A Proposed Method for Summarizing Arabic Single Document

Asmaa Awad A. Bialy
Computer department Faculty of Specific Edu.
Damietta University, Egypt

A. A. Ewees
Computer department Faculty of Specific Edu.
Damietta University, Egypt

A. F. ElGamal
Computer department Faculty of Specific Edu.
Mansoura University, Egypt

ABSTRACT

This paper proposes an automatic text summarization method, which is considered as a selective process for the most important information in the original text. It could be divided into two types extractive and abstractive. In this study, a system for single documents text summarization is introduced to be used for Arabic text that rely on extractive method. According to this, we will go three stages, which are pre-processing phase, scoring of sentence, and summery generation. The pre-processing phase starts by removing punctuation marks, stop words, unifies synonyms as well as stemming words to obtain root form. Then it measures every sentence according to a collection of features in order to get the sentences with a higher score to be included in the final summary. The system has been evaluated by comparing between manual and automatic summarizations and some measurements are used especially Rouge measure. Manual summarize is done by two human experts to check the summaries' quality in terms of the general form, content, coherence of the phrases, lack of elaboration, repetition, and completeness of the meaning. The final results proved that the proposed method achieved the higher performance than other systems.

Keywords

Text summarization, Arabic single document, Text mining.

1. INTRODUCTION

The World Wide Web contain billion of documents and it is growing at an exponential pace. Tools that provide timely access to, and digest of, various sources are necessary in order to alleviate the information overload people are facing. These concerns have sparked interest in the development of automatic summarization systems [1]. The increasing availability of online information has necessitated intensive research in the area of automatic text summarization within the Natural Language Processing (NLP) community [2]. "Document summarization is the process of generating a summary by reducing the size of input document and retaining important information of input document. There is arising a need to provide high quality summary in less time because at present, the growth of data is increasing tremendously on World Wide Web or on user's desktops so document summarization is the best tool for making summary in less time" [3].

Text summarization purpose is to reduce the length and detail of a document while retaining most important points and general meaning [4]. Automatic Text Summarization can be characterized into single document text summarization and multi document summarization [5]. It worth noting that Arabic language faces many challenges such as translation and summarization, one of the most serious semantic problems in translation and summarizing is the difference in the contextual distribution of words that appear to be synonymous in two languages, which may be synonymous with one another, although they may differ in usage applications or language contexts [6]. It also notable that grammatical confusion is a greater challenge when dealing with the Arabic language, and the previous studies of the automatic summary did not reach the accuracy of satisfactory to summarize the Arabic documents [7]. Compared to English document summarization, very few works were performed for Arabic document summarization [8].

2. RELATED WORK

Different document summarization methods have been developed in recent years. Generally, those methods can be either extractive or abstractive ones. Extractive summarization creates the summary from phrases or sentences in the input document, and the abstractive summary express the idea in the input document using different words [3].

This section discusses some of these existing summarization systems and reviews some systems of summarization Arabic texts:

I. Keskes et. al, (2012) addressed the automatic summarization of Arabic texts, it presented a new method to generate a summary. This method relied, and for the first time in Arabic language, on SDRT (Segmented Discourse Representation Theory). The method contains two main parts. The first one creates the discourse structure by extracting rhetorical relations between elementary discourse units text and drawing SDRS graph that represents the discourses structure of the text. The second part focuses on building the automatic summary depending on SDRS graph by minimizing it throw eliminating rhetorical relations not supported in the summary chosen. This method was evaluated by implementing it in the "SDRTResume" system [9]. **H. Oufaid et. al., (2014)** it proposed a novel statistical summarization system for Arabic texts. This system used a clustering algorithm and an adapted discriminant analysis method: MRMR (minimum redundancy and maximum relevance) to record terms. Through MRMR analysis, terms are ranked according to their discriminant and coverage power. Also, it proposed a novel sentence extraction algorithm, which selects sentences with top ranked terms and maximum diversity. This system uses minimal language-dependent processing: sentence splitting, tokenization and root extraction. Experimental results on EASC and TAC 2011 Multilingual datasets showed that the proposed approach was competitive to the state of the art systems [8]. **El Sherief, (2015)** introduced a hybrid system for the summary process to enhance the results of the QPM Query Processing Module, this system relied on RST and the Network Representation approach. The system used in agriculture has been applied and given good results in the summary [10]. **Froud et. al. (2016)** this paper investigated an evaluation for the impact of text summarization using the Latent Semantic Analysis Model on

Arabic Documents Clustering, thus by using five similarity/distance measures: Euclidean Distance, Cosine Similarity, Jaccard Coefficient, Pearson Correlation Coefficient and Averaged Kullback-Leibler Divergence, for two times: with and without stemming. The experimental results indicated that the proposed approach effectively solved the problems of noisy information and documents length, and thus significantly improved the clustering performance [11].

