Chapter XI
Extracting the Essence:
Automatic Text Summarization

Fu Lee Wang
City University of Hong Kong, Hong Kong

Christopher C. Yang
Drexel University, USA

ABSTRACT

As more information becomes available online, information-overloading results. This problem can be resolved through the application of automatic summarization. Traditional summarization models consider a document as a sequence of sentences. Actually, a large document has a well-defined hierarchical structure. Human abstractors use the hierarchical structure of the document to extract topic sentences. They start searching for topic sentences from the top level of the document structure downwards. Similarly, hierarchical summarization generates a summary for a document based on the hierarchical structure and salient features of the document. User evaluations that have been conducted indicate that hierarchical summarization outperforms traditional summarization.

INTRODUCTION

The explosion in the amount of information available online has resulted in a well-recognized problem of information overloading. This problem can be eased through the application of automatic summarization. Automatic summarization extracts the most important information from the source document and presents the information to the users in a condensed form. By reading these summaries, users can understand the information that is contained in the source documents in a short time, and make decisions quickly.

Human professionals produce high quality summaries; however, it is too time-consuming and labor-intensive. Automatic summarization is capable of generating summaries for a large volume of information efficiently. In recent years, there has been an increasing need for automatic summarization due to the information explosion.

Moreover, automatic summarization is indispensable in digital libraries. This chapter will review techniques in automatic summarization. We will also introduce hierarchical summarization, which is a new summarization technique (Yang & Wang, 2003a, 2003b).

The traditional summarization models consider the source document as a sequence of sentences and ignore the hierarchical structure of the document. Similar to the abstracting process of human professionals, hierarchical summarization generates a summary by exploring the hierarchical structure and salient features of the document (Yang & Wang, 2003a). Experimental results have indicated that hierarchical summarization is promising and outperforms traditional summarization techniques that do not consider the hierarchical structure of documents.

BACKGROUND

In general, automatic summarization is represented by a three-stage framework, that is, representation of source document, extraction of information, and generation of summary (Sparck-Jones, 1999). Most of the current research work focuses on the second stage. Traditionally, the summarization system calculates the significance of sentences to the document based on the salient features of the document (Edmundson, 1969; Luhn, 1958). The most significant sentences are then extracted and concatenated as a summary. The compression ratio of the summary can be adjusted to specify the amount of information to be extracted. A lot of extraction features have been proposed.

The extraction approaches are usually classified into three major groups according to the level of processing in the linguistic space (Mani & Maybury, 1999). The surface-level approaches use salient features of a document to extract the important information. The entity-level approaches build an internal representation for text units and their relationships, and use graph theories to determine the significance of units. The discourse-level approaches model the global structure of the text, and the text units are extracted based on the structure. Generally, the deeper approaches are more promising to give more informative summaries. However, the surface-level approaches are proved to be robust and reliable (Goldstein, Kantrowitz, Mittal, & Carbonell 1999). They are still widely adopted at present.

The summarization systems can be evaluated either by intrinsic or extrinsic evaluation (Sparck-Jones & Galliers, 1996). The intrinsic evaluation judges the quality of the summarization by direct analysis of the summary (Kupiec, Pedersen, & Chen, 1995). The extrinsic evaluation judges the quality of the summarization based on how it affects the completion of some other tasks (Morris, Kasper, & Adams, 1992). A number of general-purpose summarization systems have been developed. Experiments have been conducted on these systems. All the systems identify an upper bound for the precision of the summarization system, the performance of the system grows fast with addition of extraction features, and they reach their upper bound after three or four extraction features (Kupiec et al., 1995).

AUTOMATIC SUMMARIZATION

Related research has shown that human abstractors use readymade text passages from a source document for summarization (Endres-Niggemeyer, 2002). Eighty percent of the sentences in the man-made abstracts are closely matched with sentences in the source documents (Kupiec et al., 1995). As a result, selection of representative sentences is considered as a good approximation of summarization (Aminin & Gallinari, 2002). The existing automatic text summarization is mainly the selection of sentences from the source document based on their significances in the document using statistical techniques and linguistic