوم. ونائبيه في الحكم هما: شقيقه الشيخ حمدان بن راشد آل مكتوم وزير المالية والصناعة والشيخ مكتوم بن محمد بن راشد آل مكتوم. يرأس المجلس التنفيذي لحكومة دبي الشيخ حمدان بن محمد بن راشد آل مكتوم.

The Emirate of Dubai is the second Emirate of the United Arab Emirates and Dubai its capital. Dubai is the economic capital of the United Arab Emirates and has developed significantly over the past years. Emirate of Dubai is located between Abu Dhabi and Sharjah. The ruler now is Sheikh Mohammed bin Rashid Al Maktoum. His two deputies in the government are his brother Sheikh Hamdan bin Rashid Al Maktoum, Minister of Finance and Industry and Sheikh Maktoum bin Mohammed bin Rashid Al Maktoum. The Dubai Executive Council is headed by Sheikh Hamdan bin Mohammed bin Rashid Al Maktoum.

a. Reference summary (Golden standard summary)

إمارة دبي هي ثاني الإمارات المكونة لدولة الإمارات العربية المتحدة وعاصمتها مدينة دبي .تشكل هذه الإمارة مركزا هاما للمال والأعمال في العالم ، ووجهة سياحية يقصدها الملايين من السياح سنوياً .وأهل إمارة دبي ينحدرون من قبائل عربية متنوعة ، على رأسها قبيلة آل بو فلاسه التي تنحدر منها أسرة آل مكتوم الحاكمة .بينما يتولى منصب ولاية العهد بالإمارة الشيخ حمدان بن محمد بن راشد آل مكتوم رئيس المجلس التنفيذي للإمارة .يرأس المجلس التنفيذي لحكومة دبي الشيخ حمدان بن محمد بن راشد آل مكتوم .ويجمع هذا المجلس في عضويته جميع مدراء الدوائر في حكومة دبي حيث يعقدوا اجتماعاتهم الدورية لتسيير شؤون الإمارة .

The Emirate of Dubai is the second Emirate of the United Arab Emirates and Dubai its capital. This Emirate is an important financial and business center in the world and a tourist destination for millions of tourists annually. The people of the emirate of Dubai come from various Arab tribes, led by the tribe of the Al-Falasah, which descends from the ruling Al-Maktoum family. While Sheikh Hamdan bin Mohammed bin Rashid Al Maktoum, the Chairman of the Executive Council of the Emirate, is the Crown Prince. The Dubai Executive Council is headed by Sheikh Hamdan bin Mohammed bin Rashid Al Maktoum. This council brings together all the directors of the departments in the Government of Dubai, where they hold regular meetings to manage the affairs of the Emirate.

b. Summary generated by score-based approach

تشكل هذه الإمارة مركزا هاما للمال والأعمال في العالم، ووجهة سياحية يقصدها الملايين من السياح سنوياً. الاقتصاد الحر والنشط في الإمارة وعدم وجود نظام ضريبي لعب دورا كبيرا في جذب المستثمرين من جميع أنحاء العالم. وتقطنها قبائل بني كعب و آل بو فلاح وآل بو مهير والسودان والشوامس والبلوش والمناصير والرميثات والشحوح وغيرهم .وبها عوائل كثيرة من أصول أفريقية وفارسية .ودين أهالي دبي هو الإسلام على نهج أهل السنة والجماعة ، والمذهب الرسمي في دبي هو المذهب المالكي .آل مكتوم هم حكام دبي .وهم من آل بو فلاسه من بني ياس .حاكمها الآن هو الشيخ محمد بن راشد آل مكتوم .وهو أيضا نائب لرئيس الدولة و رئيس لمجلس الوزراء في الحكومة الاتحادية .ونائبيه في الحكم هما: شقيقه الشيخ حمدان بن راشد آل مكتوم وزير المالية والصناعة والشيخ مكتوم بن محمد بن راشد آل مكتوم .بينما يتولى منصب ولاية العهد بالإمارة الشيخ حمدان بن محمد بن راشد آل مكتوم رئيس المجلس التنفيذي للإمارة .يرأس المجلس التنفيذي لحكومة دبي الشيخ حمدان بن محمد بن راشد آل مكتوم .

This Emirate is an important financial and business center in the world and a tourist destination for millions of tourists annually. The free and active economy of this emirate and the absence of a tax system played a major role in attracting investors from all over the world. The tribes of Bani Ka'ab, Al Bo Falah, Al Bu Muheir, Sudan, Shawams, Baloch, Manasir, Rumaythat, Al Shahouh and others, inhabit it. With many families of African and Persian origin. The religion of the people of Dubai is Islam on the approach of the Sunnis and the community, and the official doctrine in Dubai is the Maliki school. Al Maktoum is the rulers of Dubai. They are from Al Bu Falasa from Bani Yas. The ruler now is Sheikh Mohammed bin Rashid Al Maktoum. He is also Vice President and Prime Minister of the Federal Government. His two deputies in the government are his brother Sheikh Hamdan bin Rashid Al Maktoum, Minister of Finance and Industry and Sheikh Maktoum bin Mohammed bin Rashid Al Maktoum. While Sheikh Hamdan bin Mohammed bin Rashid Al Maktoum, the Chairman of the Executive Council of the Emirate, is the Crown Prince. The Dubai Executive Council is headed by Sheikh Hamdan bin Mohammed bin Rashid Al Maktoum.

c. Summary generated by machine-learning approach

Fig. 5. Summary generated by score-based method and machine-learning method.

associated with their extractive summaries (i.e. dataset). The trained classifier will be used to predict whether to include a given sentence in the summary or not based on the values of its feature vector. Fig. 4.b shows the predicted label of each sentence of the input document shown in Fig. 3. Finally, the summary is formulated from the sentences that were predicted as Yes as shown in Fig. 5.c.

