An Automatic Extraction of Key Paragraphs Based on Context Dependency

Fumiyo Fukumoto **Yoshimi Suzuki†** **Jun'ichi Fukumoto‡**
Dept. of Electrical Engineering and Computer Science, Yamanashi University
4-3-11 Takeda, Kofu 400 Japan
{fukumoto@skye, ysuzuki@suwa†}.esi.yamanashi.ac.jp
Kansai Lab., R & D Group
Oki Electric Industry Co. Ltd.,‡
1-2-27 Shiromi, Chuo-ku, Osaka 540 Japan
fukumoto@kansai.oki.co.jp

Abstract

In this paper, we propose a method for extracting key paragraphs in articles based on the degree of context dependency. Like Luhn's technique, our method assumes that the words related to theme in an article appear throughout paragraphs. Our extraction technique of keywords is based on the degree of context dependency that how strongly a word is related to a given context. The results of experiments demonstrate the applicability of our proposed method.

1 Introduction

With increasing numbers of machine readable documents becoming available, automatic document summarisation has become one of the major research topics in IR and NLP studies.

In the field of an automatic summarisation, there are at least two approaches. One is knowledge-based approach with particular subject fields (Reimer, 1988), (Jacobs, 1990). This approach, based on deep knowledge of particular subject fields, is useful for restricted tasks, such as, for example, the construction of 'weather forecasts' summaries. However, when unrestricted subject matter must be treated, as is often the case in practice, the passage retrieval and text summarisation methods proposed have not proven equal to the need, since deep knowledge of particular subject fields is required (Paice, 1990), (Zechner, 1996).

The other, alternative strategy is the approach that relies mainly on corpus statistics (Paice, 1990), (Paice, 1993). The main task of this approach is the sentence scoring process. Typically, weights are assigned to the individual words in a text, and the complete sentence scores are then based on the occurrence characteristics of highly-weighted terms (keywords) in the respective sentences.

Term weighting technique has been widely investigated in information retrieval and lots of techniques such as location heuristics (Baxendale, 1958), rhetorical relations (Miike, 1994), and title information (Edmundson, 1969) have been proposed. These techniques seem to be less dependent on the domain. However, Salton claims that it is difficult to produce high accuracy of retrieval by using these term-weighting approaches (Salton, 1993).

The other term weighting technique is based on keyword frequency (Luhn, 1958). Keyword frequency is further less dependent on the domain than other weighting methods and therefore, well studied. Major approaches which are based on keyword frequency assume on the fact that the keywords of the article appear frequently in the article, but appear seldom in other articles (Luhn, 1958), (Nagao, 1976), (Salton, 1993), (Zechner, 1996). These approaches seem to show the effect in entirely different articles, such as 'weather forecasts', 'medical reports', and 'computer manuals'. Because each different article is characterised by a larger number of words which appear frequently in one article, but appear seldom in other articles. However, in some articles from the same domain such as 'weather forecasts', one encounters quite a number of words which appear frequently over articles. Therefore, how to extract keyword from these words is a serious problem in such the restricted subject domain.

In this paper, we propose a method for extracting key paragraphs in articles based on the degree of context dependency and show how the idea of context dependency can be used effectively to extract key paragraphs than other related work.

The basic idea of our approach is that whether a word is a key in an article or not depends on the domain to which the article belongs. Let 'stake' be a keyword and 'today' not be a keyword in the article. If the article belongs to a restricted subject domain, such as 'Stock market', there are other articles which are related to the article. Therefore, the frequency of 'stake' and 'today' in other articles are similar with each other. Let us consider further a broad coverage domain such as newspaper articles; i.e. the article containing the words 'stake' and 'today' belongs to a newspaper which consists of different subject domains such as 'Stock market' news, 'International'

news, 'Weather forecasts' news. 'Today' should appear frequently with every article even in such a domain; i.e. newspaper articles, while 'stake' should not. Our technique for extraction of keywords explicitly exploits this feature of context dependency of word: how strongly a word is related to a given context.

In the following sections, we first explain context dependency using newspaper articles, then we present our term weighting method and a method for extracting key paragraphs. Finally, we report some experiments to show the effect of our method.

