Automatic Evaluation of Texts by Using Paraphrases

Kazuho Hirahara, Hidetsugu Nanba, Toshiyuki Takezawa **Manabu Okumura**

Hiroshima City University
{hirahara, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp

Tokyo Institute of Technology
oku@pi.titech.ac.jp

Abstract

The evaluation of computer-produced texts has been recognized as an important research problem for automatic text summarization and machine translation. Traditionally, computer-produced texts were evaluated automatically by n-gram overlap with human-produced texts. However, these methods cannot evaluate texts correctly, if the n-grams do not overlap between computer-produced and human-produced texts, even though the two texts convey the same meaning. We explored the use of paraphrases for the refinement of traditional automatic methods for text evaluation. To confirm the effectiveness of our method, we conducted some experiments using the data from the Text Summarization Challenge 2. We found that the use of paraphrases created using a statistical machine translation technique could improve the traditional evaluation method.

Keywords: text summarization, machine translation, text evaluation, synonyms

1. Introduction

The evaluation of computer-produced texts has been recognized as an important research problem for text summarization and machine translation. Traditionally, computer-produced texts were evaluated by n-gram overlap with human-produced texts (Papineni, 2002; Lin and Hovy, 2003; Lin, 2004). However, these methods cannot evaluate texts correctly, if the n-grams do not overlap between the computer-produced and human-produced texts, even though the two texts convey the same meaning. Therefore, we explore the use of paraphrases for the refinement of traditional automatic methods for text evaluation.

Several evaluation methods using paraphrases were proposed in text summarization (Zhon et al., 2006) and machine translation (Kauchak and Barzilay, 2006; Kanayama, 2003; Yves and Etienne, 2005), and their effectiveness was confirmed. However, these studies did not discuss what paraphrases techniques gave more accurate text evaluation. We analyzed 318 paraphrases in texts to be evaluated, and classified them into five categories. Then we examined several paraphrase methods that covered four of those categories (about 70% of the 318 paraphrases). We evaluated texts using the data of the Text Summarization Challenge 2. We found that our method could improve a traditional evaluation method.

The remainder of this paper is organized as follows. Section 2 describes related work. Section 3 describes the benefits of paraphrases in text evaluation. Section 4 explains our evaluation method using paraphrases. To investigate the effectiveness of our method, we conducted some experiments, and Section 5 reports on these. We present some conclusions in Section 6.

2. Related Work

We describe the related studies of "automatic evaluation of texts" and "text evaluation using paraphrases" in Sections 2.1. and 2.2., respectively.

2.1. Automatic Evaluation of Texts

Several measures for evaluating computer-produced texts have been proposed (Papineni, 2001; Lin and Hovy, 2003; Lin, 2004). BLEU (Papineni, 2001) was developed as a measure of automatic evaluation for machine translation. It compares the n-grams of the candidate with the n-grams of the reference translation, and counts the number of matches. These matches are position independent. The quality of the candidate translation depends on the number of matches.

ROUGE-N (Lin and Hovy, 2003; Lin, 2004) is a standard evaluation measure in automatic text summarization. The measure compares the n-grams of the two summaries, and counts the number of matches. The measure is defined by the following equation.

$$ROUGE-N = \frac{\sum_{S \in R} \sum_{gram_N \in S} Count_{match}(gram_N)}{\sum_{S \in R} \sum_{gram_N \in S} Count(gram_N)}$$

where *N* is the length of the n-gram, *gram_N*, and *Count_{match}(gram_N)* is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. Lin examined ROUGE-N with values of N from one to four, and reported that ROUGE-N had a high correlation with manual evaluation when N was one or two. In our work, we focus on evaluation of computer-produced summaries, and use ROUGE-N as a baseline method for text evaluation.

2.2. Text Evaluation Using Paraphrases

Several evaluation methods using paraphrases were proposed in text summarization (Zhon et al., 2006) and machine translation (Kauchak and Barzilay, 2006; Kanayama, 2003; Yves and Etienne, 2005). Zhou et al. (2006) proposed a method "ParaEval" to obtain paraphrases automatically using a statistical machine translation (SMT) technique. If translations of two terms X and Y are the same term, then the terms X and Y are considered to be paraphrases. Based on this idea, they automatically obtained paraphrases from a translation model, which was created from pairs of English and Chinese sentences using the SMT technique. They then used these paraphrases for the improvement of ROUGE-N. In our work, we also use paraphrases acquired by the SMT technique as paraphrase method.