5. Experiments and results

5.1. Data set

Testing and evaluating an automatic text summarization system is a difficult process since there is no ideal summary for a given document or a set of related documents. Moreover, the lack of existing Arabic standard datasets made the evaluation process more complex and maybe subjective in certain cases since researchers tend to collect their own datasets (Al-Saleh and Menail, 2016). To the best of our knowledge, there are four publicly available Arabic extractive single document datasets (El-Haj et al., 2011; Giannakopoulos, 2013; El-Haj and Koulali, 2013; EL-Haj et al., 2010). In El-Haj et al. (2011) and Giannakopoulos (2013) summaries are automatically generated by translating the English corpus into Arabic using Google translation service. This way of dataset generation reduces the cost of building an Arabic dataset compared to the human translation. However, such a way may produce a low-quality text or affect semantics. In El-Haj and Koulali (2013), authors have previously developed Arabic summarizers to automatically generate extractive summaries which may be biased to these summarizers. Finally, in EL-Haj et al. (2010), the dataset has been made by human-generated extractive summaries. Therefore, Essex Arabic Summaries Corpus (EASC) (EL-Haj et al., 2010) have been used for testing and evaluating the proposed method. EASC corpus is a human-generated extractive summary published by a group of researchers at Essex University. It comprises 153 articles on different topics which have been collected from Arabic newspapers and Wikipedia. For each article in the EASC corpus, there are five different reference-summaries; each reference summary is generated by a different human. The unique thing about this dataset is that it is the only human-generated Arabic dataset which makes the evaluation more realistic compared to other approaches such as relying on a translated dataset or depending on the output of previously developed summarizers.

Please cite this article as: A. Qaroush, I. Abu Farha, W. Ghanem et al., An efficient single document Arabic text summarization using a combination of statistical and semantic features, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.03.010

Algorithm 1 Score-Based Method

```
INPUT : Preprocessed Document D
SET List=ϕ
SET OrderedList = ϕ
SET L as the maximum summary length
SET S=ϕ
SET Summary=ϕ list represent the summary of D

▷ Split document into their sentences
S = sentenceSplitting(D) ;

▷ Compute score for each sentence
for each S_i in Document D do
    SET TotalScore(S_i) = 0
    SET Index = position(S_i)
    for each Feature F_j do
        Score(F_j, S_i) = computeScore(F_j, S_i, D)
        TotalScore(S_i) = TotalScore(S_i) + Score(F_j, S_i)
    end for
    List.add(TotalScore(S_i))
end for

▷ Sort sentences based on their total score
OrderdList = sort(List)

▷ Extract top-scored sentences based on summary ratio
while S.lenght() < L do
    SET TR = getTopRankedSentence(OrderdList)
    S = S ∪ TR
end while

▷ Sort sentences based on their position in the document
Summary = sort(S, Index)
OUTPUT: Summary
```

Fig. 6. Score-Based Algorithm.

5.2. Evaluation measure

After generating the final summary, an evaluation process is needed to assess the quality of the proposed method. Currently, two evaluation methods are used: human-based and automatic-based (Al-Saleh and Menail, 2016; Das and Martins, 2007). In the human-based evaluation, the summary is given to people to be evaluated. The advantage of this approach is that it assesses coherence and informativity of the summary compared to the original text. However, manual evaluation is too expensive and may be subjective. On the other hand, automatic evaluation is faster and depends on some objective measures (i.e., purity, entropy, recall, precision, and F-measure) for assessment. One of the well-known automated measures used in text summarization is ROUGE which stands for Recall-Oriented Understudy for Gusting Evaluation (Lin, 2004). It includes measures like ROUGE-L, ROUGE-W, ROUGE-S, and ROUGE-N to assess the quality of the generated summary by comparing it to some reference summaries. Among these measures, ROUGE-N is considered the most popular one. It counts the number of overlapping units between the computer-generated summary and the reference summaries which can be computed using the following formula (Lin, 2004):

$$ROUGE-N = \frac{\sum_{S \in Summ_{ref}} \sum_{N-grams \in S} Count_{match}(N-gram)}{\sum_{S \in Summ_{ref}} \sum_{N-grams \in S} Count(N-gram)} \quad (12)$$

where V is the length of the N-gram, $Count_{match}(N-gram)$ is the maximum number of the common N-grams between the set of reference summaries ($Summ_{ref}$) and the generated summary, and $Count(N-grams)$ is the total number of n-grams in the reference summary. There are many variations of ROUGE-N depending on the unit size. The most used ones which are used by DEC 2007 are ROUGE-1 and ROUGE-2. Since ROUGE-N is a recall-oriented measure, Precision P, Recall R, and F-score can be defined as follows (Oufaida et al., 2014):

$$P = \frac{|grams_{ref} \cap grams_{gen}|}{grams_{gen}}, \; R = \frac{|grams_{ref} \cap grams_{gen}|}{grams_{ref}}, \; F1 = \frac{2PR}{P+R} \quad (13)$$

Where $grams_{ref}$ includes the grams of reference summary and $grams_{gen}$ includes the grams of generated candidate summary. To compute these measures automatically, we used ROUGE 2.0 API which is language independent Java package for summary tasks evaluation with updated ROUGE measures.[3]

5.3. Experiments setup and results

The goal of the proposed experiment is to achieve the following results: (i) evaluating the proposed design of the selected statistical and semantic features, (ii) evaluating the application of a statistical summarization method on the Arabic texts, (ii) and comparing our proposed method to other related works. As mentioned earlier, the EASC dataset has been used in experimenting and evaluating the proposed method. For evaluation measures, ROUGE-N (i.g ROUGE-1, ROUGE-2) is used where the precision, recall, and F-score are calculated for each of the generated summaries for both summary methods.

5.3.1. Evaluation of score-based method

Firstly, in score-based summarization, an input threshold (summary ratio) is needed to be adjusted to generate the output summary. The problem is in determining the best ratio since the corpus contains 153 documents where each document has five human reference summaries with a different ratio. To avoid such problem, the generated summaries are adjusted based on an adaptive ratio calculated based on the length of the reference summary we are comparing to it. Accordingly, the average performance when using five reference summaries using ROUGE-1 and ROUGE-2 in terms of precision, recall, and F-measure are presented in Table 2.