2 Context Dependency

Like Luhn's assumption about keywords, our method is based on the fact that a writer normally repeats certain words (keywords) as he advances or varies his arguments and as he elaborates on an aspect of a subject (Luhn, 1958). In this paper, we focus on newspaper articles. Figure 1 shows the structure of *Wall Street Journal* corpus.

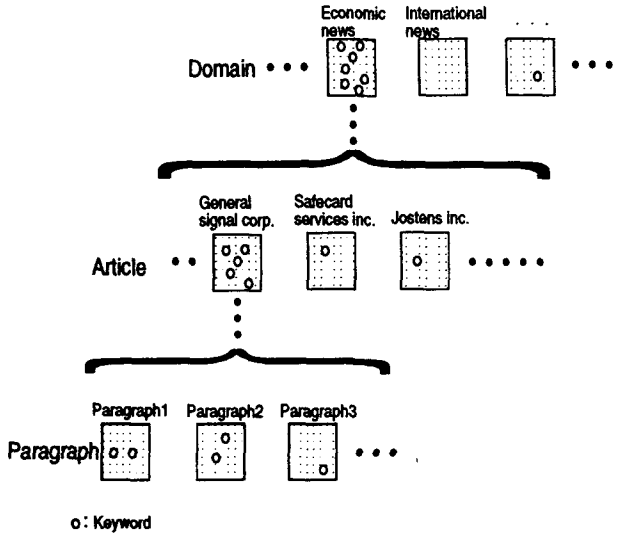

Figure 1: The structure of newspaper articles

In Figure 1, one day's newspaper articles consist of several different topics such as 'Economic news', 'International news', etc. We call this Domain, and each element ('Economic news', or 'International news') a *context*. A particular domain, for example, 'Economic news', consists of several articles each of which has different title name. In Figure 1, 'General signal corp.', 'Safecard services inc.', and 'Jostens inc.' show title names. We call this Article, and each element ('General signal corp.' etc) *context*. Furthermore, a particular article, for example, 'General signal corp.' consists of several paragraphs and keywords of the 'General signal corp.' article appear throughout paragraphs. We call each paragraph *context* in the Paragraph.

We introduce a degree of context dependency into the structure of newspaper articles shown in Figure 1 in order to extract keywords. A degree of context dependency is a measure showing how strongly each word related to a given *context*, a particular *context* of Paragraph, Article, or Domain. In Figure 1, let '○' be a keyword in the article 'General signal corp.'. According to Luhn's assumption, '○' frequently appears throughout paragraphs. Therefore, the deviation value of '○' in the Paragraph is small. On the other hand, the deviation value of '○' in the Article is larger than that of the Paragraph, since in Article, '○' appears in a particular element of the Article, 'General signal corp.'. Furthermore, the deviation value of '○' in the Domain is larger than those of the Article and Paragraph, since in the Domain, '○' appears frequently in a particular *context*, 'Economic news'. We extracted keywords using this feature of the degree of context dependency. In Figure 1, if a word is a keyword in a given article, it satisfies the following two conditions:

1. The deviation value of a word in the Paragraph is smaller than that of the Article.
2. The deviation value of a word in the Article is smaller than that of the Domain.

3 Term Weighting

Every sense of words in articles for extracting key paragraphs is automatically disambiguated in advance. This is because to disambiguate word-senses in articles might affect the accuracy of context dependent (domain specific) key paragraphs retrieval, since the meaning of a word characterises the domain in which it is used. Word-sense disambiguation (WSD in short) is a serious problem for NLP, and a variety of approaches have been proposed for solving it (Brown, 1991), (Yarowsky, 1992). Our disambiguation method is based on Niwa's method which uses the similarity between a sentence containing a polysemous noun and a sentence of dictionary-definition (Niwa, 1994). Furthermore, we linked nouns which are disambiguated with their semantically similar nouns mainly in order to cope with the problem of a phrasal lexicon. A phrasal lexicon such as *Atlantic Seaboard, New England* gives a negative influence for keywords retrieval, since it can not be regarded as units, i.e. each word which is the element of a phrasal lexicon is assigned to each semantic code (Fukumoto, 1996).

