Some Examinations of Intrinsic Methods for Summary Evaluation Based on the Text Summarization Challenge (TSC)

Hidetsugu Nanba* **and Manabu Okumura***

* Precision and Intelligence Lab.,
Tokyo Institute of Technology,
4259 Nagatsuta Midori-ku Yokohama, JAPAN, 226-8503
{nanba, oku}@pi.titech.ac.jp

Abstract

Computer-produced summaries have traditionally been evaluated by comparing them with human-produced summaries using the F-measure. However, the F-measure is not appropriate when alternative sentences are possible in a human-produced extract. In this paper, we examine some evaluation methods devised to overcome the problem, including utility-based evaluation. By giving scores for moderately important sentences that does not appear in the human-produced extract, utility-based evaluation can resolve the problem. However, the method requires much effort from humans to provide data for evaluation. In this paper, we first propose a pseudo-utility-based evaluation that uses human-produced extracts at different compression ratios. To evaluate the effectiveness of pseudo-utility-based evaluation, we compare our method and the F-measure using the data of the Text Summarization Challenge (TSC), and show that pseudo-utility-based evaluation can resolve this problem. Next, we focus on content-based evaluation. Instead of measuring the ratio of sentences that match exactly in the extract, the method evaluates extracts by comparing their content words to those of human-produced extracts. Although the method has been reported to be effective in resolving the problem, it has not been examined in the context of comparing two extracts produced from different systems. We evaluated computer-produced summaries by content-based evaluation, and compared the results with a subjective evaluation. We found that the evaluation by content-based measure matched those by subjective evaluation in 93% of the cases, if the gap in content-based scores between two summaries is more than 0.2.

1. Introduction

Recently, the issue of how to evaluate computer-produced summaries has been recognized as one of the problems that must be addressed in the field of automatic summarization. Evaluation methods for text summarization can be divided into two categories: intrinsic and extrinsic(Sparck-Jones and Galliers, 1996). The methods that evaluate system outputs (summaries) themselves are often called intrinsic methods. On the other hand, extrinsic methods evaluate the performance of a summarization system in a given task, such as information retrieval and text categorization. In this paper, we focus on intrinsic methods.

Computer-produced summaries have traditionally been evaluated by comparing them with human-produced summaries using the F-measure. Jing et al.(Jing, Barzilay, McKeown, and Elhadad, 1998), however, pointed out that the F-measure has a problem: it is not appropriate when alternative sentences are possible in a human-produced extract. For example, if a human subject extracts sentence 1 and a system extracts sentence 2, the system obtains a lower score, even when sentences 1 and 2 are interchangeable. In this paper, we examine some of the evaluation methods devised to overcome this problem.

Several such methods have been proposed, including the utility-based evaluation proposed by Radev et al.(Radev, Jing, and Budzikowska, 2000). Utility is the importance of each sentence, as scored manually on a one-to-ten scale. Utility-based evaluation measures the coverage of utility scores of the human-produced extract. By giving scores for moderately important sentences that does not appear in the human-produced extract, utility-based evaluation can resolve the problem. However, the method requires a great deal of effort for humans to assign scores manually.

Content-based evaluation(Donaway, Drummey, and Mather, 2000) is another method. Instead of measuring the ratio of sentences that match exactly in the extract, the method evaluates computer-produced extracts by comparing their content words to those of human-produced extracts. The score for content-based evaluation is obtained by computing the similarity between the term frequency (tf) vector of a computer-produced extract and the tf vector of a human-produced extract, using the cosine distance. Content-based evaluation does not require much effort for humans to make a data set for evaluation. Although the authors reported that content-based evaluation was effective for resolving the problem, the method has not been examined in the context of comparing two extracts produced from different systems.

In this paper, we first propose a pseudo-utility-based evaluation that overcomes the problems of utility-based evaluation. We can generally assume that sentences in an extract at high compression ratios are more important than those at low compression ratios. Based on this assumption, we can assign an importance to each sentence in a text, when there are human-produced extracts at different compression ratios. We can then use them for utility-based evaluation. We think that pseudo-utility-based evaluation is more practical than utility-based evaluation for making a data set for evaluation, because a number of data sets with extracts at different compression ratios have been made (e.g., (Jing, Barzilay, McKeown, and Elhadad, 1998)).

To evaluate the effectiveness of pseudo-utility-based evaluation, we compare it with the F-measure using the data of the Text Summarization Challenge (TSC)(Fukushima and Okumura, 2001:a; Fukushima and Okumura, 2001:b), a subtask of NTCIR workshop 2, and show that pseudo-utility-based evaluation can improve the F-measure.

Next, we focus on content-based evaluation. We compare content-based evaluation with subjective evaluation, and investigate the effectiveness of content-based evaluation for comparison of two computer-produced summaries.

In the following sections, we first briefly review some intrinsic methods that overcome some problems of the F-measure. In Section 3, we propose pseudo-utility-based evaluation. To reveal the effectiveness of pseudo-utility-based evaluation, we evaluated computer-produced extracts by our method and the F-measure. We report the results in Section 4. We also report an examination of content-based evaluation.

2. Related Work

Jing et al.(Jing, Barzilay, McKeown, and Elhadad, 1998) conducted some examinations on intrinsic and extrinsic methods to investigate the factors affecting evaluation results. From their results, they found that the F-measure has at least the following two problems.

- **Problem 1:**
  the F-measure is very sensitive to the compression ratio, i.e., the scores differ greatly according to the compression ratio.
- **Problem 2:**
  the F-measure is not appropriate when alternative sentences are possible in a human-produced extract; for example, if a human extracts sentence 1 and a system extracts sentence 2, the system obtains a lower score, even when sentences 1 and 2 are interchangeable.

Several methods to reduce the effect of problem 1 have been proposed. Mittal et al.(Mittal, Kantrowitz, Goldstein, and Carbonell, 1999) proposed that systems should be evaluated at a variety of compression ratios, and the results should be reported in a manner similar to the 11-point precision score that is used in information retrieval.

