A Method for Abstracting Newspaper Articles by Using Surface Clues

Hideo Watanabe
IBM Research, Tokyo Research Laboratory
1623-14, Shimotsuruma, Yamato-shi, Kanagawa-ken 242, JAPAN
watanabe@trl.ibm.co.jp

Abstract

This paper describes a system which automatically creates an abstract of a newspaper article by selecting important sentences of a given text. To determine the importance of a sentence, several superficial features are considered, and weights for features are determined by multiple-regression analysis of a hand processed corpus.

1 Introduction

The rapid expansion of the Internet enables us to easily access a lot of information sources in the world. The ability to browse information quickly is therefore a very important feature of an information retrieval and navigation system. Abstraction of a document is one useful tool for quick browsing of textual information.

Generally, an abstract can be considered to be a concise text giving an outline of the original text. Creating an abstract requires deep semantic processing with broad knowledge, and the strategy for generating an abstract depends on the type of target text. Abstracts created by humans tend to differ according to their creators' background knowledge and interests. Furthermore, as stated in [6], the same person is likely to create different abstracts of the same text at different times. Simulating this human process is clearly outside the area that can be dealt with by current computational linguistics. There are, however, some cases in which an abstract can be created by using surface clues to make conjectures as to which portions are the most important without using deep semantic processing.

The most practical way to create an abstract is thus to determine the most important portions by using surface clues. There are two lines of research based on this approach: one analyzes some aspects of a text's structure, such as the rhetorical structure [7], and selects some sentences according to this structure [5, 3]; the other analyzes surface features for each sentence in a given text and selects the most important sentences according to some heuristics [6, 1, 9]. In methods of former type, the rhetorical structure is appropriate for a relatively small set of sentences such as a paragraph, but it does not give enough information to create an abstract for a large set of sentences. In methods of the latter type, the validity of the heuristics is uncertain when the target text is changed. Therefore, this paper proposes a method for selecting important sentences by using an equation based on surface features and their weights, and a method for determining these weights by multiple-regression analysis of abstracts created by humans. The target texts of this method are Japanese newspaper articles.

2 Surface Features of a Sentence

The proposed method is to create an abstract by determining important sentences according to features extracted from each sentence. For each sentence in a given Japanese newspaper article, the following features are analyzed:

- Important Keywords:
  An important keyword is defined as a keyword that appears in another sentence or in a title. The number of points for this feature is the total number of occurrences of important keywords.

- Tense:
  The tense of a sentence is analyzed as past or present. This feature gives 1 point for present, and 0 for past.

Most of these features were proposed in the previous studies. Keywords were proposed in [6], sentence location was proposed in [1], sentence type was proposed in [1, 9], etc., and rhetorical relations were proposed in studies using rhetorical structures such as [3].
• Type of a Sentence:
  Sentence types are fact, conjecture, or insistence. This feature gives 0 points for fact, 1 for conjecture, and 2 for insistence.

• Rhetorical Relation:
  The rhetorical relations to the preceding context is analyzed as example, adverse, parallel, comparison, or connection. This feature gives 1 point for reason, 2 for example, and 0 for others.

• Distance from the beginning of a text:
  In general, sentences located near the beginning of a text tend to be important. Therefore, sentences in the first paragraph are given 5 points for this feature, sentences in the next paragraph 4, and so on.

• Distance from the end of a text:
  Sentences located near the ending of a text also tend to be important. Therefore, sentences in the last paragraph are given 5 points for this feature, sentences in the previous paragraph 4, and so on.

The tense of a sentence is simply determined to be past if it has “ta” (an inflection for the past tense) in the last phrase. The reason why tense is used is that sentences stating about the current fact seem to be more important than ones about the past fact in the context of editorial articles.

The sentence type is determined by checking special expressions in the last phrase. For instance, if the final phrase contains “bekida” (“should”) or “nakereba-naranai” (“must”), then its sentence type is insistence; if it contains “darou” (probably ...”), then its type is conjecture; otherwise, its type is fact. Examples of special expressions used to determine sentence type are as follows:

- Conjecture: kamosirenai (may), kanenai (be capable of), souda (likely to), youda (likely to), darou (probably), etc.
- Insistence: tai (want to do), hosii (want someone to do), bekida (should), nakereba-naranai (must), taisetu-dearu (important), hituyouda (necessary), etc.

