

Experimenting with Automatic Text Summarization for Arabic

Mahmoud El-Haj, Udo Kruschwitz, Chris Fox

University of Essex
School of Computer Science and Electronic Engineering
{melhaj, udo, foxcj}@essex.ac.uk

Abstract

The volume of information available on the Web is increasing rapidly. The need for systems that can automatically summarize documents is becoming ever more desirable. For this reason, text summarization has quickly grown into a major research area as illustrated by the DUC and TAC conference series. Summarization systems for Arabic are however still not as sophisticated and as reliable as those developed for languages like English. In this paper we discuss two summarization systems for Arabic and report on a large user study performed on these systems. The first system, the Arabic Query-Based Text Summarization System (AQBTS), uses standard retrieval methods to map a query against a document collection and to create a summary. The second system, the Arabic Concept-Based Text Summarization System (ACBTSS), creates a query-independent document summary. Five groups of users from different ages and educational levels participated in evaluating our systems. Each group had 300 individuals. We also performed a comparative evaluation with a commercial Arabic summarization system.

Keywords: Arabic Natural Language Processing, Automatic Text Summarization, Query-based, Concept-based

1. Introduction

The aim of this paper is to report the results of experiments with two Arabic Summarization Systems: the Arabic Query-Based Text Summarization System (AQBTS) and the Arabic Concept-Based Text Summarization System (ACBTSS). In both systems we take a document written in the Arabic language and attempt to provide a summary. The system's primary source of knowledge is a collection of Arabic articles extracted from Wikipedia, a free online encyclopaedia¹. Automatic text summarization is the process in which a computer takes a text document as an input and produces a summary of that document as an output. There are various approaches to text summarization, some of which have been around for more than 40 years (Luhn, 1958).

2. Related Work

Over time, there have been various approaches to automatic text summarization. These approaches include single-document and multi-document summarization. One of the techniques of single-document summarization is summarization through extraction. This relies on the idea of extracting what appear to be the most important or significant units of information from a document and then combining these units to generate a summary. The extracted units differ from one system to another. Most of the systems use sentences as units while others work with larger units such as paragraphs. Assessing the importance of the extracted units depends on some statistical measures. Each unit is given a score based on features such as word frequencies (Luhn, 1958), position in the text (Baxendale, 1958), and the presence of key phrases (Edmundson, 1969). Recent approaches use more sophisticated techniques for deciding which sentences to extract. These techniques include machine learning (Leite and Rino, 2008), to identify important features, and various natural language processing techniques to

identify key passages and relationships between words. Bayesian classifiers have also been used (Kupiec, 1995). Evaluating the quality and consistency of a generated summary has proven to be a difficult problem (Fizman et al., 2008). This is mainly because there is no obvious ideal summary. The use of various models for system evaluation may help in solving this problem. Automatic evaluation metrics such as ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) have been shown to correlate well with human evaluations for content match in text summarization and machine translation. Other commonly used evaluations include measuring information by testing readers' understanding of automatically generated summaries. Human evaluation provides better results than automatic evaluation methods, but on the other hand the cost is high.

Research in Arabic Natural Language Processing (ANLP) has focused on the manipulation and processing of the structure of the language at morphological, lexical, and syntactic levels. Unfortunately, semantic processing of the Arabic language has not yet received enough attention (Haddad and Yaseen, 2005). There are some aspects that slow down progress in Arabic Natural Language Processing (NLP) compared to the accomplishments in English and other European languages (Diab et al., 2007) including the complex morphology, the absence of diacritics in written text and the fact that Arabic does not use capitalization. In addition to the above linguistic issues, there is also a shortage of Arabic corpora, lexicons and machine-readable dictionaries. These tools are essential to advance research in different areas. Despite these difficulties, there has been some success in tackling the problem of Arabic syntax (e.g. Al-Shammari, 2008; Elabbas, 2007).