3. METHODS

This paper focus on extractive summarization methods. A lot of research implemented in the direction of extraction based approaches. In extractive summarization, the important task is to find informative sentences, a subpart of sentence or phrase and include these extractive elements into the summary [12].

The basic process flow of generic extractive text summarization is shown in Figure 1.

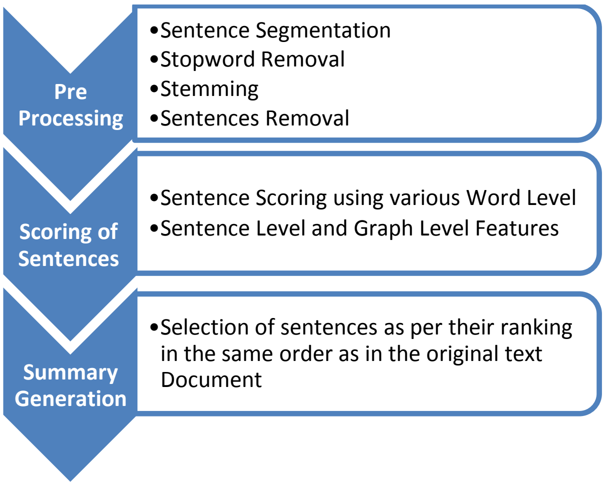
Pre Processing
•Sentence Segmentation
•Stopword Removal
•Stemming
•Sentences Removal
Scoring of Sentences
•Sentence Scoring using various Word Level
•Sentence Level and Graph Level Features
Summary Generation
•Selection of sentences as per their ranking in the same order as in the original text Document

Fig. 1. Generic Extractive Summarization Process [13]

The Frist Step is preprocessing, in which sentences are segmented using appropriate methods, punctuations symbols are used as sentence end marker. Stopwords are deleted as they do not add more information related to the text in text summarization. After that the sentences are scored using various word level and sentence level features. These sentences after scoring are selected in the same order as they appear in the source document in the final phase [13].

The proposed method consists of three steps beginning with entering of the Arabic document to be automatically summarized into the proposed method. The PHP language will be used to write the code for the proposed method as well as a set of other languages.

The proposed method has three main stages as follows:

1. The pre-processing stage.
2. The processing stage.
3. The final summary stage.

3.1 The Pre-processing Stage

The pre-processing phase aims to obtain a structured representation of the original text, and this phase includes the following:

- Sentence Boundary Identification: in Arabic, the limits of the sentence are determined by using a set of punctuation marks at the end of the sentence such as: (، ، ؟ ، .).
- Remove repeated sentence.
- Remove parentheses and quotation marks: The parentheses are removed such as ("," ( ), [ ],} {)..
- Normalization Alef by replacing (ااا , اا , آ , إ , أ) to (ا) , Normalization Yaa by replacing (ي) to (ى) , and Normalization Tah by replacing (ة) to (ه), and the deletion of the diacritical marks (ْ ، ّ ، ِ ، ٌ ، ُ ، ً ، َ ، ٌ).
- Remove the Stop word: Stop words are common words that appear in the text but carry little meaning [14]. Remove all stop words from sentences so that each sentence has only the verbs and the nouns. A stop word does not have a root, and it does not add any new information to the text (does not affect the meaning of the sentence if removed). Some of these words are (هو ، هذا ، الذي ، هي) [15].
- Remove punctuation marks: punctuation marks such as (? ، :).
- Stemming: Stem is a part of the word (with or without meaning) which are used to form new words through various linguistic methods [16]. It is possible to find the Arabic root automatically by removing the subparts of suffixes, prefixes, and infixes from the word [15]. Removing the subparts of suffixes (ها ، ان ، ين ، ون ، ات ، وا) from the end of the word and Removing the subparts of prefixes (لل ، ولل ، وتال ، وكال ، وال ، كال ، تال ، ال) from the beginning of the word from all sentences.
- Remove spaces: Spaces between words are removed if two or more spaces were exist.