To the results of WSD and linking methods, we then applied a term weighting method to extract keywords. There have been several term weighting based on word frequencies, such as TF(Term Frequency), IDF(Inverse Document Frequency), TF*IDF, WIDF(Weighted Inverse Document Frequency) (Luhn, 1957), (Sparck, 1973), (Salton, 1983), (Tokunaga, 1994). We used Watan-

abe's $\chi 2$ method for term weighting which is shown in formula (1) (Watanabe, 1996).

$$\chi^2_{ij} = \begin{cases} \frac{(x_{ij}-m_{ij})^2}{m_{ij}} & \text{if } x_{ij} > m_{ij} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Formula (1) shows the value of χ^2 of the word i in the domain j. x_{ij} in (1) is the frequency of word i in the domain j. m_{ij} in (1) is shown in formula (2).

$$m_{ij} = \frac{\sum_{j=1}^{l} x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{l} x_{ij}} * \sum_{i=1}^{k} x_{ij} \quad (2)$$

In formula (2), k is the number of different words and l is the number of the domains. A larger value of χ^2_{ij} means that the word i appears more frequently in the domain j than in the other.

4 An Extraction of Keywords

The first step to extract keywords is to calculate χ^2 for each word in the Paragraph, the Article, and the Domain. We used formula (1) to calculate the value of χP^2_{ij}, χA^2_{ij}, and χD^2_{ij}, where χP^2_{ij}, χA^2_{ij}, and χD^2_{ij} indicate which word i appears most frequently in the context j of Paragraph, Article, and Domain, respectively. For example, χP^2_{ij} is shown in formula (3) by using formula (1).

$$\chi P^2_{ij} = \frac{(x_{ij}-m_{ij})^2}{m_{ij}} \quad (3)$$

In formula (3), x_{ij} is the frequency of word i in the context j of Paragraph. m_{ij} in formula (3) is shown in (2) where k is the number of different words and l is the number of contexts in Paragraph.

The second step is to calculate the degree of word i in Paragraph (χP^2_i), Article (χA^2_i), and Domain (χD^2_i). We defined the degree of word i in Paragraph, Article, and Domain as the deviation value of k contexts in Paragraph, Article, and Domain, respectively. Here, k is the number of contexts in Paragraph, Article, and Domain, respectively. For example, the deviation value of the word i in Paragraph is defined as follows:

$$\chi P^2_i = \sqrt{\frac{\sum_{j=1}^{k}(P^2_{ij}-m_i)}{k}} \quad (4)$$

In formula (4), k is the number of contexts in Paragraph, and m_i is the mean value of the total frequency of word i in Paragraph which consists of k contexts.

The last step to extract keywords is to calculate the context dependency of word i using formula (4). We recall that if i satisfies both 1 and 2 in section 2, the word i is regarded as a keyword.

$$\frac{\chi P^2_i}{\chi A^2_i} < 1 \quad (5)$$

$$\frac{\chi A^2_i}{\chi D^2_i} < 1 \quad (6)$$

Formulae (5) and (6) shows 1, and 2 in section 2, respectively. In formulae (5) and (6), χP^2_i, χA^2_i, and χD^2_i are the deviation value of a set of Paragraph, Article, and Domain, respectively.

5 An Extraction of Key Paragraphs

The procedure for extracting key paragraphs has the following three stages:

Stage One: Representing every paragraph as a vector

The goal of this stage is to represent every paragraph in an article as a vector. Using a term weighting method, every paragraph in an article would be represented by vector of the form

$$P_i = (N_{i1}, N_{i2}, \cdots, N_{in}) \quad (7)$$

where n is the number of nouns in an article and N_{ij} is as follows;

$$N_{ij} = \begin{cases} 0 & N_j \text{ does not appear in } P_i \\ f(N_j) & N_j \text{ is a keyword and appears in } P_i \\ 0 & N_j \text{ is not a keyword and appears in } P_i \end{cases}$$

where $f(N_j)$ is a frequency with which the noun N_j appears in paragraph P_i.

Stage Two: Clustering method

Given a vector representation of paragraphs P_1, $\cdots$, P_m as in formula (7), a similarity between two paragraphs P_i, P_j in an article would be obtained by using formula (8). The similarity of P_i and P_j is measured by the inner product of their normalised vectors and is defined as follows:

$$Sim(P_i, P_j) = \frac{V(P_i) * V(P_j)}{|\, V(P_i) \,||\, V(P_j) \,|} \quad (8)$$

The greater the value of $Sim(P_i, P_j)$ is, the more similar these two paragraphs are. For a set of paragraphs P_1, $\cdots$, P_m of an article, we calculate the semantic similarity value of all possible pairs of paragraphs. The clustering algorithm is applied to the sets and produces a set of semantic clusters, which are ordered in the descending order of their semantic similarity values. We adopted non-overlapping, group average method in our clustering technique (Jardine, 1968).

