Comparison of some automatic and manual methods for summary evaluation based on the Text Summarization Challenge 2

Hidetsugu Nanba

Hiroshima City University
3-4-1 Ozukahigashi Aasaminamiku
Hiroshima Japan 731-3194
nanba@its.hiroshima-cu.ac.jp

Manabu Okumura

Tokyo Institute of Technology
4259 Nagatsuta Midoriku
Yokohama Japan 226-8503
oku@pi.titech.ac.jp

Abstract

In this paper, we compare some automatic and manual methods for summary evaluation. One of the essential points for evaluating a summary is how well the evaluation measure recognizes slight differences in the quality of the computer-produced summaries. In terms of this point, we examined 'evaluation by revision' using the data of the Text Summarization Challenge 2 (TSC2). Evaluation by revision is a manual method that was first used in TSC2, whose effectiveness has not been tested. First, we compared evaluation by revision with a ranking evaluation, which is a manual method used both in TSC1 and in TSC2, by checking the gaps of the edit distance from 0 to 1 at 0.1 intervals. To investigate the effectiveness of evaluation by revision, we also tested other automatic methods: content-based evaluation, BLEU and RED, and compare their results with that of evaluation by revision for reference. As a result, we found that evaluation by revision is effective for recognizing slight differences between computer-produced summaries. Second, we evaluated content-based evaluation, BLEU and RED by evaluation by revision, and compared the effectiveness of the three automatic methods. We found that RED is superior to the others in some examinations.

1. Introduction

How to evaluate computer-produced summaries has been recognized as a problem in the field of automatic summarization. A number of manual and automatic methods for evaluating summaries have been proposed. In this paper, we compare some automatic and manual methods for summary evaluation.

One of the essential points for such an evaluation is how well the evaluation measure recognizes slight differences in the quality of the computer-produced summaries. In our previous work (Nanba & Okumura, 2002), we compared a content-based evaluation (Donaway et al., 2000) with a ranking evaluation, which is a manual method used both in Text Summarization Challenge 1 (TSC1) and in TSC2 (Fukushima et al., 2002), by checking the gaps of the content-based score from 0 to 1 at 0.1 intervals. We found that the content-based evaluation matched the ranking evaluation in 93% of the cases, if the gap between the content-based scores of the two summaries was more than 0.2.

In the same way as our previous examination, we examine the following four methods: evaluation by revision, content-based evaluation, BLEU (Padineni et al., 2001) and RED (Lin & Hovy, 2003). Evaluation by revision is a manual method that was first used in TSC2, whose effectiveness has not been tested. BLEU is an automatic evaluation method devised for machine translation, whose effectiveness in MT has been reported recently. Lin et al. analysed BLEU scoring method and developed the corresponding evaluation methodology for summarization, which they call RED. They examined RED and BLEU using the data of Document Understanding Conference (DUC) and showed that RED could improve BLEU.

First, we test evaluation by revision by comparing with ranking evaluation in the same way as our previous work (Nanba & Okumura, 2002). To investigate the effectiveness of evaluation by revision, we also test other automatic methods and compare their results with that of evaluation by revision for reference, though it is unfair to compare automatic methods with the costly manual method. Second, we evaluate content-based evaluation, BLEU and RED, by evaluation by revision, and compare the effectiveness of three automatic methods.

The remainder of the paper is organized as follows. Section 2 describes an examination for investigating the effectiveness of evaluation by revision. Section 3 reports the comparison of some automatic methods with evaluation by revision. We discuss the results in Section 4, and conclude our work in Section 5.

2. Investigating the effectiveness of evaluation by revision

To evaluate the effectiveness of evaluation by revision, we conducted some tests using the data of the TSC2. In Section 2.1, we describe the task and evaluation in TSC2. In Section 2.2, we explain some automatic methods for comparison, and report the experimental results in Section 2.3.

2.1 Data for the evaluation

We used the data of TSC2 for our examinations. The data consists of human-produced summaries, computer-produced summaries (eight systems and a baseline system using lead-method), and the results of both evaluation by revision and ranking evaluation. All summaries were made from thirty newspaper articles, written in Japanese, and they were extracted from the Mainichi newspaper database for 1998 and 1999. Two tasks were conducted in TSC2, and we used the data of a single document summarization task. In this task, participants were asked to produce summaries in plain text at the ratios of 20% and 40%.

Summaries were evaluated in the following two ways;

- Ranking Evaluation

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

  - Summaries 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.

- Evaluation by revision

The measure evaluates summaries by measuring the degree to which computer-produced summaries are revised. The judges read the original texts and revise the computer-produced summaries in terms of their content and readability. The human revisions are made with only three editing operations (insertion, deletion, replacement). The degree of the human revision, which we call `edit distance', is computed from the number of revised characters divided by the number of characters in the original summary. If summary's quality is too law to revise more than half of the original summary, the judges stop to revise.

