Summary-passage extraction for multiple articles summarization

Chikara Hashimoto

School of Information Science,
Japan Advanced Institute of Science and Technology

February 15, 2001

Keywords: multiple articles summarization, summary-passage.

The purpose of this study is to describe how to extract summary-passage from articles which are retrieved by searching newspaper corpus and to describe how to summarize articles using summary-passage. There are articles in which a journalist asserts something about incidents of the topic of the articles, a journalist explains the background or the relation of cause and effect about incidents of the topic, or a journalist summarizes major incidents of the topic. These articles have passages in which a journalist summarizes major incidents of the topic of the articles which he thinks important. We define these passages as "summary-passage". In summarizing multiple articles automatically, summrization systems have to identify where important contents are described in the articles. Summary-passage tells the system where important contents are. It follows that it is reasonable to think that summary-passage is effective in summarizing multiple articles.

To extract summary-passage from articles collected by searching newspaper corpus, firstly we need to detect assertive articles, explanatory articles, and summary articles. The problem in this stage is that some articles have contents of different categories(e.g. assertive, explanatory, summary). So we should detect these contents of different categories in one article individually. We define these contents corresponding to one category in a article as "section". And We define sections in which a journalist asserts something about incidents of the topic of the articles as "assertive section", sections in which a journalist explains the background or the relation of cause and effect about incidents of the topic as "explanatory section", sections in which a journalist summarizes major incidents of the topic as "summary section". So, in order to extract summary-passage from the articles, we follow the following procedure.

1. Detect assertive sections, explanatory sections, summary sections from articles.

2. Extract summary-passage from sections detected in 1.

Because assertive sections, explanatory sections and summary sections all contains summary-passage, We define these sections as "section containing summary-passage". In addition, We define sections which is not "section containing summary-passage" as "other section"(e.g. sections in which a journalist reports new incident of the topic). On the basis of the formalization described above, We implemented the system which extracts summary-passage from articles retrieved by searching newspaper corpus. The whole system consists of the system which detects sections containing summary-passage and the system which extracts summary-passage from sections detected. The system which detects sections containing summary-passage consists of the section partitioning module which partitions one article to sections, and a labeling module which gives a section the label representing the category which the section belongs to. In addition, the labeling module consists of nine filters and each filter corresponds to one category. The labeling module receives one section at a time, and the module filters the section through nine filters following the following procedure.

1. summary("syukannisshi")	4. other(official document)	7. other(report)
2. asserive("syasetsu")	5. summary(list)	8. explanatory(rest)
3. explanatory("kaisetsu")	6. assertive(predicate)	9. other

Filtering policy of this module is that the module firstly filters the section which can be labeled only by checking the title of the section(1〜4), secondly filters the section which can be labeled by checking sentences in the section(5〜7), and finally filters the section whose characteristic is unknown for now(8〜9). When a section is caught by some filter, the module gives the section a label corresponding to that filter, and outputs the section immediately without filtering the section through rest filters. For example, the section which is caught by 7th filter is output immediately without filtering through 8th and 9th filter.

We evaluated the system which detects sections containing summary-passage and the system which extracts summary-passage from sections detected respectively. The system which detects sections containing summary-passage was evaluated quantitatively and the system which extracts summary-passage from sections detected is evaluated qualitatively. In evaluating the system which detects sections containing summary-passage, section partitioning was done by human. That is, the section partitioning system was not evaluated while the labeling module was evaluated. Evaluation showed that detecting sections containing summary-passage can be done 60〜80% in both recall and precision,and extracting summary-passage from sections can be done well on the whole.

We suggest new multiple articles summarization algorithm using summary-passage as follows.

1. Extract summary-passage from articles collected by searching newspaper corpus.

2. Identify the correspondance of summary-passages with the relevant content to each summary-passage in the articles.

3. Generate summary from the relevant contents to summary-passages.

The future direction of this study will be (1) to improve both recall and precision of the labeling module, (2) to evaluate the performance of the section partitioning module, (3) to elaborate the summarization algorithm using summary-passage and (4) to construct more objective data set which is used to sophisticate the system's algorithm and to evaluate the system more objectively. Such data set should be constructed by performing a experiment of judging which category a section belongs to with some subjects.

In this study, *Mainichi* newspaper corpus was used.