Automatic Generation of Digests for Meeting Announcements in the NetNews

Madoka SATO Satoshi SATO Yoichi SHINODA

July, 1994
IS-RR-94-22I

School of Information Science
Japan Advanced Institute of Science and Technology, Hokuriku
Asahidai 15, Tatsunokuchi
Nomi, Ishikawa, 923-12, JAPAN
{madoka,sato,shinoda}@jaist.ac.jp

ISSN 0918-7553

Abstract

Total amount of electronic text is already enormous now and computers provide remarkably good technologies of text editing and storing. However, as for text retrieval, there is no efficient means if the desire to information is not exact enough for the keyword retrieval. A new technology that provides navigative function and helps people acquire desired information is required now. In this study, we have developed a method of automatic generation of digests for the NetNews. A digest is a collection of summaries of original texts. Headlines of a newspaper, a table of contents of a magazine and summary on a back cover of book are all digests and work as navigators. We propose digests are the most appropriate as the navigator for the NetNews with consideration of the fact that readers are guided by above digests almost every time they reach to those printed materials and retrieve the desired information without much effort. Generation of digests for printed materials require certain degree of skills and a lot of time because all summarization, edition, and categorization processes are executed mainly by manual works. By contrast, automatic generation of digests is possible for the NetNews because the NetNews is electronic text. The main characteristics of the NetNews is made good use by automatic generation of digests. We concentrate on making digesting system for a newsgroup, fj.meetings that holds only "meeting announcement" and "call for paper" articles in both Japanese and English. The digesting system has two modules: MekeSummary and EdigDigest. In MakeSummary, we utilized expression patterns and distinctive document styles to extract information. The introduction of utilization of document styles contributes to obtaining high performance on summary extraction tasks without deep analysis of texts. MakeSummary has been evaluated on pre-examined and blind data and the result has demonstrated that the system is powerful enough in ability to accurately extract summaries. Adding more expression patterns could raise the performance easily. We believe that the experience of digesting the NetNews articles can be transplanted to other electronic text information also.

Contents

1 Introduction

There are tremendous numbers of computers in homes and offices already and many of them are already interconnected by computer networks. Under this circumstance, total amount of electronic text is already enormous and is still growing rapidly. Computers now provide remarkably good technologies of text editing and storing as a result of recent development. These technologies extended human abilities to manage text much more than people imagined in the era of paper. As for text retrieval, we can extract desired information by keyword retrieval if the desire is exact enough. However, if the desire is not exact, there is no efficient means of navigating through large information ocean. A new technology that provide navigative function and help people acquire useful information is required now.

In this study, we have developed a method of automatic generation of digests for the NetNews. The NetNews is a considerably large collection of unedited electronic text so that there is a strong need for an effective navigator. We are inspired by the fact that almost all printed materials have something that work as a navigator. For example, newspapers have headlines that show the topics with their gravity, magazines have tables of contents and books usually have their summary on the back covers. All these are digests of original text. Readers are guided by digests almost every time they reach to those printed materials and retrieve the desired information without much effort. Taking this fact into account, we propose digests are the most appropriate as the navigator for the NetNews. In the following, we will briefly explain why the digest is appropriate as the NetNews navigator and key technologies of automatic generation of digests in section 2. Section 3 describes overview of automatic generation of digests. Section 4 describes the method of summary extraction from "meeting announcement" and "call for paper" articles in both Japanese and English. Then, in section 5, results of experiments on the accuracy of the summary extraction and discussion on the results are described. Finally section 6, concluding remarks are presented.

2 The Digest of the NetNews

There is a lot of useful information from all over the world in the NetNews. The NetNews has the capability to be an active information source comparable to newspapers, magazines and other printed mass communication media. However, as the NetNews has the following characteristics, it is difficult for many people to search for information of interest.

- The amount of information is too large for a human to check all.
- The NetNews is displayed on the CRT now so that readability is not comparable to other printed information sources such as newspapers, magazines, etc.
- There is no editor in the NetNews so that articles reach readers unedited.

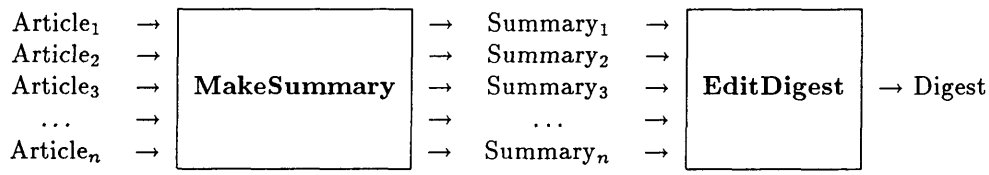

Figure 1: The Configuration of AutoDigest

Regardless of above characteristics, effective retrieval system for the NetNews articles is not introduced yet. The experiences of navigating readers in newspapers, magazines and books can be utilized when designing the NetNews digest. However, the way of making the NetNews digest is different from it of printed materials. Most of summarization, edition and categorization processes for texts on printed materials are executed mainly by manual works. These works require certain degree of skills and a lot of time. By contrast, the NetNews is in electronic form from the beginning so that it is far more suitable for the automatic generation of digests than printed materials. The main characteristics of the NetNews is made good use by automatic generation of digests.

