Journal of the Korean Society of Speech Sciences Vol. 7 No. 4 (2015.12.31) ksss.jams.or.kr
ISSN 2005-8063, pp. 41~47 http://dx.doi.org/10.13064/KSSS.2015.7.4.041

A Study on Word Vector Models for Representing Korean Semantic Information

Yang, Hejung[1] · Lee, Young-In[2] · Lee, Hyun-jung[3] · Cho, Sook Whan[4] · Koo, Myoung-Wan[5]

ABSTRACT

This paper examines whether the Global Vector model is applicable to Korean data as a universal learning algorithm. The main purpose of this study is to compare the global vector model (GloVe) with the word2vec models such as a continuous bag-of-words (CBOW) model and a skip-gram (SG) model. For this purpose, we conducted an experiment by employing an evaluation corpus consisting of 70 target words and 819 pairs of Korean words for word similarities and analogies, respectively. Results of the word similarity task indicated that the Pearson correlation coefficients of 0.3133 as compared with the human judgement in GloVe, 0.2637 in CBOW and 0.2177 in SG. The word analogy task showed that the overall accuracy rate of 67% in semantic and syntactic relations was obtained in GloVe, 66% in CBOW and 57% in SG.

Keywords: GloVe, Korean corpus, semantic similarity, vector synthesis

1. Introduction

Measuring and representing the semantic similarity of words has been one of the most fundamental issues in natural language processing. To achieve the goal, vector space models have been widely used to represent each word as a real-valued vector. The vector keeps track of the context (co-occurring words, for example) in which target terms appear in a large corpus as proxies for meaning representations, and apply geometric techniques to these vectors to measure similarity in meaning of the corresponding words (Clark, 2014; Erk, 2012; Turney and Pantel, 2010). Thus if the vectors of two words are close to each other, it can be said that they are semantically similar to each other. Due to its simplicity superiority, these vector space models involve many useful applications, such as information retrieval (Manning et al., 2008), document classification (Sebastiani, 2002), question answering (Tellex et al., 2003), named entity recognition (Turian et al., 2010), and parsing (Socher et al., 2013)[6].

There are two major model families for learning vector space representations of words: global matrix factorization methods and local context window methods. The former is latent semantic analysis (LSA) (Deerwester et al., 1990) and the latter the skip-gram model by Mikolov et al. (2013c). Despite their successful applicability in diverse fields, it is currently observed that the two leading models have problems: global matrix factorization methods show a relatively poor performance on a word analogy task, and local context window methods do not employ corpus statistics.

Along these lines, Global Vector model is suggested as a solution to the problem. GloVe proposes a specific weighted least squares model, which is designed to train with global

[1] Sogang University, mindle714@naver.com
[2] Sogang University, youngin.lee721@gmail.com
[3] Sogang University, indeed1122@gmail.com
[4] Sogang University, swcho@sogang.ac.kr
[5] Sogang University, mwkoo@sogang.ac.kr, corresponding author

This work was supported by the Sogang University Research Grant of 2012 (201210046.01), Open ICT Convergence support program (R0215-15-1001), ICT/SW Creatieve Research program (R2215-15-1003).

Received: December 3, 2015
Revised: December 20, 2015
Accepted: December 21, 2015

[6] The reviewers of this paper commented on other tests with applied field tests such as named entity recognition of dependency parsing, and we agree with them. We will, in fact, employ a new discourse-related model based on our persent study, in which we are mainly concerned with the applicability of GloVe in Korean.

word-word co-occurrence counts. In the word similarity tasks, it has been observed that the GloVe's performance is far better than prior models.

While a number of studies on English word embedding models have been conducted and still under way, there have been no attempts to examine the potential of vector space representation for Korean language so far. This is important to note in light of possibility that GloVe can be valid in Korean as well.

