Advances in Automatic Text Summarization

Inderjeet Mani and Mark T. Maybury (editors)
(MITRE Corporation)

Cambridge, MA: The MIT Press, 1999,
xv+434 pp; hardbound, ISBN
0-262-13359-8, $45.00

Reviewed by
Mark Sanderson
University of Sheffield

It has been said for decades (if not centuries) that more and more information is becoming available and that tools are needed to handle it. Only recently, however, does it seem that a sufficient quantity of this information is electronically available to produce a widespread need for automatic summarization. Consequently, this research area has enjoyed a resurgence of interest in the past few years, as illustrated by a 1997 ACL Workshop, a 1998 AAAI Spring Symposium and in the same year SUMMAC: a TREC-like TIPSTER-funded summarization evaluation conference. Not unexpectedly, there is now a book to add to this list: *Advances in Automatic Summarization*, a collection of papers edited by Inderjeet Mani and Mark T. Maybury and published by The MIT Press. Half of it is a historical record: thirteen previously published papers, including classics such as Luhn's 1958 word-counting sentence-extraction paper, Edmundson's 1969 use of cue words and phrases, and Kupiec, Pedersen, and Chen's 1995 trained summarizer. The other half of the book holds new papers, which attempt to cover current issues and point to future trends. It starts with a paper by Karen Spärck Jones, which acts as an overall introduction. In it, the summarization process and the uses of summaries are broken down into their constituent parts and each of these is discussed (it reminded me of a much earlier Spärck Jones paper on categorization [1970]). Despite its comprehensiveness and authority, I must confess to finding this opener heavy going at times.

The rest of the papers are grouped into six sections, each of which is prefaced with two or three well-written pages from the editors. These introductions contain valuable commentary on the coming papers—even pointing out a possible flaw in the evaluation part of one. The opening section holds three papers on so-called classical approaches. Here one finds the oft-cited papers of Luhn, Edmundson, and Pollock and Zamora. As a package, these papers provide a novice with a good idea of how basic summarization works. My only quibble was in their reproduction. In Luhn's paper, an article from *Scientific American* is summarized and it would have been beneficial to have this included in the book as well. Some of the figures in another paper contained very small fonts and were hard to read; fixing this for a future print run is probably worth thinking about.

The next section holds papers on corpus-based approaches to summarization, starting with Kupiec et al.'s paper about a summarizer trained on an existing corpus of manually abstracted documents. Two new papers building upon the Kupiec et al. work follow this. Exploiting the discourse structure of a document is the topic of the next section. Of the five papers here, I thought Daniel Marcu's was the best, nicely describing summarization work so far and then clearly explaining his system, which is based on Rhetorical Structure Theory. The following section on knowledge-rich approaches to summarization covers such things as Wendy Lehnert's work on breaking

down narratives into plot units and two papers on creating summaries from numerical data such as sports results and military testing logs. This is an area I know little about, but I found all papers in this section to be easily understood and clear.

After this is a section on evaluation, which contains papers that cover well the issues of this important area, with perhaps a slight let-down at the end. The first paper describes human inconsistency in summary generation. The second shows the leading paragraphs of newspaper articles being better than automatic summaries. In contrast, the third illustrates genuine utility from summarization. The fourth and final paper describes the creation and results of the TIPSTER-backed SUMMAC exercise. What is disappointing, however, is that the smaller dry run of the exercise held in 1997 is described rather than the much larger-scale version held a few months later. Unless there were some pressing editorial deadlines for this book that prevented the larger SUMMAC being written up in time, it really is a shame that this earlier paper was allowed in.

The final section covers new problem areas of summarization, for example, processing multiple documents and summarizing multimedia data, such as images or combined video and text. The multimedia papers stood out as being the most interesting, with the video and text paper of Merlino and Maybury presenting a number of interesting user evaluation studies; the image-summarization paper, while being new work with little concrete to present, was an intriguing introduction to a potentially new field. I thought there were two areas missing from this section, however: summarization of translated documents is very active these days with work from the CRL group at New Mexico State University and Philip Resnik's gisting paper (1997). The other aspects that seem little investigated are the human-factors issues of summaries: how they are presented and how users might interact with them. It would have benefited the book to have these areas covered as well.

I don't have a wide knowledge of past summary papers, so it is hard for me to think of historical ones that might be missing. After reading through the references in the book, however, it was clear that papers by Chris Paice are well thought of and perhaps one of his should have been included.

Despite these complaints, I found this book to be most informative. Reviewing it led me to conclude that it isn't intended to be read from cover to cover; instead it's something to be dipped into on occasion. It is accessible and I imagine that anyone with a basic understanding of text processing will be able to follow most of the papers in this book. As a rather nonacademic footnote, I must mention the physical feel of the book itself. The design and finish of the cover are high quality. More importantly, the type of binding used allows the book to stay at whatever page it's opened. These may seem trivial points, but they very much added to the pleasure of its reading.

References

Resnik, Philip. 1997. Evaluating multilingual gisting of Web pages. *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval—Electronic Working Notes*.

Spärck Jones, Karen. 1970. Some thoughts on classification for retrieval. *Journal of Documentation*, 26(2): 89–101.

Mark Sanderson is a lecturer in the Information Studies department at the University of Sheffield. His research interests are in information retrieval. He has recently been appointed to the Editorial Board of the *Journal of the American Society of Information Science*. Sanderson's address is: Department of Information Studies, University of Sheffield, Western Bank, Sheffield, S10 2TN, UK; e-mail: m.sanderson@shef.ac.uk