They also pointed out that differences in inherent properties of corpora affect the results. To be able to compare the performance of systems on different corpora, they suggested that a score of system performance should be normalized by a baseline score, which was defined to be the average performance of all possible extracts (randomly extracted sentences). Given a baseline score *b* and a score of system performance *p*, the adjusted score is calculated by the following equation.

$$p' = \frac{p-b}{1-b}$$

Here, the baseline score of the F-measure is equivalent to the compression ratio, and the F-score generally increases when the baseline score (i.e., compression ratio) increases (see Tables 3 and 4). Mittal's method, which adjusts the score of system performance by the baseline score, reduces the effect of problem 1.

Radev et al. improved Mittal's measure (Radev, Jing, and Budzikowska, 2000). In addition to Mittal's proposal, Radev et al. took account of inter-judge agreement *J*. When several human subjects are asked to make extracts from a text, the inter-judge agreement measures to what extent the sentences each judge extracts agree with each other. *J* is considered as an upper bound on the performance of a system. Given a baseline score *R* and a score of system performance *S*, a modified system performance *S'* is calculated using the following equation.

$$S' = \frac{S-R}{J-R}$$

Several methods for reducing the effect of problem 2 have also been proposed. Jing et al. (Jing, Barzilay, McKeown, and Elhadad, 1998) proposed an evaluation method that took account of moderately important sentences that do not appear in the human-produced extract. In this method, the agreement between a sentence in a human-produced extract and a sentence in a computer-produced extract is represented as the degree of the human subjects' agreement. For example, if three of five human subjects extract sentence 1, and two subjects extract sentence 2, a system that extracts sentence 1 will receive a score of 3/5, and a system that extracts sentence 2 will receive a score of 2/5, instead of one or zero, respectively[1].

Radev et al.(Radev, Jing, and Budzikowska, 2000) proposed a utility-based evaluation. Utility is the importance of each sentence according to a score assigned manually on a one-to-ten scale. A utility-based score is calculated by dividing the sum of utilities of the computer-produced extract by that of the human-produced extract. By giving scores for moderately important sentences from the human-produced extract, utility-based evaluation can resolve problem 2.

Donaway et al.(Donaway, Drummey, and Mather, 2000) proposed two evaluation methods. In one, both human subjects and a system are asked to rank the sentences of a text in order of their importance, and the computer-produced extract is then evaluated by comparing the ranks of both computer-produced and human-produced extracts. This method classifies all sentences in the original text in terms of their importance, instead of classifying them into two categories (important or unimportant).

Another method is content-based evaluation. Instead of measuring the ratio of sentences that match exactly in the extracts, the method evaluates extracts by comparing their content words with those of human-produced extracts. The score for content-based evaluation is obtained by computing the similarity between the term frequency (tf) vector of a computer-produced extract and the tf vector of a human-produced extract, using the cosine distance.

Donaway et al. conducted an examination for comparison of these two methods together with recall. They prepared several pairs of human-produced extracts, whose contents were highly similar, and use them for the evaluation of computer-produced extracts. The hypothesis of their examination is good evaluation method should yield similar scores by comparing a computer-produced extract with a pair of human-produced extracts, if they are highly similar. They calculated correlation coefficients of both scores for each evaluation method. As a result, they concluded that content-based evaluation was the best way to resolve problem 2.

In Document Understanding Conference 2001, computer-produced summaries were evaluated by comparing with human-produced summaries using the notion of model units (MUs) and peer units (PUs)(McKeown et

[1] If we consider sentences that more than half of the human subjects extract as correct, sentence 1 is correct, and 2 is incorrect.

al., 2001). First, the human-produced summaries were manually segmented into MUs, which are informational units that should express a self-contained fact in the ideal case. Second, computer-produced summaries were automatically segmented into PUs, which are always sentences. Third, the assessor located the PU(s) that covered the content of each MU, if any. Finally, the scores of precision were calculated for each computer-produced summary as the number of PUs matching some MU divided by the number of PUs in the peer summary. The third step in this procedure can resolve problem 2, because a PU is located by an assessor, if only the PU covers the content of the MU, though they were extracted from different parts in a text.

3. Pseudo-utility-based Evaluation

In this section, we propose a pseudo-utility-based evaluation that uses human-produced extracts at different compression ratios. When there are human-produced extracts for a text at ratios of r1%, r2%, and r3% (r1 < r2 < r3), we can classify each sentence in the text into the following four categories: (a) sentences contained in the r1% extract, (b) sentences that are not contained in the r1% extract but are contained in the r2% extract, (c) sentences that are not contained in the r2% extract but are contained in the r3% extract, and (d) other sentences[2]. If we regard these categories as a one-to-four scale, we can use them as utilities for pseudo-utility-based evaluation.

Now, we explain pseudo-utility-based evaluation using an example shown in Table 1. Table 1 shows a human-produced extract and two computer-produced extracts, all at 10%, 30%, and 50% ratios. Of the ten sentences (S1–S10) in the original text, the extracted sentences are marked as '+' in the table. Now, we define the weight of each sentence *w* as **1/(compression ratio)**.

Of the five sentences (S3, S4, S7, S9, and S10) extracted by system 1 at 50% ratio, three (S4, S7 and S10) are contained in the human-produced extracts. The F-score of system 1 at 50% ratio is 0.6 (3/5). Here, as the weights of the five sentences (S3, S4, S7, S9, S10) are 0, 1/30, 1/50, 0, 1/30, respectively, the total of the weights is 13/150 (0 + 1/30 + 1/50 + 0 + 1/30), while the total weight of the human-produced extract is 31/150 (1/10 +1/30 + 1/50 + 1/50 + 1/30). The pseudo-utility score is calculated by the total weight of a computer-produced extract divided by the total weight of a human-produced extract. The pseudo-utility score in this case is 0.419 ($\frac{13/150}{31/150}$).

The F-scores and pseudo-utility scores of systems 1 and 2 are shown in Table 2. In the table, as both systems extract S4 instead of S1, which was extracted by the human subject, the F-score is zero. Here, as S4 is contained in a human-produced extract at 30% ratio, this sentence is considered as a moderately important sentence. In this example, the only possible score of the F-measure at 10% is zero or one, while pseudo-utility-based evaluation makes it possible to evaluate extracts appropriately by taking account of moderately important sentences such as S4.

