World of Computer Science and Information Technology Journal (WCSIT)
ISSN: 2221-0741
Vol. 2, No. 3, 62 – 67, 2012

Arabic Text Summerization Model Using Clustering Techniques

Ahmad Haboush
Computer Science Department
Jerash University
Jerash, Jordan

Maryam Al-Zoubi
Computer Science Department
Jerash University
Jerash, Jordan

Ahmad Momani
Computer Science Department
Jerash University
Jerash, Jordan

Motassem Tarazi
Computer Science Department
Jerash University
Jerash, Jordan

Abstract— the current work investigates a developed automatic Arabic text summarization model. In this model, a technique of word root clustering is used as the major activity. Unlike the previously presented systems of Arabic text summarization in the extract based design field, the current model adopts cluster weight of word roots instead of the word weight itself.

The model is thoroughly illustrated through its different stages. Obviously, the general scheme follows traditional descriptive model of most of the system stages in literature with the exception of the ranking stage. This model with its developed technique has been subjected to a set of experiments. Various Arabic text examples are used for evaluation purposes. The efficiency of the summarization is calculated in terms of Precision and Recall measures. Result obtained actually is considered promising and competitive to the verb/noun categorization ranking method. This enhancement has been detected for Precision 76% and Recall 79% with the analogous values of 62% and 70% obtained in the verb/noun categorization method. The enhancement emerges in this tangible result is attributed to the implicit embedding of semantic capability of the developed model to expand the extract boundaries towards the abstract extremes of the design theme.

Keywords - Text Summarization; Clustering; Natural Language Processing Evaluation.

I. INTRODUCTION

The increasing number of documents and related sorts of informational text on web has led to various trends towards Arabic text summarization applications and model design. The early work of [1] has been followed by different proposals. Despite of all of the presented schemes in these proposals, the ranking stage is considered as the primitive processing characterizing the summarization activity. In fact, the fundamental design principles of Arabic language summarization do not differ from that of Latin language. However, these principles are classified to fall into two main categories. The first denotes the extract based design and the second is the abstract based design, [2]. In the former design, the system after its processing is supposed to give a summary that is composed of existing words extracted from the original text. Whereas, in the second design, the system is supposed to generate a summary that involves the conceptual declarations using a set of words that are not necessary be extracted from the original text but it should hold the meaning **[**3]. Hence the latter is much complex from the design prospective point of view in comparison with the former design and it needs suitable database and higher level of linguistic details and processing.

The nature of Arabic language and due to the wide range of derivations of functional word allows for higher level of grammatical investigations. And thus, similar conceptual sentences either by analogous words or dissimilar ones can be generated for expression formalization. This may give wider tolerance of investigations to adopt extract and abstract design basis conjugationally. This fact has been investigated in the current work to propose a model of automatic Arabic text summarization which depends on a low level of abstract theme driven in extract basis of design. In this model, the ranking stage is designed to assemble all the words of the same root in a distinct cluster. The words of this cluster inherit a common weight of the cluster they belong to. Therefore, individual ranking is avoided and the new ranking method seems to

justify the semantic design that approaches abstract principles of summarization.

II. FEATURES OF EXTRACT BASED TEXT SUMMARIZATION MODEL

Obviously languages differ from each other in expression styles and grammar. In literature, Latin language has been processed with various tools and applications. In text summarization, the extract based models are used widely. These models are composed of three main stages, Fig.1. They are initiated by Document feeding and terminated by text summary generation or by keywords generation in other words. These stages conduct their activities with different techniques but in general can be given as.

1) Morphological Analysis

2) Noun Phrase (NP) Extraction and Scoring

3) Noun Phrase (NP) Clustering and Scoring.

