

Korean paraphrase evaluation using morphological analysis

Sung Won Moon*, Hojin Choi**, Gahgene Gweon***

*Dep. Of Knowledge Service Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Rep. Of Korea

** Dep. Of Computer Science, Korea Advanced Institute of Science and Technology, Daejeon, Rep. Of Korea

***Corresponding Author: Dep. Of Knowledge Service Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Rep. Of Korea

augustmoon@kaist.ac.kr, hojinc@kaist.ac.kr, ggweon@kaist.ac.kr

Abstract— Paraphrase evaluation is used to determine whether two input sentences share a same meaning. The automatic analysis for paraphrase evaluation technology has a potential use in the area of information retrieval technology since correctly paraphrased sentences can be used as alternative input sentence in the retrieval process. In this paper, we suggest an evaluation method that is suitable for the Korean language by using morphological analysis. Using our method, we present preliminary results on how our automatic evaluation scores compare to the human evaluation scores.

Keywords— Morphological analysis, Paraphrase evaluation

I. INTRODUCTION

Paraphrase evaluation is used to determine whether two input sentences share a same meaning. In this paper, we suggest an automatic evaluation method given a pair of paraphrased sentences. An evaluation is a necessary step in order to use paraphrased sentences in various applications, since we need to first determine whether the paraphrased sentences are suitable for such usage. Thus far, research on paraphrase evaluation has been largely conducted for the English language [1, 2, 3]. Therefore, for other languages, such as Korean, modified versions of evaluation methods used in the English language has been used. However, such methods are not ideal considering that the two languages differ in structure. Given such need, we suggest a Korean evaluation method in paraphrase using morphological analysis.

An automatic evaluation for paraphrase technology has a potential use in the area of information retrieval technology since paraphrased sentences can be used as alternative input sentence in the retrieval process. For instance, if a user wants to know what countries U2 has held their concert, he can ask, "What nations have U2 held concert in". A google search with this query sentence yields 19,200,000 answers. However, the top ten returned pages do not contain the answer. Instead, if a paraphrased sentence "What countries have U2 played in?" is used as a query sentence, google outputs 1, 240,000 answers. Although the number of answers has decreased by tenfold, the top page among the returned pages lists the intended answer by listing all the countries that U2 has played a concert. Figure 1 shows the screen capture from the google search engine using these two paraphrase sentences.

As seen in the example above, despite the advancements in hardware and information retrieval techniques, there is room for improvement in the area of information retrieval by using alternative search query using paraphrase techniques. Obtaining a desired answer to a query is not just a matter of increased number of returned answers, but more about providing a query that captures a user's intention. Therefore, evaluation of the resulting paraphrased sentences is also an important research field that can be used in the process of information retrieval along with research on automatic paraphrase techniques. In addition, the paraphrase evaluation technique can be also used in natural language processing applications such as text-to-text generation or information extraction applications [4].

