Hiroshima City University at Evaluation Subtask in the NTCIR-8 Patent Translation Task

Hidetsugu Nanba
Graduate School of Information Sciences, Hiroshima City University
3-4-1 Ozukahigashi, Hiroshima
731-3194, Japan

Kazuho Hirahara
Graduate School of Information Sciences, Hiroshima City University
3-4-1 Ozukahigashi, Hiroshima
731-3194, Japan

Toshiyuki Takezawa
Graduate School of Information Sciences, Hiroshima City University
3-4-1 Ozukahigashi, Hiroshima
731-3194, Japan

ABSTRACT
The evaluation of computer-produced texts is 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 explore the use of paraphrases for the refinement of traditional automatic methods for text evaluation. In our previous work, we devised an evaluation method for text summarization using multiple paraphrase methods. Our goal in NTCIR-8 is to confirm the effectiveness of our method for machine translation. We evaluated 1200 computer-produced translations by six proposed methods and two baseline methods, and confirmed the effectiveness of our methods.

Categories and Subject Descriptors
H.3.3 **[Information Search and Retrieval]**: Search process

H.3.4 [**Systems and Software**]: Performance evaluation

H.3.5 [**Online Information Services**]: Data sharing

General Terms
Measurement, Performance, Experimentation

Keywords
text evaluation, paraphrase, machine translation

1. INTRODUCTION
The evaluation of computer-produced texts is 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 have been proposed in text summarization (Zhou et al., 2006) and machine translation (Kauchak and Barzilay, 2006; Kanayama, 2003; Yves and Etienne, 2005), and their effectiveness has been confirmed. In our previous work, we also proposed an evaluation method for text summarization using multiple paraphrase methods (Hirahara et al., 2009). Our goal in NTCIR-8 (Fujii et al., 2010) is to confirm the effectiveness of the method in machine translation.

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

2. RELATED WORK
We describe the related studies of "automatic evaluation of texts" and "text evaluation using para-phrases" 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, gramN, and Countmatch(gramN) 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 translations, and use ROUGE-1 as a baseline method.

2.2 Text Evaluation Using Paraphrases
Several evaluation methods using paraphrases have been proposed in text summarization (Zhou et al., 2006, Hirahara et al., 2009)

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, the researchers automatically obtained paraphrases from a translation model, the paraphrases were 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 a paraphrase method.

In addition to the SMT-based paraphrases, Hirahara (2009) examined other three paraphrase methods: distributional similarity (Lin, 1998; Lee, 1999), WordNet dictionary, and NTT GoiTaikei dictionary (Hirahara et al., 2009), and experimentally confirmed the effectiveness of their method for evaluating sum-maries written in Japanese. In our work, we applied Hirahara's method to the evaluation of computer-produced translations written in English.

3. AN AUTOMATIC EVALUATION OF TEXTS USING PARAPHRASES

In this section, we describe our text evaluation method using paraphrases based on Hirahara's method (Hirahara et al. 2009). In Section 3.1, we de-scribe the procedure for our method. In Section 3.2, we explain two paraphrase methods.

3.1 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 phrase-level or clause-level paraphrase matches.

Step 2: The non-matching fragments from Step 1 are then searched using a greedy algorithm to find 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 translation from Steps 1, 2, and 3, and output the Recall value for the reference translation as an evaluation score.

3.2 Paraphrase Methods

We used the following two paraphrase methods for evaluation of computer-produced translations.

- SMT (automatic): Paraphrases using the SMT technique.
- WN (manual): WordNet dictionary.

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

3.2.1 Paraphrases using the SMT technique

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 1,800,000 pairs of English and Japanese sentences automatically extracted from patent documents published during 1993-2000 (Fujii et al., 2008) using the translation tool Giza++[1]. 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 do not consider a noun phrase to be a paraphrase of a verb phrase.

3.2.2 WordNet dictionary (WN)

WordNet[2] is a very widely used lexical resource in natural language processing. This database links nouns, verbs, adjectives, and adverbs to sets of synonyms (synsets) that are linked in turn through semantic relations that determine word definitions. We considered a set of words linked in the same synset as paraphrases and used them for evaluation.

4. EVALUATION

4.1 Experimental Method

4.1.1 Data

We used 1200 English sentences, which were translated from 100 Japanese sentences by 12 machine translation systems (Fujii et al., 2008).

