Automated Text Summarization

Chin-Yew LIN
Information Sciences Institute
University of Southern California
4676 Admiralty Way, Marina del Rey, CA 90292-6695
cyl@isi.edu

Abstract

After lying dormant for over two decades, automated text summarization has experienced a tremendous resurgence of interest in the past few years. Research is being conducted in China, Europe, Japan, and North America, and industry has brought to market more than 30 summarization systems; most recently, a series of large-scale text summarization evaluations, Document Understanding Conference (DUC) and Text Summarization Challenge (TSC) have been held yearly in the United States and Japan.

In this tutorial, we will review the state of the art in automatic summarization, and will discuss and critically evaluate current approaches to the problem. We will first outline the major types of summary: indicative vs. informative; abstract vs. extract; generic vs. query-oriented; background vs. just-the-news; single-document vs. multi-document; and so on. We will describe the typical decomposition of summarization into three stages, and explain in detail the major approaches to each stage. For topic identification, we will outline techniques based on stereotypical text structure, cue words, high-frequency indicator phrases, intratext connectivity, and discourse structure centrality. For topic fusion, we will outline some ideas that have been proposed, including concept generalization and semantic association. For summary generation, we will describe the problems of sentence planning to achieve information compaction.

How good is a summary? Evaluation is a difficult issue. We will describe various suggested measures and discuss the adequacy of current evaluation methods including manual evaluation procedures used in DUC, the factoid and pyramid method reference summary creation procedures and fully automatic evaluation method such as ROUGE. The recently developed automatic evaluation method based on basic element (BE) will also be covered.

Throughout, we will highlight the strengths and weaknesses of statistical and symbolic/linguistic techniques in implementing efficient summarization systems. We will discuss ways in which summarization systems can interact with and/or complement natural language generation, discourse parsing, information extraction, and information retrieval systems.

Finally, we will present a set of open problems that we perceive as being crucial for immediate progress in automatic summarization.

Biography

Chin-Yew Lin is a senior research scientist at the Information Sciences Institute of the University of Southern California. He was the chief architect of SUMMARIST and NeATS. He also developed the automatic summarization evaluation package ROUGE that have been used in the DUC evaluations. He has co-chaired several text summarization and question answering workshops in ACL, NAACL, COLING.