Received: 2 December 2018 | Revised: 22 July 2019 | Accepted: 4 October 2019
DOI: 10.1111/exsy.12476

WILEY Expert Systems

ORIGINAL ARTICLE

ArA*summarizer: An Arabic text summarization system based on subtopic segmentation and using an A* algorithm for reduction

Belahcene Bahloul[1,2] | **Hassina Aliane**[3] | **Mohamed Benmohammed**[2]

[1]Department of Mathematics and Computer Science, University Djilali Bounaama of Khemis Miliana, Ain Defla, Algeria
[2]Department of Software Technologies and Information Systems, University of Abdelhamid Mehri Constantine 2, Constantine, Algeria
[3]Information Sciences R&D Laboratory, Research Center on Scientific and Technical Information, Algiers, Algeria

Correspondence
Belahcene Bahloul, Department of Mathematics and Computer Science, University Djilali Bounaama of Khemis Miliana, Rue Thniet El Had Khemis Miliana Wilaya de Ain Defla, Algeria.
Email: d.bahloul@univ-dbkm.dz

Abstract

Automatic text summarization is a field situated at the intersection of natural language processing and information retrieval. Its main objective is to automatically produce a condensed representative form of documents. This paper presents ArA*summarizer, an automatic system for Arabic single-document summarization. The system is based on an unsupervised hybrid approach that combines statistical, cluster-based, and graph-based techniques. The main idea is to divide text into subtopics then select the most relevant sentences in the most relevant subtopics. The selection process is done by an A* algorithm executed on a graph representing the different lexical–semantic relationships between sentences. Experimentation is conducted on Essex Arabic Summaries Corpus and using recall-oriented understudy for gisting evaluation, automatic summarization engineering, merged model graphs, and n-gram graph powered evaluation via regression evaluation metrics. The evaluation results showed the good performance of our system compared with existing works.

KEYWORDS

data-driven, graph theory, information extraction, mathematics, method, natural language processing, text analysis, text mining, topic identification

1 | INTRODUCTION

With the development of electronic documents, massive amounts of information are generated. This increase in volume of texts requires the production of high performance software tools whose task is to find and retrieve relevant information in a condensed form, which is known under the name of automatic summary generation. The first automatic summarization works date from 1958 (Luhn, 1958). The basic idea was to make simple statistical calculations on the distribution of terms in a text. Works in this field have not stopped developing since that date. Summarization approaches are intended to be smarter by taking into consideration the meaning of words and sentences and studying the semantic relationships that may exist between them in order to produce coherent summaries and eliminate redundancy. Whereas extensive research has targeted automatic summarization for the English language, few works have been dedicated to Arabic language, and the efforts that have been made until date are still far from the community's expectations.

As a contribution to this research field, we propose an approach, which aims to automatically produce coherent and nonredundant summaries. The approach is mainly based on a topic segmentation technique using lexical cohesion detection. The objective is to identify and weight lexical–semantic relationships that may exist between different text segments (long sentences) in order to cluster them into subtopics. The lexical cohesion relationships are then illustrated by a weighted cyclic graph on which we execute an A* search algorithm in order to select the most important segments in the subtopics deemed relevant.

The rest of this paper is organized as follows. Section 2 presents an overview on some related works and states potential interesting contributions of topic segmentation to the text summarization problems. We describe the proposed approach in Section 3. The evaluation and experimental results are discussed in Section 4 and finally, we end up with a conclusion in Section 5.