The rhetorical relation is determined by checking special expressions both in the first phrase and in the last phrase of a sentence. For instance, if “sitakarada” is found in the last phrase, then the rhetorical relation is reason, and if the conjunction “sikasi” (“but”) is found, then the rhetorical relation is adverse. Examples of special expressions used to determine rhetorical relations are listed below:

- Example: tatoeba (for instance), nado (etc.), etc.
- Adverse: sikasi (but), tokoroga (however), etc.
- Comparison: koreni-taisi (while), etc.
- Parallel: mata (further), sarani (in addition), etc.
- Reason: karada (because), tameda (because), etc.

3 Process of Creating an Abstract

The basic method for creating an abstract in most previous studies has been to analyze the sentences of a text in terms of some surface features, and a heuristic to determine the most important sentences on the basis of these features.

The method proposed in this paper formalizes the above approach so that the importance of each sentence is calculated as the sum of feature points multiplied by their feature weights. The most important sentences are then extracted as an abstract. The importance $S$ of a sentence is calculated as follows:

$$S = a + \sum_{i=1}^{n} W_i \times P_i$$

where $a$ is a constant, $P_i$ is the number of points assigned to the $i$-th feature, which is normalized to be between 0 and 1, and $W_i$ is the weight assigned to the $i$-th feature.

The steps in creating an abstract are as follows:

1. For each sentence, calculate the importance.
2. Select the sentence that has the highest importance value among the unselected sentences.
3. If the selected sentence $s_1$ has another sentence $s_2$ in the previous context that is related to $s_1$ by any rhetorical structure, then $s_2$ is also selected and marked.

In English, this expression corresponds to “because” in the first phrase.

In this method, past does not imply the past tense in a strict sense but rather the sentence is not in the present tense. In Japanese, “ta” implies the past tense, completion, and so on. Most cases are actual instances of the past tense.

It is sufficient to check in the last phrase for Japanese sentences, because a predicative phrase is always located at the end of a Japanese sentence. Therefore, another strategy is needed for languages in which a predicative phrase may be located in the middle of a sentence.

3 These checking of sentence types and rhetorical relations are based on [10].
4. If the ratio of the number of selected sentences to the number of sentences in the text exceeds the specified one, then terminate this process; otherwise, goto 2.

These steps select sentences on the basis of their importance value, but they also respect the rhetorical structure to some extent (step 3), because if the rhetorical structure is totally ignored, the output text will be awkward to read.

4 A Method for Determining the Weights of Features

Most previous systems can be considered to determine the weights of features according to human intuition. On the other hand, this paper proposes a method for determining the weights of features by multiple-regression analysis of correct examples, which are abstracts created by testers. A tester selects important sentences that should be included in an abstract. The importance value of a sentence is defined as the number of supporters (testers who selected it as an important one) divided by the total number of testers. Let this importance value be \( S \); we then get the following equation for each sentence:

\[
S = a + \sum_{i=1}^{n} W_i \cdot P_i
\]

where, \( a \) is a constant, \( P_i \) is the number of points assigned to the \( i \)-th feature which is normalized to be between 0 to 1, and \( W_i \) is the weight assigned to the \( i \)-th feature.

In this equation, \( W_i \) is the only variable. Therefore, the feature weight \( W_i \) is calculated by multiple-regression analysis.

5 Experiment

We conducted an experiment to check the validity of the proposed method.

The testers were divided into two groups, A and B, each consisting 10 people. Those in group A selected important sentences (about 1/3 of the article) in 5 editorials and 3 general articles from the Nikkei Newspaper. Those in group B selected important sentences (about 1/3 of the article) in 3 editorials and 3 general articles, which were different from those used for group A. One of the editorials and one of the general articles used for group B are shown in Figures 1 (a) and 2 (a), respectively. In each of these figures, the first number is a sentence number, the second number is the number of supporters in group B, and the last part is a rough English translation.

