13 Generating coherent summaries from documents

～Document summarization by discourse tree trimming～

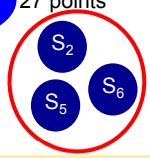
Abstract

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.

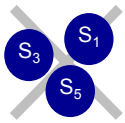
Document	Score	Words
S_1: It had been snowing during the morning.	6	7
S_2: It was a cold and windy morning in Kyoto.	10	9
S_3: The weather had improved in the afternoon.	6	7
S_4: We had been busy preparing for the meeting at the afternoon.	5	11
S_5: Our schedule had been canceled suddenly.	7	6
S_6: So, we enjoyed playing tennis.	10	5

Previous method

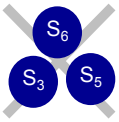
20 words, 27 points



20 words, 17 points



18 words, 23 points



Subset within 20 words that maximizes the sum of the scores

S_2: It was a cold and windy morning in Kyoto.

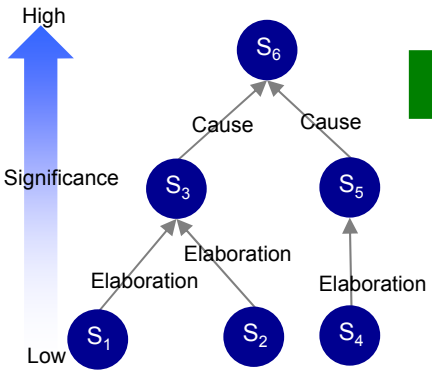
S_5: Our schedule had been canceled suddenly.

S_6: So, we enjoyed playing tennis.

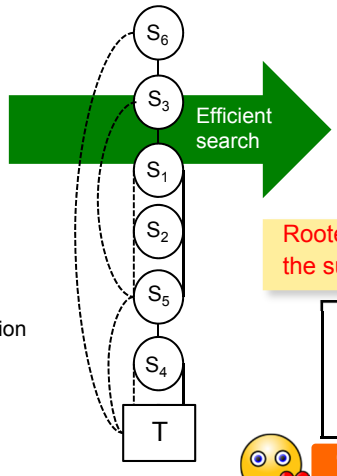
Incoherent summary: misleads readers

Our method

Dependency-based Discourse Tree

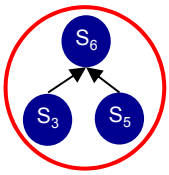


Succinct expression of the rooted subtrees

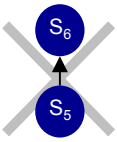


Efficient search

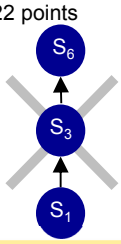
18 words, 23 points



11 words, 17 points



19 words, 22 points



Rooted subtree within 20 words that maximizes the sum of scores

S_3: The weather had improved in the afternoon.

S_5: Our schedule had been canceled suddenly.

S_6: So, we enjoyed playing tennis.

Coherent summary: does not mislead readers

Related work

[1] T. Hirao, Y. Yoshida, M. Nishino, N. Yasuda, M. Nagata, "Single-document summarization as a tree knapsack problem," in *Proc. EMNLP-2013*, pp. 1515-1520, 2013.

Contact

Tsutomu Hirao Linguistic Intelligence Research Group, Innovative Communication Laboratory

E-mail : hirao.tsutomu{at}lab.ntt.co.jp (Please replace {at} with @)