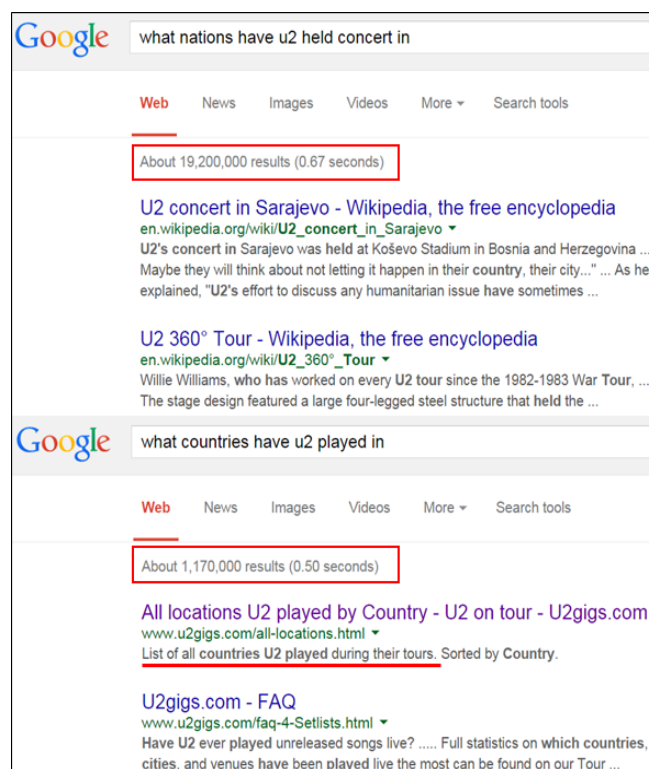


Figure 1. Example of different results using paraphrased sentences

Our proposed evaluation methodology is suitable for the Korean language, which is an agglutinative language. This method is valuable in that much of the work in paraphrase evaluation has been geared towards inflectional language, such as English. We consider two characteristics of agglutinative language in our suggested method; use of endings and postpositions. In agglutinative languages, “endings” are added at the end of a verb to convey information such as tense, mood, or social relationships between speaker and listener. In contrast, an inflectional language, such as English makes extensive use of auxiliaries to convey such information. In addition to ending, postpositions are used in agglutinative languages to determine a case of a noun or relationship between multiple words. Since agglutinative language has many variations using endings and postpositions, using the same evaluation methodology as those used in inflectional language yields evaluation results that are too strict. Therefore our proposed method discards endings and postpositions to reflect the characteristics of agglutinative languages.

In the rest of this paper, we first present existing work on paraphrase evaluation. Next, a simple experiment that is designed to assess our proposed evaluation method is presented followed by the results of the experiment.

II. BACKGROUND

The task of paraphrase evaluation has recently been receiving increased attention along with the advancement in the task of paraphrase generation [5]. The evaluation task is one of the three main subtasks of the generation process, namely extraction, recognition/ evaluation, and generation [6]. The paraphrase evaluation task can be conducted either manually or automatically. Table 1 summarizes previous research on paraphrase evaluation using both manual and automatic evaluation processes.

TABLE 1. RESEARCH ON PARAPHRASE EVALUATION

First author	Year	Method
Bangalore [7]	2000	String accuracy, Bag accuracy
Barzilay [8]	2002	Readability, Fidelity
Papineni [1]	2002	N-gram precision (BLEU)
Fujita [9]	2004	Longest Common Subsequence(LCS)
Glickman [10]	2004	A case study for verbs
Dolan [11]	2005	Correctness
Snover [2]	2006	Translation Edit Rate (TER)
Callison-burch [12]	2008	5-point Likert scale
Snover [13]	2009	TER, Stem match, Synonym match
Chen [3]	2011	Paraphrase In N-gram Changes (PINC)
Fujita [14]	2012	5-point Likert scale

For a manual evaluation, human evaluators assess the quality of paraphrased sentences according to the provided standards. The main advantage of a manual evaluation is that human judgement that is difficult to capture with rules, such as idioms or jokes, can be correctly evaluated. However, due to the nature of human judgement, different evaluators will assign varying scores even if they are using the same

standards. Therefore, a method for checking the reliability of the evaluated scores, such as a correlation analysis, should follow a manual evaluation process. In addition to variations in judgement, a manual evaluation has limitations in terms of the number of sentences that can be evaluated due to time and monetary constraints.

An automatic evaluation can address some of the disadvantages of a manual evaluation. Compared to a manual evaluation, an automatic evaluation is cheaper and can yield results that are more consistent when the system is provided with a clear set of rules. However, it is difficult to provide such clear rules and there exists limitations on the level of expressions that machines can comprehend. Despite such difficulties, as the amount of data that needs evaluation increases, it becomes infeasible for researchers to conduct a manual evaluation. Therefore, attempts in improving the accuracy of automatic evaluations are increasing.

Automatic paraphrase evaluation can be conducted at syntactic or semantic level. In this paper, we suggest a method that employs syntactic features in a sentence. One popular syntactic feature that has been used in previous work is n-gram. For example, Chen proposed an evaluation method called the “Paraphrase In N-gram Changes (PINC)”. The equation used for PINC is shown in Figure 2 for reference [3].

$$PINC(s, c) = \frac{1}{N} \sum_{n=1}^N 1 - \frac{|n\text{-gram}_s \cap n\text{-gram}_c|}{|n\text{-gram}_c|}$$

Figure 2. The equation to evaluate paraphrase in PINC

Although using n-gram as a feature is an effective and simple method, such a method is difficult to adapt for an agglutinative language like Korean. In the agglutinative language, a word can take many different syntactic forms by using various endings or postpositions. Therefore, using n-gram as a feature would be too strict of a rule for detecting words that are syntactically different but are semantically identical in Korean. Therefore a different evaluation method adapted to the agglutinative language in paraphrase needs to be developed.