Table 1 shows two weight sets; weight set 1 was created by the author in such a way that sentences located near the beginning and end are regarded as important, sentence importance is not proportional to points for rhetorical relation, and the importance of insistence-type sentences is higher in editorials than in general articles. Weight set 2, on the other hand, was calculated from the results obtained from group A by the method described in the previous section. Weight set 2 for general articles implies that sentences near the beginning are more important than ones near the end, and insistence-type sentences are less important, and so on. On the other hand, weight set 2 for editorials implies that sentences both near the beginning and near the end are important, and that insistence-type sentences are important.

To check the validity of these weight sets, we compared the abstracts created by the system, using weight set 1 and 2, from the articles supplied to group B.
the abstracts created by group B. For the general article in Figure 1 (a), the three most important sentences (roughly 1/3 of the article) determined by using the weight sets 1 and 2 are listed in Figures 1 (b) and (c), respectively. In this case, the three most important sentences selected by group B were 0, 2, and 3. Likewise, for the editorial in Figure 2 (a), the eight most important sentences (roughly 1/3 of the article) determined by using weight sets 1 and 2 are listed in Figures 2 (b) and (c), respectively. In this case, the eight most important sentences selected by group B are 0, 2, 3, 12, 15, 20, 21, 22. Here, we introduce the following metric of estrangement to check which abstract is most similar to the result of group B:

\[
\text{Estrangement} = \sum \alpha_i \left( \text{the number of supporters of a sentence } s_i \right) - \sum \alpha_j \left( \text{the number of supporters of a sentence } s_j \right)
\]

where \( s_i \) is a sentence that is included in an abstract by group B but not in an abstract created by the system, and \( s_j \) is a sentence that is not included in an abstract by group B but is included in an abstract created by the system.

The estrangements of the articles in Figures 1 and 2 are as follows: From this result, the weight set 2 calculated by multiple-regression analysis is more similar to the human selection than the weight set 1 created according to the author's intuition. For the other general articles used with group B, the estrangement values of weight set 2 are also better than those of weight set 1. In the other editorials, the estrangement values are comparable. This implies that the weight set 1 is not such a bad estimate for editorials.

### 6 Discussion

So far, most systems for creating an abstract of a text has been selected important sentences by some heuristics on the basis of surface features. However, most of these heuristics were derived from human intuition, and the validity of them are uncertain if the target text is changed. As mentioned in the introduction, the strategy of an abstraction should be changed according to the given text. Therefore, it is needed to adjust these heuristics for the given text. This paper proposed a method for this adjustment; that is, a method for determining weights of surface features by multiple-regression analysis of abstracts created by human. By using this method, a system can have an ability to be applied to a variety of texts.

### 7 Conclusion

This paper has proposed a method for creating an abstract by using surface features and their weights to select important sentences, and a method for determining these feature weights by multiple-regression analysis of abstracts created by humans. By using the proposed method to calculate feature weight, this system can be applied to other types of texts, and gives results more similar to those of a human process than a set of weights based on human intuition.

This abstract creation system is currently used in an information navigation assistance system [8] as a tool for quickly viewing the contents of newspaper articles.

### References

[1] Edmundson, H. P., “New Methods in Automatic Extracting,” Journal of the Association for Computing Machinery, Vol. 16, No. 2, pp. 264-285, April 1969.

[2] Kita, S., “A System for Summarization of an Explanatory Text” (in Japanese), Report 63-6 of NIWG of Information Processing Society of Japan (IPSJ), 1987.

[3] Oto, K., Sumita, K., and Miike, S., “Abstract Generation based on Rhetorical Structure Extraction,” Proc. of CoLLing'94, Vol.1, pp. 344-348, 1994.

[4] Tanura, T., and Tanura, N., “A Summary Generation based on The Form of Text” (in Japanese), Report 92-1 of NIWG of IPSJ, 1992.

[5] Schaak, R. and Aibelson, A., “Script Plans Goals and Understanding,” Lawrence Erlbaum Associates, Hillside, New Jersey, 1977.

