International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

A Proposed Clustering Technique for Arabic Text Summarization

Raed Muhammad Albadrani[1], Mohammed Abdullah Al-Hagery[2], Mohamed Tahar Ben Othman[3]

[1]Master Student, Department of Computer Science, College of Science and Arts in Uglat Asugour University, Qassim, KSA
r.albdrani2018[at]gmail.com

[2]Professor, Department of Computer Science, College of Computer, Qassim University, Qassim, KSA
hajry[at]qu.edu.sa

[3]Professor, Department of Computer Science, College of Computer, Qassim University, Qassim, KSA,
maathaman*[at]qu.edu.sa*

Abstract: ***Since the internet revolution, the content of information is growing quickly in this digital world. Therefore, the need for techniques to benefit from that amount of information becomes very required. Automatic Text Summarization (ATS) is one of these techniques which is derived from the Text Mining area. It aims to shorten the size of the original text as much as possible. Consequently, we propose a Clustering-Based Technique for Arabic Text Summarization. The framework for this technique combines the cosine measure and phrase to cluster phrases using the k-means algorithm. Then, to rank these phrases, the framework uses a Modified Page Rank algorithm (MPR) with the value of the position sentences. Finally, it generates a summary that includes the most important sentences. This Technique of Arabic Text Summarization could produce a complete summary of Arabic documents, including most of the ideas of the original text. The evaluation process is done using a dataset derived from different domains, like education, sport, religion, music, and environment. The experiment results prove that the proposed approach gives better performance than previous works in the same domain.***

Keywords: Text Summarization, Arabic Language, Similarity Measures, Natural Language Processing

1. Introduction

Due to intense human activity, data size is rapidly growing in the digital world. This is one of the factors boosting research motivation in the fields of text mining (TM) and natural language processing (NLP). Automatic Text Summarization is one of the crucial areas of TM and NLP (ATS). ATS is an automatic process that creates a condensed or condensed version of a text that retains as much of the original text's key information as possible[1]. The value of ATS stems from its capacity to extract significant information from a given text. This improves the management process's quality, saves resources, and cuts down on the time it takes to retrieve the necessary data. As a result, ATS is becoming increasingly sought after in a variety of industries, including the legal, news, and medical ones. Because of this, ATS has rapidly expanded as a significant study topic. Based on the criteria used for ATS categorization, such as the number of input text documents, the type of returned summary, the output form, and others, ATS methods can be broken down into many groups. For instance, ATS can be a single or document summary depending on the number of input documents. On the other hand, depending on the kind of sentences that were returned, the ATS methods [2]:

- Extractive techniques: the summary is formed by choosing a portion (sentences, paragraphs) of the original text.
- Abstractive techniques: This method uses an internal semantic model of the text to represent it, and then it employs natural language generation techniques to create a summary that is as near to what a human might produce as possible. There's a chance the abstractive summary will use words that weren't in the original text. The material of an extractive summary, however, is entirely drawn from the original text. The studies to date have mainly concentrated on extractive approaches even though generating abstractive summaries is an important and active research area, but because of its complexity restrictions of it, the research to date has focused primarily on extractive methods [3]. Additionally, due to direct sentence retrieval, extractive techniques are quicker than abstractive and provide higher accuracy.

Being a language that is spoken as a first language in more than 25 nations, Arabic is a very popular language. In addition, Arabic is a highly organized, derivational language in which morphology is crucial. The intricacy of Arabic's semantic and syntactic structures, on the other hand, makes it a difficult language to learn [4].

Arabic hence, requires specific preprocessing for use in NLP and TM applications due to the numerous difficulties related to the character and structure of the Arabic language. Therefore, Arabic ATS continues to have poor performance and little study has been done in this area of natural language processing [6]. Moreover, text redundancy removal and subtopics addressing are two more significant issues in ATS, particularly in document summarization. We believe that these difficulties can be resolved by utilizing the clustering ability to lessen the repetition and identify the text subtopics. In addition, analyzing text similarity using phrases rather than just a random selection of words.

The purpose of this study is to propose a method for summarizing Arabic documents that rely on the combination of clustering, similarity based on phrases, and MPR algorithm.

