Automatic Summarization of Arabic Texts

Dr. Mohammed M. A. Ibrahim Sakre

The Shorouk Academy
Higher Institute for Computers and Information Systems

M_sakre2001@yahoo.com

Abstract: *In this article, the linguistic and statistical approaches used in text summarization are presented. Statistical approach is adopted to build an Arabic text summarization system. Morphological analysis of words of the original text is used to support a heuristic for selecting the summary sentences. Evaluation of the proposed system is introduced. Primary subjective evaluation, based on Compression Ratio and Retention Ratio, showed that the used approach is effective and efficient, and performance of the system is promising.*

Keywords: Artificial Intelligence, Natural Language Processing, Information Retrieval, Text summarization, Arabic Morphological analyzer, Arabic language.

1- Introduction and background:

The main aim of automatic summarization is to extract and present the most important content to the user from an information source. Generally two types of summaries are generated: *extract*, i.e., a summary which contains text segments copied from the input, and *abstract*, i.e., a summary consisting of text segments which is not present in the input.

Current summarization systems can be categorized by the type of input that they handle, whether single or multiple documents, and by approach, whether extractive or abstractive. To allow summarization in arbitrary domains, most current single document summarization systems use sentence extraction, identifying and extracting key sentences from an input article using a variety of different criteria. The key sentences are then put together to form the summary. Early approaches used statistical metrics (e.g., word frequencies and key phrases) to identify important sentences [1]. Examples of these systems, for non Arabic languages, are SweSum [12] and EstSum [13] systems. SweSum is the first automatic text summarizer for the Swedish language. EstSum is a summarization tool for the Estonian language. Both SweSum and EstSum focus on extraction methods from a single document. SUMMONS [14] is an example of a multi-document summarizer

which extracts and combines information from multiple sources and passes this information to a language generation component to produce the final summary.

More recent approaches use a corpus of articles with summaries for training to identify the features of sentences that are typically included in abstracts. Other approaches use lexical chains, sentence position, discourse structure, and user features from the query to score sentences and rank them for selection.

Extractive systems tend to produce summaries with very long sentences; longer sentences score higher on metrics that rate them for importance. Abstractive approaches to single document summarization address this problem by editing the extracted sentences. They reduce a sentence by eliminating constituents which are not essential to be included in the summary. These approaches are based on the observation that the "importance" of a sentence constituent can often be determined based on shallow features, such as its syntactic role, the words it contains and their relation to surrounding sentences. Approaches for text compression have used symbolic reduction rules [2], as well as an aligned corpus of documents and their human written summaries to determine which constituents can be reduced [3, 4].

Summarization across multiple documents has also often been addressed through sentence extraction. Many approaches generate a summary that focuses on similarities found across all articles; they use clustering to find common themes within the articles [5,6] producing sets of sentences where each set, or theme, contains sentences saying roughly the same thing. Extractive approaches extracts one sentence from each set to form the summary. Other multi-document extractive approaches find and extract information about the centroid of the documents [7] or use spreading activation and graph matching to compute similarities and differences between the salient topics of two articles [8].

Only a few researchers have developed abstractive approaches for multi-document summarization. An approach based on information fusion [9] starts from the identification of themes as described above, but instead of extracting a representative sentence from the theme, it uses alignment to find phrases that occur in multiple sentences within the theme. These phrases are extracted and statistical language generation is used to fuse the phrases forming a novel sentence for the summary [11].

Text summarization plays crucial role in the development of effective and efficient information retrieval (IR) systems. Even very effective retrieval techniques can find large amounts of potentially interesting information, and it is important for a system to provide additional tools such as extraction and summarization. Progress in text summarization and extraction will not only enable the development of better retrieval systems, but will also support the access and analysis of text-based information in a number of novel ways helping to create discrete as well as

continual access systems. Integrated into existing automatic information retrieval systems they can be effectively used in the Internet search engines, providing users with summaries of documents thus enabling them to better identify relevant documents [10].

