3rd International Conference on Arabic Computational Linguistics, ACLing 2017, 5-6 November 2017, Dubai, United Arab Emirates

Arabic Single-Document Text Summarization Using Particle Swarm Optimization Algorithm

Raed Z. Al-Abdallah and Ahmad T. Al-Taani[*]

Department of Computer Science, Yarmouk University ,Irbid,,Jordan
{raedauc@yu.edu.jo; ahmadta@yu.edu.jo}

Abstract

In this research, we propose the use of Particle Swarm Optimization (PSO) algorithm for the extraction of summaries for single Arabic documents. The PSO approach is compared with evolutionary approaches that use Genetic Algorithms (GA) and Harmony Search (HS). The Essex Arabic Summaries Corpus (EASC) and the Recall-Oriented Understanding for Gisting Evaluation (ROUGE) tool are used to evaluate the proposed approach. Experimental results showed that the proposed approach achieved competitive and even higher ROUGE scores in comparison to HS and GA approaches in the state-of-the-art.

Keywords: Automatic Text Summarization; Particle swarm optimization; Extractive summarization.

* Corresponding author. Tel.: +962-799-196-333; *E-mail:* ahmadta@yu.edu.jo

1. Introduction

A summary is a shorter version of the original text that preserves the overall meaning [1]. Automatic summarization is creating a summary automatically by a computer. In the last decade, there has been a great interest in the development of automated text summary solutions because of the time and cost usually consumed in understanding large volume of information manually. Automatic Text Summarization (ATS) has attracted the attention of both the research community and commercial community as a solution for reducing information overload and helping users to scan large number of documents to identify documents of interest [2]. ATS systems are important for most layers of people in the both communities whom have different interests and different

education levels. Google search engine is an example of ATS importance for all users in retrieving the list of more relevant query results. ATS approaches can be divided into abstractive and extractive approaches. An Abstractive summarization approach attempts to develop an understanding of the main concepts in a document using linguistic methods and then express those concepts in fewer words and in clear language [3][4]. An extractive summarization approach selects all important sentences from the original document and then concatenates them exactly as they ordered in the original text based on statistical and linguistic features of sentences [5]. A good summary should pick the most relevant information from the document while also maintaining low redundancy information in the summary [6]. The main problem with the extractive approach is that it produces an inconsistent summary since it does not consider the semantic relations between sentences; it is based only on statistical features [3]. The text summary can be classified into either single-document or multi-document summarization. In multi-document systems the documents must have a topic-relation with each other, and then the summarizer generates a summary based on these documents. In single-document systems the summarization is built using that single-document only [7]. Generally, there are four main different extractive summarization approaches applied to text [8]: statistical approaches [9], semantic approaches [10], machine learning approaches [11], and meta-heuristic approaches.

Two evolutionary techniques have been proposed for the Arabic ATS, the use of HS algorithm [14] and the use of GA algorithms [15]. In this research, we propose the use of PSO algorithm for Arabic single-document text summarization. PSO is a meta-heuristic search algorithm where it is not from evolutionary methods like GA and HS but it is a swarm intelligent method. The proposed approach is compared with two evolutionary key references [14] that use HS to produce a summary and [15] that use GA to produce a summary.

2. Literature Review

In this section we present a review of the most significant research used extractive approaches on the Arabic language. A hybrid Arabic ATS system proposed by Al-Omour and Al-Taani [12] is based on both graph and statistical approaches and is evaluated using EASC corpus. The results showed that using the n-gram in the summarization process achieved results that outperform those of stem and word. Arabic ATS presented by Elghazaly & Gheith [13] extracts the most significant paragraphs as text summaries. A cosine similarity measure is used in the vector representation for ranking and extracting significant paragraphs. The precision is used to evaluate both summaries. However, the limitation of their approach appears when it is applied to large-sized documents. Jaradat [14] incorporate HS with the summarization process to obtain the near to the optimal summary of a document using the EASC corpus and the ROUGE toolkit to evaluate the proposed approach. Obtained results showed that the proposed approach outperformed other state of arts approaches. Jaradat and Al-Taani [15] presented a hybrid-based single-document extractive Arabic text summarization approach using GA. The proposed approach is evaluated using EASC corpus and ROUGE evaluation method to determine the accuracy of the proposed approach. The results showed that the proposed method outperformed many state-of-the art approaches. Al-Zahrani *et al.* [27] used PSO to evaluate the effectiveness of different features used to summarize Arabic text. The PSO is trained on the Essex Arabic summaries corpus data to determine the best particle that represents the most appropriate simple combination of eight informative structure features used regularly by Arab summarizers. Based on the elected features and their relevant weights in each PSO iteration, the input text sentences are scored and ranked to extract the top ranking sentences in the form of an output summary. The output summary is then compared with a reference summary using the cosine similarity function as the fitness function. The experimental results illustrate that Arabs summarize texts simply, focusing on the first sentence of each paragraph. Mahjoub [28] proposed the use of PSO algorithm for automatic extraction of summaries for Arabic single documents. Experimental results showed the effectiveness of the proposed approach for extracting summaries for Arabic single documents.

After reviewing the related literature on ATS, we concluded that graph-based approaches [12] guarantee the semantic relationships between sentences, so the extracted summaries of these approaches have a better flow. On the other hand, these approaches suffer from local optimum problem because of using normal graph search techniques. We also found that PSO has not been used yet by for the extraction of summaries for the Arabic documents. Also, we concluded that Al-Zahrani et al. [27] and Mahjoub [28] approaches cannot capture relevant information which is spread across sentences because their method does not consider the semantic relations between sentences, so the