The suggested strategy uses clustering as an extraction method to provide a single summary of Arabic documents. So, we call it "A Proposed Clustering-based Technique for Arabic Text Summary" (PCATS). The PCATS is based on the grouping of sentences using cosine similarity and combination equations of phrases. As shown in [7], this combination produces encouraging results for document clustering [7]. Preprocessing, feature extraction, sentence grouping, sentence ranking using Modified Google Page Rank (MPR), and summary extraction make up the PCATS' four main steps.

We anticipate the PCATS to have some benefits, including the capacity to manage redundant data and generate an exhaustive summary that covers the most crucial textual subtopics. The remaining of this paper is organized as follows: section 2 contains the related work, section 3 presents the methodology section 4 concentrates on the results and discussion, and finally, section 5 explains the conclusions.

2. Related Works

Finding the most significant sentences (which may also be the most informative) is the fundamental difficulty in the text summary process. Researchers have recently been paying more attention to the ATS. The text summarization makes several ways and strategies suggestions. Semantic-based, statistical-based, machine learning-based, cluster-based, graph-based, and discourse-based summarization are just a few of the numerous categories that these methods fall under. In this part, we will examine a few methods for ranking sentences that use machine learning, clustering, and specific datasets [8], as well as algorithms like Google Page Rank [9], MPR, and Text Rank [10].

Elbarougy et *al*. proposed in [6] an extractive Arabic text summarization approach. They represent each document as one graph. In this graph, each sentence is represented by a vertex while the value of cosine similarity between the sentences represents the edges between vertices (sentences). The ranking of sentences is achieved by applying the MPR Qaroush et *al.* [11] used a group of semantic features and the most informative statistics to produce a summary with rich information on a single document. This combination is used to detect the most important sentences and to reduce the redundancy in the resultant summary. The Essex Arabic Summaries Corpus (EASC) is used to evaluate the performance of this approach. In [12], a Graph-Based Arabic Text Summarizer based on NLP is sorted based on the ranks of nodes. The GPR combines the effect of both incoming and outgoing links of one sentence. The links between sentences are set based on the cosine similarity between sentences.

In [13], an approach called GGSDS is using clustering proposed to detect the subtopics of text. The GGSDS uses a combination of cosine and phrase equations [7] to calculate the similarity between sentences. The GGSDS exploited the Text Rank to rank the sentences to select the most important group. Experimental results showed that GGSDS generates summaries that covered most of the sub-topics of the English documents of the data set. GGSDS gives encourages design to generate a summary but it is applied to the English data set. Furthermore, using GGSDS to cluster the sentences need a lot of time, especially with big texts. As well, a proposed model was developed to identify the root of verbs [14] and another one by a software tool called RootIT to overcome the problem of verb root generation without disambiguation [15].

Besides, an approach for Arabic root generation presented by [16], is a novel technique for Arabic NLP to generate the roots of the Arabic word. likewise, Farwaneh, in [17], focused on the Levantine Arabic variety and created an account of a set of complex facts related to the inflexion of sound verbs and non-sound verbs. The account distinguishes four levels of correspondence, (input-output), and (output-output). Also, concentrated on the paradigmatic differences found in the inflection of sound verbs. This method concentrated on the stems of more than two consonants and on the non-sound verbs, whose stems comprise two consonantal realizations.

The previous methods and approaches could generate a better summary if they take into account the following points:

- Covering the text's subtopics, particularly in documents.
- Finding links between words when comparing sentence similarities.
- Using a beginning value for the sentences that is derived from the text itself rather than assuming that pre-values are either 1 or random.

We are enhancing the techniques [6] and [13] in our suggested PCATS to establish a higher performance by implementing it on the Arabic dataset and addressing the challenges they faced.

3. Methodology

The research process of PCATS includes several steps, as illustrated in Figure 1, to achieve the research objectives and address the issues raised above. These steps are as follows:

a) Making use of the EASC dataset [8], which has a variety of domains.
b) Preparing the corpus of Arabic text (Dataset), employing techniques that have been developed and tested.
c) Generation of the features of the text which are required to measure the similarity between sentences and to rank them.
d) Converting the clustering of sentences into clusters of similar sentences using the k-means algorithm. These clusters represent the sub-topics of the text.
e) Based on the importance of sentences rank them using the MPR algorithm.
f) Develop the summarization model (PCATS).
g) Evaluate the PCATS using the identified datasets, then compare its results with the results of others.