2.2 Automatic methods for comparison

For comparison of evaluation by revision, we also tested the following automatic methods;

- Content-based evaluation (Donaway et al., 2000)

The measure evaluates summaries by comparing their content words with those of human-produced extracts. The score of content-based measure is obtained by computing the similarity between the term vector using tf*idf weighting of a computer-produced summary and the term vector of a human-produced summary by cosine distance.

- BLEU (Papineni et al., 2001)

The measure compares n-grams of the candidate with the n-grams of the reference translation and count the number of matches. These matches are position-independent. The more matches there are, the better the candidate translation will be.

- RED (Lin & Hovy, 2003)

Lin and Hovy devised RED based on BLEU. They reported that simple unigram overlap scores perform better for summary evaluation than BLEU's combination of n-gram scores plus a brevity penalty.

2.3 Experiments

We compared evaluation by revision with ranking evaluation. To investigate how well the evaluation measure recognizes slight differences in the quality of the summaries, we calculated the percentage of cases where the order of edit distance of two summaries matched the order of their ranks given by the ranking evaluation by checking the gaps of the score from 0 to 1 at 0.1 intervals. We also evaluated the content-based evaluation, BLEU, and RED. To calculate the scores of the three automatic methods, we used FREE as reference summaries.

Table 1 shows the experimental results. The evaluation by revision matched the ranking evaluation better than the other methods, when the gap was less than 0.1. It indicates that evaluation by revision is effective for recognizing slight differences between computer-produced summaries.

The gap between scores	The percentage of cases that each evaluation method matched ranking evaluation			
	Evaluation by revision	Content-based evaluation	BLEU	RED
0.0-0.1	82.17 (318/387)	70.78 (155/219)	72.47 (179/247)	71.36. (147/206)
0.1-0.2	88.35 (220/249)	86.49 (256/296)	91.01 (172/189)	86.40 (216/250)
0.2-0.3	84.25 (123/146)	86.81 (204/235)	93.23 (179/192)	92.13 (164/178)
over 0.3	96.17 (402/418)	90.67 (408/450)	87.93 (503/572)	88.87 (503/566)
Total	88.38 (1063/1200)	85.25 (1023/1200)	86.08 (1033/1200)	85.83 (1030/1200)

Table 1: Effectiveness of some evaluation methods (40%)

3. Comparison of the three automatic methods with evaluation by revision

To investigate the effectiveness of the content-based evaluation, BLEU and RED, we also compared the three automatic methods with evaluation by revision. We compared them from three points of view. In Section 3.1, we explain three viewpoints for comparison of automatic methods. In Section 3.2, we report the results of comparison.

3.1 Viewpoints for comparison of the three automatic methods

We compared the three automatic methods from the following three points of view.

- Point 1 (Ranking four kinds of summaries by each evaluation method)

In the same way as we compared the ranking evaluation in Section 2, we ranked four kinds of summaries, FREE, PART, SYS and BASE by the scores of content-based evaluation, BLEU and RED, respectively. We also ranked them by their edit distances, and compared them with the results of automatic methods. We calculate the percentage of cases where the order of edit distance of two summaries matched the order of scores of each automatic method by checking the gaps of the score from 0 to 1 at 0.1 intervals.

- Point 2 (Direct comparison of computer-produced summaries by their scores by each method)

Computer-produced summaries could not be compared directly by ranking evaluation, because the method compared only among FREE, PART, SYS and BASE. On the other hand, evaluation by revision allows us to compare computer-produced summaries directly by their edit distances. We can also compare them directly with their scores calculated by each method. We therefore tested the usefulness of the three automatic methods for direct comparison of computer-produced summaries. In a way similar to point 1, we calculated the percentage of cases where the order of edit distances of the two summaries matched the order of their ranks calculated by each automatic method.

In both points 1 and 2, we excluded a pair of summaries from the counts if the judges stopped to revise both of them, because we could not identify which was better in quality.

- Point 3 (Comparison of rankings using Spearman order correlation coefficients)

Lin and Hovy (2003) ranked systems participating in DUC 2002 by some automatic methods, and compared them with a manual ranking by Spearman rank order correlation coefficients. As a result, they found that RED was superior to the others. In the same way, we compared a ranking by evaluation by revision with those by the three automatic methods by Spearman rank order correlation coefficients.

3.2 Results

- Point 1 (Ranking four kinds of summaries by each evaluation method)

The results are shown in Tables 2 (the compression ratio is 40%) and 3(20%). Generally, the results for automatic methods matched evaluation by revision better than they matched ranking evaluation (Table 1). The results of the three automatic methods are quite close, but RED is slightly superior to the others both at 40% and at 20%.