Text summarization can be combined with Information Retrieval (IR) and Question Answering (QA) to provide users with focus-based or query-based summaries which are targeted towards the users' specific needs. When the information a user is interested in is spread across multiple sources, text summarization can be used to condense facts and present a non-redundant account of the most relevant facts found across a set of documents [10].

2. Arabic Summarization Systems:

Text summarization, for non-Arabic languages, has reached a relatively mature stage; there are well established methods for summarization of a single document and many researchers are working on techniques for summarizing a set of related documents [11]. On the contrary, there is very little number of Arabic summarization systems, both commercial and under research, and still in early stages of development. It is interesting to note that many of researches which involve Arabic text summarization are in universities and research centers in western countries. An example for commercial Arabic systems is β-version Sakher summarization system[*], which is a tool that identifies the most relevant sentences within a text and displays them in the form of a short text summary. The Customer can select the generated summary according to a percentage of the input document, or according to a fixed number of phrases or size.

Another non-commercial research presents a prototype system that evaluates the relevancy of each sentence of the text to the desired domain. The system removes irrelevant sentences and keeps the most relevant sentence. The system is compared with Lakhas[17] commercial summarization tool and is evaluated by three human experts[15].

CLASSY (Clustering, Linguistics, And Statistics for Summarization Yield) is an automatic, extract-generating, summarization system that uses linguistic trimming and statistical methods to generate generic or topic (/query)-driven summaries for single documents or clusters of documents. [18]. CLASSY is considered to be a multi-lingual multi-document system.

[*] http://textmining.sakhr.com/Main.asp?Lang=1, date: 15 July 2008

Farsi Sum[16] is an attempt to create an automatic text summarization system for the Persian language that is an Arabic script-based language. It uses modules implemented in SweSum system [12].

Most of the above mentioned systems use extracting approach for summarization. Few of them make use of Arabic linguistic analysis like morphology, syntax or semantics.

3. The proposed system:

In this paper, the extraction (statistical based) approach is used to build a prototype system that make a summary for Arabic documents. This approach is supported with Arabic morphological analysis module to improve the sentence selection.

3.1 Modules of the proposed system:

The main modification of the proposed system considers improving the sentence scoring method. Then we select the sentences with the highest score for the summary. This improvement is based on two facts. The first fact is that words which are frequent in a document indicate the topic discussed [1]. The second is that Arabic words with the same root, in the same text, are most probable have semantic relations.

The proposed system gives the user the freedom to choose the summary size by selecting the percentage, he/she wants, of the original text. Also, the proposed system takes into consideration the sentence position and cue words (indicative phrases) [1].

So, the proposed system functions as following:

- *Divide the Input-Text into Words:* extract all words of the original document (including verbs, stopping words, etc).
- *Remove Stop words:* remove all words that are not significant, like من ، عن ، ما،
- *Normalize the Arabic characters:* this is by removing the diacritics like (ً ، ٌ ، ٍ) and by replacing the many forms of a single Arabic character by only one of them like replacing (أ ، إ ، آ) by (ا).
- *Get root for each word in the rest of text* and attach with it the semantic tag of its original word. The process of getting roots is discussed in the next section in more details. Give score to each root (accompanied with its semantic tag) according to the number of its appearance. The score of each root will be the same score for each word in the text of that root with the same semantic tag. For example the root (ك ت ب) will be assigned to the words (كتيب ، كتاب ، كُتب ،….) which have semantic relation, while it will not be assigned to (مكتب) because it has a different semantic tag.

- Give score to each sentence according to the following equation:

$$\text{Sentence score} = \frac{\text{Sum of scores of sentence's words}}{\text{Number of words in the sentence}}$$

- Taking the sentence position into consideration [1], increment the score of the first and last sentence by a constant value 'X'.
- Increment sentence which contains cue words (indicative phrases) by a constant value 'Y'. Cue words(phrases) like (الهدف هو) and (... الخلاصة هى)
- Extract number of sentences which satisfy the user percentage and with the highest scores. Keep the order of these sentences as their appearance order in the original text. These set of sentences represent the output summary.