3.2 The Processing Stage

The processing phase consists of the following steps:

1. **Sentence Feature Calculation:** Each sentence is given a score, which serves as a good measure of the sentence by using a set of specific features. Each preset feature score takes a value ranging from (1,0), the following set of features will be used:

- **Frequency Feature:** frequency of word play a crucial role, to decide the importance of any word or sentence in a given document [12,17,18]. In our method, the weight of the sentence is calculated on the basis of the frequency of the term or synonyms and frequency of relations by calculating the average value of the frequency of the term in each sentence as well as the synonyms. The weight of the word root is calculated by equation (1) [15].

$$W_{i.j} = log(N/n_i) * tf \qquad (1)$$

where *Wi,j* means weight of word *i* in sentence *j*, *N* the total number of words in a paragraph, n_i is the frequency of each word in text, *tf* (term frequency) = n_i/ max n_i (i.e. frequency of word *i* / max frequency in text).

Then calculate the weight of the sentence by equation (2) as following [15]:

$$S(i) = \sum(W_{i.j}) \qquad (2)$$

where S (i) the weight of the sentence, the sum of the weights of the words (i) in the sentence (j).

- **Feature of Important Words:** In Arabic, there are some words that are indicative of important information that can be included in the summary such as: (أهم الأمور، يدل ذلك)

[10,15]. Also in Arabic, the date (Hijri / Gregorian) is important information that can be included in the final summary, a new feature that has not been used before.

Based on Equation (2), the score of sentence is calculated after adding the important words through Equation (3) [15].

$$S(i) = \sum(W_{i.j}) \quad + A \qquad (3)$$

where A is the important words in the text, which are taken from some of previous works.

- **The location Feature:** introduced a feature based on "Sentence Position". Shekhar and Sharan in [12] were almost manual but, later on this measure used widely in sentence scoring, so leading sentences of an article are important, and it takes a value between 0 and 1. The model which used in this paper are using the given below, where N is the total number of sentences. The used model is: (where: 1 < i < N, and $Score\ Si$ = (0, 1) [12].

$$Score\ S(i) = 1 - \frac{i-1}{N} \qquad (4)$$

2. **Sentence Scoring:** After determining the features of the sentence, the weight of each sentence is calculated based on the summation of Equations (3) and (4).

3.3 The Final Summary Stage

This stage aims at extracting the final summary through the following:

1. **Rank calculation:** The minimum weight of the sentence that will be included in the final summary is calculated, by extracting the mean value of the sentence weights by using the following formula:

$$Rank = (\frac{\sum S(i)}{N}) \qquad (5)$$

where $\sum S(i)$ is the sum of values of sentences' weights.

2. **Final selection process:** The higher-grade sentences (extracted from the previous step) will be included in the final summary and the sentences with the lower scores will be deleted. The sentences will be included in the final summary as they exist in the original text, In addition to the first sentence in case of (weight less than the rank) because of its importance. This phase is done through the following Algorithm:

```
Algorithm: The last stage

Input: one array contains Sentences Scoring
Operation:
calculating rank for every sentence by $Rank = (\frac{\sum S(i)}{N})$
Check Frist Sentence
If (Frist Sentence score < Rank) then
        Summary = Frist Sentence
For I =2 to N
    If (sentence scoring [i] > Rank ) then
                Summary += ' ' + sentence [i]
    EndIf
Next
Output: summary (summary sentences, as given in Source)
```