4.1.2 Alternatives

We examined the following six proposed methods and two baseline methods. Here, "Tagger" indicates that all words in each translation were lemmatized by the part-of-speech tagging tool TreeTagger[3].

Our methods

- HCU-3 (S+T): ROUGE+SMT+Tagger
- HCU-4 (S): ROUGE+SMT
- HCU-5 (W+T): ROUGE+WN+Tagger
- HCU-6 (W): ROUGE+WN
- HCU-7 (SW+T): ROUGE+SMT+WN+Tagger
- HCU-8 (SW): ROUGE+SMT+WN

Baseline methods

- HCU-1 (base+T): ROUGE+Tagger
- HCU-2 (base): ROUGE

4.1.3 Evaluation

In each experiment, evaluation scores were calculated by taking the reference translation. We then ranked the 12 computer-produced translations by our methods and baseline methods, and compared them with manual ranking[4] using Spearman rank-order correlation coefficients and Pearson's correlation coefficient. The details of the data and the evaluation procedure were described in the overview paper (Fujii et al., 2010).

---

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

[2] http://wordnet.princeton.edu/

[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[4] Computer-produced sentences were ranked in terms of adequacy and fluency.

4.2 Results and Discussion

The experimental results are shown in Tables 1-4. In the following, we discuss these results.

Effect of lemmatization

As can be seen from Tables 1 and 3, the values of HCU-1, 3, 5, and 7 in the row of "ALL" are higher than those of HCU-2, 4, 6, and 8, respectively. The former four methods used lemmatization, and this indicates that lemmatization is effective in our evaluation.

Effect of paraphrasing

The WordNet dictionary is considered to be a useful paraphrase method, because HCU-5 (W+T) is the only method that performed better than two baseline methods in the evaluation of adequacy (the row of "ALL" in Tables 1 and 3). Although, SMT-based paraphrases could also improve baseline methods in the evaluations of several systems (e.g., HCU-5 for system 2 in Table 1), the overall performances of our methods using the SMT-based paraphrases (HCU3, 4, 7, and 8) was worse than that of the two baseline methods.

Effect of our methods on fluency

Our methods performed worse than the baseline methods in the evaluation of fluency, because our methods were originally developed for the evaluation of text summarization. Traditionally, the creation of extract-type summaries has been considered an important research problem in text summarization, and researchers in this field have focused on the evaluation of summaries in terms of adequacy using word-level matches. In our evaluation procedure, we also employed word-level matches in Steps 2 and 3, which we described in Section 3.1. If we employ combinations of n-gram matches, such as BLEU (Papineni, 2002) instead of word-level matches, our methods might be improved in the evaluation of fluency.

5. CONCLUSIONS

We participated in the evaluation subtask in the NTCIR-8 Patent Translation Task. We constructed six proposed methods using paraphrase methods and compared them with two baseline methods. From the experimental results, we confirmed that one of our methods HCU-5, which used the WordNet dictionary as a paraphrase method, was an improvement over the baseline methods.

6. REFERENCES

[1] Fujii, A., Utiyama, M., Yamamoto, M., Utsuro, T., Ehara, T., Echizen-ya, H., and Shimohata, S. 2010. Overview of the Patent Translation Task at the NTCIR-8 Workshop. In *Proc. 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access.*

[2] Fujii, A., Utiyama, M., Yamamoto, M., and Utsuro, T. 2008. Overview of the Patent Translation Task at the NTCIR-7 Workshop. In *Proc. 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, 389-400.

[3] Hirahara, K., Nanba, H., Takezawa, T., and Okumura, M. 2009. Automatic Evaluation of Texts by Using Paraphrase. In *Proc. 4th Language & Technology Conference*. LTC'09. 370-374.

[4] Kanayama, H. 2003. Paraphrasing Rules for Automatic Evaluation of Translation into Japanese. In *Proc. First International Workshop on Paraphrasing*, 88-93.

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

[6] Lee, L. 1999. Measures of Distributional Similarity. In *Proc. 37th Annual Meeting of the Association for Computational Linguistics*, 25-32.

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

[8] Lin, C. Y. 2004. ROUGE A Package for Automatic Evaluation of Summaries. In *Proc. ACL-04 Work-shop "Text Summarization Branches Out"*, 74-81.