3.2 Arabic morphological analysis:

The aim of using Arabic morphological analysis is to detect the list of keywords in the original text that have the same semantic or have semantic relation(s). Prolog is used as a powerful tool to accomplish this task [19], [20]. Each Arabic word in the original text is morphologically analyzed in three steps to find its root. These steps are [19] :

Stripping the prefixes off the word:

This is implemented in Prolog by defining a set of predicates for the Arabic prefixes like(.... ، إست ، ن ، سن). The general form of these predicates is:

Strip (Key_word, Rest).

Example of this predicate is: Strip(['إ','س','ت',Rest], Rest).

Stripping the postfixes off the word:

This is implemented in Prolog by defining a set of predicates for the Arabic postfixes like (.....، ت، ك، كم، هما، هم). The general form of these predicates is:

Strip_reverse (Key_word, Rest).

Example of this predicate is: Strip_reverse ([Rest,'هـ','م','ا'], Rest).

Finding the root from the rest of the word: This is implemented in Prolog by defining a set of predicates for all the possible Arabic weights of derivatives, "المشتقات ", and the corresponding roots. The general form of these predicates is: Weight(Rest_of_Arabic_word, Root).

Example of these predicates are:

Weight(['X','Y','Z'], ['X','Y','Z'])…. Three-letters root.

Weight(['ا','ن',X','Y','Z'], (['X','Y','Z'])…. Three-letters root.

Weight(['ت','X','Y','Z','W'], (['X','Y','Z','W'])….four-letters root.

A semantic dictionary, which defines a semantic tag for each Arabic word, is consulted to find the semantic tag of the original word and attach it to the root.

4. Results and evaluation:

There are two basic properties of the summary that must be measured when evaluating summaries and summarization systems [21],[22]. These are Compression Ratio and Retention Ratio.

The Compression Ratio (CR) is defined as the ratio between the summary text length (in words or sentences) and the original text length (in words or sentences). The Compression Ratio must be shorter than the original input text.

The Retention Ratio (RR) is defined as the ratio between the information in the summary and the information in the original text. Therefore, it reflects the degree of precision of the summary. It can be said that a good summary is one in which *CR* is small (tending to zero) while *RR* is large (tending to unity).

The proposed system attempts to get the summary length as close to the user's desire as possible, so the resultant CR might slightly deviate from the required percentage as shown in figures 1, 2, 3 where CR is selected by the user to be 50%, 30% and 20% respectively. The Retention Ratio (RR) is calculated by giving each sample run (the summary together with the original document) to three persons. Each one of them studies the information in both the original and summary texts. Each examiner gives a score out of ten to the summary, which represents the ratio between the information in the summary to that in the main text. The score of each of them is reported and the average score is calculated.

Another method for evaluating the proposed system is to compare its output with other systems' outputs. The output of the proposed system is compared with the output of Sakher summarization system for three input text files and the output is presented in table(1). It is noted that the proposed system is equivalent to or better than Sakher system in many runs.

4- Conclusion

The proposed system, discussed in this article, showed that using Arabic Morphological analysis has significant effects on Arabic text summarization. It reveals the need to have Arabic linguistic tools like Arabic semantic annotated dictionary. The strength of using such approach appears when it is compared with other commercial systems.