Hence, the current study aims to examine whether the Global Vector model is applicable to Korean data as a universal learning algorithm. With this purpose in mind, we will briefly review previous studies using GloVe and Word2Vec models in Section 2. Section 3 and 4 will each present the details of our experiment and its results. Section 5 will discuss main findings of the present study and draw a conclusion. We utilized the source code for the GloVe model at http://nlp.stanford.edu/projects/glove/.

2. The GloVe Model

2.1 The word2vec model

Word2vec is a simple single-layer neural network architecture consisting of two models, the continuous bag-of-words (CBOW) and skip-gram models of Mikolov et al. (2013a). CBOW model is a feedforward neural network language model sharing the projection layer for all words without the hidden layer. The model allows the context to predict missing information of the word before and after it. On the other hand, skip-gram models predict surrounding words given the current word. In other words, utilizing the current word as the input, skip-gram models predict words within a certain range before and after the current word.

2.2 Global Vectors Model

Global Vectors (henceforth GloVe) is an unsupervised learning algorithm for obtaining vector representations for words. It represents each word w $\in$ V_W and each context c $\in$ V_C as d-dimensional vectors x and c as in the equation below: We use F(w,c) to denote the number of co-occurrences of a pair (w,c).

$$\vec{w} \bullet \vec{c} + b_w + b_c = \log(F(w,c)) \ \forall \ (w,c) \in D \quad (1)$$

b_w and b_c (scalars) are word/context-specific biases, and at the same time they are parameters to be learned in addition to w and c.

Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear sub-structures of the word vector space. Pennington et al., (2014) proposed a new weighted least squares regression model as in the equation (2), where V is the size of the vocabulary, $X \in R^{V x V}$ is a word co-occurrence matrix, and X_{ij} is the frequency of word i co-occurring with word j. In the equation (3), f(x) is a weighted function.

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \widetilde{w_j} + b_i + \widetilde{b_j} - \log X_{ij})^2 \quad (2)$$

$$f(x) = \begin{cases} (x/x_{\max})^{\alpha} & \text{if } x < x_{\max} \\ 1 & otherwise \end{cases} \quad (3)$$

There have been several attempts to evaluate vector space modeling framework as a tool for representing the Korean language (Li, Ma, & Lee, 2007). Most of them, however, mainly utilized prior models and focused on its spoken language. Based on the findings and limitations in the previous approaches, the current study aims to compare GloVe with previous methods and to finally examine its applicability of the model to the Korean language.

3. Experimental Setup

3.1 Corpus Construction

We trained GloVe with our own Korean corpus which we constructed for this study. In Korea, *the Sejong Corpus* (a.k.a., The Korean National Corpus, The National Institute of Korean Language, 2007) has been widely used in various fields. However, due to its comparatively small size of vocabulary (approximately ten million tokens), it is not enough to test the reliability of GloVe in Korean. For that reason, we built a new Korean corpus with a bigger size of data for our experiment.

One hundred million sentences were collected by a web crawler from internet bulletin boards of the Korean web sites. Having been pre-processed by removing stopwords, sentence splitting, and tokenization, the corpus contains one million sentences, spanning 668,284,389 tokens, which is 66 times bigger than *the Sejong Corpus* in size (We present example sentences from our corpus for your reference in Table 1). Words that appeared less than 100 times in the corpus were ignored, resulting in vocabularies of 3,552,280. Based on these vocabulary sets, from the dictionary pulled out from the corpus, we constructed a matrix of cooccurrence counts *X*. Pennington et al. (2014) set x_{max} as 100 and a as 3/4 for their study, and we also used these parameters in our experiment. These parameters are set to be small for large values of x so that frequent co-occurrents are not overweighted.

