

The effects of analysing cohesion on document summarization

Branimir K. Boguraev and Mary S. Neff

IBM T.J. Watson Research Centre, P.O. Box 704, Yorktown Heights, NY 10598

bkb@watson.ibm.com, neff@watson.ibm.com

Abstract

We argue that in general, the analysis of lexical cohesion factors in a document can be used to drive a summarizer, as well as enable other content characterization tasks, such as topical highlights of documents, multi-threaded summaries for personal information delivery, or navigating through multiple-document spaces. More narrowly, this paper focuses on how few cohesion factors—simple lexical repetition and coreference—can enhance an existing sentence extraction-based summarizer, by enabling strategies alleviating some of the particularly jarring end-user effects in the summaries, typically due to coherence degradation, readability deterioration, and topical under-representation. Repetition, in its various manifestations, is instrumental to, among other things, the topical make-up of a text. Being sensitive to lexical chains in a document, in an environment where fine-grained distinctions can be made among different sets of inter-related topical highlights, not only offers improvements to an existing summarizer, but also makes it possible to develop other kinds of content characterization and management functionalities. The fact that summaries derived by such an enriched environment under certain conditions compare favourably with our baseline, suggests that lexical cohesion analysis would be a significant enabling factor in content characterization applications.