Abstract
The information to include in a summary varies depending on the author's intention and the use of the summary. To create the best summaries, the appropriate goals of the extracting process should be set and a guide should be outlined that instructs the system how to meet the tasks.

The approach described in this report is intended to be a basic architecture to extract a set of concise sentences that are indicated or predicted by goals and contexts. To evaluate a sentence, the sentence selection algorithm simply measures the informativeness of each sentence by comparing with the determined goals, and the algorithm extracts a set of the highest scored sentences by repeat application of this comparison.

This approach is applied in the summary of newspaper articles. The headlines are used as the goals. Also, the method to extract characteristic sentences by using property information of text is shown.

In this experiment in which Japanese news articles are summarized, the summaries consist of about 30% of the original text. On average, this method extracts 50% less text than the simple title-keyword method.

1 Introduction
Summary requirements (such as length and content) vary widely, depending on author and situation of use. For example, even several sentences may seem too long for news articles obtained from a network. Similarly, as short as possible summaries will be desirable to preview sites in a web browser, when a huge number of results are retrieved from search engines.

To extract a short summary for this kind of purpose, an extract covering all topics in the text will be too long. Using small number of sentences to extrapolate the contents of the entire text will be adequate for an efficient preview. To include the intended points and characteristic information in a short summary, the mechanism to detect the purpose of the summary and select the sentences that match the goals is needed in the summarization process.

In this report, an algorithm that helps realize such a goal and context information oriented summarization system is described. The algorithm evaluates the informativeness of each sentence in a text and selects a small number of sentences, including effective information. One of the applications of this algorithm is shown in the experiment on the sentence extraction from the newspaper articles and market surveys. The experimental system uses headlines and titles as the goals of the sentence selection, and the results are shorter and more effective than the simple title-keyword method (Pace, 90).

The results of the current simple experiment are based on the word matching that as the goal processing. However, the experiments should include processing of the following structural goals, the concept level matching that uses the thesaurus, and the topic detection from the text.

2 The Goal-Directed Summarization
Summaries in this system may differ from the general notion of a summary that covers all topics described in the original text. A summary is defined as a set of extracted sentences that gives some idea to the reader of the contents of a text, the reader is able to determine whether the text is worth reading or not based on the summary. Under this definition, a summary is effective if the extract includes the author's intention or required information of the reader by the fewest number of sentences possible. These information should be included and satisfied by extracted sentences are called the 'goals.' The summarization process is guided by the goals called 'goal-directed.' Figure 1 shows the system architecture of a general goal directed summarization system.
system consists of a goal detection and sentence selection process by informativeness evaluation.

The 'goal-directed' method may be overstated, because the current experimental system handles only the headlines, titles, and some text property expressions. However, the 'goal-directed' method is named as the first step toward realizing a context-based summarization system.

### 3 Sentence Selection Algorithm

The sentence selection algorithm calculates the 'informativeness' for each sentence in a document. The measurement represents the strength of relation between the goals, sentences, and the richness of information in a document. These variables are defined by the following three numerical values:

1. Number of different sentence expressions related to the goals
2. Total number of sentence expressions related to the goals
3. Total number of sentence expressions being not related to the goals

The order of these measurements defines their precedence. The first measurement is given the highest priority. Sentences that satisfy many of the goals are considered more informative. Both the first and second values above represent the amount of information included in a sentence. The third measurement indicates the amount of information in a sentence and roughly simulates the contained amount of explanation or description about the goal.

The sentence selection algorithm (shown in Figure 2) relates the highest scored sentences by the informativeness measurement. The measurements are repeatedly evaluated until all the goals are related to the sentences or all relations are found.

### 4 Goal Detection

This system is designed to be built into the text preview menu of a word processor or the query results listing of a document retrieval system. Thus, the contents of a document are unpredictable and the system needs to work in real time. This limitation requires the system to handle rather simple information. For example, the word list compiled from the headlines is used as the goals when processing news articles. The title words are used to extract a text from a report. These simple word lists may be too simple and a little inadequate as goals.

Goal-directed summarization includes the processing of the structural information. This includes the concept-level goal detection using a thesaurus, document structure, and structural information in the titles (section, subsection...).

### 5 Experiments

The first experiment is summary for 13,562 newspaper articles and 62 monthly market survey report articles. Both texts are in Japanese. The calculated extraction rates based on the total number of

---

**Table 1** Extraction rates of newspaper articles

| Extraction Rate | Simple title-keyword | Informativeness Selection |
|-----------------|----------------------|--------------------------|
|                 | Number of Articles   | Rate                     |
| 100%            | 2,237                | 16.5%                    |
| 90%             | 1,038                | 8.0%                     |
| 80%             | 1,758                | 13.0%                    |
| 70%             | 1,642                | 12.1%                    |
| 60%             | 1,441                | 10.6%                    |
| 50%             | 1,250                | 9.2%                     |
| 40%             | 1,027                | 7.9%                     |
| 30%             | 813                  | 6.0%                     |
| 20%             | 654                  | 4.8%                     |
| 10%             | 501                  | 3.7%                     |
| - 10%           | 218                  | 1.6%                     |
| 0%              | 938                  | 0%                       |
| Total           | 13,562               | 100%                     |
| Average         | 64%                  | 32%                      |
| Median          | 70%                  | 27%                      |
Table 2: Extraction rates of computer business survey reports