Table 1. Sample sentences from our Korean corpus

No.	Sample Sentences
1	무조건 어렵고 불편해서 피하기만 하는 것은 정말로 무책임하게 보이네요.
2	지금 배가 고파서 정신이 하나도 없네요.
3	그냥 그 새로움과 설레임 그 자체를 즐기는 사람 많아요.
4	근데 그 친구한테서 청천벽력같은 소리를 들었습니다.
5	하지만 정색을 하고 물어보는 상황에서 하얀 거짓말로 위기를 모면할 수는 없는 거 아닐까요.

3.2 Word Synthesis

In a pilot study for the experiment, we learned that it would be necessary to synthesize words with the same roots. As Lee et al. (2015) pointed out, Korean is an agglutinative language unlike English.

Table 2. Comparison of derived forms between Korean and English

Languages	original form	derived forms
English	boy	boys
Korean	*pap* (밥)	*pap-un* (밥은) *pap-man* (밥만) *pap-ul* (밥을) *pap-to* (밥도) *pap-ina* (밥이나) ......

To be specific, as shown in Table 2 above, English noun ‘boy’ can have a limited number of its derived form such as a plural form (‘boys’, for example) while Korean noun ‘pap’ can have significantly a larger number of derived forms (‘*pap-i* (밥-이)’, ‘*pap-to* (밥-도)’, or ‘*pap-ul* (밥-을)’, for example). It is because Korean allows a word to have multiple particles such as suffixes, and postpositions among others. This is important as in computational process, they will be recognized as different individual words although its semantic notion is same (or almost similar).

Based on this idea from the pilot study, we conducted a word synthesis process. As given in Figure 1, we selected a set of top five derived forms for each word and grouped them as one synthesized category. This method was used after the Korean corpus training session and raw mapping from a word to a vector were all finished. The detailed process consists of three steps:

1) construct set V with tokens *v* which has the same stem i (For instance, tokens ‘*pap* (밥)’, ‘*pap-i* (밥-이)’, and ‘*pap-ul* (밥-을)’ are treated as a same set named ‘*pap*’.);

$$V_{\text{밥}} = \{w_{\text{밥이}}, w_{\text{밥을}}, w_{\text{밥만}}, \cdots\}$$

(V = {v | v is in dictionary and v has a stem i})

2) define new vector value w_i' as summing vector values of every token *v* with there weight function. Weight function f_v of w_i is ratio of particular frequency of w_i on the corpus over overall frequency of tokens in the set V.

$$w_i' = \sum_{v \in V} f_V(w_v) \times w_v$$

$$f_V(w_i) = w_i / \sum_{v \in V} frequency(v)$$

3) update vector of w_i as w_i'.

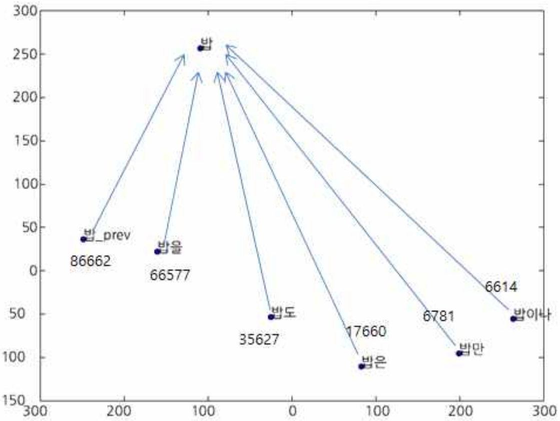

Figure 1. The result of word synthesis process. Through this process, all the words above can be grouped as one word, ‘*pap* (밥)’.

We built two types of word-vector mapping: one which underwent the vector synthesis process and one which was given the raw vector values by the corpus training. The results by two different mapping types were compared on a word similarity task.