However, building reference summary from five reference summaries will suffer from containing less important sentences due to the subjectivity and variations of these five summaries in a way that affects the interpretation of the score of the extracted features (El-Haj, 2012). For example, only 170 sentences, amongst 2360 sentences, were agreed upon to be included in the summary amongst the five reference summaries; noting that 465 sentences were excluded. This entails that the agreement ratio is 27% which is quite low. Moreover, the Kappa measure, which is a measure of inter-observer agreement (Viera and Garrett, 2005) between summary reference A and summary reference B, give 0.247 indicating a fair agreement between the two summaries (Viera and Garrett, 2005).

The disparity will prevail the constructed model explaining the reason for its low results. To enhance results and to avoid the problem of subjectivity, a majority summary so-called gold-standard summary has been constructed through a voting process amongst the five references. Therefore, if a sentence exists in most of the five references (three or more), it will be included in the gold-standard reference summary (El-Haj, 2012). Table 2 shows the results when using a gold-standard reference summary; it is

[3] http://www.rxnlp.com/rouge-2-0.

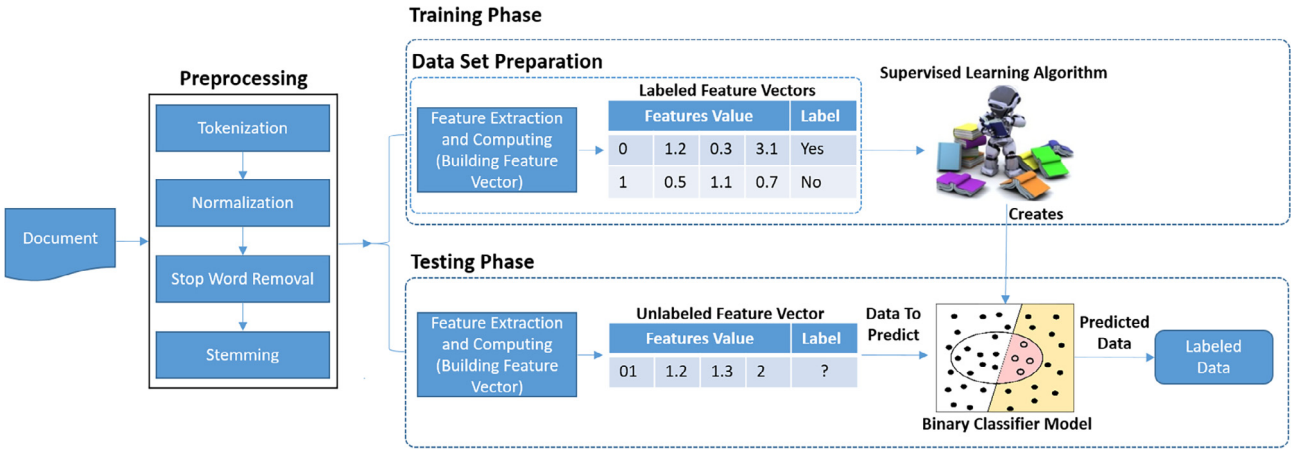

Fig. 7. Machine learning-based stages.

Table 2
Performance of Score-Based summarization method

Reference Summary	ROUGE-1			ROUGE-2		
	Recall	Precision	F-score	Recall	Precision	F-score
Five reference summaries	0.513	0.388	0.442	0.382	0.313	0.344
Gold-Standard reference summary	0.673	0.616	0.643	0.633	0.601	0.617

noticeable that the results have been improved extremely with F-score of 0.617, using ROUGE-2.

5.3.2. Evaluation of machine learning-based method

For the machine learning approach, in order to test the impact of the formulation of the chosen features, five well-known classifiers are tested including Naive Bayes, SVM (with RBF kernel), two-layer neural network, J48, and Random Forest (with 100 random trees) using WEKA tool (Hall et al., 2009). The classifiers are trained and tested using two forms of the data set. The first dataset is formed by using five reference summaries where the sentence will be labeled Yes if it appears in any of the five reference summaries. On the other hand, the second dataset is formed using the gold-standard dataset where the sentence will be labeled Yes if it appears in three or more reference summaries. After forming datasets and computing feature vector of each sentence, each dataset is split into training and testing sets respectively, 120 documents out of the 153 were used as training data to build a model while the others were used as a testing set. Table 3 shows the results for each dataset where Neural Network achieves the best results using five reference summaries. It is noticeable that all results have improved when using gold-standard reference summary bearing in mind that the best classifier is SVM with F-score of 0.524, using ROUGE-2. It is worth mentioning that the experiment has been repeated many times to have different combinations of training and testing samples taking into consideration that the presented result is the averages among these experiments. To sum up, the proposed design of the selected features is superior in terms of precision, and F-score on the score-based method with an improvement of 52% and 17% respectively. On the other hand, machine learning method proved to be better than score-based in terms of recall with an improvement of 16%.

5.4. Comparing to related works

In this section, results of the proposed method are compared against results of other related Arabic summarization methods and systems. Table 4 lists 10 related summarization methods/systems along with a brief description in terms of summary type, summarization technique, and features used. These systems have

Table 3
Performance of machine learning-based summarization technique using different well known classifiers

Reference Summary	Classifier	ROUGE-1			ROUGE-2		
		Recall	Precision	F-Score	Recall	Precision	F-Score
Five reference summaries	Naive Bayes	0.502	0.339	0.400	0.396	0.276	0.323
	SVM	0.545	0.428	0.365	0.460	0.420	0.332
	Neural Network	0.581	0.350	**0.431**	0.475	0.307	**0.367**
	J48	0.556	0.341	0.418	0.434	0.289	0.345
	Random Forest	0.546	0.353	0.412	0.460	0.312	0.355
Average		0.546	0.362	0.405	0.445	0.321	0.345
Gold-Standard reference summary	Naive Bayes	0.513	0.459	0.485	0.447	0.418	0.432
	SVM	0.738	0.414	**0.554**	0.783	0.394	**0.524**
	Neural Network	0.735	0.431	0.543	0.671	0.406	0.506
	J48	0.613	0.423	0.501	0.516	0.375	0.434
	Random Forest	0.641	0.414	0.503	0.562	0.371	0.447
Average		0.669	0.428	0.517	0.566	0.393	0.469

Please cite this article as: A. Qaroush, I. Abu Farha, W. Ghanem et al., An efficient single document Arabic text summarization using a combination of statistical and semantic features, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.03.010

Table 4
Comparison of the proposed method with other research related work and systems in terms of summarization technique, type of summary, and features.