Stage Three: Extraction of key paragraphs

The sample results of clustering is shown in Table 1.

Table 1: The sample results of clustering

Num	Cluster
1	(3,4)
2	(1,(3,4))
3	((1,(3,4)),2)

'Num' in Table 1 shows the order of clusters which we have obtained and the number shown under 'Cluster' shows the paragraph numbers. In Table 1, if the number of keywords which belonging to the third paragraph is larger than that of the fourth, the order of key paragraphs is $3 \longrightarrow 4 \longrightarrow 1 \longrightarrow 2$, otherwise, $4 \longrightarrow 3 \longrightarrow 1 \longrightarrow 2$.

6 Experiments

We have conducted three experiments to examine the effect of our method. The first experiment, **Keywords Experiment**, is concerned with the keywords extracting technique and with verifying the effect of our method which introduces context dependency. The second experiment, **Key Paragraphs Experiment**, shows how the extracted keywords can be used to extract key paragraphs. In the third experiment, **Comparison to Other Related Work**, we applied Zechner's key sentences method (Zechner, 1996) to key paragraphs extraction (we call this **method_A**), and compared it with our method.

6.1 Data

The corpus we have used is the 1988, 1989 *Wall Street Journal* (Liberman, 1991) in ACL/DCI CD-ROM which consists of about 280,000 part-of-speech tagged sentences (Brill, 1992). *Wall Street Journal* consists of many articles, and each article has a title name. These titles are classified into 76 different domains. We selected 10 different domains and used them as **Domain**. As a test data, we selected 50 articles each of which belongs to one of these 10 domains. The selected domain names and the number of articles are shown in Table 2.

Table 2: The selected data

Domain	No	Domain	No
BBK: buybacks	6	BVG: beverages	8
DIV: dividends	5	FOD: food products	5
STK: stock market	5	RET: retailing	1
ARO: aerospace	5	ENV: environment	3
PCS: stones, gold	9	CMD: farm products	3

There are 3,802 different nouns in 50 articles. As a result of **WSD** and linking methods for these articles, we have obtained 3,707 different nouns.

6.2 Keywords Experiment

Formulae (5) and (6) are applied to 50 articles which are the results of **WSD** and linking methods, and as a result, we have obtained 1,047 keywords in all. The result of keyword extraction is shown in Table 3.

Table 3: The results of keyword experiment

Paragraph	*Recall/Precision*
3(1)	88.9/81.2
4(13)	62.7/86.2
5(6)	76.7/86.2
6(6)	67.3/77.5
7(4)	83.2/86.4
8(3)	89.0/80.0
9(4)	80.3/75.4
10(2)	90.2/72.2
11(1)	80.1/87.6
12(1)	100.0/83.7
14(3)	46.5/50.2
15(2)	100.0/73.4
16(2)	89.2/82.0
17(1)	62.4/89.4
22(1)	64.3/70.0
Total(50)	78.7/78.1

In Table 3, x in '$x(y)$' of 'Paragraph' shows the number of paragraphs in an article, 'y' shows the number of articles. For example, 3(1) shows that there is one article which consists of three paragraphs. *Recall* and *Precision* in Table 3 are as follows;

$$\textit{Recall} = \frac{\text{Number of correct keywords}}{\text{Number of keywords which are selected by human}}$$

$$\textit{Precision} = \frac{\text{Number of correct keywords}}{\text{Number of keywords which are selected in our method}}$$

Recall and *Precision* in Table 3 show the means in each paragraph. The denominator of *Recall* is made by three human judges; i.e. when more than one human judged the word as a keyword, the word is regarded as a keyword.