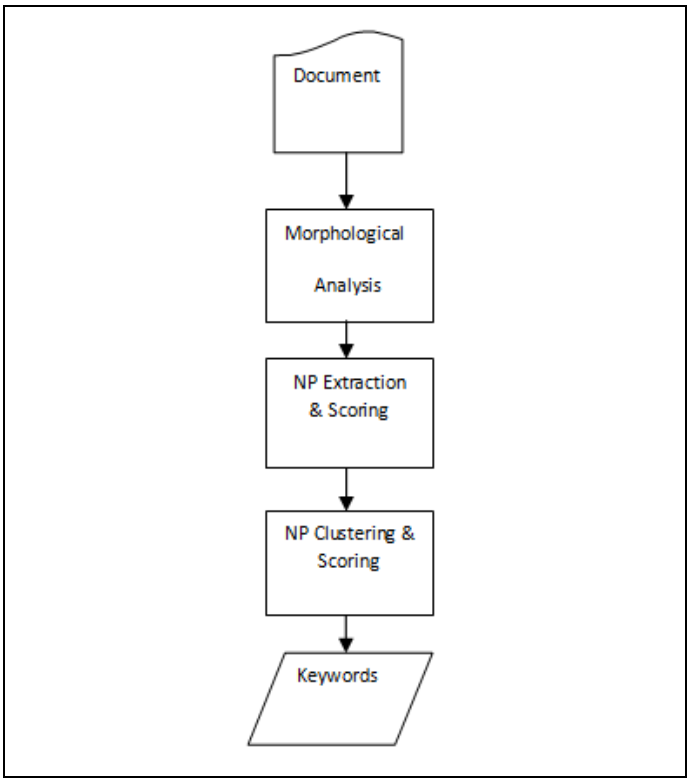

Figure 1. The main three stages in Extract Based Design Model.

The major features of this model can be explained as:

1) Content words or Keywords are usually nouns: Sentences having keywords are of greater chances to be included in summary.

2) Title word feature: Sentences containing words that appear in the title are also indicative of the theme of the document. These sentences are having greater chances for including in summary.

3) Sentence location feature: Usually first and last sentence of first and last paragraph of a text document are more important and are having greater chances to be included in summary.

4) Sentence Length feature: Very large and very short sentences are usually not included in summary.

5) Proper Noun feature: Proper noun is name of a person, place and concept etc. Sentences containing proper nouns are having greater chances for including in summary.

6) Upper-case word feature: Sentences containing acronyms or proper names are included.

7) Cue-Phrase Feature: Sentences containing any cue phrase (e.g. “in conclusion”, “this letter”, “this report”, “summary”, “argue”, “purpose”, “develop”, “attempt” etc.) are most likely to be in summaries.

8) Biased Word Feature: If a word appearing in a sentence is from biased word list, then that sentence is important. Biased word list is previously defined and may contain domain specific words.

9) Font based feature: Sentences containing words appearing in upper case, bold, italics or Underlined fonts are usually more important.

10) Pronouns: Pronouns such as “she, they, it” cannot be included in summary unless they are expanded into corresponding nouns.

11) Sentence-to-Sentence Cohesion: For each sentence s compute the similarity between s and each other sentence s’ of the document, then add up those similarity values, obtaining the raw value of this feature for s. The process is repeated for all sentences.

12) Sentence-to-Centroid Cohesion: For each sentence s as compute the vector representing the centroid of the document, which is the arithmetic average over the corresponding coordinate values of all the sentences of the document; then compute the similarity between the centroid and each sentence, obtaining the raw value of this feature for each sentence

13) Occurrence of non-essential information: Some words are indicators of non-essential information. These words are speech markers such as “because”, “furthermore”, and “additionally”, and typically occur in the beginning of a sentence. This is also a binary feature, taking on the value “true” if the sentence contains at least one of these discourse markers, and “false” otherwise.

14) Discourse analysis: Discourse level information in a text is one of good feature for text summarization. In order to produce a coherent, fluent summary, and to determine the flow of the author's argument, it is necessary to determine the

overall discourse structure of the text and then removing sentences peripheral to the main message of the text [15].

