Multi-Document Summarization Based on BE-Vector Clustering

Dexi Liu[1,2,3], Yanxiang He[1,3], Donghong Ji[3,4], and Hua Yang[1,3]

[1] School of Computer, Wuhan University, Wuhan 430079, P. R. China
[2] School of Physics, Xiangfan University, Xiangfan 441053, P. R. China
[3] Center for Study of Language and Information, Wuhan University, Wuhan 430079, P. R. China
[4] Institute for Infocomm Research, Heng Mui Keng Terrace 119613, Singapore
dexiliu@gmail.com, yxhe@whu.edu.cn, dhji@i2r.a-star.edu.sg, yh@eis.whu.edu.cn

Abstract. In this paper, we propose a novel multi-document summarization strategy based on Basic Element (BE) vector clustering. In this strategy, sentences are represented by BE vectors instead of word or term vectors before clustering. BE is a head-modifier-relation triple representation of sentence content, and it is more precise to use BE as semantic unit than to use word. The BE-vector clustering is realized by adopting the k-means clustering method, and a novel clustering analysis method is employed to automatically detect the number of clusters, K. The experimental results indicate a superiority of the proposed strategy over the traditional summarization strategy based on word vector clustering. The summaries generated by the proposed strategy achieve a ROUGE-1 score of 0.37291 that is better than those generated by traditional strategy (at 0.36936) on DUC04 task-2.

1 Introduction

With the rapid growth of online information, it becomes more and more important to find and describe textual information effectively. Typical information retrieval (IR) systems have two steps: the first is to find documents based on the user's query, and the second is to rank relevant documents and present them to users based on their relevance to the query. Then the users have to read all of these documents. The problem is that these docs are much relevant and reading them all is time-consuming and unnecessary. Multi-document summarization aims at extracting major information from multiple documents and has become a hot topic in NLP. Multi-document summarization can be classified into three categories according to the way that summaries are created: sentence extraction, sentence compression and information fusion.

The sentence extraction strategy ranks and extracts representative sentences from the multiple documents. Radev [1] described an extractive multi-document summarizer which extracts a summary from multiple documents based on the document cluster centroids. To enhance the coherence of summaries, Hardy Hilda [2]