	Type of Summary	Summarization Technique	Features
Al-Radaideh and Afif (2014)	Single document, Extractive, Generic, and Informative	Score-based	Aggregate similarity calculated using the Inner Product measure based on nouns frequencies.
Haboush et al. (2012)	Single document, Extractive, Generic, and Informative	Cluster-based	Term frequency, and remarkable words
LCEAS (AL-Khawaldeh and Samawi, 2015)	Single document, Extractive, Generic, and Informative	Semantic-based	Semantic relations (gloss relation, holonym relation, and meronym relation), and lexical chains
mRMR (Oufaida et al., 2014)	Single and multi-document, Extractive, Generic, and Informative	Statistical-based	Minimum redundancy and maximum relevance based using hierarchal clustering based on Terms' frequency
AQBTSS (El-Haj et al., 2009)	Single and multi-document, Extractive, Query-based, and Informative	Score-based	Weighting scheme based on the VSM model using term frequency and inverse document frequency
LSA-Summ (El-Haj et al., 2009)	Single document, Extractive, Generic, and Informative	Score-based	Latent Semantic Analysis (LSA) to analyse relationships between document's sentences
Gen-Summ (El-Haj et al., 2009)	Single document, Extractive, Generic, and Informative	Score-based	Weighting scheme based on the VSM model using term frequency and inverse document frequency
ESMAT (Binwahlan, 2015)	Single document, Extractive, Generic, and Informative	Score-based	TF-ISF, sentence length, sentence position(SP), sentence similarity to document, sentence concepts, and log entropy
Al-Radaideh and Bataineh (2018)	Single document, Extractive, Generic, and Informative	Optimization-Based	Term frequency, TF-IDF, Title similarity, Sentence position, and Sentence length
Al-Abdallah (2017)	Single document, Extractive, Generic, and Informative	Optimization-Based	Term frequency, TF-IDF, Title similarity, and Sentence length
Proposed ML-Based	Single document, Extractive, Generic, and Informative	Machine learning-based	Title similarity, Key-phrases, sentence location, sentence length, sentence centrality, strong words, Cue-phrases, Occurrence of non-essential information, and existence of numerical data
Proposed Score-Based	Single document, Extractive, Generic, and Informative	Score-based	Title similarity, Key-phrases, sentence location, sentence length, sentence centrality, strong words, Cue-phrases, Occurrence of non-essential information, and existence of numerical data