In addition to the SMT-based paraphrases, we examined another method for automatic acquisition of paraphrases. Lin (1998) and Lee (1999) proposed a method for calculating the similarity between terms, called "distributional similarity". The underlying assumption of their approach is that semantically similar words are used in similar contexts. Therefore, they define the similarity between two terms as the amount of information contained in the commonality between the terms, divided by the amount of information in the contexts of the terms. In our work, we use "distributional similarity" as a method for acquiring paraphrases.

3. The Benefits of Paraphrases in Text Evaluation

To investigate the benefits of paraphrases in text evaluation, we compared multiple summaries created from the same text.

3.1. Data

In this investigation, we used 30 Japanese editorials[1] from the Mainichi newspaper databases of 1998 and 1999. For each editorial, we asked 10 human subjects to create abstract-type summaries for a summarization ratio of 20%, which were produced to try to obtain the main idea of the editorial without worrying about sentence boundaries. We compared these summaries and obtained 318 paraphrases.

3.2. Paraphrases in Text Evaluation

We classified the 318 paraphrases into the following five categories.

A) Synonymous expressions with different Japanese characters

The senses of two expressions are the same, but they were expressed with different Japanese characters. There were 64 (20.1%) paraphrases in this category.

B) Word-level synonymous expressions

Two words have the same sense, but are different, such as "時間" (time) and "時" (moment). There were 78 (24.5%) paraphrases in this category. Synonym dictionaries are required for paraphrases in this category.

C) Phrase-level synonymous expressions

Two phrases have the same sense, but are expressed differently, such as "切り離せない" (cannot be divided) and "繋がっている" (linked to each other). There were 38 (11.9%) paraphrases in this category. Phrase-level synonym dictionaries are required for paraphrases in this category.

D) Clause-level synonymous expressions

Two clauses have the same sense, but are expressed differently, such as "X がなければ Y はできなかった" (If it were not for X, Y would not succeed.) and "X により Y ができた" (Y succeeded because of X). Changing voices and transitive/intransitive alternations are also classified in this category. There were 39 (12.3%) paraphrases in this category. Clause-level synonym dictionaries or more sophisticated paraphrasing techniques are required for paraphrases in this category.

E) Other paraphrases

We can recognize that two expressions are the same meaning by guessing from their contexts. There were 99 (31.2%) paraphrases in this category. Taking account of paraphrases in this category is very difficult using current natural language processing technologies.

Among these paraphrases, we focused on categories A, B, C, and D, and examined several paraphrase resources for these four categories.

4. An Automatic Method of Text Evaluation using Paraphrases

In this section, we describe our text evaluation method using paraphrases. In Section 4.1., we describe the procedure for our method. In Section 4.2., we explain several paraphrase methods for categories A, B, C, and D.

4.1. Procedure for Text Evaluation

We evaluated texts using the following procedure, which resembles Zhou's ParaEval (Zhou et al., 2006).

Step 1: Search using a greedy algorithm to find (C) phrase-level or (D) clause-level paraphrases matches.

Step 2: The non-matching fragments from Step 1 are then searched using a greedy algorithm to find (A) paraphrases using different Japanese characters or (B) word-level paraphrases or synonym matches.

Step 3: Search by literal lexical unigram matching on the remaining text.

Step 4: Count the agreed words in a reference summary from Steps 1, 2, and 3, and output the Recall value for the reference summary as an evaluation score.

4.2. Paraphrase Methods

We used the following four paraphrase methods for summary evaluation.

- **SMT** (automatic): Paraphrases using the statistical machine translation (SMT) technique.
- **DS** (automatic): Paraphrases using the distributional similarity method.
- **Word** (manual): WordNet dictionary.
- **NTT** (manual): NTT Goi-Taikei dictionary.

In the following, we explain the details of each paraphrase method.

Paraphrases using the statistical machine translation technique (SMT)

If translations of two expressions X and Y are the same expression, then the expressions X and Y are considered to be paraphrases. Therefore, we constructed a translation model from 150,000 pairs of English-Japanese sentences automatically extracted (Utiyama and Isahara, 2003) from the Yomiuri newspaper database and Daily Yomiuri using a translation tool Giza++[2]. In this translation model, we deleted English-Japanese expression pairs, in which the number of words and parts of speech of each word were different. For example, we don't consider a noun phrase and a verb phrase to be a paraphrase. From the remainder of the English-Japanese expression pairs, we obtained 85,858 pairs of paraphrases.