Next, we compare two computer-produced extracts at 50% ratio. The F-scores of both systems are 0.6, because three sentences of five in both computer-produced extracts are correct. Of the three correct sentences of both systems, S4 and S10 are common and the remaining ones are different. For their third sentences, system 1 extracts S7, and system 2 extracts S1. S1 is considered more important than S7, because S1 is contained in the human-produced extract at 10% ratio. As a result, the pseudo-utility scores of systems 1 and 2 are 0.419 and 0.806, respectively. This means that pseudo-utility-based evaluation can identify the difference between two computer-produced extracts.

4. Analysis of Evaluation Methods

To evaluate the effectiveness of pseudo-utility-based evaluation, we conducted some tests using the data of the TSC. We also discuss content-based evaluation, which was used as one of evaluation methods in the TSC.

In Section 4.1, we first explain the tasks and evaluation in the TSC. In Section 4.2, we report the analysis of both measures based on the data of the TSC.

4.1. Evaluation in the TSC

The Text Summarization Challenge (TSC) is a subtask of NTCIR Workshop 2, which was held so that researchers in the field could collect and share text data for summarization, and to make clear the issues of evaluation measures and methods for summarization of Japanese texts. Three tasks were conducted in the TSC, and we describe two of them, task A-1 and A-2, as their evaluation uses an intrinsic method (For further detail, please refer to (Fukushima and Okumura, 2001:a; Fukushima and Okumura, 2001:b)).

- **Task A-1 (Extraction of important sentences):**

  to extract important sentences at 10%, 30%, and 50% ratios. Extracts were evaluated by F-measure.

- **Task A-2 (Summaries to be compared with human-produced summaries):**

  to produce summaries in plain text at the ratios of 20% and 40%. Summaries were evaluated in two ways: content-based evaluation and subjective evaluation. In subjective evaluation, human judges were asked to evaluate and rank the computer-produced summaries in terms of coverage of important contents and readability. Judges were given four types of summaries: two human-produced summaries, a system result, and an extract produced by a lead-based method.

4.1.1. Texts

Thirty newspaper articles were extracted from the Mainichi newspaper database for 1994 and 1998. In terms of genre, editorials and articles on social issues were used. The editorials were grouped into two sets of about 1200 and 2400 characters in length, while the social issue articles were grouped into three sets with lengths about 600, 900 and 1200 or more characters.

4.1.2. Evaluation methods for each task

Task A-1

For task A-1, recall, precision, and F-measures were used, where

[2] Here, an extract at r_1% must be contained in an extract at r_2%, ($r_1 < r_2$), and an extract at r_2% must be contained in an extract at r_3% ($r_2 < r_3$).

Table 1: An example of pseudo-utility-based evaluation

	Human subject			*Importance*	*System 1*			*System 2*		
	10%	30%	50%	*w*	10%	30%	50%	10%	30%	50%
S1	+	+	+	1/10	-	-	-	-	+	+
S2	-	-	-	0	-	-	-	-	-	-
S3	-	-	-	0	-	-	+	-	-	-
S4	-	+	+	1/30	+	+	+	+	+	+
S5	-	-	-	0	-	-	-	-	-	-
S6	-	-	-	0	-	-	-	-	+	+
S7	-	-	+	1/50	-	-	+	-	-	-
S8	-	-	+	1/50	-	-	-	-	-	-
S9	-	-	-	0	-	+	+	-	-	+
S10	-	+	+	1/30	-	+	+	-	-	+

Table 2: An example of evaluation by F-measure and pseudo-utility-based measure

	System 1		*System 2*	
	F-measure	pseudo-utility evaluation	F-measure	pseudo-utility evaluation
10%	0.000 ($\frac{0}{1}$)	0.333 ($\frac{1/30}{1/10}$)	0.000 ($\frac{0}{1}$)	0.333 ($\frac{1/30}{1/10}$)
30%	0.667 ($\frac{2}{3}$)	0.400 ($\frac{2/30}{1/10+2/30}$)	0.667 ($\frac{2}{3}$)	0.800 ($\frac{1/10+1/30}{1/10+2/30}$)
50%	0.600 ($\frac{3}{5}$)	0.419 ($\frac{2/30+1/50}{1/10+2/30+2/50}$)	0.600 ($\frac{3}{5}$)	0.806 ($\frac{1/10+2/30}{1/10+2/30+2/50}$)

- Recall = the number of correct sentences marked by the system / the total number of correct sentences marked by the human subject

- Precision = the number of correct sentences marked by the system / the total number of sentences marked by the system

- F-measure = 2 × Recall × Precision / (Recall + Precision)

After calculating these scores for each article, we computed the average of them. We also evaluated the results of two baseline systems. One was based on the lead method (Lead), and the other was based on term frequency (tf).

Task A-2

(i) Subjective evaluation

The following four kinds of summaries as well as the original texts were prepared.

- Summaries produced by extracting important parts of the sentences in the text (PART)
- Freely summarized texts (FREE)
- Summaries produced by a system (SYS)
- Summaries produced by using the lead method (BASE)

First, the evaluator (one person) read the original text and its summaries (4 kinds). Then, the person evaluated and scored the summaries in terms of how readable they were, and how well the content of the text was described. The scores were 1, 2, 3, or 4, where 1 is the best and 4 is the worst, i.e., a lower score indicates a better evaluation.

(ii) Content-based evaluation

Like Donaway et al.(Donaway, Drummey, and Mather, 2000), we tried to find out how close the two summaries were by examining the content words. Morphological analysis was applied to the system and human summaries, and only content words (keitaiso) were selected for both of them. The distances between the word-frequency vector of system- and human-produced summaries were then computed. We used both PART and FREE as keys.

