Summarizing the News (Automatically)

KATHLEEN MCKEOWN

Henry and Gertrude Rothschild Professor of Computer Science

At first glance, the nine-year-old Columbia Newsblaster Web site (newsblaster.cs.columbia.edu) looks like Google News. Both feature the day's top stories plus sections on national, world, financial, and science/technology news.

The difference is their technologies. Google lists the first sentences of one news article and links to similar stories. Newsblaster publishes summaries of a dozen or more articles – all written and edited by software developed by Professor Kathleen McKeown.

"Newsblaster summarizes multiple news articles. We're using similar technology to answer questions from information on the Web. Today, users read the documents their search returns to see if they are relevant. Our software takes the next step. It looks into the documents, pulls out the relevant information, and summarizes it in a paragraph."

McKeown's software starts by scraping 25 different Web sites for news every night. It uses key words to cluster articles and categorize topics, counting the number of articles in each cluster to determine its importance.

Once classified, the software uses several approaches to generate summaries. First, it extracts sentences from important sources, such as stories from prominent newspapers and wire services.

It also pairs each sentence with every other sentence in the cluster. It analyzes their similarity and groups related themes together. "The software lines up the sentences in each group side by side and looks at where they overlap or intersect," McKeown explained. "It is looking for phrases that say the same thing, where words overlap or there is paraphrasing.

"The software parses the sentences for grammatical structure, so it knows that this phrase functioned as a noun and that phrase acts as an adjective. This helps it align similar sentences and fuse phrases to create summary sentences. It then generates the summary by ordering the sentences, using information about chronological order of the events. It also edits for coherence, substitutes proper nouns for pronouns, and adds or removes references, depending on whether a person or place is well known or not," she said.

The core technology has found other uses. A small company is using it to power smart phone applications that track and create timelines for breaking news on specific topics. Another application responds to open-ended questions, generating summaries of information about, for example, a particular event or a particular person. A third creates English summaries from news sources in other languages.

While some Newsblaster stories read like newspaper articles, others are choppier. Still, the technology could become an important tool for making sense of all the information on the Web.

B.A., Brown, 1976; M.S., Pennsylvania, 1979; Ph.D., 1982