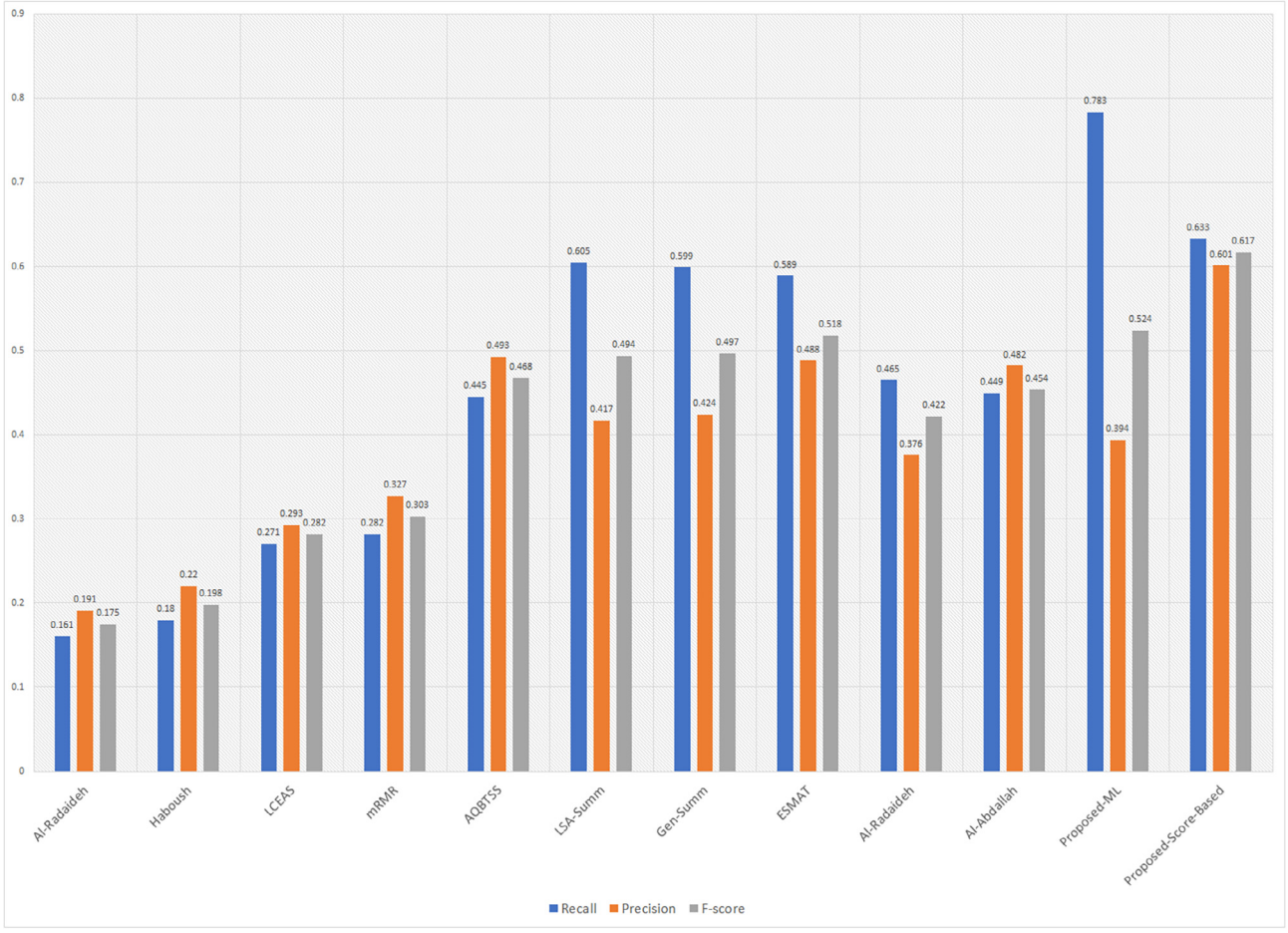

Fig. 8. Comparison of the proposed method with other related summarization methods/systems using ROUGE-2 in terms of recall, precision, and F-score under golden-standard summary reference.

Please cite this article as: A. Qaroush, I. Abu Farha, W. Ghanem et al., An efficient single document Arabic text summarization using a combination of statistical and semantic features, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.03.010

been evaluated under Essex Arabic corpus using gold-standard as a reference summary where the summarizers provide a summary with no more than 50% of the documents words count. In the evaluation process, ROUGE-N (N = 2) in terms of recall, precision, and F-score were used as automatic evaluation measures since it works better for the evaluation of single document summarization. Fig. 8 shows the performance results of the proposed summarization method compared to the performance results of the related summarization methods/systems based on their published results in terms of recall, precision, and F-Score. As shown in the figure, the proposed machine learning method outperforms the others in term of recall and F-score with an average improvement of 33% and 14% respectively. On the other hand, the proposed score-based method outperforms the others in term of recall, precision, and F-Score with an average improvement of 23%, 23%, and 24% respectively. This is due to the powerfulness/strength of the selected feature and the novelty in their formulation besides using the right and up to date Arabic NLP tools.

6. Conclusion

The phenomenal growth of Internet data increases the necessity of an automatic summarization system that solves information overloading and saves user's time. A good summary is expected to preserve key sentences, which represent the main ideas of the document in addition to reduce redundancy to provide an information rich summary. Despite the current efforts to design text summarization methods and formulating representative features, these formulations still lack the ability to provide sufficient representation of sentence's importance, coverage, and diversity. This paper proposes a generic extractive single document summarization method, in which two well-known text summarization approaches are deployed. The first approach is score-based, while the other is a machine learning based one. Both of them, utilize a set of features that were chosen and formulated through deep analysis of summarization methods, properties of Arabic text, and the writing patterns. These features vary from statistical features to semantic based ones. The adopted formulations help to measure the importance of sentences, which is crucial to process deciding whether to they are part of the summary or not, that is while taking into consideration that these sentences are diverse and covering the whole idea in the document. We evaluate the proposed method on EASC dataset. Using ROUGE-2 as a performance measure the system achieved an F-score of 0.524, and 0.617 for machine learning and score-based approaches respectively. The achieved results show that both approaches surpass the current state-of-the-art score-based systems, specifically in the precision. This is due to the informative formulation of the proposed features, which helps in capturing sentence's importance. Future studies would investigate the methods to improve the presented approach through optimizing the weights of the extracted features to reflect their contribution in the total score, using local-search methods such as a Genetic algorithm.

Conflict of interest

None.

References

Abdelkrime, A., Djamel Eddine, Z., Khaled Walid, H., 2015. Allsummarizer system at multiling 2015: multilingual single and multi-document summarization. Proceedings of the SIGDIAL 2015 Conference, pp. 237–244.

Abuobieda, A., Salim, N., Albaham, A., Osman, A., Kumar, Y., 2012. Text summarization features selection method using pseudo genetic-based model. In: International conference on information retrieval knowledge management, pp. 193–197.

Adhvaryu, N., Balani, P., 2015. Survey: Part-of-speech tagging in nlp. Int. J. Res. Advent Technol.

Al_Gaphari, G., BaAlwi, F., Moharram, A., 2013. Text summarization using centrality concept. Int. J. Computer Appl. 79 (1).

Al-Abdallah, R.Z., Al-Taani, A.T., 2017. Arabic single-document text summarization using particle swarm optimization algorithm. Procedia Computer Science 117, 30–37. https://doi.org/10.1016/j.procs.2017.10.091. URL: http://www.sciencedirect.com/science/article/pii/S1877050917321488. arabic Computational Linguistics.

Alami, N., Meknassi, M., Ouatik, S.A., 2016. Ennahnahi N. Impact of stemming on arabic text summarization. 4th IEEE International Colloquium on Information Science and Technology (CiSt). IEEE.

Alguliev, R.M., Aliguliyev, R.M., Hajirahimova, M.S., Mehdiyev, C.A., 2011. Mcmr: maximum coverage and minimum redundant text summarization model. Expert Syst. Appl. 38 (12), 14514–14522. https://doi.org/10.1016/j.eswa.2011.05.033.

Alguliev, R.M., Aliguliyev, R.M., Isazade, N.R., 2013. Multiple documents summarization based on evolutionary optimization algorithm. Expert Syst. Appl. 40 (5), 1675–1689. https://doi.org/10.1016/j.eswa.2012.09.014.

Alguliev, R.M., Aliguliyev, R.M., Isazade, N.R., 2013. Formulation of document summarization as a 0–1 nonlinear programming problem. Comput. Ind. Eng. 64 (1), 94–102. https://doi.org/10.1016/j.cie.2012.09.005.