The key technology of automatic generation of digests is the extraction of pre-specified information from the original texts. In this study we concentrate on making automatic digesting system for a newsgroup, fj.meetings. Fj.meetings holds only "meeting announcement (MA)" and "call for paper (CFP)" articles. Articles are written in Japanese or in English. Taking the aims of articles in fj.meetings into account, we specified the components of the digest as follows:

- category of the article (MA or CFP)
- title of the meeting
- date
- place
- deadline (CFP only)

We call a set of above components "summary" of the article. Summaries are extracted from the original articles by using the following method.

3 Overview of AutoDigest

Figure 1 shows the configuration of AutoDigest, a system that generates a digest of an article. The system consists of two modules: "MakeSummary" and "EditDigest ".

MakeSummary automatically extracts summaries from an article using *document styles* and *expression patterns*. This module first check whether the article is written in Japanese or English, then selects an appropriate program to the

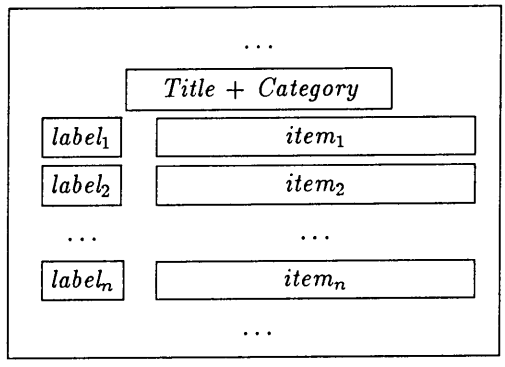

Figure 2: A Typical Style of Japanese articles

language from prepared two programs for those two languages. MakeSummary works as the key of the system. The details are described in the following sections.

EditDigest generates a digest from summaries made by MakeSummary. This is a very simple program and its fundamental function is to sort summaries by their dates.

4 Summary Extraction

Summary extraction is the key technology of generation of digests. Similar approaches are studied in the area of information extraction from texts. Related past works in this area include [1], [2] and [3]. All of them use *expression patterns* as trigger words to extract information. In our MakeSummary, in addition to using expression patterns, we utilize *document styles* to extract summaries from articles.

4.1 Document Style

We studied MA and CFP articles in fj.meetings and found distinctive document styles. These distinctive styles exist in both Japanese and English articles but in different styles.

A typical document style of Japanese articles (Figure 2) has centered lines and itemized lines below the centered lines. The centered lines consist of a title and words that represent the category of the article, e.g., "Heiretu Jinko Chino Kenkyukai no Kaisai no Oshirase (Announcement of SIG Meeting of Parallel Artificial Intelligence) ". Other information, i.e., date, place and deadline, are represented in itemized lines. An item is often with an item label, which represents what kind of information is presented in that line. For example, synonyms of *date* are used as *date* item labels, e.g., Nichiji, Kijitsu, Nittei, Kaiki, Kaisaibi.

By contrast, most of English articles consists of two blocks (Figure 3). In the first block (we call this block "primary block"), title, date and place are written in

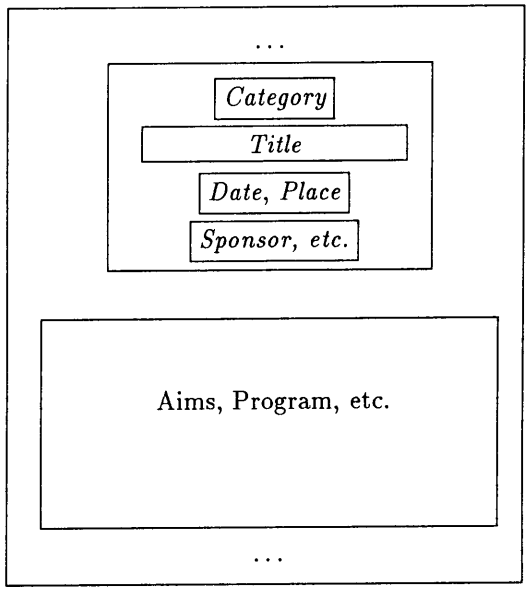

Figure 3: A Typical Style of English articles

centered or lefted lines. In the second block (we call this block "secondary block"), topic, aims, detailed description, program and deadline are written in lefted lines. All the information for a summary except deadline is placed in the primary block in most of the articles: the primary block is a natural *digest* of the article.