III. RELATED WORKS

The foregoing section presents the main features of summarization. In fact, it should be noted that summarization as a technique was characterized in its early trends by simplicity during 1950's and 60's. Recent approaches use more sophisticated techniques for deciding which sentences to extract. However a historical review can demonstrate a convenient paradigm of the current proposal with primitive capabilities. Luhn 1958 developed a system for Automatic Text Summarization. This model is considered to be an early algorithm with primitive features and it used selection - based summarization approach [4]. Michael J. Witbrock and Vibhu O. Mittal, have written a paper that represents a statistical model of the process of a summarization, which jointly applies statistical models of the term selection and term ordering process to produce brief coherent summaries in a style learned from a training corpus. This approach of summarization, is not based on sentence extraction, capable of generating summaries of any desired length, but it is considered as statistically learning models of both content selection and realization. When it is given an appropriate training corpus, it can generate summaries similar to the training ones, of any desired length [5]. Sanda M. Harabagiu_, Finley Lacatus¸U, 2002 describe a proper technique that was implemented in GISTEXTER to produce extracts and abstracts from both single and multiple documents. These techniques promote the belief that highly coherent summaries may be generated when using textual information. Such a trend is identified afterwards by the Information Extraction technology [6]. Mahmoud El-Haj, Udo Kruschwitz, Chris Fox describe two summarization systems in their work; The Arabic Query-Based Text Summarization System and the Arabic Concept-Based Text Summarization System. The first is a query-based single document summarizer system that takes an Arabic document and a query (in Arabic). This system gives a summary for the document in accordance to the organized query. Whereas the second takes a bag-of-words representing a certain concept as input to the system. In both systems the summarization is sought consistent with the sentences that best match the query or the concept [7].

IV. THE ROOTS OF ARABIC WORDS

Arabic language is one of the six official languages of the united nation, [8]. Arabic is spoken by almost 250 million people in more than twenty-two countries, but up to now the numbers of researches still few in Arabic natural language (NLP). It has been considered a challenging language for information retrieval. Such considerations are attributed to four main reasons. First, certain combinations of characters can be written in different ways and this depends on the position of letter in the word. Second, Arabic is highly inflectional and derivational, which makes morphology is a very complex task. Third, Broken plurals are common. Broken plurals are somewhat like irregular English plurals except that they often do not resemble the singular form as closely as irregular plurals resemble the singular in English. Four, Arabic words are often ambiguous due to the tri-literal root system [9].

Based on such specifications in Arabic language, natural language processing seems more sophisticated and needs much time compared with the accomplishments in English and other European languages. These languages despite of their nature they are discriminated from Arabic by their writing direction which flows from right –to- left, capitalization to identify proper names, acronyms, and abbreviations. Besides they are rich with corpora, lexicon, and machine– readable dictionaries, which are essential to advanced research in the different areas [10]. To know the original words in Arabic it is necessary to know the root of this word. Usually the root of any Arabic word consists of either three or four letters. Even though, some words may have more than four letters. On the roots of Arabic word Suffix, prefix and infix can be added to build a set of derivations [11]. It worth mentioning that it is a hard matter to determine the root of any Arabic word since it requires a detailed morphological, syntactic and semantic analysis of the text. In addition, Arabic words might not be derived from existing roots; they might have their own structures. In this work, it is considered as a basic task to find the root of each word in text, since the root can be a base of different words with informative related meaning. For example the root لعب "laaeba" is used for many words relating to "playing", including لاعب" , " laaeb", "player", ملعب " malaab" .

It is possible to find the Arabic root automatically by removing the subparts of suffixes, prefixes, and infixes from the word. These auxiliary subparts might be positioned in beginning, middle or last locations of words. In order to remove these subparts the word first is matched to the existing basic structures as rhythms, called as "tafaaelat" giving the meaning of derivations. Whenever the basic structure is found, one can then removes the subparts and abstracts the word to its root. Table .1 gives an example for this process of removal. Thus, in this example the root of all the noted words (المدارس، دارسون ، مدرسات) after removing subparts is the unique root of "dares" (درس) .

TABLE I. DIFFERENT WORDS HAVE VARIOUS SUBPARTS AND A SAME ROOT

Derivation (التفعيلة)	Suffixes	Infixes	Prefixes
المدارس	-	ا	ا + ل + م
دارسون	و + ن	ا	-
مدرسات	ا + ت	-	م

V. THE PROPOSED SUMMARIZATION MODEL

The main stage of processing in the presented model is oriented towards finding the root of each sentence. Based on the roots found in a text, words can be grouped in distinct clusters. It is thought that important words in any text appear more than once. This fact is considered as the main principle

to summarize a given text into an outcome of a summary using the words of high frequencies. For the purpose of explanation a common root set of words are given in Table.2.