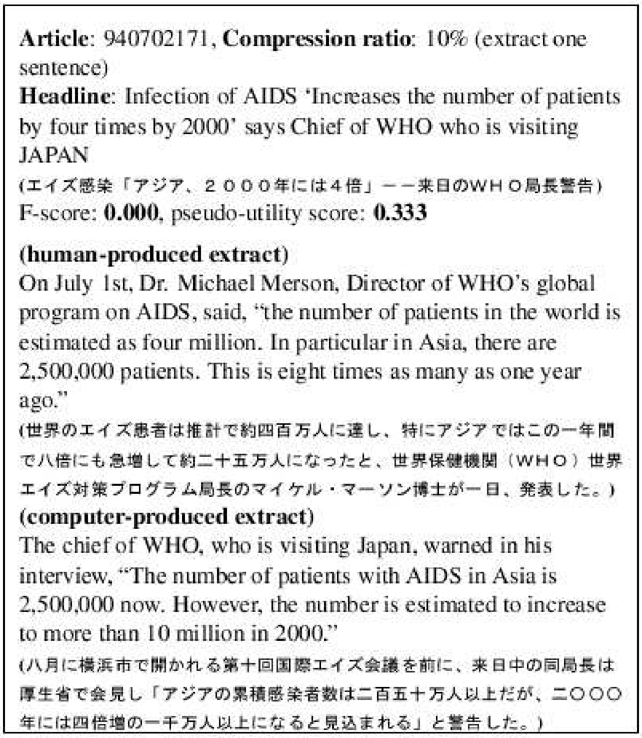
4.2. Analysis of Evaluation Methods

To show the effectiveness of pseudo-utility-based evaluation, two examinations are necessary. One is a comparison between pseudo-utility-based evaluation and the F-measure, and the other is a comparison between utility-based evaluation and pseudo-utility-based evaluation. However, as it requires a great deal of effort for humans to make a data set for utility-based evaluation, we did not conduct the latter examination.

In this section, we first compare pseudo-utility-based evaluation with the F-measure, and the results are reported in Section 4.2.1. We also conducted an examination on content-based evaluation, and the results are reported in Section 4.2.2.

4.2.1. Comparison of pseudo-utility-based evaluation and F-measure (task A-1)

First, we investigated the effectiveness of pseudo-utility-based evaluation by comparing human-produced extracts with computer-produced extracts. Figure 1 shows a typical example in which pseudo-utility was applied effectively. The theme of the original newspaper article is ‘the infection of AIDS in Asia’. Both a human-produced extract and a computer-produced extract at 10% ratio (the task was to extract only one sentence from article 940702171) are shown in the figure. As can be seen from the figure, the extracts (sentences) do not match each other, so the F-score is zero. Both sentences, however, describe almost the same topic, ‘the number of patients has been increasing recently in Asia’. Now, as the computer-produced extract is contained in the human-produced extracts at 30% ratio, the pseudo-utility score is 0.333 ($\frac{1/0.3}{1/0.1}$), which seems more ap-



Article: 940702171, **Compression ratio**: 10% (extract one sentence)
Headline: Infection of AIDS 'Increases the number of patients by four times by 2000' says Chief of WHO who is visiting JAPAN
(エイズ感染「アジア、２０００年には４倍」――来日のＷＨＯ局長警告)
F-score: **0.000**, pseudo-utility score: **0.333**

(human-produced extract)
On July 1st, Dr. Michael Merson, Director of WHO's global program on AIDS, said, "the number of patients in the world is estimated as four million. In particular in Asia, there are 2,500,000 patients. This is eight times as many as one year ago."
(世界のエイズ患者は推計で約四百万人に達し、特にアジアではこの一年間で八倍にも急増して約二十五万人になったと、世界保健機関（ＷＨＯ）世界エイズ対策プログラム局長のマイケル・マーソン博士が一日、発表した。)
(computer-produced extract)
The chief of WHO, who is visiting Japan, warned in his interview, "The number of patients with AIDS in Asia is 2,500,000 now. However, the number is estimated to increase to more than 10 million in 2000."
(八月に横浜市で開かれる第十回国際エイズ会議を前に、来日中の同局長は厚生省で会見し「アジアの累積感染者数は二百五十万人以上だが、二〇〇〇年には四倍増の一千万人以上になると見込まれる」と警告した。)

Figure 1: An example showing how pseudo-utility evaluation was applied effectively (1)

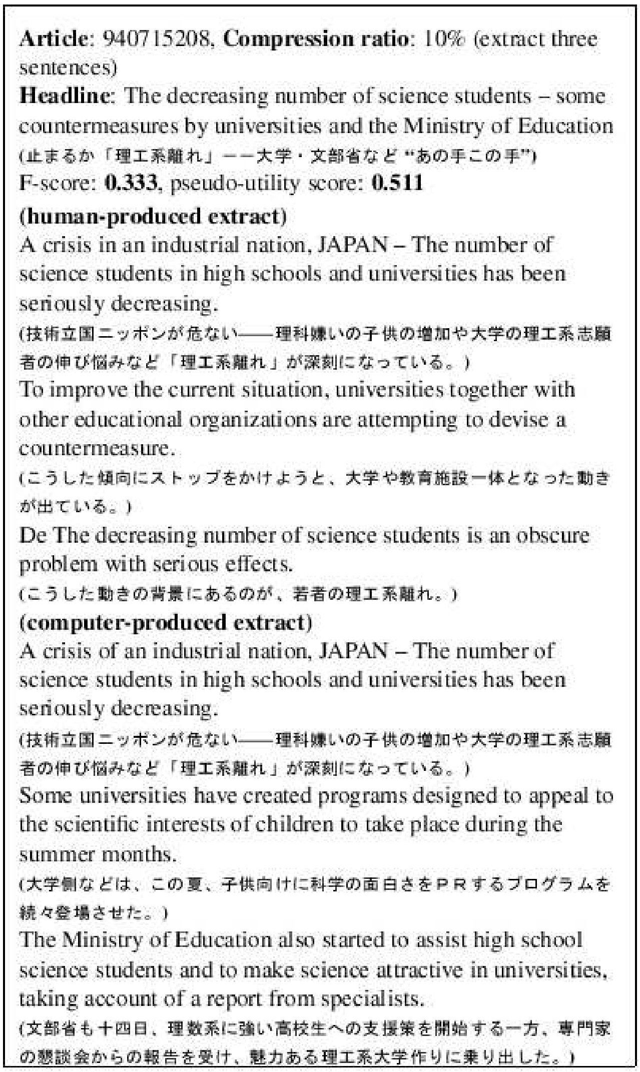
propriate than the F-score of zero.