Al-Hashemi, R., 2010. Text summarization extraction system (tses) using extracted keywords. Int. Arab J. of e-Technol. 1 (4).

AL-Khawaldeh, F., Samawi, V., 2015. Lexical cohesion and entailment based segmentation for arabic text summarization (lceas). World Comput. Sci. Inf. Technol. J. 5 (3), 51–60.

Al-Radaideh, Q., Afif, M., 2014. Arabic text summarization using aggregate similarity. In: The International Arab Conference on Information Technology.

Al-Radaideh, Q.A., Bataineh, D.Q., 2018. A hybrid approach for arabic text summarization using domain knowledge and genetic algorithms. Cognitive Comput.

Al-Saleh, A., Menai, M.E.B., 2018. Solving multi-document summarization as an orienteering problem. Algorithms 11 (7).

Al-Saleh, A., Menail, M., 2016. Automatic arabic text summarization: a survey. Artif. Intell. Rev. Arch. 45 (2), 203–234.

Al-Taani, A., Al-Omour, M., 2014. An extractive graph-based arabic text summarization approach. The International Arab Conference on Information Technology.

Althobaiti M., Kruschwitz U., Poesio M. 2014. Aranlp: A java-based library for the processing of arabic text.

Attia, M., 2007. Arabic tokenization system. Proceedings of the 2007 workshop on computational approaches to semitic languages: Common issues and resources, pp. 65–72.

Ayedh, A., Tan, G., Alwesabi, K., Rajeh, H., 2016. The effect of preprocessing on arabic document categorization. Algorithms 9 (2).

Azmi, A., Al-Thanyyan, S., 2012. A text summarizer for arabic. Comput. Speech Lang. 26 (4), 260–273.

Barrera, A., Verma, R., 2012. Combining syntax and semantics for automatic extractive single-document summarization. In: Proceedings of the 13th international conference on computational linguistics and intelligent text processing. Springer-Verlag, pp. 366–377.

Barzilay, R., Elhadad, M., 2015. Using lexical chains for text summarization. Adv. Autom. Text Summarization, 111–121.

Baxendale, P., 1958. Machine-made index for technical literature: An experiment. IBM J. Res. Dev. 2 (4), 354–361.

Belkebir, R., Guessoum, Ahmed, 2015. A supervised approach to arabic text summarization using adaboost. New contributions in information systems and technologies, advances in intelligent systems and computing, vol. 353, pp. 227–236.

Binwahlan, M., 2015. Extractive summarization method for arabic text – esmat. Int. J. Computer Trends Technol. (IJCTT) 21 (2).

Bossard, A., Généreux, M., Poibeau, T., 2008. Description of the lipn systems at tac 2008: summarizing information and opinions. Text Analysis Conference.

Boudabous, M.M., Belguith, L., 2010. Digital learning for summarizing arabic documents. Advances in natural language processing, lecture notes in computer science;, vol. 6233. Springer, Berlin, pp. 79–84.

ChoSeoung, S., Kim, B., 2015. Summarization of documents by finding key sentences based on social network analysis. International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems IEA/AIE, pp. 285–292.

Das, D., Martins, A., 2007. A survey on automatic text summarization. Literature Survey for the Language and Statistics II.

Doko, A., Stula, M., Stipanicev, D., 2013. A recursive tf-isf based sentence retrieval method with local context. Int. J. Mach. Learn. Comput. 3 (2).

Edmundson, H.P., 1969. New methods in automatic extracting. J. ACM 16 (2).

El-Beltagy, S., Rafea, A., 2009. Kp-miner: a keyphrase extraction system for english and arabic documents. Inf Syst 34 (1), 132–144.

El-Gedawy, M., 2014. Comparing pmi-based to cluster-based arabic single document summarization approaches. Int. J. Eng. Trends Technology (IJETT) 11 (8).

Elghazaly, T., Ibrahim, A., 2012. Arabic text summarization using rhetoricalstructure theory. International Conference on INFOrmatics and Systems.

Please cite this article as: A. Qaroush, I. Abu Farha, W. Ghanem et al., An efficient single document Arabic text summarization using a combination of statistical and semantic features, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.03.010