Table 1. An example of the final summary of the paragraph

Original text	Generated summary
هي دراسة تصميم هندسة البرمجيات وتنفيذ وتعديل البرمجيات بما يضمن توفر هذه البرمجيات بجودة عالية وتكلفة معقولة متاحة للجميع وقابلة للتطوير فيما بعد وسريعة للبناء. وهندسة البرمجيات تقوم على أسس ونظريات من الهندسة وعلوم الحاسب كمبدأ ال Functional Structure من الهندسة، والذي يعتمد على مبدأ تصميم أجزاء صغيرة تتجانس في العمل مع بعضها لتشكل عمل الكل. ومن علوم الحاسب يأخذ مبادئ كثيرة لعل من أبرزها ال Object Oriented Design والذي يتعامل مع كل الأجزاء في البرمجيات ككائنات تتفاعل مع بعضها لتشكل عمل النظام بالكامل، وهي تختلف عن علوم الحاسب حيث أنها تعد فرع مهم من فروع علوم الحاسب.	هندسة البرمجيات هي دراسة تصميم وتنفيذ وتعديل البرمجيات بما يضمن توفر هذه البرمجيات بجودة عالية وتكلفة معقولة متاحة للجميع وقابلة للتطوير فيما بعد وسريعة للبناء. وهندسة البرمجيات تقوم على أسس ونظريات من الهندسة وعلوم الحاسب كمبدأ ال Functional Structure من الهندسة، والذي يتعامل مع كل الأجزاء في البرمجيات ككائنات تتفاعل مع بعضها لتشكل عمل النظام بالكامل،

4. EXPERIMENT AND RESULTS

In this experiment, 33 Arabic articles are used, they are added one by one to summarize them through the proposed method, and the summary ratios (per document) were calculated by Equation (6):

$$SR = \left(1 - \frac{Number\ of\ words\ in\ the\ summary\ (S)}{Number\ of\ words\ in\ the\ document\ (D)}\right) * 100 \qquad (6)$$

4.1 Dataset description

The experiment sample consisted of a collection of articles collected from Wikipedia (https://ar.wikipedia.org/). Where a random sample of articles in Arabic was selected in various fields such as: astronomy, biology, chemistry, etc. The volume of articles ranged from long articles containing three or more paragraphs or medium containing two paragraphs or small containing one.

4.2 Performance measures

To evaluate the quality and efficiency of the proposed method, the following measures were used:

$$Recall = \frac{TP}{TP+FN} \qquad (7)$$

$$Precision = \frac{TP}{TP+FP} \qquad (8)$$

$$Fmeasure = \frac{2*Recall*Precision}{Recall+Precision} \qquad (9)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (10)$$

where: TP is the number of sentence pairs in the human expert summary and system summary.

TN is the number of pairs of sentences not found in the expert summary and system summary.

FP is the number of sentences in the system summary that are not in the expert summary.

FN is the number of sentences in the expert summary that are not in the system summary.

$$ROUGE-N = \frac{\sum S\epsilon\{ReferencesSummaries\} \sum gram_n \,\epsilon S\, Count_{match}(N-gram)}{\sum S\epsilon\{ReferencesSummaries\} \sum gram_n \,\epsilon S\, Count(N-gram)} \quad (11)$$

where: N stands for the length of the N-gram, Count (N-gram) is the number of N-grams present in the reference summaries, and the maximum number of N-grams co-occurring in the system summary, the set of reference summaries is $Count_{match}$ (N-gram) ROUGE measures generally gives three basic score Precision, Recall, and F-Score [12].

4.3 Evaluation of the overall summary ratio Subsubsections:

The summary of the proposed method has reached (44.5 %), while the ratio of the first human expert (46.7 %) and the second human expert (48.6 %). The differences in summary ratios are due to differences in the method of summarization between the proposed method and each of the two human experts. These results show the superiority of the proposed method. Table 2 shows the summary ratio for each human expert compared to the proposed method.

Table 2. Comparison between the summary ratios of the proposed method and two human experts

	The first expert	**The second expert**	**The proposed method**
The overall summary ratio	46.7	48.6	44.5

4.4 Evaluation of the proposed method results by human experts in the following four levels:

the general form and content, the coherence of the phrases, lack of elaboration or repetition, completeness of the meaning, and the results were as following:

The degree of evaluation of human experts of the proposed method (automatic summary) has generally come in accordance with the pentagram of Carter, as the arithmetic mean of the responses of the arbitrators was in the four axes as a whole (4) and Table 3 presents the statistical results of the expert responses in the four axes according to the pentagram of Carter.