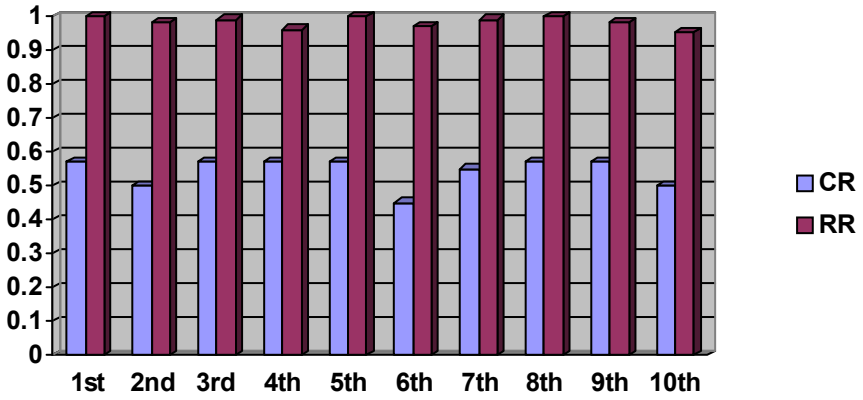

Figure 1 : CR and RR for ten summarized documents with CR = 50%

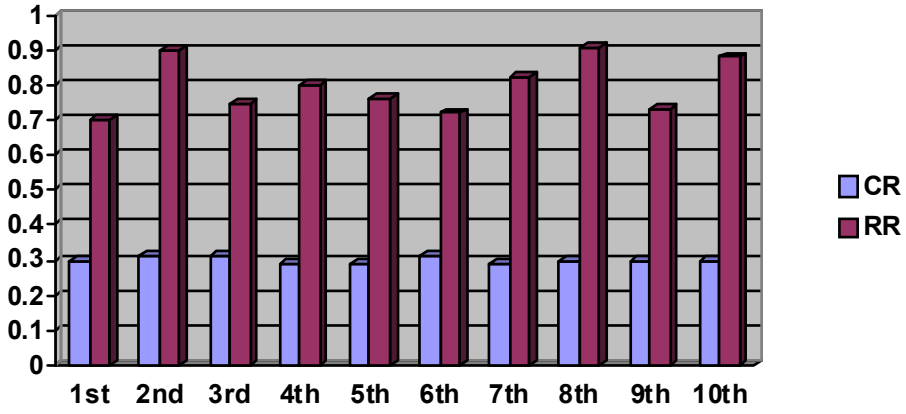

Figure 2 : CR and RR for ten summarized documents with CR = 30%

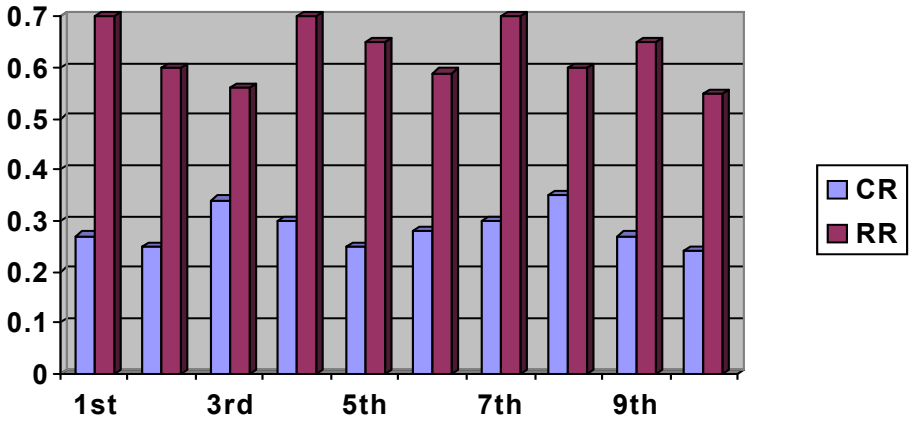

Figure 3 : CR and RR for ten summarized documents with CR = 20%

Table 1 : Comparing the summarization with Sakher system

Sample Run	The original text size words & sentences	Results of the proposed system			Results of the Sakher system	
		CR	text size words & sentences	RR	text size words & sentences	RR
1	1166 / 29	20% 30%	333 / 5 403 / 8	8 9	369 / 6 -	8 -
2	541 / 16	30%	172 / 4	8.5	147 / 4	4.6
3	469 / 7	50%	227 / 4	8.5	205/4	8.1

Referances

[1] I. Mani And M. T. Maybury, Editors. "Advances In Automatic Summarization" The Mit Press, Cambridge, Massachusetts, 1999.

[2] G. Grefenstette "Producing Intelligent Telegraphic Text Reduction To Provide An Audio Scanning Service For The Blind" In Proceedings Of The Aaai Spring Workshop On Intelligent Text Summarization, Pages 111–115, 1998.

[3] H. Jing And K. Mckeown "Cut And Paste Based Summarization" In Proceedings Of The First Naacl, Pages 178–185, Seattle, Washington, 2000.