[1] These editorials were used in Text Summarization Challenge (Fukushima et al., 2002), which is an evaluation workshop of text summarization, conducted in the NTCIR workshop.

[2] http://www.fjoch.com/GIZA++.html

Category	Paraphrase level	Number of cases	SMT	DS	WordNet	NTT	Required techniques
A	character	20.1% (64/318)	△	△	△	◎	character-level paraphrases
B	word	24.5% (78/318)	○	◎	◎		word-level paraphrases
C	phrase	11.9% (38/318)	○	△		△	phrase-level paraphrases
D	clause	12.3% (39/318)	△				clause-level paraphrases
E	others	31.2% (99/318)					paraphrases based on context analysis

Table 2: The classification of the paraphrases and necessary correspondence

Paraphrases using distributional similarity

We automatically collected paraphrases using distributional similarity in the following procedure.

1. Analyze the dependency structures of all sentences in a total of 56 years of Japanese newspapers from the Mainichi, Yomiuri, and Nikkei newspaper databases using the Japanese parser CaboCha[3].
2. Extract noun-verb pairs that have dependency relations from the dependency trees obtained in Step 1.
3. Count the frequencies of each noun-verb pair.
4. Collect verbs and their frequencies for each noun, creating indices for each noun.
5. Calculate the similarities between two indices of nouns using the SMART similarity measure (Salton, 1971).
6. Obtain a list of synonymous nouns[4].

In Step 2, we also extracted noun-phrase-verb pairs, instead of noun-verb pairs, and obtained a list of synonymous noun phrases using the same Steps 3 to 6.

As well as collecting verbs for each noun in Step 4, we similarly collected nouns for each verb, and obtained a list of synonymous verbs.

WordNet dictionary

WordNet (Bond et al., 2009) is a most widely used lexical resource in natural language processing. This database links nouns, verbs, adjectives, and adverbs to sets of synonyms (synsets) that are in turn linked through semantic relations that determine word definitions. We considered a set of words linked in the same synset as paraphrases and used them for text evaluation.

NTT Goi-Taikei dictionary

NTT Goi-Taikei is a Japanese thesaurus produced by NTT Communication Science Laboratories. In this dictionary, a list of synonymous expressions of nouns, adjectives, and verbs with different Japanese characters is included.
The four paraphrase methods are summarized in Table 1. Table 2 shows the relations between the four paraphrase methods and four categories of paraphrases in Section 3.2.

Paraphrase method	Target POS	Automatic/ Manual
SMT	All	automatic
DS	Noun, Noun Phrase, Verb	automatic
Word	Noun, Verb	manual
NTT	Noun, Verb, Adjective	manual

Table 1: Paraphrases for text evaluation

5. Experiments

To investigate the effectiveness of our method, we conducted several experiments.

5.1. Experimental Settings

Correct data sets

In addition to the 300 abstract-type summaries created from 30 editorials by 10 human subjects, we prepared another 300 extract-type summaries, which were produced by extracting important parts of the original texts. Three human subjects assigned evaluation scores manually on a one-to-four scale to each of these 600 summaries.

Alternatives

We conducted examinations using 16 proposed methods and a baseline method ROUGE-1, shown in Table 3. [5] The proposed methods used different combinations of the four kinds of paraphrases: SMT, DS, NTT, and Word.

Experimental method

We used the top extract-type summary and the top abstract-type summary in each topic as reference summaries. Then we conducted the following experiments for each topic.

[3] http://chasen.org/~taku/software/cabocha/

[4] For each noun, we extracted the top 20 similar nouns, and used them for text evaluation.

[5] We employed ROUGE-1 as a baseline method, because ROGUE-1 obtained the best performance among a series of ROUGE family in this dataset (Nanba and Okumura, 2006).

	Combination of Paraphrases	**SMT (S)**	**DS (D)**	**Word (W)**	**NTT (N)**
Our method	S	○			
	D		○		
	W			○	
	N				○
	SD	○	○		
	SW	○		○	
	SN	○			○
	DW		○	○	
	DN		○		○
	WN			○	○
	SDW	○	○	○	
	SDN	○	○		○
	SWN	○		○	○
	DWN		○	○	○
	SDWN	○	○	○	○
Baseline method	ROUGE-1				

Table 3: List of 16 proposed methods and a baseline method.

