Chapter 2
Multi-Document Summarization by Extended Graph Text Representation and Importance Refinement

Uri Mirchev
Ben Gurion University of the Negev, Israel

Mark Last
Ben Gurion University of the Negev, Israel

ABSTRACT

Automatic multi-document summarization is aimed at recognizing important text content in a collection of topic-related documents and representing it in the form of a short abstract or extract. This chapter presents a novel approach to the multi-document summarization problem, focusing on the generic summarization task. The proposed SentRel (Sentence Relations) multi-document summarization algorithm assigns importance scores to documents and sentences in a collection based on two aspects: static and dynamic. In the static aspect, the significance score is recursively inferred from a novel, tripartite graph representation of the text corpus. In the dynamic aspect, the significance score is continuously refined with respect to the current summary content. The resulting summary is generated in the form of complete sentences exactly as they appear in the summarized documents, ensuring the summary’s grammatical correctness. The proposed algorithm is evaluated on the TAC 2011 dataset using DUC 2001 for training and DUC 2004 for parameter tuning. The SentRel ROUGE-1 and ROUGE-2 scores are comparable to state-of-the-art summarization systems, which require a different set of textual entities.

DOI: 10.4018/978-1-4666-5019-0.ch002
1. INTRODUCTION

The amount of information on the web is huge and it continues to increase dramatically, causing the effect of data overload. The purpose of **multi-document summarization** is extracting important information from an input collection of topic-related documents and representing it in a concise and usable form. Since one of the reasons for data overload is the fact that many documents share the same or similar topics, automatic multi-document summarization has drawn much attention in recent years. Text summarization is challenging because of its cognitive nature and interesting because of its practical applications. For example, every day many news websites publish articles discussing the same hot topic of the day. One can read all these articles to achieve the complete understanding of the news topic. Alternatively, multi-document summarization can be used, giving the reader one exhaustive story covering the topic. Summarization can also be applied to information retrieval. We can run a summarizer on a search engine output, generating a unified summary of the information contained in result pages, hence letting the user save the time spent on viewing these pages.

Manual summarization of large document collections is a time-consuming and difficult task, which requires a significant intellectual effort. Therefore, automation of the summarization process is required. McKeown, et al. (2005) conducted experiments to determine whether multi-document summaries measurably improve the user performance and experience. Four groups of users were asked to perform the same fact-gathering tasks by reading online news under different conditions: no summaries at all, single-sentence summaries drawn from one of the articles, automated summaries, and human summaries. The results showed that the quality of submitted reports was significantly better and the user satisfaction was higher using both automated and human multi-document summaries rather than relying on the source documents only.

The automated text summarization area has been extensively explored during the last decade, mostly due to DUC and TAC annual competitions. Thousands of research works have been conducted and published on the subject of multi-document generic summarization. However, despite the significant efforts dedicated to design of novel summarization approaches, the automated summary quality is still far from being perfect. Thus, in the TAC 2011 competition (Text Analysis Conference www.nist.gov/tac) on English dataset, the best summarization system (ID2) achieved performance of 0.46 in terms of ROUGE-1 recall score vs. the upper bound of 0.52 obtained by the topline system based on human summaries.

In this chapter, we offer a fresh look at the summarization process by enhancing the graph representation of a document collection. We also propose that decision about including a sentence in a summary should be influenced by the previously selected sentences. This feature is expressed in the continuous refinement of the sentence importance score. We introduce an algorithm called SentRel (Sentence Relations) for automated summarization of a topic-related document collection. The algorithm copes with the generic summarization task, where the goal is to reflect the most important information described by the input collection. To achieve this goal, the proposed extractive summarization algorithm distills the most relevant sentences from the collection into a short extract, which can be quickly digested by the end-user.

Our summarization approach is based on the mutual reinforcement principle used to compute global importance of the sentences, representing the text corpus as a tripartite graph. In addition, the importance scores of textual entities (i.e., documents and sentences) are iteratively updated by the current summary content. The ordinal and chronological dependencies of sentences in a multi-document summary are calculated beforehand from a training dataset(s) accompanied by gold standard summaries. The SentRel algorithm is greedy, since it iteratively chooses the most