Table 3. statistical results of the experts' responses in the four axes according to the pentagram of Carter

	Form and content	**Phrases Coheren.**	**Lack of Elaborat.**	**Meaning Compl.**
Mean	3.95	3.97	4.05	4.04
STD	0.39	0.25	0.26	0.34

Table 3. shows the following: The degree of evaluation of the human experts of the proposed system (automatic summary) has generally been appropriate. The arithmetic average of the arbitrators' responses to the questionnaire reached the four axes as a whole (4.00). The " Lack of elaboration " axis in the first order obtained the highest mean (4.05), the " Meaning Completeness " axis in the second order with an arithmetic mean (4.04), and the " Phrases Coherence" (3.97), and the axis of " The form and content " came in fourth place with an average of (3.95).

4.5 General comparison between the proposed method and related studies:

In order to judge the efficiency and accuracy of the proposed method in the light of previous studies, the results of the proposed method were compared with other automatic summary systems for Arabic texts. The following measurements were used: Recall, Precision, F-measure, Accuracy, ROUGE1, it has been considered that ROUGE is an effective approach to measure document summarizes so widely accept. ROUGE measures, overlap words between the system summary and standard summary (gold summary/human summary) [12]. Table 4 compares the arithmetic mean of the following measures: Recall, Precision, F-measure of the proposed method with some automatic summary systems for Arabic texts [9, 19] as follows:

Table 4. shows the Comparison between the arithmetic mean of the following measures: Recall, Precision, F-measure of the proposed method with some automated summary systems for Arabic texts

Automated Summary Systems	**Recall**	**Precision**	**F-measure**
Proposed method	0.68	0.78	0.71
system SDRTResume	0.85	0.56	0.65
system R.I.A	0.42	0.69	0.60
ARSTResume system	0.45	0.63	0.50

It is clear from the previous table that the proposed method is superior to the three systems and is ranked first on the Precision and F-measurement scales of 78 % and 71 %, respectively; while in the second order according to the Recall scale of 68 %; this result shows the efficiency of the proposed method in summarizing the articles. In addition, results from the application of proposed system on 15 articles and the averages of the performance measures (i.e. Precision, Recall, F-measure, Accuracy, and ROUGE) are listed on Table 5. It shows that the whole accuracy and Rouge measures of the proposed method equal to 0.68 and 0.47 respectively.

Table 5. Precision, Recall, F-measure, Accuracy, and ROUGE measures of the proposed method

Doc. Num	**Prec.**	**Recall**	**F-measure**	**Acc.**	**Rouge**
1	0.67	0.67	0.67	0.67	0.33
2	0.67	0.75	0.71	0.64	0.44
3	1	0.5	0.67	0.6	0.4
4	0.67	0.5	0.57	0.57	0.4
5	1	0.6	0.75	0.67	0.33
6	0.63	0.71	0.67	0.58	0.44
7	1	0.75	0.86	0.8	0.8
8	0.6	0.6	0.6	0.6	0.44
9	0.8	0.8	0.8	0.82	0.44
10	0.8	1	0.89	0.85	0.67
11	0.83	0.63	0.71	0.64	0.33

12	0.67	0.8	0.73	0.63	0.33
13	0.75	0.5	0.6	0.64	0.33
14	0.8	0.67	0.73	0.73	0.5
15	0.8	0.67	0.73	0.73	0.86
Avg	**0.78**	**0.68**	**0.71**	**0.68**	**0.47**

In terms of ROUGE scale, the following table compares the proportions of this scale to the proposed method with some automatic text summarization systems [20], as shown in Table 6:

Table 6. Comparison between the ROUGE1 arithmetic averages of the proposed method with some automatic summary systems for Arabic texts

Automatic summary systems	**ROUGE**
The proposed method	0.47
EMDG system	0.44
LSA system	0.34
Random system	0.31

From the results in Table 6, the proposed method is outperformed the previous systems according to the ROUGE measure. The first order was 47 % followed by EMDG. LSA came in the third order; whereas, Random came in the last order. These results show the efficiency of the proposed method in summarizing the Arabic articles.

5. CONCLUSION

In this paper, a method for the automatic summary of Arabic texts was presented and discussed to summarize a single document. This method is based on the extraction method, where the summary consists of a set of important sentences from the original text, and depends on the selection of sentences on the weight of each sentence based on a set of features, the processing is on the roots itself not the words, and then the semantic similarity between the sentences is measured to select the most important sentences in the final summary after rearranging them in the same order in the original text. The proposed method has been evaluated, where the ratio of the arbitrators results in summarizing reached (80 %) to measure the quality of the summary, which is a positive rate in favor of the proposed method, in addition to the superiority of the proposed method to the previous systems in the F-measure and ROUGE1 scale. In the future, we will improve the proposed method to work with long documents and enhance the accuracy ratio.