4. Results and Analysis

4.1 Evaluation methods

To evaluate whether GloVe was successfully trained with our Korean data, we administered an experiment on a word similarity task and a word analogy task. Based on the findings by Lee et al. (2015), we wanted to assure the representability of the Korean corpus discussed in Section 3.1, Hence, we obtained the testing data by the following process: 1) choose target words *N* with high frequency values from a corpus for the prevention of noisy words such as misspellings (e.g. *him-isnun* (힘-잇는), *mence* (먼저)) or slang words (e.g. *nwunkkal* (눈깔), *akali* (아가리); 2) find the closest context words *M* for each selected target word in the prior stage by certain vector similarity calculating methods; and 3) choose randomly word pairs from N*M word pair matrix. With each word pair, we wanted to calculate human judgements and to find their average score. Thus we constructed our testing data with N=70, M=50 and compared the values obtained by the cosine-similarity method with human judgements of similarity. In the currrent study, the human judgments were gained by two graduate students majoring in linguistics.

Based on Mikolov et al. (2013a), we conducted a word analogy task in addition to a word similarity task. The task consisted of questions such as "*a* is to *b* as *c* is to ___?" For example, the relationship between '*namca* (남자)' and '*yeca* (여자)' pair is shown to be same as the one between '*namphyeon* (남편)' and '*anay* (아내)' pair due to the gender difference. Hence, in this relationship we removed one of the four words and asked our trained vector model to fill in the blank.

A word similarity task was evaluated by comparing the human judgements with vector similarity values. By comparing their statistical correlations, we can justify the existence of positive relationship between human scores and automatic vector scores.

If a word analogy query containing three words *a*, *b*, *c* is given, we will find the word *d,* which has the closest vector value to $w_b - w_a + w_c$, where w_i is the vector value of the word *i*, and will then examine whether word *d* is the answer as predicted.

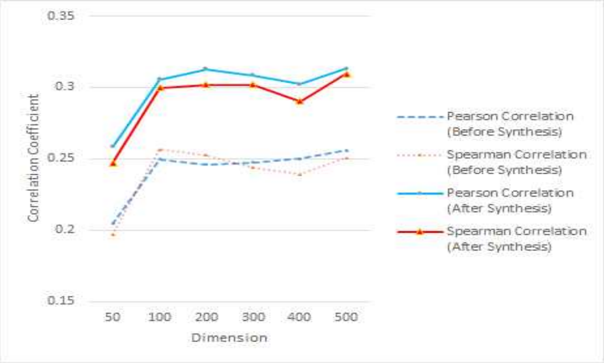

Figure 2. Two correlation coefficients on various dimensions before and after the word synthesis (p=2.7E-133; p=2.8E-123)

4.2 Word similarity task

We present results on the word similarity task in Figure 2. Before the word synthesis process was conducted, the Spearman rank correlation coefficient was the highest on dimension 100 and the Pearson rank correlation coefficient was the highest on dimension 500; whereas both Pearson and Spearman rank correlation coefficients were highest on dimension 500 after the word synthesis process.

In the current study, the Pearson rank correlation coefficient was higher than Spearman in all word categories except for the entailment category. It appears that it is due to the lack of divergence in human judgement. Despite a variety in cosine-similarity value, human scores could not reflect the subtle differences between words. This results in a monotonous variance in human judgement and several ties.

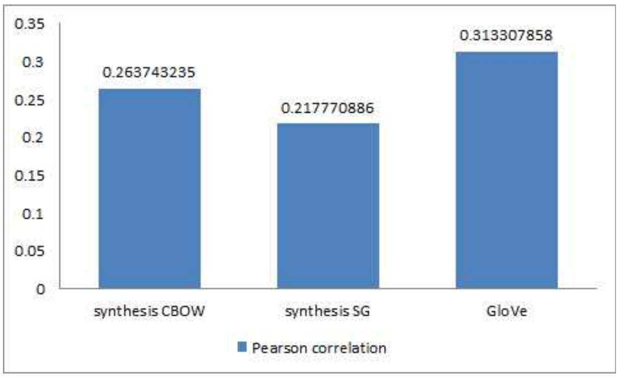

Figure 3. Pearson correlation coefficient of word2vec models

Best performance was achieved from the Pearson correlations in dimension 500 with vectors synthesized. Figure 3 shows the results of word2vec model, SG and CBOW, under that circumstance. As found in Pennington et al., (2014), GloVe outperformed SG and CBOW model.