4.2 Expression Patterns

Other triggers for summary extraction are *expression patterns*. For example, we have distinctive expression patterns to describe date in English, i.e., "*Month Day* [, *Year*]" and "*Day Month* [, *Year*]". Therefore the system can easily find the date by using pattern matching with these expression patterns. Table 1 shows the examples of expression patterns we use.

4.3 The Algorithm of Summary Extraction

We use distinctive document styles and expression patterns to find prespecified information. Deep analysis of an article is not necessary because we do not use recognition of concepts in an article.

We developed two programs of summary extraction: for Japanese articles and for English articles. These two programs have different details but the same skeleton that consists of four phases as listed below.

1. Extracting the category.
2. Locating the starting point of the extraction.
3. Extracting the components of the summary.

Table 1: Examples of Expression Patterns

	Japanese Articles	English Articles
Category (M.A.)	Kaisai, Sanka Bosyu, etc.	*Announcements, Call for Participation*
Category (CFP)	Ronbun Bosyu, Happyo Bosyu, etc.	*Call for Paper*
Title	including Gakkai, Kenkyukai, Dai N Kai, etc.	including *Conference*, *Workshop*, *N-th*, etc.
Date	"[*Year*] *Month Day* [-[*Month*] *Day*]", etc.	"*Month Day* [-[*Month*] *Day*][, *Year*]", etc.
Place	none	the names of the country, the state, the city, etc.
Deadline	"[*Year*] *Month Day*" with Shimekiri, Hicchaku, etc.	"*Month Day* [, *Year*]" with *submission*, *due to*, etc.

4. Rescanning for unextracted information.

The first phase extracts the category of an article. The system scans the whole article to find the phrase that matches expression patterns of categories. If the system found the phrase that matched an expression pattern, the category of the article is determined. If not, the system infers the category of the article is MA.

The second phase locates the starting point of the extraction process. It is often the case that extra texts that represent a short introduction of the meeting or Japanese introduction for the announcement in English is placed in front of the *body* of the announcement. Therefore the system has to skip this extra part. As for Japanese articles, the system searches the title line by using title expression patterns. As for English articles, the system searches the date line in the primary block by using date expression patterns. These lines are the starting points of the following extraction processes.

The third phase starts extracting components of the summary from the starting point we found in the second phrase. As for Japanese articles, after extracting the title, the system searches itemized lines and extract other components by using item labels. As for English articles, the system determines the first and the last lines of the primary block by using several heuristics, which utilize several style information such as centered lines, blank lines, separation lines (e.g., ——), decoration by symbols (e.g., ******) and so on. Then the system extracts title, date and place from the primary block.

If all components are extracted by the first three phases, the extraction process is completed. Otherwise the system rescans the whole article to find unextracted components. Extraction of this phase mainly relies on expression patterns. For example, deadline is extracted if there is a keyword of deadline, such as "submission", "due to", "deadline", and date expressions around the keyword.

Table 2: Experimental Results of Summary Extraction for Japanese Articles

The Number of Correct Answers and Accuracy						
	Known Data			Unknown Data		
	MA	CFP	total	MA	CFP	total
	90articles	24articles	114articles	73articles	24articles	97articles
categories	89 (98.9%)	22 (91.7%)	111 (97.4%)	70 (95.9%)	21 (87.5%)	91 (93.8%)
titles	84 (93.3%)	24 (100%)	108 (94.7%)	63 (86.3%)	21 (87.5%)	84 (86.6%)
dates	89 (98.9%)	24 (100%)	113 (99.1%)	70 (95.9%)	20 (83.3%)	90 (92.8%)
Places	85 (94.4%)	23 (95.8%)	108 (94.7%)	70 (95.9%)	23 (95.8%)	93 (95.9%)
deadline		22 (91.7%)			19 (79.2%)	
total	77 (85.6%)	21 (87.5%)	98 (86.0%)	57 (78.1%)	13 (54.2%)	70 (72.2%)

4.4 Examples

Figure 4 shows a typical example of an English CFP article. The article consists of the primary block and the secondary block. Figure 5 shows the summary extracted from the above article by MakeSummary. It shows that all the components for a summary are correctly extracted.

5 Experiments and Discussion

We evaluated the subsystem "MakeSummary" on data that are examined beforehand and blind data. The result has demonstrated high performance. Table 2 shows the number of correct answers and its accuracy for both the pre-examined and blind data. We obtained the following accuracy from the experiments.

- 86% for the pre-examined 114 Japanese articles.
- 72% on the blind test of 97 Japanese articles.
- 91% for the pre-examined 46 English articles.
- 53% on the blind test of 60 English articles.

