Automatic Summarization

Inderjeet Mani
(MITRE Corporation and Georgetown University)

Amsterdam: John Benjamins (Natural language processing series, edited by Ruslan Mitkov, volume 3), 2001, xi+285 pp; hardbound, ISBN 1-58811-059-1, $82.00; paperbound, ISBN 1-58811-060-5, $29.95

Reviewed by
Chris D. Paice
Lancaster University

Though the first attempts at automatic abstracting were made about 45 years ago, this area was until quite recently a rather obscure specialism. No doubt, this was due to the nonavailability of machine-readable texts: It seemed absurd to go to the trouble and expense of keypunching complete texts and then use a program to throw away 90% of them! Since about 1990, with the urgent need to manage the mass of information available on the World Wide Web, the field of automatic summarization has exploded into prominence.

Automatic Summarization is the first monograph on the subject, and as such it is sure to be widely read and cited. The author, Inderjeet Mani, is currently one of the field's most active researchers, and it is therefore no surprise that the book has an authoritative feel. In general, it provides good, balanced coverage of the field (except that cross-lingual summarization is not covered) and describes a great deal of the published research in the area, with a bibliography of well over 200 items. Mani succeeds in conveying a sense that this is a most challenging field for system designers. With a grim topicality, many of the illustrations used refer to the summarization of reports on terrorist incidents.

The book is structured sensibly into nine main chapters. Chapter 1, "Preliminaries," introduces a range of fundamental terms and ideas. A distinction is made between *extracts*, consisting of material (typically sentences) taken straight from the source, and *abstracts*, "at least some of whose material is not present in the input." Chapter 2 looks at professional (i.e., nonautomatic) summarizing.

Chapter 3 is about extraction, which involves assigning an importance score to each sentence in a text and outputting those with the highest score. Various importance clues may be used, including the presence of concentrations of content words, use of cue expressions such as *we see that* and *certainly*, and the location of the sentence in the text. A long section describes and discusses the work by Edmundson and by later workers pursuing similar approaches. Coming up to date, the chapter continues with a discussion of corpus-based approaches to sentence extraction. Chapter 4 is entitled "Revision" and discusses how to deal with problems of cohesion in extracted sentences that arise from the occurrence of anaphors and other awkward features. Chapter 5 discusses the potential use of discourse-level information in approaches such as cohesion graphs, text tiling, lexical chains, and rhetorical structure theory.

In Chapter 6, the author turns to abstraction, as opposed to sentence extraction. Typical approaches here have involved the use of sketchy scripts or frames tailored to a particular domain. Analysis of a text results in instantiation of the appropriate

frame, and when this process is complete, output can be generated, either as stylized "canned text" or by the use of proper text generation techniques. A major problem with this approach is its inflexibility: A text can be summarized only if the correct frame is available and can be identified.

Chapter 7 deals with multidocument summarization, which is now seen as a most important area (terrorism news reports, again). This is not simply a matter of merging a number of separate descriptions of the same issue or event: There is a need to highlight discrepancies between different reports and to deal with time-related issues (e.g., changing estimates of numbers of casualties). Another nontrivial problem is handling cross-document coreferences—for example, ensuring that mentions of *Johnson* in two different documents actually refer to the same person.

Chapter 8, "Multimedia Summarization," is in my view the weakest in the book. Mani starts by acknowledging that this area is evolving rapidly, so that "no clear principles have emerged" (p. 209). But he then takes this as an excuse for a rather haphazard collection of topics (summarization of dialog, of video, and of diagrams, and multimedia briefing generation) dealt with in varying degrees of detail. The chapter is only 12 pages long (nine pages of actual text), whereas to begin to tackle this area, at least twice the space is surely required. To give a proper structure, the chapter should begin with a discussion of possibilities for the direct analysis of various nontext media, followed by an examination, with examples, of how different media (especially text) can support one another in building useful systems.

Chapter 9, on the evaluation of summarization systems, is fully three times as long as the previous chapter, and there is no space here to go into details. Though the exposition gets a bit tangled in places, the account is generally authoritative and thorough and provides a valuable overview of this problematic area.