Newspaper articles in the TSC generally consist of 10 to 20 sentences. The human-produced extract at 10% ratio therefore consists of only one or two sentences. In this case, even though a system can extract a moderately important sentence, it will not be reflected in the F-score. Pseudo-utility-based evaluation, however, takes account of such sentences, so more appropriate evaluation is possible in cases such as the example shown in Figure 1.

Another example is shown in Figure 2. The number of sentences extracted at 10% ratio from article 940715208 was three. In this example, the F-score is 0.333, because the first sentence of the three in the computer-produced extract is contained in the human-produced extract. Here, one of the other two sentences in the computer-produced extract is contained in the human-produced extracts at 30% ratio, and another one is in 50%. As a result, the pseudo-utility score is 0.511 ($\frac{1/0.1+1/0.3+1/0.5}{3/0.1}$). Comparing the computer-produced extract with the human-produced extract, sentences (2) and (3) in the computer-produced extract are examples of the description 'universities, together with other educational organizations, are attempting to devise a countermeasure' in sentence (2) in the human-produced extract. From these results, pseudo-utility-based evaluation is effective in assigning scores when the system extracts moderately important sentences.

From the result of this analysis, we can conclude that pseudo-utility-based evaluation can reduce the effect of problem 2, as pointed out by Jing et al..

We then calculated the average scores for pseudo-utility-based evaluation and F-measure for each system. The results are shown in Table 3 for F-measure and 4 for pseudo-utility-based evaluation[3]. Ten systems from seven groups attempted task A-1. The system IDs are shown from I to IX in the tables. Different systems from the same group are shown by dash (VII' and IX').



Article: 940715208, **Compression ratio**: 10% (extract three sentences)
Headline: The decreasing number of science students – some countermeasures by universities and the Ministry of Education
(止まるか「理工系離れ」――大学・文部省など "あの手この手")
F-score: **0.333**, pseudo-utility score: **0.511**
(human-produced extract)
A crisis in an industrial nation, JAPAN – The number of science students in high schools and universities has been seriously decreasing.
(技術立国ニッポンが危ない――理科嫌いの子供の増加や大学の理工系志願者の伸び悩みなど「理工系離れ」が深刻になっている。)
To improve the current situation, universities together with other educational organizations are attempting to devise a countermeasure.
(こうした傾向にストップをかけようと、大学や教育施設一体となった動きが出ている。)
De The decreasing number of science students is an obscure problem with serious effects.
(こうした動きの背景にあるのが、若者の理工系離れ。)
(computer-produced extract)
A crisis of an industrial nation, JAPAN – The number of science students in high schools and universities has been seriously decreasing.
(技術立国ニッポンが危ない――理科嫌いの子供の増加や大学の理工系志願者の伸び悩みなど「理工系離れ」が深刻になっている。)
Some universities have created programs designed to appeal to the scientific interests of children to take place during the summer months.
(大学側などは、この夏、子供向けに科学の面白さをＰＲするプログラムを続々登場させた。)
The Ministry of Education also started to assist high school science students and to make science attractive in universities, taking account of a report from specialists.
(文部省も十四日、理数系に強い高校生への支援策を開始する一方、専門家の懇談会からの報告を受け、魅力ある理工系大学作りに乗り出した。)

Figure 2: An example showing how pseudo-utility evaluation was applied effectively (2)

Systems I and II are ranked first and second in Table 3, but second and first in Table 4. The ranks of many other systems also changed; in particular the rank of system V changed from 9 by F-measure to 5 by pseudo-utility-based evaluation. To investigate the reliability of these ranks by pseudo-utility-based evaluation, we focused on systems I and II, and compared the extracts of both systems.

Among 90 pairs (30 texts × three ratios(10%, 30%, 50%)) of computer-produced extracts, we chose 16 pairs with the same F-scores and different pseudo-utility scores. Among the 16 pairs, the pseudo-utility scores of system I are larger than those of system II in 10 cases. A typical example of these cases is shown in Table 5. We show several sentences extracted by systems I and II at the 10% ratio from article 980500136 and the weights of each sen-

[3] We eliminated four articles (940701189, 940702187, 940716331, 980203053) from this examination, as they did not meet the condition in footnote 2.

Table 3: F-score of each system

SYSTEM	10%	30%	50%	total (rank)
I	0.363	0.435	0.589	0.463 (2)
II	0.337	0.452	0.612	0.467 (1)
V	0.251	0.447	0.574	0.424 (9)
VI	0.305	0.431	0.568	0.435 (6)
VI’	0.282	0.435	0.572	0.429 (8)
VII	0.305	0.474	0.586	0.455 (3)
VII’	0.241	0.497	0.578	0.439 (5)
VIII	0.199	0.399	0.590	0.396 (11)
IX	0.358	0.420	0.571	0.450 (4)
IX’	0.268	0.409	0.570	0.416 (10)
TF	0.284	0.433	0.586	0.434 (7)
Lead	0.276	0.367	0.530	0.391 (12)
Ave!%	0.289	0.433	0.577	0.433

Table 4: Pseudo-utility score of each system

SYSTEM	10%	30%	50%	total (rank)
I	0.518	0.559	0.664	0.581 (1)
II	0.450	0.603	0.673	0.569 (2)
V	0.410	0.546	0.641	0.527 (5)
VI	0.444	0.537	0.608	0.521 (8)
VI’	0.420	0.516	0.607	0.504 (9)
VII	0.433	0.560	0.651	0.541 (3)
VII’	0.401	0.556	0.636	0.525 (6)
VIII	0.330	0.515	0.654	0.495 (11)
IX	0.463	0.544	0.616	0.535 (4)
IX’	0.388	0.509	0.612	0.498 (10)
TF	0.406	0.526	0.657	0.525 (6)
Lead	0.401	0.481	0.549	0.468 (12)
Ave!%	0.422	0.537	0.630	0.530

tence used for pseudo-utility-based evaluation. Here, all the sentences, which have weights of 1/10, show a human-produced extract at 10% ratio. Among the five sentences extracted by system I, two sentences (S44 and S54) were contained in the human-produced extract, so the F-score is 0.4. System I also extracted S30, which occurs in human-produced extracts at 30%, and both S3 and S4, which occur in human-produced extracts at 50%, so the pseudo-utility score is 0.547.