The gap between scores	The percentage of cases that each evaluation method matched evaluation by revision		
	Content-based evaluation	BLEU	RED
0.0-0.1	88.83 (334/376)	88.30 (332/376)	88.83 (334/376)
0.1-0.2	91.14 (216/237)	93.25 (221/237)	93.67 (222/237)
0.2-0.3	96.58 (113/117)	95.73 (112/117)	96.58 (113/117)
over 0.3	97.40 (75/77)	97.40 (75/77)	96.10 (74/77)
Total	91.45 (738/807)	91.70 (740/807)	92.07 (743/807)

Table 2: Comparison of some automatic methods based on the evaluation by revision (40%)

The gap between scores	The percentage of cases that each evaluation method matched evaluation by revision		
	Content-based evaluation	BLEU	RED
0.0-0.1	80.42 (152/189)	82.01 (155/189)	81.48 (154/189)
0.1-0.2	88.60 (101/114)	82.46 (94/114)	85.09 (97/114)
0.2-0.3	96.92 (63/65)	98.46 (64/65)	100.00 (65/65)
Over 0.3	96.23 (51/53)	98.11 (52/53)	100.00 (53/53)
Total	87.17 (367/421)	86.70 (365/421)	87.65 (369/421)

Table 3: Comparison of some automatic methods based on the evaluation by revision (20%)

- Point 2 (Direct comparison of computer-produced summaries by their scores by each method)

The results are shown in Tables 4 (40% compression) and 5 (20%). As can be seen from Table 4, the percentages of cases with gaps less than 0.1 among all cases are higher than in Tables 2. This indicates that the difference in quality between computer-produced summaries is smaller than those between FREE, PART and BASE, when the compression ratio is 40%. As a whole, the percentages by which each method matched the ranking evaluation in Tables 4 and 5 are lower than in Tables 2 and 3.

The gap between scores	The percentage of cases that each evaluation method matched ranking evaluation		
	Content-based evaluation	BLEU	RED
0.0-0.1	58.92 (294/499)	57.51 (287/499)	59.92. (299/499)
0.1-0.2	62.93 (129/205)	67.32 (138/205)	66.34 (136/205)
0.2-0.3	72.84 (59/81)	70.37 (57/81)	67.90 (55/81)
over 0.3	62.00 (31/50)	68.00 (34/50)	70.00 (35/50)
Total	61.44 (513/835)	61.80 (516/835)	62.87 (525/835)

Table 4: Comparison of some automatic methods based on the evaluation by revision (40%) (using computer-produced summaries only)

The gap between scores	The percentage of cases that each evaluation method matched ranking evaluation		
	Content-based evaluation	BLEU	RED
0.0-0.1	64.22 (140/218)	66.97 (146/218)	69.72 (152/218)
0.1-0.2	66.19 (92/139)	64.03 (89/139)	64.75 (90/139)
0.2-0.3	69.35 (86/124)	73.39 (91/124)	71.77 (89/124)
over 0.3	68.45 (115/168)	73.81 (124/168)	74.40 (125/168)
Total	66.72 (433/649)	69.49 (451/649)	70.26 (456/649)

Table 5: Comparison of some automatic methods based on the evaluation by revision (20%) (using computer-produced summaries only)

- Point 3 (comparison of rankings using Spearman order correlation coefficients)

Tables 6 and 7 show the rankings, average scores and Spearman rank order correlation coefficients of eight systems, FREE, PART and BASE, calculated by automatic methods and by evaluation by revision. The ranks of FREE, PART and BASE by evaluation by revision and by automatic methods are about the same in both Tables 6 and 7[1]. The ranks of the eight systems by automatic methods are close to those for evaluation by revision (except for systems 7 and 8) at 40% compression. The ranks of the systems, FREE, PART and BASE by the three automatic methods are close to each other. We compared the rankings by the three automatic methods by Spearman rank order correlation coefficients. As a result,

[1] As we used FREE as referent summaries for automatic evaluations, average scores of FREE by automatic methods are one, and PART could not exceed FREE.

we obtained more than 0.9 of coefficients in most cases (except for the comparison between RED and content-based evaluation at 40% compression).

