Team develops software for automatic summarization of long texts

August 7 2014

While long-form writing, epic cinematic tales and hefty tomes have had something of a renaissance recently, the continued popularity of the so-called microblogging platform Twitter and other such tools highlights the fact that many people still like to be very succinct. The terse commentary, the abstract, the executive summary: all still favored by many of us at some time or in some context.

Moreover, who has the time to read long texts when a chunk of pithy sound bites is needed. Thankfully, researchers in India are developing new software that can make longwinded prose short and sweet.

Esther Hannah of St. Joseph's College of Engineering, in Sholinganallur and Saswati Mukherjee of Anna University, Guindy Campus, both in Chennai, have developed a classification-based summarization model that performs automatic summarization of text. The direct application of the software will be to remove extraneous noise sentences from bulk text allowing much more efficient and faster text mining to be carried out. Of course, the same summarization would allow a reader to extract the salient points from any given text too. The team suggests that the automatic summarization is comparable to that which might be carried out by an expert editor in terms of removing the redundancy and irrelevance.

The team "trained" their software with 60% of the 105 English-language documents from the Document Understanding Conference (DUC-2002), checking and correcting errors the algorithm makes along the way and

thus teaching the software what would be an appropriate summarization and what would not. They then tested the remainder at various levels of summarization – 10%, 20% and 30%. The system works well with grammatically well-constructed, even very long, documents, has problems if there are extensive mathematical and scientific symbols in the text.

Precision was optimal between 20 and 30% percent, which is a significant reduction in text length for parsing by text-mining software. They obtained a precision value of about 0.65, which is significantly better than fuzzy logic summarization software, which scores around 0.47 and far better than Microsoft Word 2007 inbuilt summarization tool, which is a little over 0.46.

More information: Esther Hannah, M. and Mukherjee, S. (2014) 'A classification-based summarization model for summarising text documents', *Int. J. Information and Communication Technology*, Vol. 6, Nos. 3/4, pp.292–308.

Provided by Inderscience