System II extracted S26 and S43, so the F-score is 0.4 (the score is same as system I). Among the other three sentences extracted by system II, S3 and S4 are common to system I and the weight of the other sentence (S31) is zero. Consequently, the pseudo-utility score of system II is 0.480, which is lower than that of system I (Table 6).

S22 brings up an important question that is a main theme of this article. S22 is contained in a human-produced extract at 10%, though neither system extracted it. However, system I extracted S50, which is a solution to the question posed in S22. This sentence is actually contained in a human-produced extract at 30%. It is therefore appropriate that pseudo-utility-based evaluation differentiates systems I and II, because of the extraction of S50 by system I.

4.2.2. Comparison of content-based evaluation and subjective evaluation (task A-2)

First, we describe the results of subjective evaluation and content-based evaluation in the TSC. Second, we compare them and discuss the effectiveness of content-based evaluation.

Table 8: Percentage of cases where the order of content-based scores of two summaries matched the order of their ranks by subjective evaluation (whole data)

	FREE-based	PART-based
20% summary	91.4% (1371/1500)	88.6% (1329/1500)
40% summary	89.3% (1339/1500)	90.0% (1350/1500)

Average: 89.8%

Table 9: Percentage of cases where the order of content-based scores of two summaries matched the order of their ranks by subjective evaluation (SYS vs BASE)

	FREE-based	PART-based
20% summary	64.3% (193/300)	58.0% (174/300)
40% summary	58.7% (176/300)	63.7% (191/300)

Average: 61.2%

Results of subjective evaluation

Table 7 shows the ratios for four types of summaries (FREE, PART, SYS, BASE) ranked first, second, third and fourth, in terms of coverage of important contents (CONT) and readability (READ)[4]. As can be seen from the table, FREE is higher than the others (73.5%) in the number of cases that it ranked first, with PART, SYS, and BASE following. However, the difference between SYS and BASE is very small compared with that between FREE and PART. The quality of the four types of summaries can be ordered as follows:

(1)FREE > (2)PART > (3)SYS and BASE

Comparison of subjective evaluation and content-based evaluation

We compared the results of subjective evaluation and content-based evaluation. We calculated the percentage of cases where the order of content-based scores of two summaries matched the order of their ranks by subjective evaluation. The possible combinations of the four types of summaries are ‘FREE-PART’, ‘FREE-SYS’, ‘FREE-BASE’, ‘PART-SYS’, ‘PART-BASE’, and ‘SYS-BASE’. Among them, we eliminate ‘FREE-PART’ from our examination, because both FREE and PART are used as keys for evaluation in the TSC.

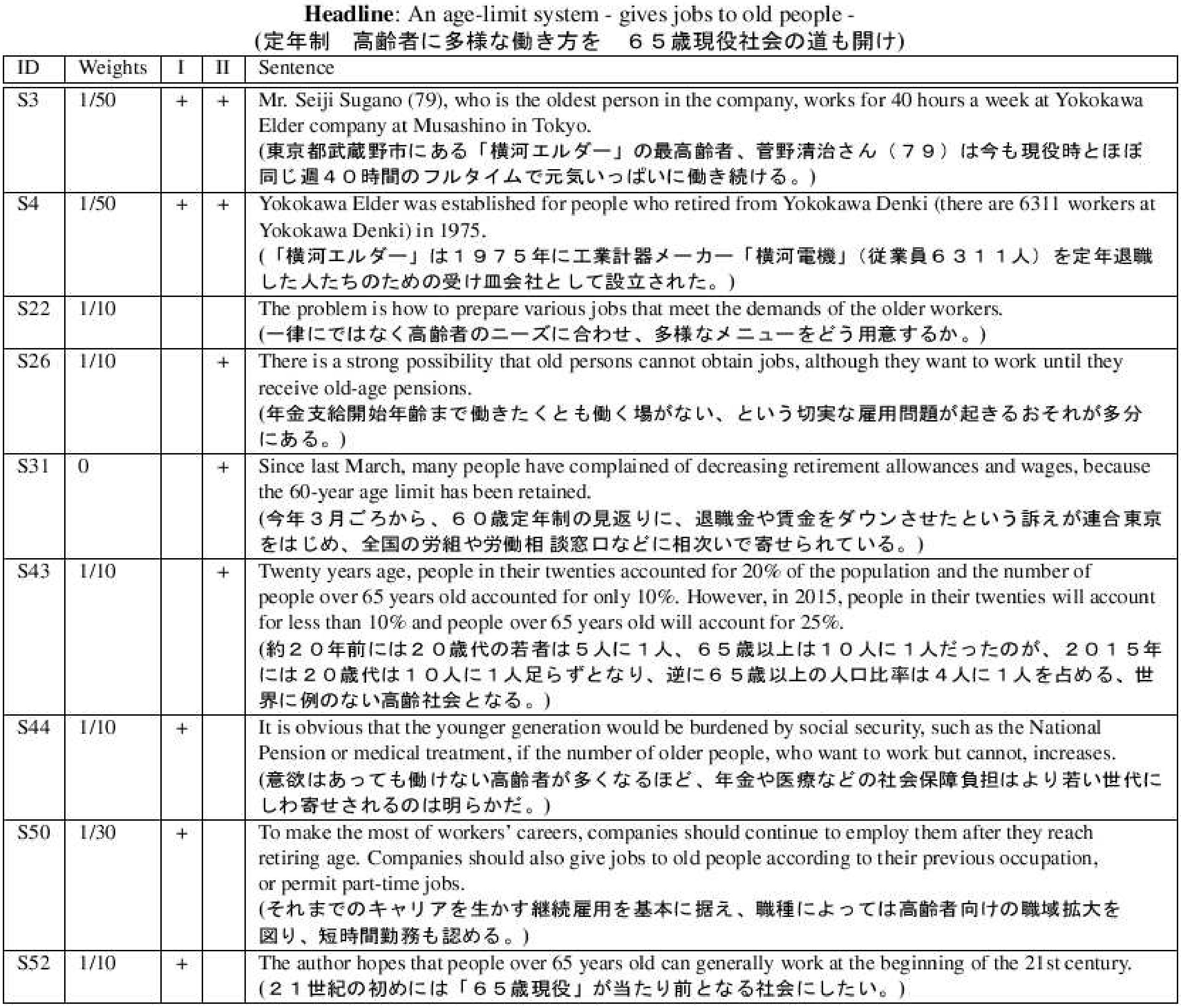
The results are shown in Table 8. As can be seen from the table, the evaluation by content-based measure matched the subjective evaluation in 90% of cases at compression ratios of both 20% and 40%.