TABLE II. A COMMON ROOT SET OF WORDS

Word (English)	Word (Arabic Voice)	Arabic Form (الكلمة)
Sciences	Aaloom	علوم
The Learners	Almotaalemon	المتعلمون
Learning	Yataalem	يتعلم
Scientists	Alolamaa	علماء

Obviously the first step in this investigation is to find the root of the set given above of words. The root is (Eaalm, علم). When the root is specified, all the words then are put in one cluster. Each word in this cluster thus holds a frequency value which represents the number of words in the cluster. In the example of Table.2 the frequency of each word is 4, since the number of words in this cluster is 4.

Root	الجذر
Eaalm	علم

When the summarization processing is run, the document involving this set of words would be decided as if it is oriented towards the (Eaalm, العلم) Science Topic because any word of this cluster will take a higher score

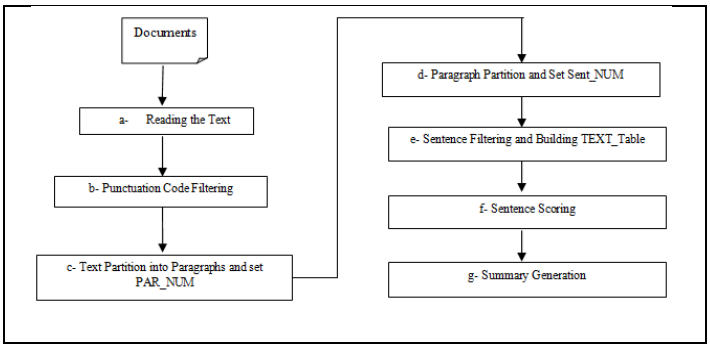

Figure 2. Model Main Processing Stages

The general flow of processing manipulating the document along the different stages is summarized in Fig. 2. C# is used for the coding purposes of the different stages of the presented model of summarization. The functional characteristics of each stage are explained as follows:

1 First, the document of the type Txt/MS-Word is fed into the model. These formats represent the most common used formats in documentation purposes.

2 Then the model divides the original text into a number of paragraphs, paragraphs to sentences, sentences to words. This process achieved by building a table that contains three fields: the first one for paragraph number, the second for sentence number, and last one for the body sentences. This stage includes the following:

a) Divide text into numbered paragraphs and save them in the table.

b) Divide the paragraphs into numbered sentences and save them in the table.

c) 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 ....).

3 The next stage is to implement stemmer that finds the root of each word in each sentence of the original text. This means that word subparts (suffixes, prefixes, and infixes) must be removed. After that, the words with the same root will be in the same cluster, the number of words in that cluster will determine the weight of each word in the cluster.

4 Finding the weight of each word in the sentence using the following equation:

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

- Where $W_{i,j}$ 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 which is obtained from step c
- *tf* (term frequency) = n_i / *max* n_i *(i.e* frequency of word *i/ max* frequency in document)

5 Then the model calculates the score of each sentences using following equation:

$$s(i) = \sum(W_{i,j}) \quad (2)$$

6 Now, in Arabic language there are remarkable words that increases the importance of the sentence, such as: (this indicates that: يدل ذلك, the most important thing: اهم الامور ,…etc). Such words are saved in the database. Thus, the sentence score increases if it has one or more of these words according to the equation

$$s(i) = sum\ (Wi,j) + A \quad (3)$$

- Where s (i): score of sentence I
- A is a constant given for the important key word.

This step may increase the probability of the sentence to appear in the summary. Moreover, the type of these key words used in the system is not necessary to be single, it can be a phrase.

7 Finally, the model takes the sentences with the highest scores and considers these sentences as a summary of the paragraph. The number of the sentences that will be taken depends on the size of document. After that, the model re-arranges the selected sentences according to their score and combines them into one paragraph.