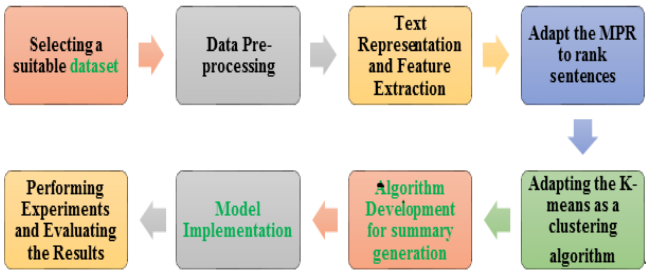

Figure 1: Methodology of PCATS approach

4. Results and Discussion

We selected one text document from the EASC dataset to apply the approach Fig 2. and Fig.3 shows, respectively, original text and the generated summary after applying the proposed approach; this summary is generated based on the proposed methodology. Figure 4 shows the accuracy of PCATS after considering only the most similar standard summary with the generated one. Due to the nature of our approach which mainly depends on the extractive method, thus, the overlap between single words is superior to others, which depends on the overlap between phrases (more than two words). Consequently, we notice the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) with ROUGE-1 overcome ROUGE-2 and ROUGE-L.

الناي آلة نفخية تعد بحق أقدم آلة موسيقية في التاريخ (إذا استثنينا الآلات الإيقاعية)وللناي عدة أسماء تعرف بها منها الناي القصبة الشبابة المنجيرة والمزمار . والناي كلمة فارسية تعني المزمار وهي آلة موسيقية مشهورة في التاريخ . ولأنها من أقدم آلات الموسيقي فهي تعتمد على مواد بدائية لتصنيعها، فهي قصبة مفتوحة الطرفين يعزف عليها بواسطة وضع الفم على أحد طرفيها مع إمالته قليلا بزاوية مما يجعل الهواء يصطدم بجدارها الداخلي مصدرا اللحن المطلوب وهو أقرب الأصوات وأجملها بالنسبة للإنسان. وللناي ستة ثقوب (وأحيانا سبعة) وثقب في منتصف القصبة من الأسفل. وتسد هذه الثقوب وتفتح حسب درجة الصوت ولإخراج العلامات بتسلسل يستطيع معه العازف إخراج العلامات الموسيقية لإخراج اللحن المطلوب وهو أقرب وأجمل أصوات للإنسان. والثقب الخلفي يسد بالإبهام ويستخدم لإظهار جواب العلامة الدنيا التي تظهر في البداية.

Figure 2: The original text

الناي آلة نفخية تعد بحق أقدم آلة موسيقية في التاريخ (إذا استثنينا الآلات الإيقاعية) وللناي عدة أسماء تعرف بها منها الناي القصبة الشبابة المنجيرة والمزمار. والناي كلمة فارسية تعني المزمار وهي آلة موسيقية مشهورة في التاريخ.

Figure 3: The generated summary

In contrast, the result of ROUGE-L overcomes ROUGE-2, because ROUGE-L is depending on the sequence of words throughout the whole sentence while the opposite concept is used in ROUGE-2 relying on shared two words. Therefore, this interprets that the probability of finding sequence words through the whole sentence is greater than the one of shared words.