6. REFERENCES

[1] Nenkova, A., Mckeown, K. (2011). Automatic Summarization, USA, p. 1.

[2] Suneetha, S. (2011). Automatic Text Summarization: The Current State of the art. International Journal of Science and Advanced Technology, ISSN 2221-8386 Volume 1, No 9 November.

[3] Mol, R., Sabeeha. (2016). An Automatic Document Summarization System Using A Fusion Method. International Research Journal of Engineering and Technology (IRJET), ISSN: 2395 -0056 Vol 3, July.

[4] Rajput, Y., Saxena, P. (2016). A Combined Approach for Effective Text Mining using Node Clustering. International Journal of Advanced Research in Computer and Communication Engineering, ISSN: 2319 5940, Vol. 5, No. 4 321-324, April.

[5] Bhatia, N., Jaiswal, A. (2015). Literature Review on Automatic Text Summarization: Single and Multiple Summarizations. International Journal of Computer Applications (IJCA), 0975 – 8887 Vol 117, No. 6 May.

[6] Tafiqe, M., Farag, Y., Younis, M. (2014). Comparative and Contrastive Linguistics, Cairo University, p126.

[7] Basiony, A. (2011). Computer for extracting knowledge and opinion mining, Dar El Kotb El-elmia for publishing, p 96, Cairo-Egypt.

[8] Oufaida, H., Noualib, O., Blache, P. (2014). Minimum redundancy and maximum relevance for single and multi-document Arabic text summarization. Journal of King Saud University- Computer and Information Sciences, 450–461, September.

[9] Keskes, I., Lhioui, M., Benamara, F., Belguith, L. (2012). Automatic Summarization of Arabic Texts Biased on Segmented Discourse Representation Theory. International Computing Conference in Arabic, ICCA, 26-28 December, Egypt.

[10] El Sherief, F. G. (2015). Towards A Hybrid Framework for Automatic Arabic Summarizer, Unpublished PhD's thesis, Faculty of Computer and Information, Cairo University.

[11] Froud, H., Lachkar, A., Ouatik, S. (2016). Arabic Text Summarization Based on Latent Semantic Analysis To Enhance Arabic Documents Clustering. Colloquium in Information Science and Technology (CIST) 22-24 October.

[12] Shekhar, Y. C., and sharan, A. (2015). Hybrid Approach for Single Text Document Summarization using Statistical and Sentiment Features. International Journal of Information Retrieval Research (IJIRR), 46-70.

[13] Menna, Y. K., and gopalani, D. (2015). Feature Priority Based Sentence Filtering Method for Extractive Automatic Text Summarization.

[14] El-Fishawy, N., Hamouda, A., Attiya, G., and Atef, M. (2014). Arabic summarization in Twitter social network", Ain Shams Engineering Journal, Vol 5, No 2 411–420 June.

[15] Haboush, A., Momani.A., Al-Zoubi,M., Tarazi,M.(2012) . Arabic Text Summarization Model Using Clustering Techniques. World of Computer Science and Information Technology Journal WCSIT, Vol 2 No 3 62 – 67.

[16] Singh, J. and Gupta, V. (2016). A systematic review of text stemming techniques. p158.

[17] Refaat, M. M., Ewees, A. A., Eisa, M. M., & Sallam, A. A. (2012). Automated Assessment of Students' Arabic Free-Text Answers. International Journal of Intelligent Computing and Information Science, 12(1), 213-222.

[18] Ewees, A. A., Eisa, M., & Refaat, M. M. (2014). Comparison of cosine similarity and k-NN for automated essays scoring. *cognitive processing*, *3*(12).

[19] Boudabous, M., Maaloul, M., Keskes, I., Belguith, L. (2012). Automatic Summarization of Arabic Texts Between Digital learning theory and Rhetorical Structure Theory. January.

[20] Dief, N. (2016). An Improved Text Mining Technique, Unpublished master's thesis. Computer and Systems Engineering Dep. Faculty on Engineering. Mansoura University.