APEM: Automatic paraphrase evaluation using morphological analysis for the Korean language

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

*Dep. Of Knowledge Service Engineering, KAIST, Daejeon, Rep. Of Korea

** Dep. Of Computer Science, KAIST, Daejeon, Rep. Of Korea

*** Knowledge Mining Research Team, ETRI, Daejeon, Rep. Of Korea

augustmoon@kaist.ac.kr, hojinc@kaist.ac.kr, ggweon@kaist.ac.kr, jeonghur@etri.re.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 sentences in the retrieval process. In this paper, we suggest an automatic paraphrase evaluation method using morphological analysis (APEM), which is suitable for the Korean language. Using APEM and its variations, we present preliminary results on how our automatic evaluation scores compare to the existing method of bilingual evaluation understudy (BLEU).**

***Keyword*— 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 have been used. However, such methods are not ideal considering that the two languages differ in structure. Given such need, we suggest an automatic paraphrase evaluation method using morphological analysis (APEM) for the Korean language. Automatic evaluation for paraphrase technologies has potential use in the area of information retrieval technology since paraphrased sentences can be used as alternative input sentences in the retrieval process. For instance, if a user wants to know in what countries U2 has held concerts, he can ask, "What nations have U2 held concerts 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 where U2 has played a concert. Figure 1 shows the screen capture from the Google search engine using these two paraphrased sentences.

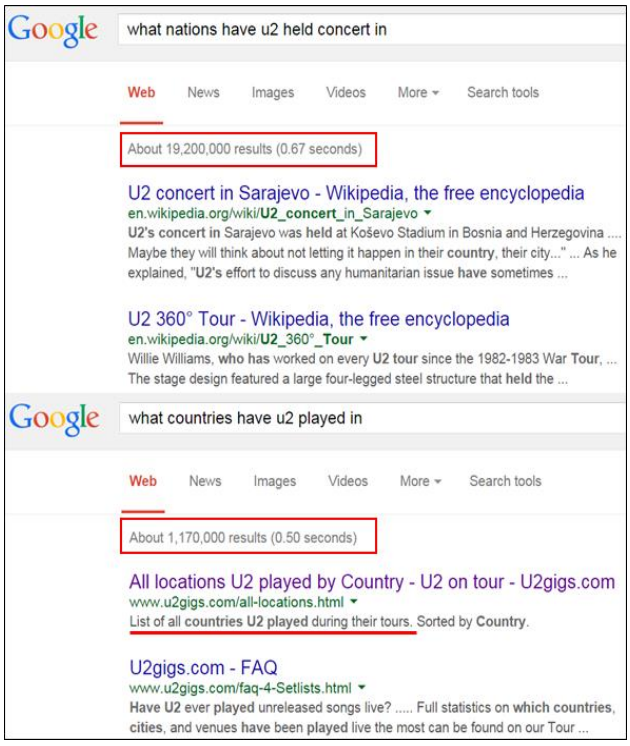

Fig 1. Example of different results using paraphrased sentences

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 phrases. Obtaining a desired answer to a query is not a matter of increased number of returned answers,

Manuscript received June 10, 2015. This work was supported in part by ICT R&D program of MSIP/IITP, Grant ID : R0101-15-0062, Development of Knowledge Evolutionary WiseQA Platform Technology for Human Knowledge Augmented Services.

S. Moon is with the Department of Knowledge Service Engineering, Korean Advanced Institute of Science and Technology, Daejeon, Rep, of Korea (phone: 82-10-9028-8724; e-mail: augustmoon@kaist.ac.kr).

G. Gweon is with the Department of Knowledge Service Engineering, Korean Advanced Institute of Science and Technology, Daejeon, Rep, of Korea (corresponding author phone: 82-42-350-1618; fax: 82-42-350-1610; e-mail: ggweon@kaist.ac.kr).

H. Choi is with the Department of Computer Science, Korean Advanced Institute of Science and Technology, Daejeon, Rep, of Korea (phone: 82-42-350-3561; fax: 82-42-350-3510; e-mail: hojinc@kaist.ac.kr).

J. Heo is with the Automatic Speech Translation and Knowledge Analytics Research Center, Electronics and Telecommunications Research Institute (ETRI), Daejeon, Rep. of Korea (phone:82-42-860-6870; fax:82-42-860-4889; e-mail: jeonghur@etri.re.kr)

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 used in natural language processing applications such as text-to-text generation or information extraction applications [4].