6.3 Key Paragraphs Experiment

For each article, we extracted 10 $\sim$50 % of its paragraphs as key paragraphs. The results of key paragraphs experiment are shown in Table 4.
In Table 4, 10 $\sim$ 50 % indicates the extraction ratio used. 'Para.' shows the number of paragraphs which humans judged to be key paragraphs, and 'Correct' shows the number of these paragraphs which the method obtained correctly. Evaluation is performed by three human judges. When more than one human judges a paragraph as a key paragraph, the paragraph is regarded as a key paragraph. '$*$' in Table 4 shows that the number of the correct data is smaller than that of an extraction ratio. For example, in Table 4, the number of paragraphs of 20 % out of 22 is 4. However, the number of paragraphs that more than one human judged the paragraph

Table 4: The results of Key Paragraphs Experiment

Paragraph (Article)	Percentage(%)										Correct
	10		20		30		40		50		
	Para.	Correct	Para.	Correct	Para.	Correct	Para.	Correct	Para.	Correct	%
3(1)	1	1	1	1	1	1	1	1	2	2	100.0
4(13)	13	12	13	12	13	12	13	12	26	21	88.4
5(6)	6	5	6	5	*11	8	*10	9	18	14	96.0
6(6)	6	6	6	6	*9	9	12	10	18	14	88.2
7(4)	4	4	4	4	8	8	12	8	16	11	79.5
8(3)	3	3	6	6	6	6	*8	6	12	7	80.0
9(4)	4	4	8	8	*8	8	16	11	*18	9	74.0
10(2)	2	2	4	2	*4	2	8	6	10	7	67.8
11(1)	1	1	2	2	3	3	4	3	6	4	81.2
12(1)	1	1	2	2	*2	2	*3	3	6	3	78.5
14(3)	3	2	4	3	*6	4	*14	7	*19	10	56.5
15(2)	*3	*2	*3	2	*3	2	*8	6	*14	10	70.9
16(2)	*3	*3	*5	5	5	5	12	8	*16	10	75.6
17(1)	2	2	3	3	*3	3	*7	4	*8	4	69.5
22(1)	2	2	*2	2	*2	2	*4	2	*8	4	66.6
Total(50)	54	50	69	63	84	75	132	96	215	130	
%	92.5		91.3		89.2		72.7		60.4		

as a key paragraph was only two. Therefore, 2 is marked with a '*'.

6.4 Comparison to Other Related Work

Zechner proposed a method to extract key sentences in an article by using simple statistical method; i.e. TF*IDF term weighting method. In order to show the applicability of our method, we applied Zechner's key sentences method to key paragraphs extraction and compared it with our method. In Zechner's method, the sum over all TF*IDF values of the content words for each sentence are calculated, and the sentences are sorted according to their weights. Finally a particular number of sentences are extracted as key sentences. The data we used consists of 1.92 sentences par a paragraph and was not so many sentences within a paragraph. Then, in order to apply his method to key paragraphs extraction, we calculated the sum over all sentences for each paragraph, and sorted the paragraphs according to their weights. From these, we extracted a certain number of paragraphs (method_A). In our method, every sense of words in articles for extracting key paragraphs is disambiguated in advance and linking method is performed. In order to examine where the performance comes from, we also compared our method to the method which WSD and linking method are not applied. The result is shown in Table 5.

In Table 5, '%' shows the extraction ratio, $10 \sim 50\%$ and 'Para.' shows the number of paragraphs corresponding to each 'Percentage'. 'Our method', 'not WSD', and 'method_A' shows the results using our method, the method which WSD and linking are not applied, and method_A, respectively.

Table 5: The results of comparative experiment

%	Para.	Our method(%)	not WSD	method_A
10	54	50(92.5)	43(79.6)	31(57.4)
20	69	63(91.3)	55(79.7)	35(50.7)
30	84	75(89.3)	66(78.5)	41(48.8)
40	132	96(72.7)	80(60.6)	63(47.7)
50	215	130(60.4)	112(52.8)	99(46.0)
Total	554	414(74.7)	356(64.2)	269(48.6)

7 Discussion

7.1 Keywords Experiment

Effectiveness of the Method

According to Table 3, *Recall* and *Precision* values range from 46.5/50.2 to 100.0/89.4, the mean being 78.7/78.6. This shows that our method is effective even in a restricted domain such as financial articles, e.g. *Wall Street Journal*, although the test set was small (50 articles). Furthermore, the correct ratio does not depend on the number of paragraphs in an article. This shows that our context dependency model is applicable for different size of the samples.

