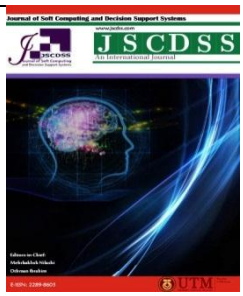
An International Journal
JSCDSS
http://www.jscdss.com
Vol.3 No.3 June 2016: 35-41
Article history:
Accepted 5 April 2016
Published online 5 April 2016

Journal of Soft Computing and Decision Support Systems



E-ISSN: 2289-8603

Sentence Similarity Techniques for Automatic Text Summarization

Yazan Alaya AL-Khassawneh [a,*], Naomie Salim [a], Adekunle Isiaka Obasae [a]
[a] Faculty of Computing, Universiti Technologi Malaysia, Johor, Malaysia

*** Corresponding author email address**: yakhassawneh@yahoo.com

Abstract

The technology of summarizing documents automatically is increasing rapidly and may give an answer for the information overload quandary. These days, document summarization is assumed an imperative part of information retrieval. With expansive amounts of documents, giving the user a short version of every document incredibly encourages the errand of discovering required documents. Text summarization is a procedure for making a packed form of a particular document that gives the users utilizable info, and summarization of multi document is engender summary distributing the meaning of the most info either explicitly or implicitly from a group of documents about main topic. In text summarization, resemblance among several sentences in a text has a major role. As such, development of methods of summarization has taken into consideration the aspect of similarities between several sentences in a text. This paper seeks to investigate different techniques of automatic summarization based on the element of sentence resemblance. Comparison is also developed for functionalities of various techniques with respect to recall, precision and F-measure values.

Keywords: Text summarization, Extractive summarization, Abstractive Summarization, Sentence similarity

1. Introduction

When an extracted or generated text carries information that is a vital segment of the primary document, it is deemed as summary for the main text. Moreover, when this occurs mechanically with involvement of a computerized program, it is known as an Automatic Text Summarization (ATS). In brief, a summary ought to sustain the mainstay of the document which paves the way for quick detection of pertinent information. Radev et al. (2002), opined that a summary could be defined as "a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that". Such definition suggests that summaries, which can be generated from one or more documents, should be reasonably brief and hold significant information derived from the primary text(s). Automatic text summarization is categorized under two procedures in accordance with the input. In a circumstance where the input involves a sole document, the procedure is termed a Single Document Summarization. An input that involves several documents of a similar nature, the procedure is known as a Multiple Document Summarization.

The aim of summarizing text is to display the ultimate essential information using smaller form from the first content while saving its principle substance and assists the user to rapidly see huge amounts of info. Text summarization studies the quandary of choosing the paramount parts from the text and the quandary of creating cohesive and reasonable summaries. The automatic procedure is fundamentally not quite the same as that of summaries generated by human, since human can catch and connect profound implications and subjects of text documents, while automatic programs' ability of such adeptness is very hard.

Programmed text summarization dated back to 1958- (Luhn, 1958). Since this time, efforts are being made by scholars to suggest systems for producing summaries. Several of scholars have suggested strategies for programmed text summarization, which could be classified into two: extractive and abstractive.

Determination of highest scored sentences or paragraphs from the pristine text and put them together to create shorter text while keeping the main meaning of the source text is known as Extractive summary. While in Abstractive summary scheme the linguistic means are used to inspect and explicate the text. Extraction systems are mostly used nowadays, to engender summary.

With organized documents, automatic text summarization works better, for example, scientific papers, reports, news and articles. The initial phase in extractive summarization is the determination of critical components, for example sentence location (Fattah and Fuji Ren, 2008), sentence length, number of numerical data (Lin, 1999), term frequency (Salton, 1989), number of opportune entities (Kupiec et al., 1995) and number of words occurring in denomination (Salton and Buckley, 1997).