El-Haj, M., 2012. Multi-document Arabic Text Summarisation Ph.D. thesis. University of Essex.
El-Haj, M., Koulali, R., 2013. Kalimat a multipurpose arabic corpus. The 2nd workshop on Arabic corpus linguistics (WACL-2).
El-Haj, M., Kruschwitzc, U., Fox, C., 2009. Experimenting with automatic text summarisation for arabic. In: Proceedings of the 4th conference on Human language technology: challenges for computer science and linguistics LTC'09. Poznan, Poland, pp. 490–499.
EL-Haj M., Kruschwitz U., Fox C. 2010. Using mechanical turk to create a corpus of arabic summaries.
El-Haj, M., Kruschwitz, U., Fox, C., 2011. Multi-document arabic text summarisation. Computer science and electronic engineering conference (CEEC). IEEE, pp. 40–44.
El-Khair, I., 2006. Effects of stop words elimination for arabic information retrieval: a comparative study. Int. J. Comput. Inform. Sci. 4 (3).
El-Shishtawy, T., El-Ghannam, F., 2012. Keyphrase based arabic summarizer (kpas). In: International Conference on Informatics and Systems (INFOS). IEEE.
El-shishtawy T., Al-sammak A. 2012. Arabic keyphrase extraction using linguistic knowledge and machine learning techniques. arXiv preprint arXiv:12034605.
Erkan, G., Dragomir, R., 2004. Computed lexical chains as an intermediate representation for automatic text summarization. J. Artif. Intell. Res. 22, 457–479.
Erkan, G., Radev, D., 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res. 22, 457–479.
Fattah, M., Ren, F., 2009. Ga, mr, ffnn, pnn and gmm based models for automatic text summarization. Computer Speech Language 23 (1), 126–144.
Fejer, H., Omar, N., 2014. Automatic arabic text summarization using clustering and keyphrase extraction. In: International Conference on information technology and multimedia (ICIMU), pp. 293–298.
Ferreira, R., Cabral, L., Dueire, R., Silva, G., Freitas, F., Cavalcanti, G., et al., 2012. Assessing sentence scoring techniques for extractive text summarization. Expert Syst. Appl. 40, 5755–5764.
Froud, H., Lachkar, A., Ouatik, S.A., 2013. Arabic text summarization based on latent semantic analysis to enhance arabic documents clustering. Int. J. Data Mining Knowl. Manage. Process (IJDKP) 3 (1).
Gholamrezazadeh, S., Salehi, M., Gholamzadeh, B., 2009. A comprehinsive survay ontext summarization systems. 2nd International Conference on Computer Science and its Applications.
Giannakopoulos G. 2013. Multi-document multilingual summarization and evaluation tracks in acl 2013 multiling workshop, p. 20–28.
Gomaa, W., Fahmy, A., 2013. A survey of text similarity approaches. Int. J. Computer Appl. 68 (13).
Gupta, V., Lehal, G., 2010. A survey of text summarization extractive techniques. J. Emerging Technol. Web Intell. 2 (3).
Gupta, P., Pendluri, V., 2011. Vats I., Summarizing text by ranking text units according to shallow linguistic features. In: 13th International conference on advanced communication technology, pp. 1620–1625.
Gupta, V., Chauhan, P., Garg, S., 2012. An statistical tool for multi-document summarization. Int. J. Sci. Res. Publ. 2.
Haboush, A., Al-Zoubi, M., Momani, A., Tarazi, M., 2012. Arabic text summerization model using clustering techniques. World Computer Sci. Inform. Technol. J. (WCSIT) 2 (3).
Hall, M., Frank, E., Holmes, G., Fahringer, B., Reutemann, P., Witten, I., 2009. The WEKA data mining software: an update. SIGKDD Explorations 11 (1), 10–18.
Hasan, K., Ng, V., 2014. Automatic keyphrase extraction: a survey of the state of the art. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. USA, pp. 1262–1273.
Hovy, E., Lin, C.Y., 1998. Automated text summarization and the summarist system. Proceedings of a Workshop on Held at Baltimore, Maryland: October 13–15, 1998; TIPSTER '98. Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 197–214.
Hua, Y.H., Chenb, Y.L., Chou, H.L., 2017. Opinion mining from online hotel reviews a text summarization approach. Inf. Process Manage 53 (2), 436–449.
Hu, M., Liu, B., 2004. Mining and summarizing customer reviews. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; KDD '04. New York, NY, USA. ACM, ISBN 1-58113-888-1, pp. 168–177.
Imam, I., Nounou, N., Hamouda, A., AbdulKhalek, H., 2013. Query based arabic text summarization. IJCST 1, 2.
John, A., Premjith, P.S., Wilscy, M., 2017. Extractive multi-document summarization using population-based multicriteria optimization. Expert Syst. Appl. 86, 385–397.
Kallimani, J., Srinivasa, K., Reddy, E., 2012. Summarizing news paper articles: experiments with ontology-based, customized, extractive text summary and word scoring. Cybern. Inform. Technol. 12 (2), 34–50.
Kanan, R.A.S.G., Jaam, J., Hailat, E., 2004. Stop-word removal algorithm for arabic language. In: International Conference on Information and Communication Technologies: From Theory to Applications. IEEE.
Keskes, I., 2015. Discourse Analysis of Arabic Documents and Application to Automatic Summarization Ph.D. thesis. Toulouse III-Paul Sabatier.
Khan, Atif, 2014. A review on abstractive summarization methods. J. Theor. Appl. Inform. Tech. 59, 64–72.
Kiyoumarsi, F., 2014. Evaluation of automatic text summarizations based on human summaries. 2nd Global Conference on Linguistics and Foreign Language Teaching, LINELT-2014. Elsevier, pp. 83–91.
Kupiec, J., Pedersen, J., Chen, F., 1995. A trainable document summarizer. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR. ACM, pp. 68–73.
Lagrini, S., Redjimi, M., Azizi, N., 2017. Automatic arabic text summarization approaches. Int. J. Computer Appl. 164 (5).
Lakshmi, S., Deepthi, K., Suresh, C., 2015. Text summarization basing on font and cue-phrase feature for a single document. Emerging ICT for Bridging the Future – Proceedings of the 49th Annual Convention of the Computer Society of India CSI, vol. 2, pp. 537–542.
Larkey, L., Ballesteros, L., Connell, M., 2007. Light stemming for arabic information retrieval. Arabic Computational. Morphology.
Lin C.Y. 2004. Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, http://www.aclweb.org/anthology/W04-1013.
Lin, C.Y., Hovy, E., 1997. Identifying topics by position. In: Proceedings of the Fifth conference on Applied natural language processing. USA, pp. 283–290.
Litvak, M., Vanetik, N., Last, M., Churkin, E. 2016. Museec: A multilingual text summarization tool. p. 73–78.
Louis, A., Joshi, A., Nenkova, A., 2010. Discourse indicators for content selection in summarization. In: Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue; SIGDIAL '10. Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 147–156. ISBN 978-1-932432-85-5.
Meena, Y., Gopalani, D., 2014. Analysis of sentence scoring methods for extractive automatic text summarization. Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies, vol. 53.
Meena, Y., Gopalani, D., 2016. Efficient voting-based extractive automatic text summarization using prominent feature set. IETE J. Res. 62 (5).
Meena, Y., Dewaliya, P., Gopalani, D., 2015. Optimal features set for extractive automatic text summarization. In: Proceedings of the 2015 Fifth International Conference on Advanced Computing and Communication Technologies. IEEE.
Mendozaab, M., Bonillaa, S., Nogueraa, C., Cobosab, C., Leónc, E., 2014. Extractive single-document summarization based on genetic operators and guided local search. Expert Syst. Appl.
Mihalcea, R., Tarau, P., 2004. Textrank: Bringing order into texts. Proceedings of EMNLP. Barcelona, Spain. Association for Computational Linguistics, pp. 404–411.
Modaresi, P., Gross, P., Sefidrodi, S., Eckholf, M., Conrad, S., 2017. On (commercial) benefits of automatic text summarization systems in the news domain: a case of media monitoring and media response analysis. CoRR. abs/1701.00728.
Mosa, M.A., Hamouda, A., Marei, M., 2017. Graph coloring and aco based summarization for social networks. Expert Syst. Appl. 74, 115–126. https://doi.org/10.1016/j.eswa.2017.01.010.
Mustafa, M., Salah-Eldeen, A., Bani-Ahmad, S., Elfaki, A., 2017. A comparative survey on arabic stemming: approaches and challenges. Intell. Inform. Manage. 9 (2).
Najadat, H., Al-Kabi, M., Hmeidi, I., Issa, M., 2016;7(2).. Automatic keyphrase extractor from arabic documents. Int. J. Adv. Computer Sci. Appl. 7 (2).
Neto, J., Freitas, A., Kaestner, C., 2002. Automatic text summarization using a machine learning approach. In: SBIA '02. In: Proceedings of the 16th Brazilian Symposium on Artificial Intelligence: Advances in Artificial Intelligence.
Nobata, C., Sekine, S., Murata, M., Uchimoto, K., Utiyama, M., Isahara, H., 2009. Sentence extraction system assembling multiple evidence. Proceedings of the Second NTCIR Workshop Meeting, p. 2001.
Oufaida, H., Nouali, O., Blache, P., 2014. Minimum redundancy and maximum relevance for single and multi-document arabic text summarization. J. King Saud. Univ. Comput. Inf. Sci. 26 (4), 450–461. https://doi.org/10.1016/j.jksuci.2014.06.008.
Ozsoy, M., Alpaslan, F., Cicekli, I., 2011. Text summarization using latent semantic analysis. J. Inform. Sci. 37, 405–417.
Patil, V., Krishnamoorthy, M., Oke, P., Kiruthika, M., 2011. A statistical approach for document summarization. Int. J. Adv. Comput. Technol. (IJACT) 2 (2).
Prasad, R., Kulkarni, U., 2010. Implementation and evaluation of evolutionary connectionist approaches to automated text summarization. J. Comput. Sci. 6 (11), 1366–1376.
Prasad, R., Uplavikar, N., Wakhare, S., Yedke, V.J.T., 2012. Feature based text summarization. International Journal of Advances in Computing and Information Researches.
Qassem, L.M.A., Wang, D., Mahmoud, Z.A., Barada, H., Al-Rubaie, A., Almoosa, N.I., 2017. Automatic arabic summarization: a survey of methodologies and systems. In: Procedia Comput. Sci. 117, 10–18. https://doi.org/10.1016/j.procs.2017.10.088. URL: http://www.sciencedirect.com/science/article/pii/S1877050917321452, arabic Computational Linguistics.
Qazvinian, V., Hassanabadi, L., Beheshti, S., 2008. Summarising text with a genetic algorithm-based sentence extraction. Int. J. Knowl. Manage. Stud. (IJKMS) 4, 426–444.
Radev, D., Hovy, D., McKeown, K., 2002. Introduction to the special issue on summarization. Comput. Linguist 28 (4), 399–408.
Radev, D., Jing, H., Budzikowska, M., 2004. Centroid-based summarization of multiple documents. Inf. Process. Manage. 40, 919–938.
Radev D., Teufel S., Saggion H., Lam W., Blitzer J., Celebi A., et al. Evaluation of text summarization in a cross-lingual information retrieval framework 2011.
Rautray, R., Balabantaray, R.C., 2018. An evolutionary framework for multi document summarization using cuckoo search approach: Mdscsa. Appl. Comput. Inform. 14 (2), 134–144.

