Text Summarization: News and Beyond

Kathleen McKeown
Department of Computer Science
Columbia University

Redundancy in large text collections, such as the web, creates both problems and opportunities for natural language systems. On the one hand, the presence of numerous sources conveying the same information causes difficulties for end users of search engines and news providers; they must read the same information over and over again. On the other hand, redundancy can be exploited to identify important and accurate information for applications such as summarization and question answering.

Columbia's Newsblaster system for online news summarization exploits online redundancy to generate a summary, at the same time creating a concise synopsis of recent events for end users. Newsblaster crawls the web nightly for news articles, clusters news on the same event and generates a summary of each event. In this talk, I will present the current capabilities of Newsblaster, with some focus on its ability to generate and edit text. I will then turn to our ongoing work which goes beyond summarization of English news. Our research on summarization of multilinguual news requires us to deal with noisy input; we rely on state of the art machine translation systems and use information that is available at the time of summarization to improve the fluency of the summary. We are also moving to summarization of other media, including email and meetings. Both of these media also require the ability to handle noisy input, but add an additional challenge to handle features of dialog.

Proceedings of the Australasian Language Technology Workshop 2005, pages 4–4,
Sydney, Australia, December 2005.