3. Summarizers for Arabic: AQBTS and ACBTSS

AQBTS is a query-based single document summarizer system that takes an Arabic document and a query (in Arabic) and attempts to provide a reasonable summary

¹<http://www.wikipedia.org/>

for the document around this query. ACBTSS is a concept-based summarizer system that takes a bag-of-words representing a certain concept as the input to the system instead of a user's query. The summary will consist of those sentences in the documents that best match the words in the query, or concept. Figure 1 depicts the general flow diagram of our systems. Both systems consist of two modules: the first module is the *Document Selection*. In this phase the user searches the document collection to find documents that satisfy his/her query and then selects a document for summarization. The selection is performed using a simple concordance system. The second module is the *Document Summarization*. In this phase, the system starts by splitting the documents into sentences. Up to this phase both systems share the same work-flow. The difference between the two systems starts at the subsequent *Sentence Matching* phase. In AQBTS each sentence is compared against the user query to find relevant sentences. This is the same query that was used in the document selection module. The ACBTSS sentence matcher ignores the user query that was used to select the documents. Instead, each sentence is matched against a set of keywords that represent a given concept.

For the summarizer we have adopted the vector space model (VSM), which has been used successfully in the field of Information Retrieval (IR) (Salton, 1975, 1983 and 1989). The weighting scheme based on the vector space model makes use of two measures: the term frequency (TF) and the inverse document frequency (IDF).

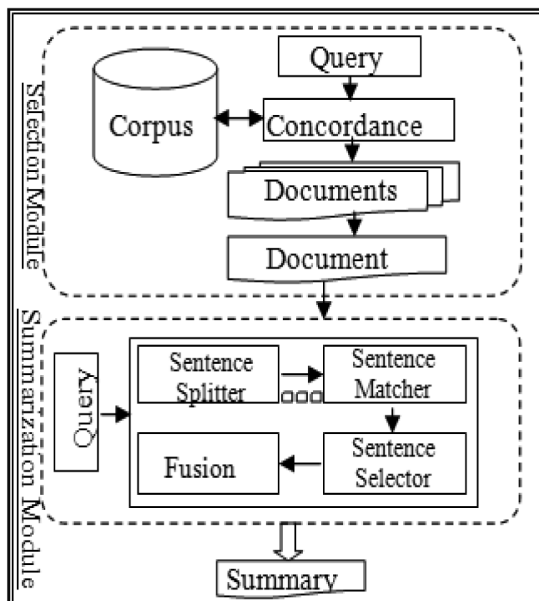


Fig. 1: AQBTS and ACBTSS diagram

To experiment with our system, we have collected 251 Arabic articles. The articles were downloaded from the Wikipedia website after obtaining their permission to use the articles for testing. The selected articles cover different topics in Arabic.

The set of concepts used in our concept-based summarizer include: art and music, environment, politics, sports, health, finance and insurance, science and technology, tourism, religion and education. Khreisat (2006) listed a set of concepts used for Arabic document

classification, in addition to these we added some more concepts that are commonly used by many Arabic newspapers. In our system the words used to represent a concept were selected based on running statistical experiments, where we processed 10,250 Arabic articles from different Arabic newspapers. The extracted articles fall in the above mentioned concepts, each concept with around 850 documents. Essentially we selected the most frequent terms for each category and subsequently deleted stopwords.

4. Experimental Design

We tested our system using a set of forty queries. Each query returned a set of documents that were then summarized by the two systems to give two summaries for each document. A group of 1,500 users participated in evaluating the readability of the generated summaries.

4.1. Document Collection Characteristics

The set of 251 articles used in our experiments were identified by asking a group of students to search Wikipedia website for articles using their own queries. The results of the process were a set of articles and their associated queries. The reason behind choosing different topics from TREC² is that we wanted the testers to select topics and articles that fall within their interests; we did not want the system to be biased to any predetermined topics. The total size of the collection was 95,933 words. The average size of each article was 378 words.

4.2. Evaluation Metrics

Each participant was handed a document and two summaries for the same document, the first summary is generated by AQBTS and the second by ACBTSS. The user cannot tell which summary came from which system as the papers were given to the participants in random order. Each participant was asked to read the document and its summaries and then to evaluate each summary based on a five-point Likert scale (Hoa, 2007). The scale measures, their corresponding scores and our interpretations are given in Table 1.

Table 1: Evaluation scale used to evaluate the systems.

