JAIST Repository
https://dspace.jaist.ac.jp/

Title	長さの異なる要約の自動生成
Author(s)	伊良波，隆
Citation	
Issue Date	1999-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1228
Rights	
Description	Supervisor:佐藤 理史，情報科学研究科，修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Automatic Generations of Brief Summaries and Detailed Summaries

Takashi Iraha

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

February 15, 1999

Keywords: information extraction, automatic summarization.

As the Internet grows, new digital media such as World Wide Web and USENET News appeared. People can access various information via these digital media without geographical restrictions and time restraint. Now, convenient and useful network society is formed. However, it is getting difficult for network users to find the necessary information, because the amount of available information increases rapidly. There are a lot of information that is useful but not usable —*sleeping information*— in the network society.

To obtain the desired information, users must do information retrieval and select from a lot of retrieval outputs. One way to support the selection of desired information is to provide simple and appropriate summaries to users; uses can select the desired information easily without reading the whole outputs of retrieval.

In this research, I implemented the system that generates summaries of netnews articles automatically. The target text is the newsgroup fj.sys.sun of USENET News; this newsgroup consists of the question articles about Sun workstations and answer articles to them. The automated editing project presents the question and answer package **SUN QA-PACK** on WWW [1]. Sun QA-PACK is generated by the system that classifies question articles hierarchically, generates summaries from question articles, and present three types of pages to assist users finding the desired information easily. The system that generates SUN QA-PACK generates one type of summaries. In this research, I implemented the system that generates two different types of summaries: the brief summary that consists of one sentence, and the detailed summary that consists of multiple sentences.

First, I investigated the target newsgroup fj.sys.sun. I determined ten sentence types using the distinctive expression patterns that often appear in the question articles:

1. greetings, self-introduction 2. software environment 3. goal 4. failure 5. question 6. error message 7. analysis 8. commentary 9. signature 10. others

Copyright © 1999 by Takashi Iraha

[1] http://www-sato.jaist.ac.jp:8000/faq/

The central information of the question article of fj.sys.sun is what he wants to do, what his problem is, and what the question is. The type of the sentence that describes what he wants is *type 3* (goal). Auxiliary verb "～たい" (want) is often used in the sentence of this type. The type of the sentence that describes what his problem is *type 4* (failure). In this type, the sentence is negation, or the verb that expresses negative meaning at the end of the sentence. The type of sentence that describe what the question is *type 5* (question). Most sentences are question sentences in this type. The system extracts the sentences of these three types as the important sentences of fj.sys.sun.

The followings are the steps of summary generation.

1. Decide the sentence type of each sentence in the question article by using string pattern matching.
2. Filter out the meaningless sentences.
3. Extract the first type 3, 4, or 5 sentence as the brief summary.
4. Extract the all sentences of type 3, 4, or 5 as the detailed summary.

At the step 2, the system filters out every sentence that does not contain the object of the main verb; two types of words can be this object. One type is technical terms of computer science, which is described by alphabets or Katakana. Another type is general nouns such as method and books. Listing up all these nouns is not practical. Therefore, I listed up meaningless sentences from question articles of fj.sys.sun by hand previously, listed up all nouns in these sentences, and regard them as meaningless nouns and put them into the meaningless-noun list. If the sentence has the noun that is described by alphabets or Katakana, or the noun that is not in the meaningless-noun list, the sentence has the meaning. If not, the sentence is meaningless.

I evaluated this summary generation system. The accuracy of the type-decision is 91 percent. The accuracy of generation of brief summary is 87 percent; the accuracy of generation of detailed summary is 93 percent.