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

[10] Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics*, 311-318.

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

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

Table 1. Pearson's correlation coefficient in adequacy

	system 1	system 2	system 3	system 4	system 5	system 6	system 7
HCU-1 (base+T)	0.3622	0.1091	0.3224	0.2592	0.1718	0.2752	0.4695
HCU-2 (base)	0.2993	0.0788	0.2551	0.2461	0.1854	0.2322	0.3462
HCU-3 (S+T)	0.3203	**0.1116**	**0.3897**	0.1669	0.1312	0.1993	0.4424
HCU-4 (S)	0.2718	0.0497	0.2556	0.1742	0.1405	0.1934	0.3431
HCU-5 (W+T)	0.3586	**0.1170**	0.2993	**0.2637**	**0.1933**	**0.2760**	0.4181
HCU-6 (W)	0.2880	0.0766	0.2179	0.2556	**0.1949**	0.2365	0.3112
HCU-7 (SW+T)	0.3043	**0.1244**	**0.3600**	0.1635	0.1795	0.1923	0.3948
HCU-8 (SW)	0.2538	0.0458	0.2057	0.1718	0.1651	0.1786	0.3185

	system 8	system 9	system 10	system 11	system 12	Avg.	All
HCU-1 (base+T)	0.2400	0.3564	0.2905	0.3759	0.3583	0.2992	0.2463
HCU-2 (base)	0.2546	0.4485	0.1979	0.2457	0.3710	0.2634	0.1977
HCU-3 (S+T)	0.1870	0.3634	0.2143	0.2858	0.3315	0.2619	0.2211
HCU-4 (S)	0.1666	**0.4771**	0.1539	0.1781	0.3484	0.2294	0.1802
HCU-5 (W+T)	0.2228	0.3828	**0.2933**	0.3436	**0.3935**	0.2968	**0.2507**
HCU-6 (W)	0.2364	**0.4645**	0.1933	0.2187	**0.3902**	0.2570	0.1990
HCU-7 (SW+T)	0.1607	0.3669	0.2240	0.2628	0.3770	0.2592	0.2217
HCU-8 (SW)	0.1502	0.4683	0.1543	0.1608	0.3895	0.2219	0.1772

Table 2. Pearson's correlation coefficient in fluency

	system 1	system 2	system 3	system 4	system 5	system 6	system 7
HCU-1 (base+T)	0.3451	0.1682	0.3751	0.1434	0.2711	0.1642	0.4429
HCU-2 (base)	0.3074	0.0346	0.3208	0.2295	0.3108	0.1558	0.3947
HCU-3 (S+T)	0.2802	0.1635	**0.3811**	0.0412	0.2541	0.0953	0.3998
HCU-4 (S)	0.2621	0.0644	0.2572	0.1395	0.2865	0.1281	0.3784
HCU-5 (W+T)	0.3199	0.1472	0.3559	0.1225	0.2724	0.1578	0.3931
HCU-6 (W)	0.2834	0.0546	0.2882	0.2083	0.3100	0.1572	0.3591
HCU-7 (SW+T)	0.2538	0.1640	0.3419	0.0384	0.2807	0.0945	0.3558
HCU-8 (SW)	0.2400	0.0602	0.2131	0.1302	0.2933	0.1165	0.3474

	system 8	system 9	system 10	system 11	system 12	Avg.	All
HCU-1 (base+T)	0.1153	0.2735	0.1594	0.2945	0.3775	0.2608	0.2285
HCU-2 (base)	0.1685	0.2861	0.1357	0.2311	0.2941	0.2391	0.1976
HCU-3 (S+T)	0.0637	0.2604	0.0664	0.2018	**0.3803**	0.2156	0.1949
HCU-4 (S)	0.1095	**0.3015**	0.0544	0.1585	0.2890	0.2024	0.1711
HCU-5 (W+T)	0.1149	**0.3208**	**0.1626**	0.2515	**0.3936**	0.2510	0.2243
HCU-6 (W)	0.1541	**0.3111**	0.1268	0.1937	0.2956	0.2285	0.1898
HCU-7 (SW+T)	0.0696	0.2908	0.0813	0.1577	**0.3972**	0.2105	0.1923
HCU-8 (SW)	0.1068	**0.3013**	0.0718	0.1195	0.3028	0.1919	0.1639