Problem of the Method

According to Table 3, the worst results of *Recall* and *Precision* was (46.5/50.2) when the number of

paragraphs was 14. As a result, the result of the extraction of key paragraphs shown in Table 4 was also worst (56.5%). The possible causes of the error were summarised the following two points:

(1) The formulae of context dependency

The sample results of keywords of the article, 'Abermin sues Geanges in Effort to rescind Joint Gold Venture' is shown in Table 6.

Table 6: Keywords and their χ^2 values in the article

Keyword	Paragraph	Article	Domain
Abermin	0.582	10.835	663.605
Belzberg	1.468	1.548	94.801
flin	1.468	1.548	94.801
gold5	1.770	2.496	52.865
Granges	0.680	15.478	948.007
Manitoba	1.468	1.548	94.801
mill1	1.706	4.925	94.801
ounces	1.765	5.064	284.402
reserves	2.912	3.060	94.801
suit2	1.099	3.096	189.601
supreme1	1.468	1.548	94.801
tartan1	0.251	6.191	379.203
word237	4.633	5.132	362.887
word238	1.468	1.548	94.801
others 15	...	...	...
Total average	1.772	2.383	78.161

In Table 6, each value of 'Paragraph', 'Article', and 'Domain', shows each χ^2 value. 'Total average' shows the mean of all keywords. 'word237' and 'word238' are representative words which are the result of linking noun with their semantically similar nouns. According to Table 6, we can observe that in 'Paragraph', for example, some words whose χ^2 values are slightly higher than the average (1,772) exist. For example, the χ^2 value of 'word237' is 4.633 and slightly higher than 1.772. However, 'word237' satisfies the formulae of context dependency. As a result, 'word237' is regarded as a keyword, while this is not. When the extracted ratio was 10%, there were four articles whose correct ratio did not attained 100%. Of these, three articles are classified into this type of the error.

From the above observation, we can estimate that the formulae of context dependency are weak constraints in some domains, while they are still effective even in a restricted domain. In order to get more accuracy, some other constraints such as location heuristics (Baxendale, 1958) or upper-case word feature (Kupiec, 1995) might be necessary to be introduced into our framework.

(2) The error of WSD

When the extracted ratio was 10%, there was one article out of four articles which could not be extracted correctly because of the error of WSD. The test article and the results of it was shown in Figure 2.

In Figure 2, the headline shows the title name. The numbers show the paragraph number, and the underlined words are keywords which are extracted in our method. The bottom shows the result of key paragraphs extraction. According to Figure 2, when the extraction ratio was 50%, the paragraphs 3 and 4 were extracted and the paragraph 1 was not extracted, although it is a key paragraph. The keywords and their frequencies of appearance in paragraph 1, 3, and 4 are shown in Table 7.

Table 7: The words and their frequencies

Para. 1		Para. 3		Para. 4	
Fr.	Word	Fr.	Word	Fr.	Word
1	crystal4	1	concern2	1	american2
1	oil4	1	crystal2	1	crystal2
5	word237	1	energy4	1	oil3
1	word78	1	oil3	5	word237
		1	rate5	1	word78
		5	word237		

word78: Nov., yesterday2
word237: exchange1, offer4, notes, shares, stock5, amount4, trading1, stock1, cents

According to Table 7, 'crystal' and 'oil in paragraph 1 are disambiguated incorrectly and were replaced by 'crystal4' and 'oil4', respectively, while 'crystal' should have been replaced by 'crystal2' and 'oil' with 'oil3'. Therefore, the number of words which appear in both paragraph 3 and 4 was larger than any other pair of paragraphs. As a result, paragraph 3 and 4 are the most semantically similar paragraphs and 1 was not extracted as a key paragraph.

In our method, the correct ratio of key paragraphs extraction strongly depends on the results of WSD. The correct ratio of our WSD was 78.4% (Fukumoto, 1996). In order to get higher accuracy, it is necessary to improve our WSD method.

7.2 Key Paragraphs Experiment

Effectiveness of the Method

In Key Paragraphs Experiment, the overall results were positive, especially when the ratio of extraction was 10~30%. The ratios of correct judgements in these cases were significantly high; i.e. 92.5%, 91.3%, and 89.2%, respectively. This demonstrates the applicability of the degree of context dependency.

