Summarising: Where are we now? Where should we go?

Karen Sparck Jones

Computer Laboratory, University of Cambridge
New Museums Site, Pembroke Street
Cambridge CB2 3QG, England
Karen.Sparck-Jones@cl.cam.ac.uk

Abstract

Summarising covers a range from text extraction to content condensation. Its essential features are picking important concepts from, and reducing, source text or information, to deliver summary information or text. General strategies for doing this are clearly preferable to application-specific ones. So far, we have found that statistically-based sentence extraction and concatenation does not produce effective summaries. But we have not yet found general methods of content analysis and condensation. We can only identify key source content and present it in summary with heavy domain and goal guidance. The most pressing need is to develop 'sufficient to the day' techniques that do more than surface sentence extraction without depending, MUC-like, on prior specifications for sought content. These needed intermediate techniques include passage extraction and linking, deep phrase selection and ordering, entity identification and relating. Such strategies benefit from, or require, shallow text analysis and do or can exploit statistical data. They may be enhanced by modern display resources. They are applicable to individual source texts or to data sets as wholes. Most importantly, we can tackle this level of summarising because current robust parsing technology may succeed, given source redundancy, in getting enough of value from sources to help users, and because current text production methods can deliver usable summary texts. We should push this line hard, seeking to minimise application-specific domain knowledge, to take advantage of discourse structure, and to address summary function for the user.