System ID	Evaluation by revision	Content-based evaluation	BLEU	RED
1	4 (0.132)	5 (0.685)	3 (0.192)	3 (0.585)
2	3 (0.124)	4 (0.687)	4 (0.190)	4 (0.584)
3	8 (0.173)	8 (0.659)	7 (0.165)	7 (0.558)
4	6 (0.154)	7 (0.661)	8 (0.163)	6 (0.561)
5	10 (0.212)	9 (0.646)	10 (0.139)	9 (0.537)
6	7 (0.165)	3 (0.694)	5 (0.171)	8 (0.552)
7	5 (0.146)	10 (0.645)	9 (0.155)	10 (0.536)
8	9 (0.191)	6 (0.667)	6 (0.168)	5 (0.568)
BASE	11 (0.331)	11 (0.582)	11 (0.111)	11 (0.498)
PART	1 (0.022)	2 (0.809)	2 (0.350)	2 (0.733)
FREE	2 (0.023)	1 (1.000)	1 (1.000)	1 (1.000)
Spearman R		0.745	0.827	0.781

Table 6: Manual and automatic rankings, average scores of each method and Spearman rank order correlation coefficients (40%)

System ID	Evaluation by revision	Content-based evaluation	BLEU	RED
1	4 (0.206)	3 (0.519)	4 (0.085)	3 (0.443)
2	3 (0.191)	4 (0.514)	3 (0.086)	4 (0.437)
3	9 (0.337)	10 (0.440)	10 (0.048)	11 (0.346)
4	6 (0.311)	9 (0.442)	8 (0.051)	9 (0.358)
5	.5 (0.309)	7 (0.464)	9 (0.050)	7 (0.374)
6	7 (0.311)	5 (0.507)	5 (0.070)	6 (0.387)
7	10 (0.344)	8 (0.447)	7 (0.056)	8 (0.362)
8	8 (0.324)	6 (0.485)	6 (0.068)	5 (0.403)
BASE	11 (0.429)	11 (0.394)	11 (0.045)	10 (0.347)
PART	1 (0.055)	2 (0.678)	2 (0.213)	2 (0.622)
FREE	2 (0.058)	1 (1.000)	1 (1.000)	1 (1.000)
Spearman R		0.864	0.818	0.836

Table 7: Manual and automatic rankings, average scores of each method and Spearman rank order correlation coefficients (20%)

4. Discussion

To investigate the characteristic features of the three automatic methods, we compared them in the same way as point 2. The results at 20% compression are shown in Table 8. As can be seen from the table, the percentages by which each method matched are much higher than the results in Table 5. Among the three methods, BLEU and RED are most similar even when the gap between scores is less than 0.1. RED and content-based evaluation are the second, both of which calculate scores based on unigram. When the gap between scores is more than 0.2, the results by the three methods matched in almost all cases. In other words, the difference between the three automatic methods mainly comes out when the gap between scores is less than 0.2. In terms of this point, we could recognize the superiority of RED in Table 5. However, we could not find the significant superiority of RED from other results.

The gap between scores	The percentage of cases that each evaluation method matched		
	BLEU vs. RED	BLEU vs. content	RED vs. content
0.0-0.1	87.70 (392/447)	71.79 (313/436)	77.18 (345/447)
0.1-0.2	99.59 (241/242)	94.27 (214/227)	94.21 (228/242)
0.2-0.3	100.00 (118/118)	99.12 (113/114)	98.31 (116/118)
over 0.3	100.00 (33/33)	98.39 (61/62)	100.00 (33/33)
Total	93.33 (784/840)	83.57 (702/840)	85.95 (722/840)

Table 8: Comparison of the three automatic methods (20%)(using computer-produced summaries only)

Conclusions

In this paper, we compared some automatic and manual methods for summary evaluation using the data of TSC2. We first investigate the effectiveness of evaluation by revision. We tested evaluation by revision by comparing with a ranking evaluation, by checking the gaps of the edit distance from 0 to 1 at 0.1 intervals. We also tested BLEU, RED and content-based evaluation, and compared their results with that of evaluation by revision. As a result, we found that evaluation by revision is effective for recognizing slight differences between computer-produced summaries. Second, we evaluated content-based evaluation, BLEU and RED by evaluation by revision, and compared the effectiveness of the three automatic methods. We found that RED is superior to the others in some examinations.

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*, (pp.69-78).

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

Fukushima, T., Okumura, M., and Nanba, H. (2002). Text Summarization Challenge 2 / Text Summarization Evaluation at NTCIR Workshop3, *Working Notes of the Third NTCIR Workshop Meeting, PART V,* (pp.1-7).

Lin, C. and Hovy, E. (2003). Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics, *Proceedings of the Human Language Technology Conference 2003.*

Nanba, H. and Okumura, M. (2002). Some Examinations of Intrinsic Methods for Summary Evaluation Based on the Text Summarization Challenge (TSC), *Proceedings of the third International Conference on Language Resources and Evaluation*, (pp.739-746).

Papineni, K., S. Roukos, T. Ward, W.-J. Zhu. (2001). BLEU: a Method for Automatic Evaluation of Machine Translation, *IBM Research Report,* RC22176 (W0109-022).