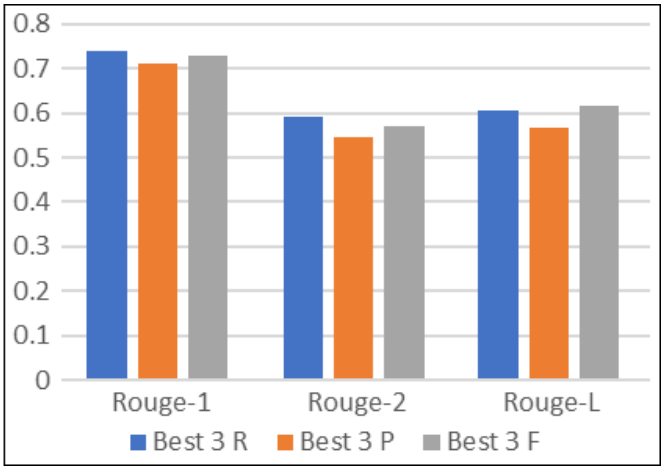

Figure 4: Results of the PCATS approach

5. Conclusion

In this paper, we proposed a Clustering-Based Technique for Arabic Text Summarization (PCATS). It is relying on a combination of cosine and phrase similarity to clustering sentences using the k-means algorithm. Then, using the Modified Page Rank algorithm (MPR) with the value of sentence position to rank the sentences. Finally, generate the final summary which contains the most important sentences. This Technique of Arabic Text Summarization could produce a complete summary of Arabic documents that includes most of the ideas of the original text. The dataset with various domains is used to evaluate this technique. The experiment results prove that the proposed approach gives better performance than previous works in the same domain.

References

[1] L. M. Al Qassem, D. Wang, Z. Al Mahmoud, H. Barada, A. Al-Rubaie, and N. I. Almoosa, "Automatic Arabic Summarization: A survey of methodologies and systems," Procedia Comput. Sci., vol. 117, pp. 10–18, 2017, doi: 10.1016/j.procs.2017.10.088.

[2] S. H. B. Sri and S. R. Dutta, "A survey on automatic text summarization techniques," J. Phys. Conf. Ser., vol. 2040, no. 1, pp. 121–135, 2021, doi: 10.1088/1742-6596/2040/1/012044.

[3] K. S. A. L. Harazin, "Multi-document Arabic Text Summarization," no. April, 2015, [Online]. Available: https://iugspace.iugaza.edu.ps/handle/20.500.12358/19174.

[4] S. Abdulateef, N. A. Khan, B. Chen, and X. Shang, "Multidocument Arabic text summarization based on clustering and word2vec to reduce redundancy," Inf., vol. 11, no. 2, 2020, doi: 10.3390/info11020059.

[5] Q. Al-Radaideh, "Applications of Mining Arabic Text: A Review," in Recent Trends in Computational Intelligence, IntechOpen, 2020.

[6] R. Elbarougy, G. Behery, and A. El Khatib, "Extractive Arabic Text Summarization Using Modified PageRank Algorithm," Egypt. Informatics J., vol. 21, no. 2, pp. 73–81, 2020, doi: 10.1016/j.eij.2019.11.001.

[7] M. Hussin, … M. F.-2008 I. I., and U. 2008, "Extending the growing hierarchal som for clustering

documents in graphs domain," ieeexplore.ieee.org.
[8] "EASC (Essex Arabic Summaries Corpus) - Browse /EASC at SourceForge.net." and T. W. L. Page, S. Brin, R. Motwani, "The PageRank Citation Ranking: Bringing Order to the Web," 1998.
[9] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts In Proceedings of EMNLP," 2004.
[10] A. Qaroush, I. Abu Farha, W. Ghanem, M. Washaha, and E. Maali, "An efficient single document Arabic text summarization using a combination of statistical and semantic features," J. King Saud Univ. - Comput. Inf. Sci., no. xxxx, 2019, doi: 10.1016/j.jksuci.2019.03.010.
[11] E. Elfarra and M. Mikki, "Automatic Arabic Text Summarization," 2015.
[12] M. Alfarra, A. M. Alfarra, and A. Salahedden, "Graph-based growing self-organizing map for single document summarization (GGSDS)," 2019, doi: 10.1109/PICECE.2019.8747236.
[13] M. T. B. Othman, M. A. Al-Hagery and Y. M. E. Hashemi, "Arabic Text Processing Model: Verbs Roots and Conjugation Automation," in IEEE Access, vol. 8, pp. 103913-103923, 2020, doi: 10.1109/ACCESS.2020.2999259
[14] B. Azman, "Root Identification Tool for Arabic Verbs," IEEE Access, vol. 7, pp. 45866–45871, 2019.
[15] M. O. Hegazi, "An Approach for Arabic Root Generating and Lexicon Development," IJCSNS Int. J. Comput. Sci. Netw. Secur., vol. 16, no. 1, pp. 9–15, 2016.
[16] S. Farwaneh, "Non-sound' verb Inflection in Arabic: Allomorphic variation and paradigmatic uniformity," Morphology, vol. 30, no. 1, pp. 61–89, 2020.