These results demonstrate that the system is powerful enough in ability to accurately extract summaries. These high performance on summary extraction tasks is obtainable with little computational effort.

Achieving higher accuracy on summary extraction in English articles is left for further study. The cause of this relatively low accuracy is that the number of articles we examined beforehand is too small. We believe that adding more expression

Ninth Annual IEEE Symposium on
LOGIC IN COMPUTER SCIENCE
July 4-7, 1994, Paris, France

CALL FOR PAPERS

The LICS Symposium aims to attract high quality original papers covering theoretical and practical issues in computer science that relate to logic in a broad sense, including algebraic, categorical and topological approaches.

Suggested, but not exclusive, topics of interest include:
abstract data types, automated deduction, concurrency, constructive mathematics, data base theory, finite model theory, knowledge representation, lambda and combinatory calculi, logical aspects of computational complexity, logics in artificial intelligence, logic programming, modal and temporal logics, program logic and semantics, rewrite rules, logical aspects of symbolic computing, problem solving environments, software specification, type systems, verification.

DATES:
Submission deadline: December 13, 1993
Notification: February 25, 1994
Final papers due: April 15, 1994
Conference: July 4-7, 1994

PAPER SUBMISSION:
10 hard copies of a detailed abstract (not a full paper) and 20 additional copies of the cover page should be received by December 13, 1993 by the program chair. This is a FIRM DEADLINE: late submissions will not be considered. Authors without access to duplication facilities may submit a single copy of each. Authors will be notified of acceptance by February 25, 1994. Accepted papers (in a specified proceedings format) will be due April 15, 1994. expected to present the paper at the conference.

.....

Figure 4: An Example of an English CFP Article

TYPE	CFP
TITLE	Ninth Annual IEEE Symposium on LOGIC IN COMPUTER SCIENCE
DATE	July 4-7, 1994
PLACE	Paris, France
DEADLINE	December 13, 1993

Figure 5: The Summary Extracted from the article in Figure 4

Table 3: Experimental Results of Summary Extraction for English Articles

The Number of Correct Answers and Accuracy						
	Known Data			Unknown Data		
	MA	CFP	total	MA	CFP	total
	24articles	22articles	46articles	31articles	29articles	60articles
categories	24 (100%)	22 (100%)	46 (100%)	31 (100%)	27 (93.1%)	58 (96.7%)
titles	23 (95.8%)	21 (95.5%)	44 (95.7%)	21 (67.7%)	26 (89.7%)	47 (78.3%)
dates	23 (95.8%)	21 (95.5%)	44 (95.7%)	27 (87.1%)	28 (96.6%)	55 (91.7%)
places	24 (100%)	21 (95.5%)	45 (97.8%)	26 (83.9%)	23 (79.3%)	49 (81.7%)
deadline		22 (100%)			23 (79.3%)	
total	23 (95.8%)	19 (86.4%)	42 (91.3%)	17 (54.3%)	15 (51.7%)	32 (53.3%)

patterns as triggers could raise the performance up to 80%. Studies on exceptional document styles would contribute to even higher performance.

6 Concluding Remarks

In this paper, we developed the system that generates digests of a meeting announcement newsgroup of the Net-News. It extracts important information from unedited articles by automatic process. In the method of summary extraction, we utilized not only expression patterns but document styles. The utilization of both document styles and expression patterns contributes to the following advantages of this system.

- It is simple. Because it does not perform deep analysis of texts.
- It is accurate enough for practical use. Additional study on distinctive expressions can bring higher performance easily.
- It has fast runtime.

Digests generated by this system work as a navigator and help people select useful information.

Digesting the NetNews articles is, in another word, editing. Edition by automatic process is a radical idea and is successfully realized because all information source, the NetNews articles, were already in electronic form, which is the fundamental characteristic of the NetNews. We believe that the experience of digesting the NetNews articles can be transplanted to other Internet information sources also.

References

[1] Douglas E. Appelt, Jerry R. Hobbs, John Bear, David Israel and Mabry Tyson. 1993. FASTUS: A Finite-state Processor for Information Extraction from Real-world Text. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1172–1178, Chambery, France.

[2] Megumi Kameyama and Isao Arima. 1993. A Minimalist Approach to Information Extraction From Spoken Dialogues. In *Proceedings of International Symposium on Spoken Dialogue*, pages 137–140, Tokyo, Japan.

[3] Sadao Kurohashi, Makoto Nagao, Satoshi Sato, and Masahiko Murakami. 1992. A Method of Automatic Hypertext Construction from an Encyclopedic Dictionary of a Specific Field. In *Proceedings of Third Conference on Applied Natural Language Processing*, page 239–240, Tokyo, Japan.