Arabic Snippet Generation Using Extractive Query-based Summarization

Mariam Safieldin, Supervisor: Dr. Shady Elbassiouni
Email: mas177@mail.aub.edu

Introduction

Summaries are commonly used in a variety of settings and at all times. Paper abstracts, book reviews, and search results snippets are all examples of summaries. With the extensive use of search engines and the increase of resources on the web, the need of summarization and information need specific search results became very important. The goal of query-based document summarization is to provide a brief summary of the source material that includes relevant information for the user's information need. Automatic summarizing is an active area, particularly in the Arabic language, where the amount of research on automatic Arabic text summary is limited and growing slowly in comparison to the amount of literature on other languages [2].

The main objective of this project is to extract the sentences related to the search query in Arabic taken as an input to the system to produce a snippet of the document gathering the related sentences in it to the query.

Figure 1 below clarifies our target case showing the input and output of the proposed approach.

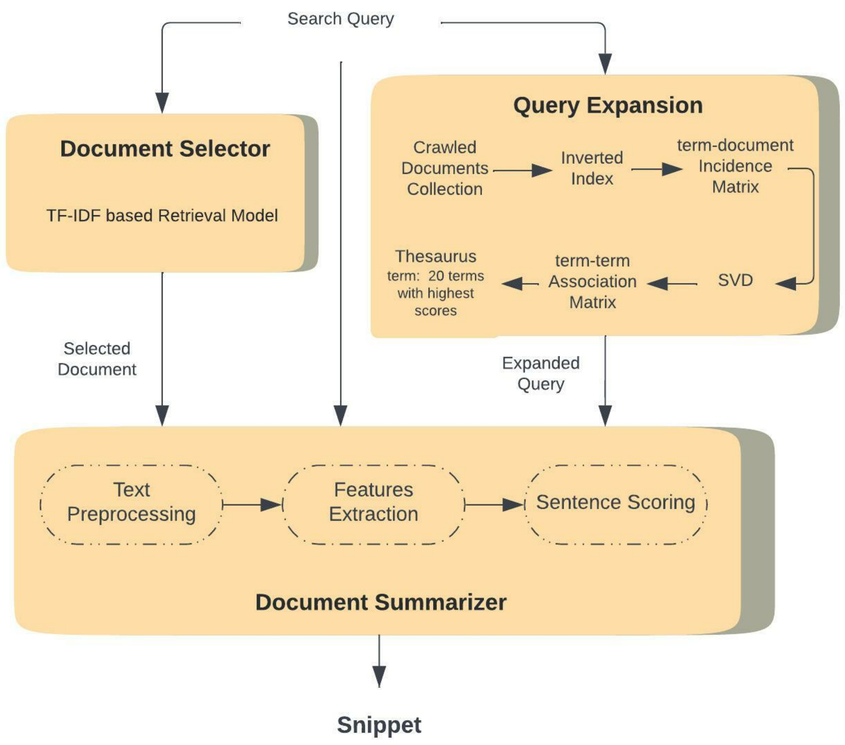

Figure 1: Arabic Query-based Summarization System

Methodology

Text Preprocessing

First, text goes through the following preprocessing steps:

- Removing the punctuation, the Arabic diacritics, and the stop words are removed from the text using NLTK's stop words list.
- The text is stemmed using Tashaphyne library's stemmer which was used to extract the root of each word.

Document Selection

Secondly, a TF-IDF vector space retrieval model, which uses a TF-IDF index built using the term frequency (TF) and inverse document frequency (IDF) to score the benchmark documents. Given an input search query, the terms of the query go through the text pre-processing steps in, then the TF-IDF scores of every term within each document are used to compute scores for the available documents in our benchmark.

Query Expansion

The query was expanded using a thesaurus which includes terms and their related terms. The thesaurus was build through crawling 30K+ documents, generating their inverted index, and using the first 8K values in it to generate a term-document incidence matrix A. Afterwards, the Singular Value Decomposition (SVD) of matrix A is computed, which was then used to compute the term-term association matrix U. Finally, matrix U was used to extract the 20 terms with highest scores for every present term [11].

Document Summarizer

Four query based features were extracted for every sentence in the selected document. Each feature and sub-feature was assigned a weight which was used to contribute in scoring the sentences. Query Sentence Unigram Overlap, w=1, Query Sentence Bigram Overlap, w=0.75. Part-of-speech Tagging with Query Comparison [5], w=0.65. Similarity Measures, w=1.