[4] K. Knight And D. Marcu. "Statistics-Based Summarization-Step One: Sentence Compression" In Proceeding Of Aaai-01, Pages 703–710, Austin, Texas, 2001.

[5] J. Carbonell And J. Goldstein "The Use Of Mmr, Diversity Based Reranking For Reordering Documents And Producing Summaries" In Proceedings Of Sigir, Pages 335–336, 1998.

[6] V. Hatzivassiloglou, J. Klavans, And E. Eskin "Detecting Text Similarity Over Short Passages: Exploring Linguistic Feature Combinations Via Machine Learning" In Proceedings Of The Joint Sigdat Conference On Empirical Methods In Natural Language Processing And Very Large Corpora, 1999.

[7] D. Radev, H. Jing, And M. Budzikowska. "Centroid-Based Summarization Of Multiple Documents: Sentence Extraction, Utility-Based Evaluation, And User Studies". In Proceedings Of The Anlp/Naacl 2000 Workshop On Automatic Summarization, Pages 165–172, 2000.

[8] I. Mani And E. Bloedorn. "Multi-Document Summarization By Graph Search And Matching". In Proceedings Of Aaai-97, Pages 622–628, Providence, Rhode Island, 1997.

[9] R. Barzilay. "Information Fusion For Multi-Document Summarization: Paraphrasing And Generation. Phd Thesis, Columbia University, 2003.

[10] Chowdhury, G. " Introduction To Modern Information Retrieval", Second Edition, Facet Publishing, 2004.

[11] Mckeown, K.; Hirschberg, J.; Galley, M.; Maskey, S., "From Text To Speech Summarization" IEEE International Conference On Acoustics, Speech, And Signal Processing, 2005. Proceedings. (ICASSP Apos;05). Volume 5, Issue , 18-23 March 2005 Page(S): V/997 - V1000 Vol. 5. Digital Object Identifier 10.1109/ICASSP.2005.1416474.

[12] Dalianis, H. 2000. Swesum - A Text Summarizer For Swedish, Technical Report, TRITA-NA-P0015, Iplab-174, NADA, KTH, October 2000.

[13] Mutso, P.; Müürisep, K. "ESTSUM - Estonian Newspaper Texts Summarizer". Second Baltic Conference On Human Language Technologies; Tallinn, Estonia; Aptil 4-5, 2005. (Toim.) Langemets, M.; Penjam, P. Tallinn: Institute Of Cybernetics At Tallinn Technical University, 2005, 311 - 316.

[14] Dragomir R. Radev And Kathleen R. Mckeown. "Generating Natural Language Summaries From Multiple On-Line Sources". Computational Linguistics, 24(3):469-500, September 1998.

[15] Abdel Rahman Galal & Ibrahim F. Imam, "An Arabic Text Summarization System", The 4th International Conference On Information And Communications Technology (ICICT 2006).

[16] Martin Hassel And N. Mazdak , "Farsisum – A Persian Text Summarizer", COLING2004, Pages 82-84,2004

[17] Fouad Soufiane Douzidia And Guy Lapalme, "Lakhas, An Arabic Summarization System", Proceedings Of DUC, 2004.

[18] Judith D. Schlesinger, Dianne P. O'Leary, And John M. Conroy, "Arabic/English Multi-Document Summarization With CLASSY - The Past And The Future," Conference On Intelligent Text Processing And Computational Linguistics (Cicling), Haifa, Israel, February 17-23, 2008.

[19] Ibrahim M. M. "Information Retrieval Of Arabic Text Using A.I. Techniques", M.Sc. Thesis, Military Technical College, Cairo, Egypt, 1987.

[20] Gerald Gazder And Chris Mellish, " Natural Language Processing In Prolog ", Addison – Wesley Publisher Company, 1989.

[21] Firmin, T. And M. J. Chrzanowski.(1999). 'An Evaluation Of Text Summarization Systems.' In Mani And Maybury (1999).

[22] Hovy, E,' TEXT SUMMARIZATION '.In Mani And Maybury. (2000).