- **EX-1:** Evaluate nine extract-type summaries using the top extract-type summary as reference summaries.
- **EX-2:** Evaluate nine extract-type summaries using the top abstract-type summary as reference summaries.
- **EX-3:** Evaluate nine abstract-type summaries using the top extract-type summary as reference summaries.
- **EX-4:** Evaluate nine abstract-type summaries using the top abstract-type summary as reference summaries.

In each experiment, evaluation scores were calculated by taking the reference summary. We then ranked summaries by our methods and ROUGE-1, and compared them with a manual ranking by Spearman rank-order correlation coefficients.

5.2. Experimental Results

We show the experimental results in Tables 4 and 5, which show the Spearman rank-order correlation coefficients for the 17 methods using two extract-type reference summaries and two abstract-type reference summaries, respectively.

As can be seen from Table 4, all of our methods could evaluate abstract-type summaries more accurately than ROUGE-1. Of our 16 methods, the combination of "SDW" gave the best performance in abstract-type summaries evaluations. This combination could improve ROUGE-1 by 0.047 (15%).

In Table 5, the combination of SN was better than ROUGE-1 by 0.113 (36%) when evaluating extract-type summaries, while the combination of "SDW" was better by 0.047 (12%) when evaluating abstract-type summaries.

	Combination of Paraphrases	Extract (EX-1)	Abstract (EX-2)
Our method	S (SMT)	0.296	0.332
	D (DS)	0.368	0.328
	W (WordNet)	0.357	0.344
	N (NTT)	0.358	0.329
	SD	0.319	0.355
	SW	0.331	0.353
	SN	0.316	0.339
	DW	0.365	0.346
	DN	**0.370**	0.331
	WN	0.358	0.348
	SDW	0.333	**0.357**
	SDN	0.332	0.337
	SWN	0.321	0.342
	DWN	0.368	0.348
	SDWN	0.314	0.332
Baseline	ROUGE-1	0.358	0.310

Table 4: Evaluation results using an extract-type reference summary

	Combination of Paraphrases	Extract (EX-3)	Abstract (EX-4)
Our method	S (SMT)	0.346	0.386
	D (DS)	0.322	0.367
	W (WordNet)	0.327	0.397
	N (NTT)	0.323	0.400
	SD	0.377	0.401
	SW	0.330	0.427
	SN	**0.426**	0.394
	DW	0.302	0.364
	DN	0.317	0.368
	WN	0.321	0.402
	SDW	0.329	**0.436**
	SDN	0.394	0.419
	SWN	0.346	0.432
	DWN	0.303	0.371
	SDWN	0.358	0.426
Baseline	ROUGE-1	0.313	0.389

Table 5: Evaluation results using an abstract-type reference summary

5.3. Discussion

Effectiveness of paraphrases in text evaluation

More than half of our methods performed worse than ROUGE-1 in the experiment Ex-1, which indicates that paraphrases were not effective for the evaluation of extract-type summaries using extract-type reference summaries. On the other hand, most of our methods improved ROUGE-1 in the experiments Ex-2, 3, and 4. We considered that they were valid experimental results, because paraphrases are generally used in abstract-type summaries.

Effectiveness of SMT

In the experiment Ex-4, the combinations of "DW" and "SDW" obtained 0.364 and 0.436 of Spearman rank order correlation coefficients, respectively. This indicates that the SMT-based paraphrases contributed to improve a score from 0.364 to 0.436. We can also confirm the effectiveness of SMT-based paraphrases from "W" and "SW". To confirm the effectiveness of the SMT-based paraphrases more precisely, we calculated Spearman rank-order correlation coefficients for each topic and counted the number of topics that "SDW" and "SW" were superior to "DW" and "W", respectively. The result is shown in Table 6. As can be seen from Table 6, the SMT-based paraphrases are useful to improve the combination of "DW", because the number of topics that the combination of "SDW" improved "DW" was much larger than the opposite cases. On the other hand, the combination of "SW" impaired "W" in 13 topics. In the SMT-based paraphrases, there were cases that a term X can be paraphrased into Y, but Y cannot be paraphrased into X. A pair of "判決" (adjudication) and "敗訴" (unsuccessful litigation) is one of such paraphrases. Recently, detecting such paraphrases has studied in the field of textual entailment recognition. In future, the SMT-based paraphrases may be improved using techniques in the field.