As described above, the difference between SYS and BASE is very small. We therefore focused on them and calculated the percentage of cases where the order of content-based scores of the two summaries matched the order of their ranks by subjective evaluation. The results are shown in Table 9. We can conclude that the content-based measure cannot detect the slight difference of quality between two summaries, such as SYS and BASE. However, it is reliable at detecting larger differences between two summaries.

To investigate the reliability of content-based evaluation, we calculated the percentage of cases where the order

[4] two ratios (20% and 40%) × 30 texts × 10 systems = 600.

Table 5: Extracts produced by computer systems I and II (10% ratio)



Headline: An age-limit system - gives jobs to old people -
(定年制　高齢者に多様な働き方を　６５歳現役社会の道も開け)

ID	Weights	I	II	Sentence
S3	1/50	+	+	Mr. Seiji Sugano (79), who is the oldest person in the company, works for 40 hours a week at Yokokawa Elder company at Musashino in Tokyo. (東京都武蔵野市にある「横河エルダー」の最高齢者、菅野清治さん（７９）は今も現役時とほぼ同じ週４０時間のフルタイムで元気いっぱいに働き続ける。)
S4	1/50	+	+	Yokokawa Elder was established for people who retired from Yokokawa Denki (there are 6311 workers at Yokokawa Denki) in 1975. (「横河エルダー」は１９７５年に工業計器メーカー「横河電機」（従業員６３１１人）を定年退職した人たちのための受け皿会社として設立された。)
S22	1/10			The problem is how to prepare various jobs that meet the demands of the older workers. (一律にではなく高齢者のニーズに合わせ、多様なメニューをどう用意するか。)
S26	1/10		+	There is a strong possibility that old persons cannot obtain jobs, although they want to work until they receive old-age pensions. (年金支給開始年齢まで働きたくとも働く場がない、という切実な雇用問題が起きるおそれが多分にある。)
S31	0		+	Since last March, many people have complained of decreasing retirement allowances and wages, because the 60-year age limit has been retained. (今年３月ごろから、６０歳定年制の見返りに、退職金や賃金をダウンさせたという訴えが連合東京をはじめ、全国の労組や労働相 談窓口などに相次いで寄せられている。)
S43	1/10		+	Twenty years age, people in their twenties accounted for 20% of the population and the number of people over 65 years old accounted for only 10%. However, in 2015, people in their twenties will account for less than 10% and people over 65 years old will account for 25%. (約２０年前には２０歳代の若者は５人に１人、６５歳以上は１０人に１人だったのが、２０１５年には２０歳代は１０人に１人足らずとなり、逆に６５歳以上の人口比率は４人に１人を占める、世界に例のない高齢社会となる。)
S44	1/10	+		It is obvious that the younger generation would be burdened by social security, such as the National Pension or medical treatment, if the number of older people, who want to work but cannot, increases. (意欲はあっても働けない高齢者が多くなるほど、年金や医療などの社会保障負担はより若い世代にしわ寄せされるのは明らかだ。)
S50	1/30	+		To make the most of workers' careers, companies should continue to employ them after they reach retiring age. Companies should also give jobs to old people according to their previous occupation, or permit part-time jobs. (それまでのキャリアを生かす継続雇用を基本に据え、職種によっては高齢者向けの職域拡大を図り、短時間勤務も認める。)
S52	1/10	+		The author hopes that people over 65 years old can generally work at the beginning of the 21st century. (２１世紀の初めには「６５歳現役」が当たり前となる社会にしたい。)

Table 6: F-scores and pseudo-utility scores of systems I and II (article 980511036 at 10% ratio)

	I	II
F-measure	0.400	0.400
pseudo-utility	0.547	0.480

of content-based scores of two summaries matched the order of their ranks by subjective evaluation, using gaps in the content-based score from zero to one at 0.1 intervals. The results are shown in Table 10. We found that the evaluation by content-based measure matched the subjective evaluation in 93% of cases, if the gap in the content-based scores between two summaries was more than 0.2.

We examined the cases when the gap in the content-based scores between SYS and BASE was more than 0.2 in Table 9. The results are shown in Table 11. The cases with gaps larger than 0.2 make up only 14.5% between SYS and BASE, while 59.5% in Table 10[5]. From these results, we reconfirm that the difference between SYS and BASE is small.

Next, we calculated the percentage of cases where the order of content-based scores of two summaries matched the order of their ranks by subjective evaluation, when the gap was more than 0.2. The result is shown in Table 12.

As can be seen from the table, the reliability of content-based evaluation increases by more than 10%. The ratio of 71.3%, however, is still lower than the 92.8% in Table 10.

