

# Lexical cohesion, discourse segmentation, and 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** Summaries automatically derived by sentence extraction are known to exhibit some coherence degradation, readability deterioration, and topical under-representation. We propose a strategy for improving upon these problems, aiming to generate more cohesive summaries by analyzing the lexical cohesion factors in the source document texts. As an initial experiment, we have looked at one particular factor, lexical repetition, which is instrumental to the topical make-up of a text. We have developed a framework for integrating a lexical repetition-based model of discourse segmentation capable of detecting shifts in topic, with a linguistically-aware summarizer which utilizes notions of salience and dynamically-adjustable size of the resulting summaries. We show that even by utilizing lexical repetition alone, summaries are of comparable, and under certain conditions better, quality than those delivered by a state-of-the-art sentence-based summarizer. This is encouraging for a broad platform of research which seeks to position a framework for the recognition and use of a number of cohesive devices in text as instrumental in the development of a wide range of content characterisation and document management tasks.