VI. EXPERIMENTS AND RESULTS

The presented model of summarization has been applied on 10 Arabic different documents. An amount of about 2700 words are involved in each document with diverse paragraph structures. Obviously in summarization, efficiency measure is not of a deterministic characteristic but it is so far been considered as one of the significant dilemma obstacles efforts of validity comparison. Despite of the way being manual or automatic in summarization, there is no explicit referenced quality of output can be used for the relative measures of any comparative study. A text can have different summary when being subjected to different human efforts or programming activities. However in literature there are a number of developed evaluative techniques for summarization efficiency measures. They are typically classified into two categories: Intrinsic and extrinsic evaluation [13]. Both methods require preliminary human efforts to attribute a referenced measure.

To evaluate the efficiency of the presented model of the current work, a technique of [11] is applied. Four different people are requested to read the documents and later their summarizations are overlapped. The common sentences only of the four summaries are collected to build the reference summarization structure. With the resulting structure, two measures of Precision and Recall are evaluated as:

$$\text{Precision} = \frac{\text{The number of retrieve and relevant sentences extracted by the system}}{\text{Total number of sentences extracted by the system}}$$

$$\text{Recall} = \frac{\text{The number of retrieve and relevant sentences extracted by the system}}{\text{Total number of sentences extracted manually}}$$

Actually in both evaluations, human decision is needed to specify the number of logical useful sentences in each case of the measured criteria. A conceptual definition of "Precision" as a measure gives the ratio of the number of the representative logical sentences that is decided by human logic and extracted by the model to the total number of sentences extracted by the model. Whereas the second measure "Recall" indicated by the ratio of the number of those sentences found suitable by human decision and extracted by the model to the total number of sentences extracted by human. In other words, Precision estimates the efficiency of model power of filtering useful expressions from self generated raw expressions, whereas the second gives logic comparison between artificial efficiency to natural human logic.

Table .3 gives the obtained results of the experimental view of the work. As it is mentioned previously, 10 different structures of documents are tested and the related measures of Recall / Precision are recorded and compared with a presented work which depended noun/verb categorization method [14]. These measures have different scores along the tested documents. This in fact is attributed to many factors. The most important ones denote sentence length, existence of key words in sentences, number of roots that exist in each cluster besides document length

TABLE III. RECALL / PRECISION MEASURES OF THE TESTED 10 DOCUMENTS

Document No.	**Recall / Precision**
1	0.85 / 0.82
2	0.84/0.87
3	0.78/0.78
4	0.69/0.72
5	0.76/0.73
6	0.83/0.68
7	0.78/0.69
8	0.79/0.76
9	0.77/0.72
10	0.78/0.8
Average	0.787/0.757

VII. CONCLUSION

In this paper, a new automatic Arabic text summarization model is presented and discussed. The major attribute of this model is the word rooting capability. This consideration made the model closer to the semantic foundations rather being of a syntax based. Arabic language depends on multi derivations of the wording structures. Throughout these derivations, meanings are formulated to suit the actions and their associated environment whether regarding actors, action receivers or even the circumstances concerned with the actions. Those modalities of derivations made the variations much wider than other languages. In this work, a trend of collecting all the possible modalities of any word into a specified cluster. Such common meaning effectively eliminates the structures and abstract them into unique word. As the results show, a convenient summarization levels have been scored with an average of Recall 0.787 to Precision of 0.757. Results of a similar study adopted Arabic articles gave a scores of 0.62 to 0.70 for the concerned factors respectively. The latter work depends on verb/ noun categorization technique.

REFERENCES

[1] Attia, M. , "A Large-Scale Computational Processor of The Arabic Morphology, and Applications", MSc. thesis, Dept. of Computer Engineering, Faculty of Engineering, Cairo University, 2000.

[2] Jiang Xiao-yu, "Chinese Automatic Text Summarization Based on Keyword Extraction ", First International Workshop on Database Technology and Applications , pp: 225-228 ,2009.