Limitations of the Method

When the ratio of extraction was higher than 30%, the results was 72.7% and 60.4%. Furthermore, the more paragraphs are in an article, the smaller the number of correct judgements. One possible cause of these results is that the clustering method might have a negative effect on extracting key paragraphs.

Crystal Oil Co. Extends Offer

1 Crystal4 oil4 co. said it extended to Nov. 17 the exchange1 offer4 for all of its non-interest-bearing convertible secured notes, due 1997, for shares of its common stock5.
2 The offer4 had been set to expire yesterday1.
3 The company1 said about 65.89% of the notes outstanding have been tendered. under the plan5, the notes will be exchanged at a rate5 of 65 shares of crystal2 oil3 common for each $1,000 principal amount4 of the notes, the energy4 concern2 said.
4 In composite trading1 on the american2 stock1 exchange1 yesterday2, crystal2 oil3 shares closed at $2.875, up 12.5 cents.

The results of Key Paragraph Extraction: 3 $\longrightarrow$ 4 $\longrightarrow$ 1 $\longrightarrow$ 2

Figure 2: The sample of the article

In the field of text summarisation, a vector model was often used for extracting key sentence or key paragraph (Tokunaga, 1994), (Zechner, 1996). In this model, the sentences with term weighting are sorted according to their weights and this information is used to extract a certain ratio of highest weighted paragraph in an article. We implemented this model and compared it with our clustering technique. The results are shown in Table 8.

Table 8: Our method and a vector model

%	Para.	Our method(%)	Vector model(%)
10	54	50(92.5)	48(88.9)
20	69	63(91.3)	58(84.1)
30	84	75(89.3)	68(78.6)
40	132	96(72.7)	91(69.0)
50	215	130(60.4)	128(60.6)

In Table 8, '%' shows the extraction ratio, $10 \sim 50\%$ and 'Para.' shows the number of total paragraphs corresponding to each '%'. 'Our method', and 'Vector model' shows the results of our method, and using vector model, respectively.

Table 8 shows that the results using our method are highly than those of using the vector model. In our method, when the extraction ratio was more than 30%, the correct ratio decreased. This phenomena is also observed in the vector model. From the observation, we can estimate that the cause of the results was not our clustering technique. Examining the results of human judges, when the number of paragraphs was more than 14, the number of paragraphs marked with a '*' is large. This shows that it is too difficult even for a human to judge whether a paragraph is a key paragraph or not. From the observation, for these articles, there are limitations to our method based on context dependency.

Other Heuristics

As we discussed in Keywords Experiment, it might be considered that some heuristics such as location of paragraphs are introduced into our method to get a higher accuracy of keywords and key paragraphs extraction, even in these articles. Table 9 shows the location of key paragraphs extracted using our method and extracted by humans. The extraction ratio described in Table 9 is 30%.

Table 9: The location of key paragraphs

	Articles	
	Hum.	Method
(a)First	39	37
(b)First and Last	4	4
(c)First, Mid-position, and Last	1	1
(d)First and Mid-position	4	4
(e)Mid-position	0	1
(f)Otherwise	2	3
Total	50	50

In Table 9, each paragraph (First, Mid-position, and Last paragraph) includes the paragraphs around it. According to Table 9, in human judgement, 39 out of 50 articles' key paragraphs are located in the first parts, and the ratio attained 78.0%. This shows that using only location heuristics (the key paragraph tends to be located in the first parts) is a weak constraint in itself, since the results of our method showed that the correct ratio attained 89.2%. However, in our method, 2 articles are not extracted correctly, while the key paragraph is located in the first parts of these articles. From the observation, in a corpus such as *Wall Street Journal*, utilising a location heuristics is useful for extracting key paragraphs.

7.3 Comparison to Other Related Work

According to Table 5, the average ratio of our method and method_A was 74.7%, and 48.6%, respectively. This shows that method_A is not more effective than our method. This is because most of nouns do not contribute to showing the characteristic of each domain for given articles. In the test data which consists of 3,802 different nouns, 2,171 nouns appeared in only one article and the frequency of each of them is one. We recall that in method_A,

when word i appears in only one article and the frequency of i is one, the value of TF*IDF equals to log50. There are 2,955 out of 3,802 nouns whose TF*IDF value is less than log50, and the percentage attained at 77.7%. This causes the fact that most of nouns do not contribute to showing the characteristic of each domain for given articles.