[6] Luhn, H. P., “The Automatic Creation of Literature Abstracts,” IBM Journal of Research and Development, Vol. 2, No. 2, pp. 159-165, 1958.

[7] Mann, W. C., “Rhetorical Structure Theory: Description and Construction of Text Structure,” In G. Kempen, editor, Natural Language Generation, pp. 279-300, Martinus Nijhoff Publishers, 1987.

[8] Morohashi, M., Takeda, K., and et. al., “Information Outlining - Filling the Gap between Visualization and Navigation in Digital Libraries,” Proc. of Int. Symposium on Digital Libraries 1995, pp. 151-158, 1995.

[9] Yamamoto, K., Masuyama, S., and Naito, S., “GIEEN: An Experimental System Generating Summary of Japanese Editorials by Combining Multiple Discourse Characteristics” (in Japanese), Report 99-3 of NIWG of IPSJ, 1994.
Watanabe, H., Tsujii, J., and Nagao, M., "A Method for Analyzing Text Structure by Using Surface Clues of Sentence" (in Japanese), Proc. of 32nd Convention of IPSJ, pp.1633-1634, 1986.

Title: 最新銘鋭MPU搭載パソコン、IBM、来夏発売——低価格・高速処理を実現。（IBM to release PC equipped with the latest MPU, featuring low cost and fast processing）

0 (10) 【ニュース10月30日松本裕】IBM市は来秋、最新銘鋭マイクログロプロセッサー（MPU、超小型演算処理装置）「パワーPC」を搭載したパソコンを来夏に発売することを明らかにした。（IBM Corporation announced on the 10th of this month that a personal computer equipped with the latest "PowerPC" microprocessor will be released next summer.)

1 (3) まずノート型を発売、続いてデスクトップ型二機種を投入する。（First, a notebook PC will go on sale; this will be followed by two types of desktop PC.)

2 (5) 低価格で処理速度の高いパワーPCは、IBM再生のカギを握るものとされているコンピュータの高機能話。（The "PowerPC" is used as a central part of a computer. It is cheap and has high processing power, and is said to be a key to IBM’s recovery.）

3 (9) 同社がパワーPC搭載パソコンの製品計画を明らかにしたことにより、他の日米欧のパソコンも対抗策を講じることになりそうだ。（Since IBM announced its plan to sell personal computers equipped PowerPCs, other PC makers in the world are likely to take countermeasures.）

4 (3) 発売する三機種はCD-ROM、マイクロ、ステレオオーディオ、音声認識機能を標準装備、マルチメディア機能も含め計画を進める計画である。（The above three types of PC will provide additional multimedia features by including CD-ROM, microphone, stereo audio, and voice recognition functions as standard features.）

5 (0) OS（基本ソフト）はIBMの「OS/2」のほか、マイクロソフトのウィンドウズNT、サン・マイクロシステムズの「ソラリス」などに対応できるようにする。（IBM’s OS/2, Microsoft’s Windows NT, and Sun Microsystems’ Solaris will be installed as operating systems.）

6 (0) PowerPCはIBM、アップルコンピュータ、モトローラの三社が共同開発のRISC型MPUで、IBM、Apple, and Motorola.）

7 (2) パソコン用ビデオ市場で実現への標準機種になっているインテル製MPUに対抗するための新製品で、低価格・高速度が特徴だ。（It is intended to compete with Intel CPUs, which are de-facto standards in the PC microprocessor market. Its main advantages are low price and fast processing.）

8 (1) パソコン業界二位のアッパルがパワープC搭載パソコンを来夏発売する計画を発表。（The second largest PC maker, IBM, has already released a PowerPC-based workstation, but has not announced any corresponding plan for PCs.）

9 (0) 同じく、IBMはパワーPC搭載パソコンのワークステーションをすでに発売しているが、パソコンにおける製品計画を明らかにしていないかった。（The largest PC maker, IBM, has already released a PowerPC-based workstation, but has not announced any corresponding plan for PCs.）

