Abstract

Text summarization is the process of distilling the most important information from a source to produce an abridged version for a particular user and task. When this is done by means of a computer, i.e. automatically, it calls as Automatic Text Summarization. Summarization can be classified into two approaches: extraction and abstraction. Extraction based summaries are produced by concatenating several sentences taken exactly as they appear in the texts being summarized. Abstraction based summaries are written to convey the main information in the input and may reuse phrases or clauses from it. This paper focuses on extraction approach. The goal of text summarization based on extraction approach is sentences selection. One of the methods to obtain the sentences is to assign some feature terms of sentences for the summary called ranking sentences and then select the best ones. The first step in summarization by extraction is the identification of important features. In our approach 1000 computer science related research papers are used as test documents. Each document is prepared by preprocessing process: sentence segmentation, tokenization, stop word removal, case folding, lemmatization, and stemming. Then, using important features, sentence filtering features, data compression features and finally calculating score for each sentence. The proposed text summarization is based on HMM tagger to improve the quality of the summary. Here, comparing our results with the existing summarizers which are Copernicus summarizer,
Great summarizer and Microsoft Word 2007 summarizers etc. The proposed system is also tested with four types’ similarities: Cosine, Jaccard, Jaro-winkler and Sorenson similarities. The results show that the best quality for the summaries was obtained by feature terms method.

References

- ISC &quot;ISC Internet Domain Survey&quot;, Available at: http://ftp.isc.org/www/survey/reports/current/
- C. H. Chang, M. Kayed, M. R. Girgis, and K. Shaalan, &quot;A survey of web information extraction systems.&quot; IEEE Transactions on Knowledge and Data Engineering, Vol. 18, 2006, pp. 1411-1428
- Kathleen R. McKeown, &quot;Discourse Strategies for Generating Natural Language Text&quot;, Department of Computer Science, Columbia University, New York, 1982
A Feature Terms based Method for Improving Text Summarization with Supervised POS Tagging

in Information Retrieval, University of California, Berkeley, August.

Index Terms
Computer Science Information Sciences

Keywords
Term Frequency Term Weight Inverse Sentence Frequency Noun And Verb
Chunking Position Sentence
Sentence
Verb Featured Sentences
Compression Ratio
Retention Ratio
Data Compression Features
Hmm Tagger