ABSTRACT

Nowadays, the amount of Arabic documents has increased significantly in different domains, such as news articles, emails, business summary, biomedical, and social media documents. Some databases have increased in its size to a terabyte. Multi-document text summarization is the method of creating a summary of a group of interrelated text documents. Therefore, the desire for Arabic multi-documents text summarization (at instant possible rates, coherent, grammatical and meaningful sentences) is increased. Consequently, this research proposes an Arabic Multi-Document Text Summarization model using two stages of Bisect K-means clustering and features extraction base method. The proposed Arabic Multi-Document Summarization System using Clustering and Feature Extraction (AMDSS-CF) model concentrates on summarizing Arabic multi-documents effectively through various aspects of linguistic text quality and relying on a linear complexity clustering algorithm to get a fast, fully coherent, chronological and accurate summary. A conceptual framework is proposed based on 27 published researches dealing with text summarization techniques of Arabic language. The dataset that is used in the investigation stage is derived from different domains, such as education, sports, and politics. This dataset contains texts of various sizes. The experiments are then designed to be in a specific domain (news domain). The proposed AMDSS-CF model covers the deficiency of Arabic Automatic Summarization Systems (ASS) by enhancing the final summary using a new component called summary builder to generate a summary quite similar to the gold summary. The results were in precision, recall, and F-measure equaled to 0,685, 0.751 and 0,708 respectively, and the AMDSS-CF achieved an accuracy using ROUGE-n of 69.4 %.