Our proposed evaluation methodology of APEM is suitable for the Korean language, which is an agglutinative language. This method is valuable in that much of the existing 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 endings, postpositions are used in agglutinative languages to determine a case of a noun or relationship between multiple words. Since an agglutinative language has many variations using endings and postpositions, using the same evaluation methodology as that used in inflectional language yields evaluation results that are too strict. Therefore APEM discards morphemes that do not carry meanings, such as endings and postpositions, to reflect the characteristics of agglutinative languages.

We also examine whether two types of APEM variation methods would improve the performance of paraphrase evaluation, namely APEM+synonym dictionary (APEM_SD) and APEM + synonym dictionary +Google distance (APEM_SDGD). 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 paraphrase generation process, i.e. 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 I
RESEARCH ON PARAPHRASE EVALUATION

FirstAuthor	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 judgment, which 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 exist limitations on the level of expressions that machines can comprehend. Despite such difficulties of automatic evaluation, 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 using features at the surface or semantic level. The main surface level feature used for paraphrase evaluation is n-gram, which is a simple yet powerful feature in that the methods using n-gram yield paraphrase evaluation results with high performance. Popular methods that use n-gram as a main feature are Bilingual evaluation understudy (BLEU), Translation edit rate (TER), and Paraphrase in n-gram changes (PINC) [1], [2], [3]. However, using n-gram without any processing is not suitable for an agglutinative language such as Korean. In an agglutinative language, a word can take many different surface 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 different on the surface but are semantically identical in Korean. Therefore a different evaluation method adapted to the agglutinative language in paraphrase needs to be developed. Thus, in this paper we suggest automatic paraphrase evaluation using morphological analysis (APEM), which accounts for the variations of surface forms in agglutinative languages.

In addition to APEM, which mainly uses surface level features for evaluation, we also suggest two additional variations of APEM by considering semantic level features as used in existing methods [13], [15], [16]. In particular, a popular semantic tool used for paraphrase evaluation is synonym dictionaries as used in the Translation edit rate – plus (TERp) method. We suspect that APEM + synonym dictionary (APEM_SD) could improve APEM's performance since words that are paraphrased with a different syntactic form, yet carry the same semantic meaning can be evaluated as having the same meaning. However, APEM_SD is a very naïve approach in that it uses data from synonym dictionaries without addressing word sense disambiguation (WSD). For example, consider the following sentence: "The power went out last night". In this example, although 'power' and 'force' are synonyms in a dictionary,

they cannot be replaced with each other due to inappropriate context. Therefore, we also propose a third method of APEM + synonym dictionary + Google distance (APEM_SDGD) that considers context by using semantic similarity of words in a sentence. In the next section, we explain how the three methods of APEM, APEM_SD, and APEM_SDGD are implemented.

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 [17].

B. Gold standard using manual evaluation

To validate the quality of the paraphrased Korean sentences generated by the automatic system, we conducted a manual evaluation. Three human evaluators were asked to assess one hundred sentences using the guidelines proposed by Callison-Burch et al [12]. Figure 2 shows the specific guidelines. The interclass correlation (ICC) score of 0.87 suggests that our manual evaluations results are reliable.

The gold standard value was determined by calculating the mean value of the three human evaluators' scores [3]. The mean value of the gold standard evaluation scores for the one hundred sentences is 2.89 (s.d=1.08). Since the scales for the manual and automatic evaluation score are difficult to compare due to the difference in the scales, we normalized each score by converting it into a standardized score. The mean value of the manual evaluation scores using standardization is 0.47 (s.d=0.27).

	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

Fig 2. Evaluator rated paraphrases along a 5-point Likert scale

C. Bilingual evaluation understudy (BLEU)

BLEU is a popular method used for various tasks in the area of machine translation, including automatic paraphrase evaluation [1]. Since BLEU was developed originally for conducting evaluations of foreign language translations, the original sentence is in a different language than the candidate and reference sentences that are used for paraphrase evaluations. Here the candidate sentence is the target sentence that will be evaluated in terms of paraphrase accuracy compared to a set of reference sentences. Typically BLEU calculates an evaluation score between the candidate and four reference sentences. The final evaluation score is the maximum score of these four scores. Note that both the candidate and reference sentences are translated versions of the original sentence. To calculate the evaluation score, the number of identical words in the candidate and reference sentences is divided by the total number of words in the candidate sentence.