Scale Measure	Score Value	Interpretation of the Measure
V. Poor	0	The summary is very poor and is not related to the document at all.
Poor	1	The summary is poor as the core meaning of the document is missing.
Fair	2	The user is somehow satisfied with the result, but expected more.
Good	3	The summary is readable and it carries the main idea of the document.
V. Good	4	The summary is of much readable and focuses more on the core meaning of the document and the user is happy with the results.

² <http://trec.nist.gov/>

4.3. Subjects

Five groups each of 300 individuals were involved in evaluating our system. The participants vary in their ages and educational levels. The selected groups were: students studying Arabic literature; students studying humanities; K-12 school teachers; K-12 school students and computer science students.

The variation of ages between participants helped us to understand the differences of their linguistic skills, while the variation of their backgrounds and degree subjects helped us to interpret their expectations from an Arabic summarization system; some of the groups are much more familiar with computer aspects than others.

The user groups in detail:

- Group 1 and 2: Arabic Literature and Humanities students.**
 These are third and fourth year students majoring in Arabic literature and Humanities at the University of Jordan.
- Group 3: Computer Science Students.**
 The members of this group are students at various levels majoring in Computer Science studying at King Abdullah School for Information Technology at the University of Jordan.
- Group 4: K-12 School Students.**
 The members of this group were from the 9th and 10th grade form private schools in Jordan.
- Group 5: K-12 School Teachers.**
 Our last group was K-12 school teachers from different specialties attending a one-year training session on ICT in education at the University of Jordan.

5. Results

We will first report the overall performance of the systems. Later, we discuss and explain the results we obtained from each individual group. Then we compare the results of some of the groups to identify any significant differences.

We also report results from an experiment to compare our query-based system with a commercial product by Sakhr³. This time we only used one group of 300 participants (Computer Science Students) and asked them to evaluate the same documents, but this time using the Sakhr summarizer system.

To determine significance we performed standard t-tests ($p < 0.05$), by testing each group (300 observations) on both systems.

5.1. AQBTTSS versus ACBTSS

In the case of AQBTTSS the queries used to select the documents are used again to summarize them. For ACBTSS the concepts' words are those described in section 3.1.

Each member of the five participating groups evaluated a summary generated by AQBTTSS and by ACBTSS. Table 2 depicts the results of the five groups of evaluators for AQBTTSS. The results are reproduced from (El-Haj, 2008). The results for ACBTSS are given in Table 3.

The results of significance testing (Table 4) show that all user groups apart from the humanities students gave

significantly higher ratings for the query-based system than the query-independent system.

Table 2: Overall gradings of the AQBTTSS system.

Group	Scale Measures and Scores					Good + V. Good
	(0) V. Poor	(1) Poor	(2) Fair	(3) Good	(4) V. Good	
K-12 Teachers	0.00%	2.00%	7.67%	47.33%	43.00%	90.33%
Arabic Lit. Students	0.00%	4.00%	11.67%	46.33%	38.00%	84.33%
Humanities Students	0.33%	5.00%	14.00%	57.67%	23.00%	80.67%
K-12 Students	0.67%	3.33%	19.33%	39.33%	37.33%	76.67%
CS Students	1.67%	7.00%	24.00%	44.00%	23.33%	67.33%
Overall Performance	0.53%	4.20%	15.40%	46.93%	32.93%	79.87%

Table 3: Overall gradings of the ACBTSS system.

Group	Scale Measures and Scores					Good + V. Good
	(0) V. Poor	(1) Poor	(2) Fair	(3) Good	(4) V. Good	
K-12 Teachers	0.67%	5.00%	21.33%	38.67%	34.33%	73.00%
Arabic Lit. Students	1.00%	7.33%	29.67%	33.33%	28.67%	62.00%
Humanities Students	1.00%	4.67%	18.00%	49.00%	27.33%	76.33%
K-12 Students	0.67%	6.33%	24.33%	42.00%	26.67%	68.67%
CS Students	2.33%	16.00%	35.67%	30.33%	15.67%	46.00%
Overall Performance	1.13%	7.87%	25.80%	38.67%	26.53%	65.20%

Table 4: t-test results.