Comparing the difference ratio of 'Our method' and 'not WSD' to that of 'not WSD' and method_A, the former was 10.5% and the latter was 15.6%. Therefore, our context dependency model contributes the extraction of key paragraphs, although WSD and linking are still effective.

8 Conclusion

We have reported an experimental study for extracting key paragraphs based on the degree of context dependency for a given article and showed how our context dependency model can use effectively to extract key paragraphs, each of which belongs to the restricted subject domain. In order to cope with the remaining problems mentioned in section 7 and apply this work to practical use, we will conduct further experiments.

9 Acknowledgments

The authors would like to thank the reviewers for their valuable comments.

References

P. B. Baxendale, "Man-made index for technical literature - an experiment", *IBM J. Res. Develop.*, 2(1958)4, pp. 354-361, 1958

E. Brill, "A simple rule-based part of speech tagger", In *Proc. of the 3rd conference on applied natural language processing*, pp. 152-155, 1992

P. F. Brown et al., "Word-Sense Disambiguation Using Statistical Methods", In *Proc. of the 29th Annual Meeting of the ACL*, pp. 264-270, 1991

H. P. Edmundson, "New methods in automatic abstracting", *Journal of ACM*, 16(1969)2, pp. 264-285, 1969

F. Fukumoto and Y. Suzuki, "An Automatic Clustering of Articles Using Dictionary Definitions", In *Proc. of the 16th COLING*, pp. 406-411, 1996

N. Jardine and R. Sibson, "The construction of hierarchic and non-hierarchic classifications", *Computer Journal*, pp. 177-184, 1968

P. S. Jacobs and L. F. Rau, "SCISOR: Extracting information from on-line news", *Communications of the ACM*, 33(1990)11, pp. 88-97, 1990

J. Kupiec et al., "A trainable document summarizer", In *Proc. of SIGIR'95*, pp. 68-73, 1995

M. Liberman, "CD-ROM I Association for Computational Linguistics Data Collection Initiative", *University of Pennsylvania*, 1991

H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information", *IBM journal*, 1(1957)4, pp. 307-319, 1957

H. P. Luhn, "The Automatic Creation of Literature Abstracts", *IBM journal*, 2(1958)1, pp. 159-165, 1958

S. Miike et al., "A full-text retrieval system with a dynamic abstract generation function", In *Proc. of SIGIR'94*, pp. 152-161, 1994

M. Nagao et al., "An Automatic Method of the Extraction of Important Words from Japanese Scientific Documents (in Japanese)", *IPS Japan*, 17(1976)2, pp. 110-117, 1976

Y. Niwa and Y. Nitta, "Co-occurrence vectors from corpora vs. distance vectors from dictionaries", In *Proc. of the 15th COLING*, pp. 304-309, 1994

C. D. Paice, "Constructing literature abstracts by computer:Techniques and prospects", *Information Processing and Management*, vol. 26, pp. 171-186, 1990

C. D. Paice and P. A. Jones, "The identification of important concepts in highly structured technical papers", In *Proc. of SIGIR'93*, pp. 69-78, 1993

U. Reimer and U. Hahn, "Text condensation as knowledge base abstraction", *IEEE Conference on AI Applications*, pp. 338-344, 1988

G. Salton and M. J. McGill, "Introduction to Modern Information Retrieval", *McGraw-Hill*, 1983

G. Salton et al., "Approaches to passage retrieval in full text information systems", In *Proc. of SIGIR'93*, pp. 49-58, 1993

K. J. Sparck, "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation*, 28(1973)1, pp. 11-21, 1973

T. Tokunaga and M. Iwayama, "Text Categorization based on Weighted Inverse Document Frequency", *SIG-IPS Japan*, 100(1994)5, pp. 33-40, 1994

Y. Watanabe et al., "Document Classification Using Domain Specific Kanji Characters Extracted by χ^2 Method", In *Proc. of the 16th COLING*, pp. 794-799, 1996

D. Yarowsky, "Word sense disambiguation using statistical models of Roget's categories trained on large corpora", In *Proc. of the 14th COLING*, pp. 454-460, 1992

K. Zechner, "Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences", In *Proc. of the 16th COLING*, pp. 986-989, 1996