10 (1) IBMはパワーPCを外販するだけでなく、搭載パソコンの技術仕様を外部に有料で公開、パワーPC搭載パソコンのフィーチャ作りを進める計画。（IBM plans to sell PowerPCs to other vendors, license the technology, and create a family of PowerPC-based PCs.）

11 (3) アップルなどと合わせたパワーPc搭載パソコン全体で、世界的市場に占めるシェアを最低20%程度まで押上げていく計画。（Together with Apple and others, IBM aims to gain at least a 20% share in the PC market for PowerPC-based PCs.）

(a) Original Article

2 (5) 低価格で処理速度の高いパワーPCは、IBM再生のカギを握るものとされているコンピュータの高機能話。（courtesy of the computer industry.）

(b) Abstract by Weight Set 1

0 (10) 【ニュース10月30日松本裕】IBM市は来秋、最新銘鋭マイクログロプロセッサー（MPU、超小型演算処理装置）「パワーPC」を搭載したパソコンを来夏に発売することを明らかにした。（IBM Corporation announced on the 10th of this month that a personal computer equipped with the latest "PowerPC" microprocessor will be released next summer.)

2 (5) 低価格で処理速度の高いパワーPCは、IBM再生のカギを握るものとされているコンピュータの高機能話。（The "PowerPC" is used as a central part of a computer. It is cheap and has high processing power, and is said to be a key to IBM’s recovery.）

(c) Abstract by Weight Set 2

3 (7) 七月から実施する庇護措置は、『迫害のない国』（ルーマニア、アルメニア、ハンガリーなど）からの亡命者は例外を除き受け入れず、庇護を求める国で定められている庇護の機会を拒絶することになる。（The logic behind this is that there cannot be political refugees from countries that have been converted into democratic nations by the East European Revolution, and so on.）

4 (1) 東欧革命などで民主主義体制に転換した国々から政治庇護が求める国は出るはずがないという論理である。（It is also said that this clause was created out of strong consideration for socialist states.）

6 (0) これは、ナチ時代の排外主義や"民族圧力に赤潮を加える."（Clause 2, Article 16 of the Basic Law established after WWI in 1949, was generous to refugees stating that people persecuted politically had a right to be protected.）

7 (6) 社会主義国家の存在を強く意識して、この議論的な用意をつくったという。"（It is also said that this clause was created out of strong consideration for socialist states.）
number arriving in Germany, which has loose restrictions, was 440,000, over 60% of the total for Europe.

10 (9) The number of refugees arriving in Germany, which has loose restrictions, was 440,000, over 60% of the total for Europe.

11 (9) Applicants from the former Eastern bloc countries, such as Romania and the Baltic states, are being detained in proportion to their numbers.

12 (0) of these, 410,000 were interviewed but only 700 were accepted.

13 (5) This shows that there is a constant flow of entrants.

14 (0) Germany is now said to be suffering the worst recession since World War II.

15 (2) The unemployment rate in April was 7.3% in former West Germany, and 14.7% in former East Germany.

16 (4) Refugees are allowed to stay in government-provided accommodation and live there until their interviews are completed, with their expenses borne by states and cities.

17 (2) From April, most regional governments have appealed for the numbers of refugees to be restricted.

18 (0) Reasons for this include anxiety that masses of immigrants will take native people's jobs.

19 (5) Refugees enter government-provided accommodation and live there until their interviews are completed, with their expenses borne by states and cities.

20 (0) It is not appropriate to apply these examples to Japan, because Japan's circumstances are different from those in Europe.

(a) Original Article

(0) The number of refugees arriving in Germany, which has loose restrictions, was 440,000, over 60% of the total for Europe.

(1) The number of refugees arriving in Germany, which has loose restrictions, was 440,000, over 60% of the total for Europe.

(2) Germany is now said to be suffering the worst recession since World War II.

(3) The unemployment rate in April was 7.3% in former West Germany, and 14.7% in former East Germany.

(4) Refugees are allowed to stay in government-provided accommodation and live there until their interviews are completed, with their expenses borne by states and cities.

(5) Reasons for this include anxiety that masses of immigrants will take native people's jobs.

(6) It is not appropriate to apply these examples to Japan, because Japan's circumstances are different from those in Europe.