All the main chapters from Chapter 3 onward finish with *(1)* a table summarizing the various approaches discussed, with their strengths and weaknesses, and *(2)* a glossary giving a brief definition of the various terms introduced in the chapter. These are both valuable features, though it seems structurally inconsistent that the former are included as ordinary numbered tables and the latter as separate "Review" sections.

Since this is the only textbook in its field, it will probably become required reading for new researchers, who will assume that it provides a full and balanced view. Its very patchy presentation of historical research therefore seems regrettable. Surely, the seminal 1958 paper by H. P. Luhn deserves a section of its own? Apart from its historical importance, this work provides a clear and clean introduction to the sentence extraction paradigm. Luhn's paper is in fact mentioned twice in the book, but in neither case is it very prominent. Neither is there any mention of James E. Rush and his work on the ADAM system (Rush, Salvador, and Zamora 1971) (though the successor system for the Chemical Abstracts Service is discussed). Chapter 4, on sentence revision, is short and should certainly have included a summary of the pioneering efforts by Mathis, Rush, and Young (1973). So little research was carried out prior to the 1980s that it seems a pity to ignore so much of it! (By contrast, the ideas of Skorochod'ko [1972] on cohesion graphs are—quite properly—discussed at some length in Chapter 5.)

Although (Chapter 8 apart) I have no problems with the general structure and content of the book, there were various minor defects that for me rather spoiled the overall impression. First, although on the whole the book is readable enough, there are a few places where the description of processes or ideas becomes so convoluted or condensed that even repeated rereading failed to reveal the meaning. In some places, intelligibility would have been much improved by proper use of examples (p. 82 is a glaring example). I also felt that some of the formulae could have been justified and

explained a bit more fully. The diagrams vary between clear and pertinent to obscure or curious (two figures in Chapter 7 seem to show a succession of aquaria with various documents suspended inside them).

A book like this is certainly going to be used very much for reference purposes and thus the reader should be able to find her way around it. Here, in fact, is my biggest gripe. In a short space of time, I was able to compile a list of 10 phrasal terms that were discussed in the text but did not appear either in the index or in the relevant Review section. Moreover, cross-referencing within the text is minimal. In some cases, there are references to equations or diagrams that are many pages away, necessitating a tedious search. Finally, neither the section numbers nor the page headers include the chapter number, making it hard to find one's way about (though this was probably not the author's decision).

I have actually found this quite a hard book to review. As a textbook, it provides a valuable map of the field, and yet, as far as I can detect, there is nothing here that is notably innovative or controversial; there is very little to focus on except the minor flaws. But taken as a whole, the book certainly provides a timely and informative overview of automatic summarization, and all researchers with an interest in this important field will wish to have it on their bookshelves.

References

Luhn, H. P. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165. Reprinted in Inderjeet Mani and Mark T. Maybury, editors, *Advances in Text Summarization*, MIT Press, 1999, pages 15–21.

Mathis, B. A., J. E. Rush, and C. E. Young. 1973. Improvement of automatic abstracts by the use of structural analysis. *Journal of the American Society for Information Science*, 24(2):101–109.

Rush, J. E., J. R. Salvador, and A. Zamora. 1971. Automatic abstracting and indexing. II. Production of indicative abstracts by application of contextual inference and syntactic coherence criteria. *Journal of the American Society for Information Science*, 22(4):260–274.

Skorochod'ko, E. F. 1972. Adaptive method of automatic abstracting and indexing. In C. V. Freiman, editor, *Information Processing 71: Proceedings of the IFIP Congress 71, Ljubljana, Yugoslavia*, North Holland, 1179–1182.

Chris Paice teaches in the Computing Department at Lancaster University, U.K. He has research interests in various aspects of information retrieval and text processing and has published several papers on automatic abstracting, the first appearing in 1981. Paice's address is: Computing Department, Lancaster University, Bailrigg, Lancaster LA1 4YR, U.K.; e-mail: cdp@comp.lancs.ac.uk.