Table 3. Spearman's rank correlation coefficient in adequacy

	system 1	system 2	system 3	system 4	system 5	system 6	system 7
HCU-1 (base+T)	0.3154	0.1128	0.3380	0.2231	0.1491	0.2296	0.4561
HCU-2 (base)	0.2404	0.1276	0.2367	0.2040	0.1375	0.1701	0.3282
HCU-3 (S+T)	0.2811	0.1112	**0.4012**	0.1632	0.1063	0.1652	0.4338
HCU-4 (S)	0.2150	0.0593	0.2347	0.1373	0.0692	0.1315	0.3155
HCU-5 (W+T)	**0.3281**	0.1121	0.3121	**0.2314**	**0.1577**	**0.2342**	0.3946
HCU-6 (W)	0.2334	0.1155	0.1964	**0.2371**	0.1328	0.1696	0.2977
HCU-7 (SW+T)	0.2783	0.1101	**0.3618**	0.1456	0.1448	0.1364	0.3738
HCU-8 (SW)	0.1923	0.0410	0.1818	0.1357	0.0838	0.1086	0.2934

	system 8	system 9	system 10	system 11	system 12	Avg.	All
HCU-1 (base+T)	0.2324	0.2439	0.2427	0.3535	0.3578	0.2712	0.2234
HCU-2 (base)	0.2878	0.3745	0.1665	0.2267	0.3538	0.2378	0.1654
HCU-3 (S+T)	0.1631	0.2865	0.2115	0.2456	**0.3616**	0.2442	0.2065
HCU-4 (S)	0.1581	**0.4012**	0.1199	0.1575	0.3475	0.1955	0.1411
HCU-5 (W+T)	0.2097	0.2963	**0.2632**	0.2989	**0.4072**	0.2705	**0.2274**
HCU-6 (W)	0.2786	**0.4050**	0.1727	0.2013	**0.3719**	0.2343	0.1673
HCU-7 (SW+T)	0.1544	0.3046	0.2238	0.2271	**0.4182**	0.2399	0.2042
HCU-8 (SW)	0.1438	**0.4027**	0.1183	0.1449	**0.3777**	0.1853	0.1363

Table 4. Spearman's rank correlation coefficient in fluency

	system 1	system 2	system 3	system 4	system 5	system 6	system 7
HCU-1 (base+T)	0.3071	0.1750	0.3783	0.0904	0.2516	0.1370	0.4445
HCU-2 (base)	0.2442	0.0078	0.2581	0.2167	0.2418	0.1192	0.3655
HCU-3 (S+T)	0.2328	**0.1875**	**0.3846**	0.0343	0.2269	0.0682	0.4009
HCU-4 (S)	0.2132	0.0296	0.1861	0.1162	0.2132	0.0844	0.3356
HCU-5 (W+T)	0.2931	0.1481	0.3676	0.0944	0.2317	0.1269	0.3925
HCU-6 (W)	0.2129	0.0282	0.2305	0.2098	0.2361	0.0931	0.3287
HCU-7 (SW+T)	0.2061	**0.1853**	0.3406	0.0294	0.2356	0.0525	0.3448
HCU-8 (SW)	0.1925	0.0310	0.1371	0.1205	0.2024	0.0507	0.2969

	system 8	system 9	system 10	system 11	system 12	Avg.	All
HCU-1 (base+T)	0.1328	0.2364	0.1101	0.3039	0.4155	0.2486	0.2126
HCU-2 (base)	0.2383	0.2432	0.1287	0.2206	0.3134	0.2165	0.1705
HCU-3 (S+T)	0.0776	0.2325	0.0571	0.1948	0.4235	0.2101	0.1868
HCU-4 (S)	0.1465	**0.2460**	0.0417	0.1485	0.3167	0.1731	0.1400
HCU-5 (W+T)	0.1299	**0.2968**	**0.1344**	0.2494	0.4142	0.2399	0.2064
HCU-6 (W)	0.2162	**0.2782**	0.1261	0.1948	0.2853	0.2033	0.1596
HCU-7 (SW+T)	0.0847	**0.2573**	0.0836	0.1443	**0.4222**	0.1989	0.1786
HCU-8 (SW)	0.1257	**0.2516**	0.0689	0.1129	0.2960	0.1572	0.1265