4.3 Word analogy task

The word analogy task was conducted in two ways. One was to compare results of semantic-syntactic tasks between word2vec and GloVe. The other was comparing the similarity calculation methods, 3COSADD and 3COSMUL in the GloVe-chosen circumstance. For the 3COSADD method, we answered the question “*a* is to *b* as *c* is to __?” by finding the word *d* whose representation w_d is closest to w_b – w_a + w_c according to the cosine similarity. We additionally used 3COSMUL, the modified version of 3COSADD introduced by Levy et al. (2014). 3COSADD and 3COSMUL have the following equations:

$$\arg\max_{w_d \in V} \cos(w_d, w_b) - \cos(w_d, w_a) + \cos(w_d, w_c) \quad (4)$$

$$\arg\max_{b^* \in V} \frac{\cos(b^*, b)\cos(b^*, a^*)}{\cos(b^*, a) + \epsilon} \quad (5)$$

(ϵ = 0.001 is used to prevent division by zero)

For the semantic analogy task, 3COSADD method in dimension 1000 resulted in the highest score, and for syntactic analogy task, 3COSADD method in dimension 50 obtained the highest score. These results are given in Figure 4 and 5[7].

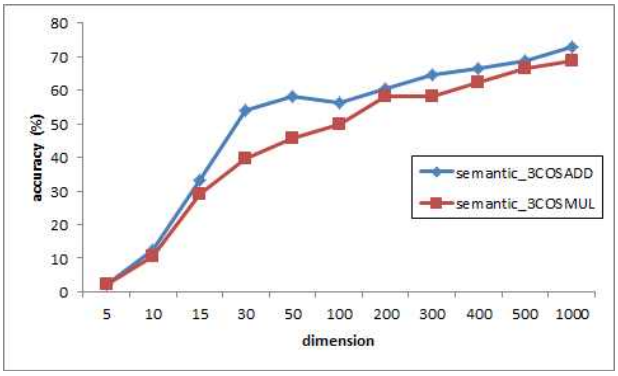

Figure 4. Accuracy on the semantic analogy task as a function of vector dimension and similarity method.

In the syntactic analogy task, it was observed that the optimal number of dimensions was 50. In case the number of dimensions was less or more than the optimal one (15 and 200, for example), wrong results were gained. To be specific, as shown in Table 4, while higher dimension succeeded in getting correct stems and failed in choosing the correct particles, lower dimensions failed in finding correct stems and further produced words with intuitively similar meanings as outputs.

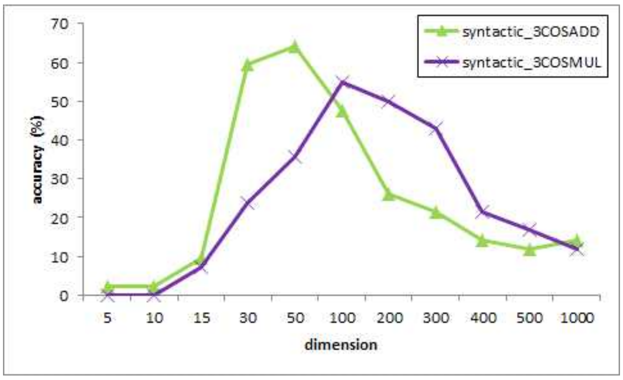

Figure 5. Accuracy on the syntactic analogy task as a function of vector dimension and similarity method.