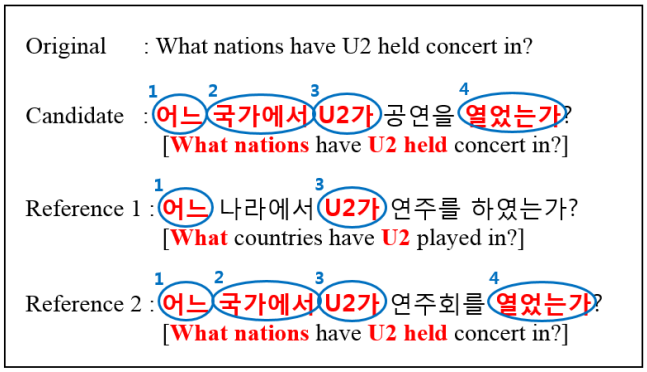

Fig 3. Example of paraphrased sentences. Circles are words that are identical across the given sentences.

Figure 3 shows sample sentences with two reference sentences as an example. The evaluation score for the candidate and reference 1 is 2/5, whereas the score for the candidate and reference 2 is 4/5. Thus the final BLEU score is 4/5, which is the maximum score between these two scores.

D. Automatic paraphrase evaluation using morphological analysis(APEM)

In this section, we introduce APEM along with its two variations, APEM_SD and APEM_SDGD.Figure 4 shows an example pair of sentences that will be used to explain the three evaluation processes. Sentence 1 is the original input sentence, and sentence 2 is the paraphrased sentence.

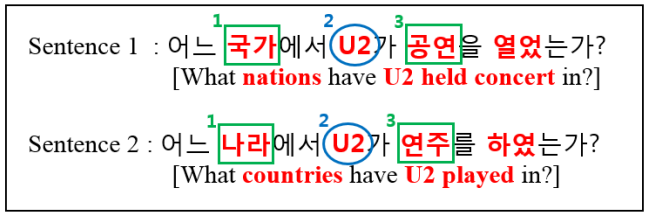

Fig 4. Example of paraphrased sentences. Circles are words that are identical whereas rectangles are words in synonym relationships.

To implement APEM, we first extracted morphemes for an input sentence using ETRI morpheme analyser [18]. Next, we removed morphemes that are labeled as endings and postpositions and kept the ones that carry 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 the essential meaning of the sentence, but merely function to make a sentence grammatically correct.

After identifying the critical morphemes that carry meanings, we calculate the evaluation score according to the equation (1). The numerator is the number of identical words

in sentence 1 and sentence 2. In our example sentences, the word pairs "U2/U2" are identical words. Thus, the numerator is one. 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 0.25. Note that according to the formula presented in equation (1), the evaluation score can range from 0 to 1, unlike the manual evaluation score, which ranges from 1 to 5.Using APEM, the mean value of the 100 sentences is 0.76 (s.d=0.17).

$$\text{Score} = \frac{Number\ of\ matched\ morphemes}{Number\ of\ meaningful\ morphemes\ in\ one\ sentence} \quad (1)$$

For APEM_SD, the method for calculating the numerator differs from APEM. In APEM_SD, if paired words are synonyms, they are regarded as "matched" whereas in AEPM, only identical words were counted as matched morphemes. Therefore, in our example sentences, the word pairs of "countries/ nations", "U2/U2", "held concert/ played" are pairs that are either identical or share synonyms. Thus, the numerator is three. The denominator is the number of meaningful morphemes in sentence 1, which is four. Therefore, the APEM_SD evaluation score for the paraphrased sentences in our example is 0.75.

For APEM_SDGD, we conduct an additional step from APEM_SD to calculate the numerator of the equation (1) by using an API from the Mechanical Cinderella project [19]. Since using APEM_SD yields "matched" words that are not in synonym relationships when considering context, we used Google distance to measure semantic similarity of words. The main assumption of the Google distance measure is that words that are used in a similar context would have a higher probability of appearing together in a document. Therefore, if a paired set of words has a high level of semantic similarity, the number of web pages that are returned in the Google search engine using those words would be higher compared to words with a low level of semantic similarity. Based on this idea, the equation uses four main numbers to calculate Google distance between two words, x and y as shown in equation (2). In the equation, f(x) and f(y) are the number of pages returned by the Google search engine using each term. F(x, y) is the number of pages where both x and y occurs, and M is the total number of web pages returned by the Google search engine.

