Abstract

Previous studies on extractive document summarization regard a document as a set of sentences and they select a subset of the sentences as a summary. However, summaries generated by these approaches may lack coherence. Incoherent summaries mislead readers. To generate a coherent summary, (1) we regard a document as a discourse tree whose nodes represent sentences and the edges represent rhetorical relationship between them, and (2) extract a rooted subtree as a summary. We formulate the procedure as a Tree Knapsack Problem and develop an efficient algorithm to solve the problem. In future, we will extend our method by integrating it with sentence compression to generate short and coherent summaries.

Related work


Contact

Tsutomu Hirao
Linguistic Intelligence Research Group, Innovative Communication Laboratory
E-mail: hirao.tsutomu(at)lab.ntt.co.jp (Please replace {at} with @)