Table 4. Comparison of results among various dimensions in a syntactic analogy task

target word	Variation in prediction of ‘-hako (-하고)’ by the number of dimension: 15, 50, and 200		
	dimension 15	dimension 50	dimension 200
전화	끊고	전화하고	전화를
연락	헤어져	연락하고	연락을
이야기	진지하게	이야기하고	이야기를
이해	객관적으로	이해하고	이해를

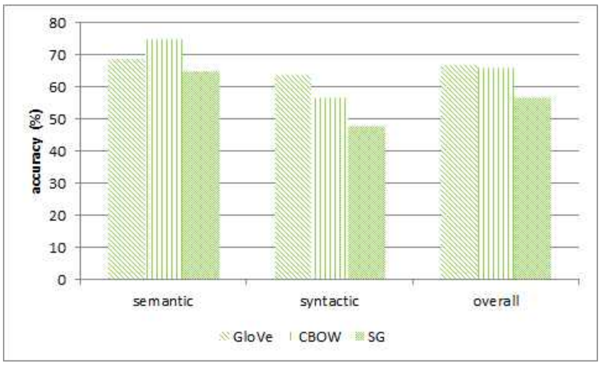

Figure 6. Accuracy on the word analogy task of word2vec and GloVe model.

Results of the word2vec model in a semantic and syntactic task are shown in Figure 6: the former was administered in dimension 500 and the latter in dimension 50. The 3COSADD method was used for both tasks. The highest scores were obtained in CBOW for the semantic task and in GloVe for the syntactic task. Overall, the accuracy of three models (GloVe, CBOW, and SG) were 67%, 66%, and 57%, respectively.

[7] A reviewer of this paper has commented that vector dimensions are correlated with the size of corpus and vocabularies. We agree with the reviewer, and will conduct an experiment along the lines in near future.

5. Conclusion

In the current study, we conducted experiments on GloVe, continuous bag-of-words, and skip-gram models with the aim to examine their applicabilities to Korean. In designing the experiments, we first built our own Korean corpus with the sufficient size of vocabulary. Based on the distinctive properties of the Korean language, we also administered a word synthesis process. As a result, for a word similarity task, it was observed that the Pearson correlation coefficent was 0.3133 in GloVe, 0.2637 in CBOW, and 0.2177 in SG. For a word analogy task, GloVe resulted in 67% accuracy as compared to 66% in CBOW and 57% in SG. Hence, these results show that GloVe model outperformed word2vec model in both word similarity and word analogy tasks. These results indicate a possibility that GloVe model can be utilized as an effective tool for calculating semantic similarity of Korean words and further discovering linear relationships among them. Based on this study, it is further speculated that Glove, as the count-based model, may capture global statistics better than word2vec model.

References

[1] Stephen Clark. (2013). Vector space models of lexical meaning. In Shalom Lappin and Chris Fox, editors, Handbook of Contemporary Semantics, 2nd ed. Blackwell, Malden, MA. In press; http://www.cl.cam.ac.uk/~sc609/pubs/sem_handbook.pdf.

[2] Katrin Erk. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compas*, 6(10):635-653.

[3] Peter Turney and Patrick Pantel. (2010). From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research, 37:141-188.

[4] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1, p. 496). Cambridge: Cambridge university press.

[5] Fabrizio Sebastiani. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1), 1-47.

[6] Tellex Stefanie, et al. (2003). Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of the 2th annual international ACM SIGIR conference on Research and development in information retrieval*, 41-47, ACM.

[7] Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. *In Proceedings of the 48th annual meeting of the association for computational linguistics*, 384-394. Association for Computational Linguistics.

[8] Socher, Richard, et al. "Recursive deep models for semantic compositionality over a sentiment treebank." *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*. Vol. 1631. 2013.

[9] Deerwester, Scott C., et al. (1990). "Indexing by latent semantic analysis." *JAsIs* 41(6):391-407.

[10] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. (2013a). Efficient estimation of word representations in vector space. http://arxiv.org/abs/1301.3781/.

[11] Jeffrey Pennington, Richard Socher, and Christopher Manning. (2014). Glove: Global vectors for word representation. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 12*:1532-1543.