$$\text{NGD}(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}} \quad (2)$$

We selected the threshold of Google distance score as 0.35 for filtering the synonyms. The threshold is heuristically calculated based on the normalized Google distance scores for a set of randomly selected words from the sentences in our dataset. We calculated approximately 600 pair of words and observed that, in general, a score of 0.35 or larger is calculated for correctly paired synonyms. For example, for the words used in Figure 4, the Google distance score between 'nations' and 'countries', which are commonly used together, is 0.352. The Google distance score for the words that do not appear frequently together, such as 'power' and 'force'is 0.163.

IV. RESULTS AND DISCUSSION

In this section, we present the correlation scores among the gold standard, BLEU, and the three variations of APEM. The positive medium correlation between the gold standard and the comparison method suggests that our automatic evaluation score captures some part of human evaluation scores. However, we acknowledge that at the current stage, the correlation is not accurate enough for use in place of human evaluation.

TABLE Ⅱ
PERFORMANCE OF THE PARAPHRASE EVALUATION

Method	Correlation
BLEU	0.57
APEM	0.62
APEM_SD	0.30
APEM_SDGD	0.60
BLEU_APEM(2:8)	0.64
BLEU_APEM(5:5)	0.64
BLEU_APEM(8:2)	0.61

As seen in Table 2, BLEU_APEM, which mixes the scores of both BLEU and APEM, performed best. In addition, our suggested method of APEM and APEM_SDGD performed slightly better than the existing method of BLEU. It was surprising to see that APEM_SD scores were much lower at 0.3 compared to other methods. Thus, our experiment shows that using synonyms without considering context has a high penalty.

V. CONCLUSION

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 higher correlation of APEM compared to BLEU shows promise in that using morphemes in agglutinative languages is helpful. However, APEM_SD, which is our naïve approach for including a semantic feature that simply utilizes a synonym dictionary, hurts performance. Efforts for considering contextual information, such as in APEM_SDGD, are necessary if semantic features are to be used.

In future work, we plan to extract more features of agglutinative languages 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, the inclusion of advanced measurements that consider context would increase the accuracy of our technique. For example, compared to using synonyms from dictionaries, the use of words from sentences with varying expressions in Wikipedia could be helpful.

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: pp. 341-387, 2010.

[5] Liu, Chang, Daniel Dahlmeier, and HweeTou 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 ProdromosMalakasiotis, "A survey of paraphrasing and textual entailment methods," *Journal of Artificial Intelligence Research*: pp. 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*, *Association for Computational Linguistics*, pp.9-16, 2004.

[10] Glickman, Oren, and I. Dagan, "Acquiring lexical paraphrases from a single corpus," *Recent Advances in Natural Language Processing III*, *John Benjamins Publishing*, *Amsterdam, Netherlands*, pp. 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), pp. 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.

[15] Callison-Burch, Chris, Trevor Cohn, and Mirella Lapata, "Parametric: An automatic evaluation metric for paraphrasing," *Proceedings of the 22nd International Conference on Computational Linguistics* -Volume 1, *Association for Computational Linguistics*, 2008.

[16] Miller, George A, "WordNet: a lexical database for English," *Communications of the ACM* 38.11: pp. 39-41, 1995.

[17] Hancheol Park, GahgeneGweon, Ho-Jin Choi, JeongHeo, and Pum-Mo Ryu, "Sentential Paraphrase Generation for Agglutinative Languages Using SVM with a String Kernel," *In Proceedings of PACLIC* (*The 28th Pacific Asia Conference on Language, Information and Computing*), pp. 650-657, 2014.

[18] Lee, Changki, Soojong Lim, and Myung-Gil Jang, "Large-margin training of dependency parsers using Pegasos algorithm," *ETRI Journal* 32.3: pp. 486-489, 2010.

[19] Šlerka. J. (n.d.). Mechanical Cinderella. [Online]. Available: http://www.mechanicalcinderella.com

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. Since 2010, he participates inSystems 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.

GahgeneGweon 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.

Jeong Heo is a senior researcher in Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea, He received his MS degree in computer science for the University of Ulsan(2001), His research interests include natural language processing, text mining, social big data analytics, and question answering.