International Journal of Natural Computing Research
Volume 10 • Issue 2 • April-June 2021

A Quantum-Inspired Genetic Algorithm for Extractive Text Summarization

Khadidja Chettah, MISC Laboratory, NTIC Faculty, Constantine 2 University-Abdelhamid Mehri, Constantine, Algeria
Amer Draa, NTIC Faculty, Constantine 2 University-Abdelhamid Mehri, Constantine, Algeria

ABSTRACT

Automatic text summarization has recently become a key instrument for reducing the huge quantity of textual data. In this paper, the authors propose a quantum-inspired genetic algorithm (QGA) for extractive single-document summarization. The QGA is used inside a totally automated system as an optimizer to search for the best combination of sentences to be put in the final summary. The presented approach is compared with 11 reference methods including supervised and unsupervised summarization techniques. They have evaluated the performances of the proposed approach on the DUC 2001 and DUC 2002 datasets using the ROUGE-1 and ROUGE-2 evaluation metrics. The obtained results show that the proposal can compete with other state-of-the-art methods. It is ranked first out of 12, outperforming all other algorithms.

KEYWORDS

Extractive Text Summarisation, Metaheuristics, Natural Language Processing, Quantum-Inspired Genetic Algorithm, Unsupervised Learning

INTRODUCTION

A huge amount of textual data is being created every single day. This data often needs to be summarized to allow users a better access to the contained information. Automating such a task is a must once huge amounts of data need to be treated. Actually, text summarization could be defined as the process of extracting or collecting significant information from the original text, and presenting that information in the form of a summary. Automatic Text Summarization (ATS) systems could be classified into several kinds, based on their input types (single or multi- document), purpose (generic, domain-specific, or query-based) and output types (extractive or abstractive) (Yogan et al., 2016).

Extractive text summarization methods aim at selecting the most important sentences as they appear in the original text, and then combining them to form a summary. In contrast, abstractive text summarization tries to understand the overall meaning of the text and re-write it in a shorter and new way to produce the summary, while keeping its meaning and maintaining its semantic correctness.

Many research works have been published in the two types of summarization systems; with special attention given to extractive summarization, mainly for its simplicity and promising results. Works on extractive summarization are mainly classified into: unsupervised or supervised learning methods.

DOI: 10.4018/IJNCR.2021040103

Recent research mostly relies on unsupervised methods especially those based on evolutionary algorithms.

Quantum-inspired Genetic Algorithms (QGA) have widely proven their efficiency to solve optimization problems, especially those with binary representation of solutions. Yet, they have never been used for solving the problem of extractive single document summarization.

In the present work, we devise a quantum inspired genetic algorithm to solve the problem of extractive single document summarization. The key idea is to use the QGA metaheuristic to search for the best combination of sentences among a whole set of possible solutions presenting all possible combination of the original document sentences. Cohesion and readability are two features that are used as components of the objective function to guide the search process.

This paper is organised as follows. Sections 2 defines the problem of extractive text summarization and introduces the notions related to quantum-inspired evolutionary algorithms. In Sections 3, we review some works related to both text summarization and quantum-inspired evolutionary approaches. The problem formulation and the proposed QGA for text summarization are described in Section 4. Section 5 details the setup of our experiments, including: the used dataset, the pre-processing phase, parameter tuning, the evaluation metric and the analysis of our findings. Some conclusions and future perspectives of research are presented in Section 6.

BACKGROUND

Computational Linguistics (CL) is an interdisciplinary field of research covering knowledge, mainly, from three areas: linguistics, computer science and mathematics (i Prat, 1994). It can be regarded in two respects: on the one side, the application of different techniques and outcomes from computer science to linguistics in order to investigate issues as how human beings acquire and produce language or, for instance, how language changes over time. On the other side, it could be defined as the application of linguistics rules and methods to computer science to devise practical engineering systems that involve the automatic processing of natural languages (Tsujii, 2011). The latter field of study is usually referred to as Natural Language Processing (NLP).

NLP research has focused on tasks such as machine translation (Wu et al., 2016; Koehn et al., 2007), information retrieval (Gysel et al., 2018; Borkar and Patil, 2013; Radwan et al., 2006), question answering (Xiong et al., 2016; Dong et al., 2015), sentiment Analysis (Severyn and Moschitti, 2015), topic modelling and text summarization. In the rest this section, we first introduce the problem of automatic text summarization, especially its extractive variant. Then, some concepts related to quantum-inspired evolutionary computing are provided.

Text Summarization

Automatic text summarization systems, regardless their type, aim at producing a concise, fluent and accurate summary that conveys the key information in the input text. They are greatly needed to address the ever-growing amount of textual data to better help discover and access relevant information faster. There exist various types of text summarization. However, there is no single universal taxonomy of summary types, such categorisation changes depending on the angle of observation. On the basis of the number of input texts, summarization can be divided into single and multi-document summarization (Fattah and Ren, 2009). Single-document summarization techniques create a short outline of a single input document; while multi-document summarization algorithms extract a summary from multiple texts written about the same topic (Uçkan and Karcı, 2020; Mendoza et al., 2014). Summaries can also be seen as either generic or query-based.

In the former class, the summary presents the general meaning of the information contained in the document (Lee et al., 2009; Kruengkrai and Jaruskulchai, 2003; Gong and Liu, 2001). The latter category includes only a query-related content, it is also known as topic-based or user-based