	Improve	Same	Impair
"SDW" vs. "DW"	17 (0.56)	4 (0.13)	9 (0.30)
"W" vs. "SW"	11 (0.37)	6 (0.20)	13 (0.43)

Table 6: The Number of Topics that the combinations of SDW and SW Improved DW and W

Effectiveness of distributional similarity

Distributional similarity did not contribute to improve ROUGE-1. For example, the combinations of "SW" and "SDW" in Ex-4 obtained 0.427 and 0.436 of Spearman rank order correlation coefficients, respectively. In this case, the distributional-similarity-based paraphrases contributed to improve ROUGE-1 by only 0.009. In another case, "SDWN" impaired "SWN" from 0.432 to 0.426. The distributional similarity collected more related terms rather than synonyms. A pair of "イギリス" (England) and "フランス" (France) is one of such paraphrases. As the method expresses the senses of each noun or noun phrase with a set of verbs having dependency relations in texts, it tends to collect terms that have the same properties.

6. Conclusions

We explored the use of paraphrases for the refinement of traditional automatic methods for text evaluation. We analyzed 318 paraphrases in texts to be evaluated, and classified them into five categories. Then we examined several paraphrase methods that covered four of those categories (about 70% of the 318 paraphrases). To confirm the effectiveness of our method, we conducted some experiments using the data from the Text Summarization Challenge 2. We found that the use of the combination of three kinds of paraphrases (SMT, distributional similarity, and WordNet) improved the traditional evaluation method ROUGE-1 from 0364 to 0.436.

References

Bond, F., Isahara, H., Uchimoto, K., Kuribayashi, T. and Kanzaki, K. (2009). Extending the Japanese WordNet. *Proc. 15th Annual Meeting of the Association for Natural Language Processing*, pp. 80-83.

Fukushima, T., Okumura, M. and Nanba, H. (2002). Text Summarization Challenge 2 / Text Summarization Evaluation at NTCIR Workshop 3. *Working Notes of the 3rd NTCIR Workshop Meeting, PART V*, pp. 1-7.

Nanba, H. and Okumura, M. (2006). An Automatic Method for Summary Evaluation Using Multiple Evaluation Results by a Manual Method. *Proc. COLING/ACL 2006 Main Conference Poster Sessions*, pp. 603-610.

Kanayama, H. (2003). Paraphrasing Rules for Automatic Evaluation of Translation into Japanese. *Proc. First International Workshop on Paraphrasing*, pp. 88-93.

Kauchak, D. and Barzilay, R. (2006). Paraphrasing for Automatic Evaluation. *Proc. 2006 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 455-462.

Lee, L. (1999). Measures of Distributional Similarity. *Proc. 37th Annual Meeting of the Association for Computational Linguistics*, pp. 25-32.

Lin, C. Y. and Hovy, E. (2003). Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. *Proc. 4th Meeting of the North American Chapter of the Association for Computational Linguistics and Human Language Technology*, pp. 150-157.

Lin, C. Y. (2004). ROUGE A Package for Automatic Evaluation of Summaries. *Proc. ACL-04 Workshop "Text Summarization Branches Out"*, pp. 74-81.

Lin, D. (1998). Automatic Retrieval and Clustering of Similar Words. *Proc. 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pp. 768-774.

Papineni, K., Roukos, S., Ward, T. and Zhu, W. J. (2002) BLEU: a Method for Automatic Evaluation of MachineTranslation. *Proc40 thAnnual Meeting ofthe Association for Computational Linguistics*, pp. 311-318.

Salton, G. (1971). The SMART Retrieval System - Experiments in Automatic Document Processing. *Prentice-Hall, Inc., Upper Saddle River, NJ.*

Utiyama, M. and Isahara, H. (2003). Reliable Measures for Aligning Japanese-English News Articles and Sentences. *Proc. 41st Annual Meeting of the Association for Computational Linguistics*, pp. 72-79.

Yves, L. and Etienne, D. (2005). Automatic Generation of Paraphrases to be used as Translation References in Objective Evaluation Measures of Machine Translation. *Proc. Third International Workshop on Paraphrasing.*

Zhou, L., Lin, C. Y., Munteanu, D. S. and Hovy, E. (2006). ParaEval Using Paraphrases to Evaluate Summaries Automatically. *Proc. 2006 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 447-454.