Group	Mean (ACBTSS)	Mean (AQBTTSS)	p
Humanities Students	2.970	2.980	0.440403
K-12 Students	2.877	3.093	0.001405
K-12 Teachers	3.010	3.313	2.69E-06
Computer Science	2.410	2.803	4.59E-07
Arabic Students	2.813	3.183	1.95E-07
ACBTSS VS AQBTTSS	2.816	3.0747	1.76E-15

5.2. Sakhr Summarization System

The Sakhr Text Summarization System is a commercial online Arabic text summarization system available on the web. It should be noted that the system was only a beta release at the time we performed our experiments. The summarizer consists of a set of text-mining tools to identify the most relevant sentences within a document and displays them in the form of a prioritized list of key sentences.

We ran the following experiment. First, we used the same set of forty documents we used throughout all our experiments and obtained their summaries from the Sakhr summarization system. We asked the Computer Science students group to evaluate the results obtained from Sakhr without telling them the source of the new summaries. Figure 2 shows the results of evaluation

³ <http://www.sakhr.com/>

obtained from Sakhr compared to those we observed for AQBTS (see Table 2). We do not include any of the results obtained from the other user groups as this experiment was only performed with the Computer Science students. This group was the one that assigned the lowest average score to both AQBTS and ACBTS.

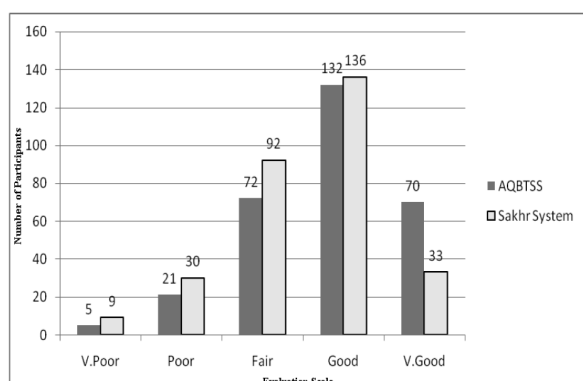


Fig. 2: AQBTS vs. Sakhr

6. Discussion of Results

6.1. The Groups

For AQBTS, Table 2, the group of K-12 Teachers gave on average the highest gradings followed by the group of students majoring in Arabic literature. The K-12 Teachers' gradings on average were significantly higher than those of any other group. The lowest gradings were awarded by the Computer Science students.

In the case of ACBTS, Table 3, the group of Humanities Major Students gave on average the highest gradings, followed by the group of K-12 Teachers. Unlike in the query-based system, significant differences between the group giving the highest gradings and the other groups can only be shown for two groups: Computer Science and Arabic Major Students. As before, the lowest gradings came from the Computer Science students.

6.2. The Systems

As shown in Table 4, overall the query-based system performed significantly better than the concept-based summarizer. This preference is perhaps not surprising as a summary is created for a specific query by AQBTS whereas the concept-based system creates a summary where the query is replaced by the set of conceptual terms (representing the particular category under which this document was classified) before a summary is created.

If we analyze each user group separately, we find that only the Humanities students did not show a significant difference in preference over one or the other system, although the average rating for the query-based summary (2.98) was also higher than for the concept-based system (2.97). All other groups appear to strongly prefer summaries coming back from the query-based summarizer.

When comparing our query-based summarizer AQBTS with the Sakhr summarizer we found that the "most critical" user group, i.e. the one that gave our system the lowest average with a score of 2.81, considered the commercial system to be performing significantly worse, with an average score of 2.52. We

hypothesize that the same results would be obtained with any of the other user groups. The most remarkable observation is that our system resulted in a significantly higher average rating by the subjects than the rating for the commercial baseline.

7. Conclusions and Future Work

The overall conclusion we draw from these experiments is that our query-based summarizer performs much better than the concept-based one and it even outperforms a sensible baseline.

The research described here accomplished several goals. We built a concept-based and a query-based text summarization system that processes and summarizes Arabic natural language documents. Because of the lack of public-domain tools for Arabic compared to what is available for English, we developed a set of useful tools such as a stemmer, tokenizer and stopwords removal to carry out our experiments and to conduct future research in Arabic NLP. Finally, we carried out experiments to evaluate the system. The evaluation results of the summaries and the way they were interpreted by each group helped us to identify some directions to improve the performance of our system and opened some research avenues for the future. We believe that the results, and our assumptions, need more investigation, and merit more theoretical analysis.