| Extraction Rate | Simple title-keyword | Informativeness Selection |
|-----------------|----------------------|---------------------------|
| Rate            | Number of Articles   | Rate                      |
|                 |                      | Number of Articles        | Rate                      |
| 100%            | 3                    | 3.3%                      | 0                         | 0%                        |
| 90%             | 2                    | 3.3%                      | 0                         | 0%                        |
| 80%             | 6                    | 9.7%                      | 0                         | 0%                        |
| 70%             | 5                    | 8.1%                      | 0                         | 0%                        |
| 60%             | 7                    | 11.3%                     | 0                         | 0%                        |
| 50%             | 3                    | 4.8%                      | 1                         | 1.6%                      |
| 40%             | 11                   | 17.7%                     | 3                         | 4.3%                      |
| 30%             | 13                   | 21.0%                     | 0                         | 0%                        |
| 20%             | 8                    | 13.0%                     | 5                         | 8.0%                      |
| 10%             | 4                    | 6.5%                      | 10                        | 16.1%                     |
| -10%            | 0                    | 0%                        | 42                        | 67.7%                     |
| 0%              | 1                    | 1.6%                      | 1                         | 1.6%                      |
| Total           | 62                   | 100%                      | 62                        | 100%                      |
| Average         | 49%                  |                           | 11%                       |                           |
| Median          | 43%                  |                           | 7%                        |                           |

Table 3: Average Extraction rates of English news articles

| Method            | Average extraction rates |
|-------------------|--------------------------|
| Informativeness selection | 8%                    |
| Simple title-keyword       | 41%                    |
| Simple frequency-keyword    | 33%                    |

Characters listed in Table 1 are a summary of the text by this system shows 50% of the length by the simple title-keyword method. The most frequent compression rate in the results of the simple title-keyword method is 100% (the entire text). By using the informative selection, the rate falls between 20% to 30%

Table 2 lists the results of the computer business survey reports. In this case, the differences between the rates are larger than the newspaper results. The text of these business reports is longer than the newspaper articles.

These experiments are mostly of Japanese documents. Only a few results for English documents are available. Table 3 lists the results of the extracting summaries of English news articles. In this case, the extraction rates are calculated based on the total number of words. The nature of this system makes evaluating the contents difficult and no clear solution can be obtained.

Table 4 shows the results of this experiment using small number of the specifications documents of hard disk drives.

As shown in Figure 3(a), the high tdf words determine the sentences describing the distinctive features of the hard disk that are to be selected. Figure 3(b) shows that the words with high document frequencies are used to select the common information about the general specifications.
(a) **Extraction by tf idf property**

**Words with high tf idf scores**

DEs, DMs, F6632A, H, path configuration, MB, GB, path, RANK, F6493, F6429G

**Summary by the high tf idf words**

Flexible configuration The F1700B has a four path configuration (connection path to a magnetic disk) as a standard feature.

In addition, in the F1700B, the path to the channel and the paths to the magnetic disk unit can be increased independently, so a flexible configuration can be found to suit the system environment.

High speed data transfer Data transfer rate between host is high speed 3.0 MB/sec or 4.5 MB/sec. F1700B + F6425G/H, or F6427G/H, or F6429G/H has to be sold as a subsystem.

(b) **Extraction by document frequency property**

**Words with the highest document frequency**

- table, page, m3, contents, width, weight, temperature, power consumption, KVA, height, heat dissipation, frequency, dimension, depth, air flow

**Summary by the high df words**

- Width 1,040
- Dimension(mm) Depth 815 Height 1,690
- Weight (Kg)
- Frequency 50/60Hz +/- 10
- 16(22) Heat dissipation (°C) includes 512MB cache
- 780(1,240) 1,240(1,700) 1,320(1,780)
- 930(1,600) 1,240(1,700) Air flow(m3/min)
- Temperature 15 - 32 degrees celsius (When controlled) Environment

**Figure 3** Summary examples using the properties for the text classification.

6 **Discussion**

This experiment only demonstrates a small part of goal-directed summarization. Many subjects still need to be tested.

1 **Using of the thesaurus**

Most failures in processing news articles were caused by synonyms (such as 'corpse' and 'dead body', 'fishery' and 'fisherman') to be matched. Most of these errors can be corrected by using the thesaurus.

2 **Processing the structured goals**

To summarize structured documents (such as manuals) the hierarchical structure of the sections and subsections can be used to create goals. These goals may control the inheritance of sub-goals to be satisfied in the substructure (such as, the 'preface' section).

3 **Resolving the anaphonic expression**

Fewer problems than the English sentence extraction occurred, because Japanese text was mostly the subject of experiment and the text less contains the anaphonic expression.

However, person and company names in news articles are often abbreviated and shortened. Resolving these abbreviated and shortened expressions is needed to increase readability.

4 **Control of the summary length**

Because the main purpose of this system is to offer concise information for previewing document contents, the length of output cannot be directly controlled. If the length needs to be varied, some methods to extend the results may be added as post-processing. The method to find sentence relations (such as lexical cohesion) may be suitable to find sentence chains with related topics.

5 **Evaluation method**

The evaluation of extracts cannot be simply defined. Extracts cannot be evaluated without context. For objective evaluation, measuring the effect (e.g., the time of previewing) may be realistic.

7 **Conclusion**

This report is about the sentence extraction experiment using the 'informativeness' evaluation method. The evaluation of the extracted summaries shows the system selects smaller sets of sentences than the simple title-keyword method without losing information content. Enough information is extracted for previewing document contents.

The current system may be too simple to be regarded as a 'goal directed'. However, this experiment shows, the efficiency of the generated summaries is improved, even when a simple words list is used as the goal of the selection process in the system.

**References**

Julian Kupiec, Jan Pedersen and Francine Chen 1995

A Trainable Document Summarizer, In ACM SIGIR’95, pages 68–73

Chris D Pace 1990 Constructing Literature Abstracts by Computer Techniques and Prospects Information Processing & Management, Vol 26, No 1, pages 171-186

Gerard Salton and James Allan 1993 Selective Text Utilization and Text Traversal In Hypertext'93, pages 131–143