The Similarity Measures [10] score is derived using the following 4 sub-features: Manhattan Distance, w=0.7. Jaccard Similarity, w=0.75. Jaro Similarity, w=0.95. Longest Common Substring (LCS), w=0.85.

Results

Benchmark

The data collection used for evaluating the system is a collection of 26 articles, 33 queries and their snippets. 14 articles were crawled from different websites on Google through the *people also ask for* set of questions provided by which was used to form a query based on the question asked and take the answer as the snippet. 12 articles were used from the EASC dataset, and the corresponding snippet for a certain query was manually extracted.

Rouge for Evaluation

The built system was evaluated using the benchmark. The system's was evaluated with and without expanding the queries. The results are shown in tables 1 and 2. The results show that the features extracted were capable of ranking the sentences in accordance to the given search query. The query expansion component did not show a significant improvement to the system.

Metric	Recall	Precision	F1 Score
ROUGE-1	0.843	0.438	0.538
ROUGE-2	0.810	0.405	0.489
ROUGE-L	0.834	0.431	0.531

Table 1: System's ROUGE Evaluation Results on The Benchmark

Metric	Recall	Precision	F1 Score
ROUGE-1	0.845	0.424	0.523
ROUGE-2	0.809	0.378	0.477
ROUGE-L	0.838	0.418	0.522

Table 2: System's ROUGE Evaluation Results on The Benchmark with Query Expansion

Conclusions

In this project, an Arabic Query-based Text Summarization System was implemented. The target of the system is to produce an extractive summary that includes the relevant parts of a document to a given user information need in Arabic. To satisfy this objective, the system encompassed three main components, a query expansion component, a document selector component, and a document summarizer component which ranks document sentences based on extracted query-based features.

References

[1] Al-Taani, A. T. (2021). Recent Advances in Arabic Automatic Text Summarization. International Journal of Advances in Soft Computing & Its Applications, 13(3).

[2] Al-Saleh, A. B., & Menai, M. E. B. (2016). Automatic Arabic text summarization: a survey. Artificial Intelligence Review, 45(2), 203-234.

[3] Bialy, A. A., Gaheen, M. A., ElEraky, R. M., ElGamal, A. F., & Ewees, A. A. (2020). Single arabic document summarization using natural language processing technique. In Recent Advances in NLP: The Case of Arabic Language (pp. 17-37). Springer, Cham.

[4] Qaroush, A., Farha, I. A., Ghanem, W., Washaha, M., & Maali, E. (2021). An efficient single document Arabic text summarization using a combination of statistical and semantic features. Journal of King Saud University-Computer and Information Sciences, 33(6), 677-692.

[5] Afsharizadeh, M., Ebrahimpour-Komleh, H., & Bagheri, A. (2018, April). Query-oriented text summarization using sentence extraction technique. In 2018 4th international conference on web research (ICWR) (pp. 128-132). IEEE.

[6] Ahuja, R., & Anand, W. (2017). Multi-document text summarization using sentence extraction. In Artificial Intelligence and Evolutionary Computations in Engineering Systems (pp. 235-242). Springer, Singapore.

[7] El-Haj, M. O., & Hammo, B. H. (2008, October). Evaluation of query-based Arabic text summarization system. In 2008 International Conference on Natural Language Processing and Knowledge Engineering (pp. 1-7). IEEE.

[8] Al-Taani, A. T., & Al-Omour, M. M. (2014). An extractive graph-based Arabic text summarization approach. In The International Arab Conference on Information Technology.

[9] Ma, Y., & Wu, J. (2014, December). Combining n-gram and dependency word pair for multi-document summarization. In 2014 IEEE 17th International Conference on Computational Science and Engineering (pp. 27-31). IEEE.

[10] Wang, J., & Dong, Y. (2020). Measurement of text similarity: a survey. Information, 11(9), 421.

[11] Efron, M. (2008). Query expansion and dimensionality reduction: Notions of optimality in rocchio relevance feedback and latent semantic indexing. Information processing & management, 44(1), 163-180.

[12] Imam, I., Nounou, N., Hamouda, A., Allah, H., & Khalek, A. (2013). Query Based Arabic Text Summarization 1.

[13] Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).