III. METHODS

This section describes the data source as well as the details on how the manual and automatic evaluations are conducted.

A. Data source

As stated in the introduction, one application of the paraphrased sentences is using them as alternative input sentences in question and answering (QA) systems. Thus, we selected one hundred question sentences from a QA system, which is a Korean quiz show called “Janghak quiz”. These hundred sentences were given as inputs to an automatic paraphrase generation system to produce paraphrased pairs that were evaluated using manual and automatic evaluation methods.

B. Manual evaluation

To validate the quality of the paraphrased Korean sentences generated by the automatic system, we conducted manual evaluation. Three human evaluators were asked to assess one hundred sentences using the guidelines proposed by Callison-burch et al [12]. Figure 3 shows the specific guidelines. A correlational analysis of the scores from the three evaluators was conducted to assess the consistency between them. The resulting r-value is 0.699, $p < 0.1$. Thus, the correlation value suggests that our manual evaluations results are reliable.

MEANING	
5	All of the meaning of the original phrase is retained, and nothing is added
4	The meaning of the original phrase is retained, although some additional information may be added but does not transform the meaning
3	The meaning of the original phrase is retained, although some information may be deleted without too great a loss in the meaning
2	Substantial amount of the meaning is different
1	The paraphrase doesn't mean anything close to the original phrase

Figure 3. Evaluator rated paraphrases along 5-point Likert scale

C. Automatic evaluation

In this section, we explain how we calculate the scores used in automatic evaluation given a pair of paraphrase sentences. Figure 4 shows an example pair of sentence that will be used to explain the evaluation process. Sentence 1 is the original input sentence and sentence 2 is the paraphrased sentence.

To implement automatic evaluation of paraphrased sentences using morphological analysis, in the first step, we extracted morphemes for an input sentence using ETRI morpheme analyser. Next, we removed morphemes that are labeled as endings and postpositions and kept the ones that carries real meaning in a sentence, such as nouns or predicates. The highlighted words in Figure 4 indicate morphemes that are kept after removing the unnecessary morphemes. Note that the unhighlighted words are discarded because these do not involve essential meaning of the sentence, but merely function to make a sentence grammatically correct.

Sentence 1: 어느 국가 에서 U2 가 공연 을 열었 는가?
[what nations have U2 held concert in?]
Sentence 2: 어떤 나라 에서 U2 는 연주 를 하 였는가?
[what countries have U2 played in?]

Figure 4. Example of paraphrased sentences

After identifying the critical morphemes that carry meanings, we calculate the evaluation score according to the equation in Figure 5. The numerator is the number of synonyms shared in sentence 1 and sentence 2. In our example

sentences, the word pairs “countries/ nations”, “U2/U2”, “held concert/ played” are the shared synonyms. Thus, the numerator is three. The denominator is the number of meaningful morphemes in sentence 1, which is four in our example sentence. Therefore, the evaluation score for the paraphrased sentences in our example is 0.75. Note that according to the formula presented in figure 5, the evaluation score can range from 0 to 1, unlike the manual evaluation score, which ranges from 1 to 5.

$$\text{Score} = \frac{\text{Number of matched morphemes}}{\text{Number of meaningful morphemes in one sentence}}$$

Figure 5. Equation of the paraphrase evaluation

IV. RESULTS AND DISCUSSIONS

The mean value of the manual evaluation scores for the one hundred sentences is 2.89. (s.d=1.21). The mean value of the automatic evaluation scores is 0.76 (s.d=0.17). Since the scales for the manual and automatic evaluation score are difficult to compare due to the difference in scales, we normalized each score by converting them into a z-score. The correlation between the normalized manual score and automatic score is $r = 0.332$, $p < .002$. The positive medium correlation between the two scores suggest that our automatic evaluation score captures some part of human evaluation scores. However, we acknowledge that at the current state, the correlation is not accurate enough for use in place of human evaluation.