Please cite this article as: A. Qaroush, I. Abu Farha, W. Ghanem et al., An efficient single document Arabic text summarization using a combination of statistical and semantic features, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.03.010

Saggion, H., Poibeau, T., 2013. Automatic text summarization: Past, present and future. In: Multi-source, Multilingual Information Extraction and Summarization. Theory and Applications of Natural Language Processing. Springer, Berlin, Heidelberg.

Sanchez-Gomez, J.M., Vega-Rodríguez, M.A., Pérez, C.J., 2017. Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach. Knowledge-Based Syst. https://doi.org/10.1016/j.knosys.2017.11.029.

Sarkar, K., 2014. A keyphrase-based approach to text summarization for english and bengali documents. Int. J. Technol. Diffusion 5 (2), 28–38.

Schlesinger, J., O'Leary, D., Conroy, J., 2008. Arabic/english multi-document summarization with classy: the past and the future. In: Proceedings of the 9th International Conference on Computational linguistics and Intelligent Text Processing, CICLing'08. Springer-Verlag, pp. 568–581.

Shareghi, E., Hassanabadi, L., 2008. Text summarization with harmony search algorithm-based sentence extraction. Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology. Cergy-Pontoise, France. ACM, pp. 226–231.

Thomas, S., Beutenmüller, C., de la Puente, X., Remus, R., Bordag, S., 2015. Exb text summarizer. In: Proceedings of the SIGDIAL 2015 Conference, pp. 260–269.

Tseng, Y.H., Wang, Y.M., Yu-I Lin, C.J.L., Juang, D.W., 2007. Patent surrogate extraction and evaluation in the context of patent mapping. J. Inform. Sci. 33 (6), 718–736.

Turney, P., 2000. Learning algorithms for keyphrase extraction. Inf. Retrieval 2 (4), 303–336.

Viera, A., Garrett, J., 2005. Understanding interobserver agreement:the kappa statistic. Family Med. 37 (5), 360–363.

Please cite this article as: A. Qaroush, I. Abu Farha, W. Ghanem et al., An efficient single document Arabic text summarization using a combination of statistical and semantic features, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.03.010