Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 89 (2016) 333 – 340

Procedia Computer Science

Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016)

Knowledge Based Summarization and Document Generation using Bayesian Network

Shrikant Malviya* and Uma Shanker Tiwary

Indian Institute of Information Technology, Allahabad 211 012, India

Abstract

In this paper an approach of Semantic Knowledge Extraction (SKE), from a set of research papers, is proposed to develop a system Summarized Research Article Generator (SRAG) which would generate a summarized research article based on the query given by a user. The SRAG stores the semantic knowledge extracted from the query relevant papers in the form of a semantic tree. Semantic Tree stores all the textual units with their score in nodes organized at different levels depending on their type such as at the bottom leaf nodes keep the words with its probability, the upper level of it represent sentences with its score, next to it paragraphs, segments and so on. Scores of all the entities are calculated in bottom to up manner, first score of words are calculated, based on words sentences are ranked and similarly all the higher levels of the knowledge tree would be scored. A method of Bayesian network is used to generate a probabilistic model which would extract the relevant information from the knowledge tree to generate a summarized article. To maintain coherency, the summarized document is generated segment-wise by combining the most relevant paragraphs. Abstract of a generated summary is shown as a sample result. To show the effectiveness of the algorithm, an intrinsic evaluation strategy, degree of representativeness (DOG) is used. DOG gives on average 50% of relevance of the summary with the source. It's been observed that the proposed approach generates a comprehensive and precise papers.

Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016

Keywords: Bayesian Network; Extractive Summarization; Information Retrieval; Multi Document Summarization (MDS); Semantic Knowledge; Text Classification.

1. Introduction

The task of Multi-Document Summarization (MDS) is to produce concise and comprehensive summary to provide the major information for a set of document (e.g. news articles, research articles)[1]. Multi Document Summarization would be useful for the users to quickly understand the central idea of document collection, and it has been shown that multi document summarization could also be used to improve the performance of information retrieval systems[2]. MDSs can be categorized in two classes. First one is query-biased MDSs where the generated summary is biased according to the given query. On the other hand If summary is focused on a particular topic or concept, they come under the category of topic-focused MDS.

Based on the past implementations of MDS, there are two main fundamental approaches for the summarization: extractive approaches and abstractive approaches[3]. Extractive approaches are statistical in nature and most widely

*Corresponding author. Tel: +91-9451851905.
E-mail address: shrikant.iet6153@gmail.com

Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016
doi:10.1016/j.procs.2016.06.080

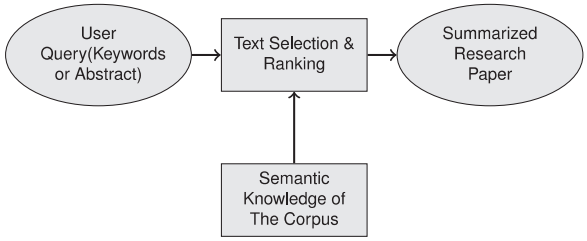

Fig. 1. Basic Architecture of Proposed Multi Document Summarizer.

used for the summarization. These approaches rank the sentences or paragraphs based on some importance measure to compose the summary. On other hand abstractive approaches promise to produce a summary which is more like human generated summaries. But the abstractive approaches are restricted by the development in the area of natural language understanding and generation. In this paper, we discuss the generic extractive approach to generate summaries from cluster of related documents.

A summary should be properly ordered to avoid the obscurity and to improve the quality/reliability of the content.[4] has shown empirically that proper order of sentences improve the readability significantly.[5] had proved experimentally that time consumption in reading is highly correlated with arrangement of paragraphs and sentences in the summary. Hence the content of the summary has to be coherent and consistent. It is possible only when constituent of summary might point towards a single concept.

In this paper, a graph based approach has been studied for constructing SRAG system. Indeed, lots of graph-based methods have been proposed for extractive summarization in the past.[6] introduced a novel stochastic graph-based, LexRank, which ranks the textual units relatively for multi-document summarization. LexRank ranks the sentences based on the concepts of eigenvector centrality obtained from the graph representation of sentences. Authors in[7] have also purported the the ability of graph based eigenvector centrality algorithm for multi-document summarization. A baysian network based summarization approach has been evaluated in this paper. Our method of MDS has divided the task into two separate subtask. First one is the extraction of semantic knowledge in tree structure from the corpus. In second part, this semantic knowledge is going to be used for the summarization through drawing a Bayesian network of terms and paragraphs. All components of the proposed MDS are described in detail in the next section of Proposed Approach. Section 3, described the used dataset and abstract of a sample paper generated by the implemented model. Finally, some conclusive remarks with future possibilities are given.

2. Proposed Approach

2.1 Overview

The architecture of the approach is shown in Fig. 1. Initially, a corpus has been formed by collecting various research articles, papers and journals related to different topics. A semantic knowledge is constructed consisting of weighted textual content in the form of a tree of all the research papers found relevant to the given user query (keywords/a short abstract). The semantic knowledge is described briefly in the next subsection *B. knowledge representation*. The relevant papers (cluster of similar documents) are selected on the basis of ranking their abstract with reference to the given user query using Naive Bayes classification. In short, Summary is going to be generated by applying a proposed extractive approach on the collected dataset basically based on the user's query which could be either a set of keywords or a small abstract.

Formally, SRAG's approach could be imitated with the help of provided pseudo code in the Fig. 2. SRAG takes the user query q and complete set of papers P as input. Using the Naive Bayes NB_{10} approach, all the papers are ranked and top 10 would be selected in the summarization. Paper's abstract is used in the process of ranking. Semantic Knowledge K, stored in the form of tree, is being generated from the selected relevant papers, is described in detail in the next subsection B. Score in the SK tree is measured by a Bayesian network approach described in C subsection. Next, $RankParSeg()$ will rank the paragraphs segment-wise and top few of them would be considered to construct