The comparison between AQBTS and Sakhr is based on assessors' results; it does not compare the techniques and tools used in both systems. At the time we performed this experiment, Sakhr was still in its beta version. The performance of Sakhr may now have been improved by the use of morphological analysis.

In future work we intend to produce improved query-based and concept-based text summarization systems and further advance research in Arabic NLP. We have plans to increase the number of queries and documents in the test collection. We plan to use a categorized document collection and summarize a document according to a certain category and measure the effectiveness on the system. We also propose to apply Latent Semantic Analysis (LSA) in an attempt to increase the system's ability to select relevant documents and sentences. We plan to automatically evaluate our systems by using metrics such as ROUGE and BLEU.

Acknowledgment

We would like to extend our gratitude to Prof. Nadim Obeid and Dr. Bassam Hammo from the CIS Department at King Abdullah II School for Information Technology/University of Jordan for their valuable suggestions and participation in the accomplishment of the evaluation process. We also want to thank the anonymous reviewers for useful comments.

References

- Al-Shammari, E.T. and Lin, J. (2008). Towards an error-free Arabic stemming. *In Proceeding of the 2nd ACM Workshop on Improving Non English Web Searching*. iNEWS '08. ACM, California, USA.
- Baxendale, P.B. (1958). Man-Made Index for Technical Literature - An Experiment. *IBM Journal of Research and Development*, Vol. 2, (pp. 346-361).

- Diab, M., Jurafsky, D., Hacıoglu, K. (2007). Automatic Processing of Modern Standard Arabic Text. *In the Book of Arabic Computational Morphology* (Vol. 38). Chapter 9, (pp. 159-179). Springer Netherlands.
- Edmundson, H.P. (1969). New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery*, 16(2), (pp. 264–285).
- Elabbas, B. (2007). Perspectives on Arabic Linguistics XIX: *Papers from the Nineteenth Annual Symposium on Arabic Linguistics*, Urbana, April 2005. John Benjamin's Publishing Company.
- El-Haj, M. and Hammo, B. (2008). Evaluation of Query-Based Arabic Text Summarization System. *In Proceeding of the NLP-KE 2008*, IEEE, Beijing, China.
- Fiszman, M., Demner-Fushman, D., Kilicoglu, H., Rindfleisch, T. C. (2008). Automatic summarization of MEDLINE citations for evidence-based medical treatment: *A topic-oriented evaluation*. *Journal of Biomedical Informatics*.
- Haddad, B. and Yaseen, M. (2005). A Compositional Approach towards Semantic Representation and Construction of ARABIC. *Logical Aspects of Computational Linguistics*, (pp. 147–161). Berlin / Heidelberg: Springer.
- Hoa, T.D. 2007, Overview of DUC (2007). *In Proceedings of the Seventh Document Understanding Conference (DUC)*. New York, USA.
- Khreisat, L. (2006). Arabic text classification using N-gram frequency statistics: A comparative study. *In Proceedings of the 2006 international conference on data mining*, (pp. 78–82).
- Kupiec, J., Pedersen, J., and Chen, F. (1995). A trainable document summarizer. *In Proceedings of the 18th Annual international ACM SIGIR Conference on Research and Development in information Retrieval*, Seattle, Washington, United States.
- Leite, D.S. and Rino, L.H. (2008). Combining Multiple Features for Automatic Text Summarization through Machine Learning. *In Proceedings of the 8th international Conference on Computational Processing of the Portuguese Language*, Aveiro, Portugal.
- Lin, C. (2004). Rouge: A package for automatic evaluation of summaries. *In Workshop on Text Summarization Branches Out at ACL*, pages (pp. 74–81).
- Luhn, H.P. (1958). The Automatic Creation of Literature Abstracts. *In IBM Journal of Research and Development*, vol. 2. no. 2, (pp. 159–162).
- Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (Philadelphia, Pennsylvania, July 07 - 12, 2002). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ
- Salton, G. (1989) Automatic Text Processing – *The Transformation Analysis and Retrieval of Information by Computer*. Addison Wesley, Reading.
- Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, New York, USA.
- Salton, G., Wong A., and Yang, S. 1975. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, vol. 18, no. 11, (pp. 613–620).