Korean paraphrase evaluation is an important research area that has much room for improvement. Research from this area can help improve the effectiveness in information retrieval technology. Although the Korean evaluation method that we suggested in this paper is a beginning step towards automatic paraphrase evaluation, the positive correlation shows promise. For the future work, we plan to extract more features of agglutinative language for use in evaluation. For example, our current method uses a single morpheme as a feature. A more advanced approach would be considering multiple morphemes that constitute a meaningful unit as a feature. Additionally, building a more comprehensive dictionary for synonyms using a larger corpus would increase the accuracy of our technique.

ACKNOWLEDGMENT

This work was supported by ICT R&D program of MSIP/IITP. [R0101-15-0062, Development of Knowledge Evolutionary WiseQA Platform Technology for Human Knowledge Augmented Services]

REFERENCES

- [1] Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.
- [2] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. A study of translation edit rate with targeted human annotation. In

- Proceedings of association for machine translation in the Americas, pp. 223-231, 2006.
- [3] Chen, David L., and William B. Dolan. "Collecting highly parallel data for paraphrase evaluation." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011.
 - [4] Madnani, Nitin, and Bonnie J. Dorr. "Generating phrasal and sentential paraphrases: A survey of data-driven methods." Computational Linguistics 36.3: 341-387, 2010.
 - [5] Liu, Chang, Daniel Dahlmeier, and Hwee Tou Ng. "PEM: A paraphrase evaluation metric exploiting parallel texts." Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010.
 - [6] Androutsopoulos, Ion, and Prodromos Malakasiotis. "A survey of paraphrasing and textual entailment methods." Journal of Artificial Intelligence Research: 135-187, 2010.
 - [7] Bangalore, Srinivas, and Owen Rambow. Corpus-based Lexical Choice in Natural Language Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2000.
 - [8] Barzilay, Regina, and Lillian Lee. Bootstrapping Lexical Choice via Multiple-Sequence Alignment Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.
 - [9] Fujita, A., Furihata, K., Inui, K., Matsumoto, Y., & Takeuchi, K. Paraphrasing of Japanese light-verb constructions based on lexical conceptual structure. In Proceedings of the Workshop on Multiword Expressions: Integrating Processing (pp. 9-16). Association for Computational Linguistics, 2004.
 - [10] Glickman, O., & Dagan, I. Acquiring lexical paraphrases from a single corpus. Recent Advances in Natural Language Processing III. John Benjamins Publishing, Amsterdam, Netherlands, 81-90, 2004.
 - [11] Dolan, W. B., & Brockett, C. Automatically constructing a corpus of sentential paraphrases. In Proc. of IWP, 2005.
 - [12] Callison-Burch, Chris. "Syntactic constraints on paraphrases extracted from parallel corpora." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008.
 - [13] Snover, M. G., Madnani, N., Dorr, B., & Schwartz, R. TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation*, 23(2-3), 117-127, 2009.
 - [14] Fujita, A., Isabelle, P., & Kuhn, R. Enlarging paraphrase collections through generalization and instantiation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 631-642). Association for Computational Linguistics, 2012



Sungwon Moon is a master student in the department of knowledge service engineering, Korea Advanced Institute of Science and Technology (KAIST). He received his Bachelor's degree from Handong Global University, Korea (2013). His current interests are in the area of natural language processing, data mining and computer supported collaborative learning.



Ho-Jin Choi is an associate professor in KAIST Computer Science Dept. In 1995, he received a PhD in artificial intelligence (AI) from Imperial College London. For 1982-89 he worked for DACOM, for 1997-2002 as a faculty member at Korea Aerospace University, and for 2002-09 at Information and Communications University. For 2003-10, he was an adjunct professor of Carnegie Mellon University for Master of Software Engineering program. For 2006-08, he was the Director of Institute for IT Gifted Youth. Since 2010, he participates in Systems Biomedical Informatics Research Center at Seoul National University Medical School. He is in the boards of directors for Software Engineering Society of Korea, for Artificial Intelligence Society of Korea, and for Korean Society of Medical Informatics. His current interests include AI, knowledge engineering, data mining, topic mining, intelligent personal assistant, natural language QA, and biomedical informatics.



Gahgene Gweon is an assistant professor in the department of knowledge service engineering, Korea Advanced Institute of Science and Technology (KAIST). She received her PhD in human computer interaction from Carnegie Mellon University, United States (2012). Her current interests are in the area of human computer interaction, computer supported collaborative learning, and natural language processing.