5. Conclusion and Future Research

In this paper, we first proposed pseudo-utility-based evaluation, and conducted an examination to investigate the effectiveness of our method. We evaluated computer-produced extracts by the F-measure and by pseudo-utility-based evaluation, and found that pseudo-utility-based evaluation could diminish the problems of the F-measure.

We also focused on content-based evaluation. We conducted tests in the context of comparing two summaries produced from different systems. We evaluated

[5] The number of cases with gaps smaller than 0.2 is 2430(1242(0.0–0.1)+1188(0.1–0.2)) among 6000 cases. So, the ratio that the gaps is larger than 0.2 is 0.595 $(1-\frac{1242+1188}{6000})$.

Table 7: The ratios of four types of summaries ranked to first, second, third and fourth

		1st	2nd	3rd	4th
FREE	CONT	69.8% (419/600)	28.7% (172/600)	1.5% (9/600)	0.0% (0/600)
	READ	77.7% (466/600)	19.0% (114/600)	3.2% (19/600)	0.2% (1/600)
	TOTAL	73.5% (885/1200)	23.8% (286/1200)	2.3% (28/1200)	0.1% (1/1200)
PART	CONT	49.0% (294/600)	49.0% (294/600)	1.8% (11/600)	0.2% (1/600)
	READ	40.6% (244/600)	47.5% (285/600)	8.5% (51/600)	3.0% (18/600)
	TOTAL	44.8% (538/1200)	48.3% (579/1200)	5.3% (64/1200)	1.6% (19/1200)
SYS	CONT	2.3% (14/600)	3.3% (20/600)	68.0% (408/600)	26.3% (158/600)
	READ	11.2% (67/600)	10.3% (62/600)	43.3% (260/600)	38.8% (233/600)
	TOTAL	6.6% (79/1200)	6.8% (82/1200)	55.7% (668/1200)	32.6% (391/1200)
BASE	CONT	0.0% (0/600)	0.8% (5/600)	38.2% (229/600)	61.0% (366/600)
	READ	6.5% (39/600)	8.0% (48/600)	52.7% (316/600)	32.8% (197/600)
	TOTAL	3.2% (39/1200)	4.4% (53/1200)	45.4% (545/1200)	46.9% (563/1200)

Table 10: Effectiveness of content-based measure

Gap between content-based scores	Percentage of cases in which content-based evaluation matched with subjective evaluation (%)
0.0 - 0.1	57.8 (718/1242)
0.1 - 0.2	77.1 (916/1188)
0.2 - 0.3	92.8 (966/1041)
0.3 - 0.4	95.9 (805/839)
0.4 - 0.5	96.4 (588/610)
0.5 - 0.6	98.8 (589/596)
0.6 - 0.7	99.4 (336/338)
0.7 - 0.8	99.0 (103/104)
0.8 - 0.9	100.0 (26/26)
0.9 - 1.0	100.0 (16/16)

Table 11: Percentage of cases when the gap between two summaries is more than 0.2 (SYS vs BASE)

	FREE-based	PART-based
20% summary	17.0% (51/300)	23.0% (69/300)
40% summary	10.0% (30/300)	8.0% (24/300)

Average: 14.5%

Table 12: Percentage of cases where the order of content-based scores of two summaries matched the order of their ranks by subjective evaluation and the gap between two summaries is more than 0.2 (SYS vs BASE)

	FREE-based	PART-based
20% summary	74.5% (38/51)	73.9% (51/69)
40% summary	60.0% (18/30)	70.8% (17/24)

Average: 71.3%

computer-produced summaries by content-based evaluation, and compared the results with a subjective evaluation. We found that the evaluation by content-based measure matched the subjective evaluation in 93% of cases, if the gap in the content-based scores between two summaries was more than 0.2.

We showed that human-produced extracts at different compression ratios could be used for utility-based evaluation. Although we used human-produced extracts at 10%, 30%, and 50% ratios for evaluation, we should look for optimal combinations of compression ratios.

In this paper, we used '1/compression ratio' as the weights for sentences for pseudo-utility-based evaluation. In future, we should also study the optimal weights of each sentence used for pseudo-utility-based evaluation.

6. References

Donaway, R.L., Drummey, K.W., and Mather, L.A. 2000. A Comparison of Rankings Produced by Summarization Evaluation Measures. *Proceedings of the ANLP/NAACL 2000 Workshop on Automatic Summarization*, pages 69–78.

Fukushima, T. and Okumura, M. 2001. Text Summarization Challenge Text Summarization Evaluation at NTCIR Workshop2. *Proceedings of the Second NTCIR Workshop Meeting*, pages 45–51.

Fukushima, T. and Okumura, M. 2001. Text Summarization Challenge Text Summarization Evaluation in Japan. *Proceedings of NAACL 2001 Workshop Automatic Summarization*, pages 51–59.

Jing, H., Barzilay, R., McKeown, K., and Elhadad, M. 1998. Summarization Evaluation Methods: Experiments and Analysis. *Technical Report SS-98-06, Intelligent Text Summarization, AAAI Press*, pages 51–59.

Mani, I. 2001. Automatic Summarization. *John Benjamins Pub Co.*

McKeown, K.R., Barzilay, R., Evans, D., Hatzivassiloglou, V., Kan, M. Y., Schiffman, B. and Teufel, S. 2001. Columbia Multi-Document Summarization: Approach and Evaluation. *Proceedings of Document Understanding Conference 2001*, pages 43–63.

Mittal, V., Kantrowitz, M., Goldstein, J., and Carbonell, J. 1999. Selecting Text Spans for Document Summaries: Heuristics and Metrics. *Proceedings of the 16th National Conference on Artificial Intelligence*, pages 467–473.

Radev, D.R., Jing, H., and Budzikowska, M. 2000. Centroid-based Summarization of Multiple Documents: Sentence Extraction, Utility-based Evaluation, and User Studies. *Proceedings of the ANLP/NAACL2000 Workshop on Automatic Summarization*, pages 21–29.

Sparck-Jones, K. and Galliers, J. 1996. Evaluating Natural Language Processing Systems: An Analysis and Review. *Lecture Notes in Artificial Intelligence 1083*, Berlin: Springer.