[3] Ohm Sornil and Kornnika Gree-ut," An Automatic Text Summarization Approach Using Content – Based and Graph Based Characteristics", Conference on Cybernetics and Intelligent Systems pp: 1-6 , 2006.

[4] H. Saggion, K. Bontcheva and H. Cunningham, " Robust Generic and Query-Based Summarization", In proceedings of the European chapter of computational linguistics (EACL), Research notes and Demos, 2002.

[5] Witbrock M. J. and Mittal, V. O.," Ultra-Summarization: A statistical Approach to Generating Highly Condensed Non-Extractive Summaries", In proceeding of the 22nd annual international ACM-SIGIR conference on research and development in information retrieval, pp: 314-315, 1999.

[6] Harabagiu, S., and Lacatusu, F., "Generarting Singleb and Multi-Document Summaries with GISTexter. In proceedings of the DUC, pp:30-39,2002.

[7] Mahmoud O.El-Haj and Bassam H. Hammo," Evaluation of Query-Based Arabic Text Summarization System", International Conference on Natural Language Processing and Knowledge Engineering, pp**:** 1-7 , 2008.

[8] Mohammed Albared, Nazlia Omar and Mohd J. Ab Aziz, "Classifier Combination to Arabic Morphosyntactic Disambiguation", International conference on electrical engineering and informatics, pp: 163-171, 2004.

[9] Xu, J., Fraser, A., and Weischedel, R., "Empirical Studies in Strategies for Arabic Retrieval", In Sigir ACM, 2002.

[10] Hammo, B., Abu-Salem, H., Lytinen, S., Evens, M., "QARAB: A Question Answering System to Support the Arabic Language", Workshop on computational approaches to semitic languages, ACL, pp: 55-65, 2002.

[11] Aqil Azmi, Suha Al- Thanyyan, "Ikhtasir- A User Selected Compression Ratio Arabic Text Summarization System", International Conference on Natural Language Processing and Knowledge Engineering. pp:: 1-7, 2009.

[12] Ricardo Baeza- Yates, Berihier Riberio Neto, "Modern Information Retrieval" Addison Wesley, A division of the association for computing machinery,Inc. ISBN 0-201-39829-X, 1999.

[13] Te-Min Chang and Wen-Feng Hsiao, "A Hybrid Approach to Automatic Text Summarization", 8th IEEE International Conference on Computer and Information Technology, pp: 65-70, 2008.

[14] Qasem A. Al-Radaideh and Mohammad Afif, "Arabic Text Summarization Using Aggregate Similarity", The international Arab conference on information Technology, 2011.

[15] Vishal Gupta, Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive Techniques", Journal of Emerging Technologies in Web Intelligence, Volume: 2, Issue: 3, PP: 258-268, 2010.

AUTHORS PROFILE

Dr. Ahmad Haboush is an assistant professor in the Department of Computer Science, Jerash Private University, Jerash, Jordan. He received his BS, MS and PhD degree in Computer Engineering from Kharkov State Poly-technical University, Kharkov, Ukraine. His research interest includes security, parallel processing, artificial intelligence, information retrieval and software engineering.

Maryam F. Al-zoubi received her M. Sc. in Computer Science, from Yarmouk University in 2004. Her main area of research is Natural language processing. Currently, she is an active researcher in the field Automatic text summarization. Also, she is interests in the field of e-learning and producing software education programs for children. Now she is Full-Time Instructor in Jerash Private University, Jordan

Motassem Y. Al-Tarazi is a graduate student at the computer science department, Iowa State University. He received his MSc. Degree from Jordan University of Science and Technology in computer Science. Also he received his BSc. Degree in computer information systems from the same university. His research interests include: ad hoc networks, routing, wireless sensor networks

Ahmad A. Momani received his M. Sc. in Computer Science, from Jordan University of Science and Technology, Jordan in 2011. His main area of research is Computer Networks. Currently, he is an active researcher in the field of Mobile Ad Hoc Networks. More specifically, his research on MANETs is focused on developing MAC layer protocols. Now he is a part time lecturer in Jordan University of Science and Technology and Jerash Private University. Moreover, he is a teacher in the Ministry of Education since 2008.