[12] Haizhou Li, Bin Ma, & Chin-Hui Lee. (2007). A vector space modeling approach to spoken language identification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(1), 271-284.

[13] The National Institute of Korean Language (2007). *The Sejong Corpus.*

[14] Young-In Lee, Hyun-jung Lee, Myoung-Wan Koo, & Sook Whan Cho. (2015). Korean Semantic Similarity Measures for the Vector Space Models. *Phonetics and Speech Sciences*, 7(4): 1-7.

[15] Levy et al. (2014). Linguistic regularities in sparse and explicit word representations. *CoNLL-2014*, 171.

• **Yang, Hejung**
Department of Computer Science, Sogang University
35 Baekbeom-ro, Mapo-gu, Seoul 04107
Tel: 02-704-8485 Fax: 02-704-8273
Email: mindle714@naver.com
Areas of interest: Software Engineering

• **Lee, Young-In**
Department of English, Sogang University
35 Baekbeom-ro, Mapo-gu, Seoul 04107
Tel: 02-705-8290 Fax: 02-715-0705
Email: youngin.lee721@gmail.com
Areas of interest: Psycholinguistics, Discourse Analysis, Pragmatics

• **Lee, Hyun-jung**
Department of English, Sogang University
35 Baekbeom-ro, Mapo-gu, Seoul 04107
Tel: 02-705-8290 Fax: 02-715-0705

Email: indeed1122@gmail.com
Areas of interest: Syntax, Morphology, Semantics

- **Cho, Sook Whan**
  Department of English, Sogang University
  35 Baekbeom-ro, Mapo-gu, Seoul 04107
  Tel: 02-705-8300 Fax: 02-715-0705
  Email: swcho@sogang.ac.kr
  Areas of interest: Language Acquisition, Psycholinguistics, Cognitive Science

- **Koo, Myoung-Wan, corresponding author**
  Department of Computer Science, Sogang University
  35 Baekbeom-ro, Mapo-gu, Seoul 04107
  Tel: 02-705-8935 Fax: 02-704-8273
  Email: mwkoo@sogang.ac.kr
  Areas of interest: Speech recognition, Natural language understanding, Dialogue Modeling

Appendix

Appendix 1. Word Analogy Test Sets

(1) Semantic Analogy Task

Word *a*	Word *b*	Word *c*	Word *d*
초등학교	초등학생	중학교	중학생
초등학교	초등학생	고등학생	고등학생
여자	남자	여성	남성
여자	남자	아내	남편
여자	남자	엄마	아빠
여자	남자	할머니	할아버지
여자	남자	여왕	왕
여자	남자	공주	왕자

(2) Syntactic Analogy Task

Word *a*	Word *b*	Word *c*	Word *d*
결혼	결혼하고	이야기	이야기하고
결혼	결혼하고	이해	이해하고
결혼	결혼하고	전화	전화하고
결혼	결혼하고	잠	잠들고
결혼	결혼하고	연락	연락하고
말	말하면	결혼	결혼하면
말	말하면	전화	전화하면
말	말하면	연락	연락하면

Appendix 2. Word Similarity Test Sets

(1) ‘엄마’

Word *a*	Word *b*	Cosine Similarity
엄마	아빠	0.978456
엄마	어머니	0.894823
엄마	아버지	0.865324
엄마	할머니	0.858674
엄마	아들	0.855033
엄마	딸	0.819675
엄마	동생	0.818473

(2) ‘한국’

Word *a*	Word *b*	Cosine Similarity
한국	독일	0.688514
한국	영국	0.716407
한국	북한	0.75566
한국	미국	0.888825
한국	일본	0.88414
한국	러시아	0.723647
한국	프랑스	0.730826

(3) ‘아침’

Word *a*	Word *b*	Cosine Similarity
아침	오전	0.68289
아침	오후	0.78428
아침	저녁	0.899646
아침	밤	0.820669