Extension of Semantic Lexicon Using Word Embeddings and Synonyms

Da-sol Park* and Jeong-Won Cha**

** Department of Computer Engineering, Changwon National University*
Changwon, South Korea
E-mail: dasol_p@changwon.ac.kr

*** Department of Computer Engineering, Changwon National University*
Changwon, South Korea
E-mail: jcha@changwon.ac.kr

Abstract

In this paper, we propose a novel method to extend semantic lexicon by using word embeddings and synonyms. We compare the embedding values between the new word and the known words. We also compare the embedding values of synonyms between the new word and the known words. The system performance of semantic category assignment for words not appearing in the Sejong electronic dictionary is 32.19%, and the system performance of extended semantic category assignment is 51.14%. We proved that it is helpful to extend the semantic category of words in the Sejong electronic dictionary by assigning semantic categories to the new words for which semantic categories have not been assigned.

Key Words: Word embeddings, Sejong Electronic dictionary, Semantic category

1. Introduction

Semantic analysis is the technique of distinguishing the meaning of the words that constitute a sentence, and identifying the semantic meaning of the sentence by logically clarifying the semantic relationship between the sentence constituents.

Semantic analysis is the upper level of the natural language processing layer on top of morphological analysis and parsing [1]. It can be divided into two aspects of solving the ambiguity and determining the semantic role in a sentence.

Decision on the semantic role should be based on the semantic role and semantic category information. However, the data from the Sejong electronic dictionary, which contains the semantic role and semantic category information, needs to be extended to handle unlimited Korean sentences. In this paper, we try to extend the Sejong electronic dictionary using word embeddings and synonyms.

In this paper, we introduce the Sejong electronic dictionary in chapter 1.2. Chapter 2 introduces the related researches. The proposed method is described in chapter 3, and finally, we describe the conclusion in chapter 4.

1.2. The Sejong electronic dictionary

The electronic dictionary of the 21st century Sejong plan contains comprehensive and vast information about the modern Korean language, and it is an essential and practical electronic dictionary that can be used universally for automatic processing of Korean language [2].

The Sejong electronic dictionary contains diverse syntactic and semantic information about the headwords in XML form and it also contains frame information used to determine the semantic role. The Sejong electronic dictionary contains 25,458 nouns, 15,181 verbs, 4,398 adjectives, 645 noun semantic subcategories, and 631 predicate semantic subcategories.

2. Related research

There have been studies to extract a hypernym and a hyponym from a corpus. Hearst first proposed a method for automatically extracting hypernym and hyponym semantic relations using pattern recognition and semantic relation in text [3]. After generating various grammatical patterns, it extracts relation triples if the given sentence type is the same as the pattern type. There is a problem with this method that it is impossible to filter out the modifier if the same pattern contains the modifier.

Cederberg, Widdows used “Latent Semantic Analysis (LSA)” to automatically extract hypernym and hyponym relation in text [4]. LSA improves precision and recall.

Verginica studied the hypernym and hyponym co-occurrence pattern, and pattern generation of hypernym and hyponym relations [5]. However, there is a problem that sentences containing adjectives and articles are not applied to the co-occurrence pattern.

Erik Tjong Kim Sang, Katja Hofmann and Maarten de Rijke extracted a hypernym and a hyponym using the lexical pattern and the dependency pattern of the corpus [6].

Marco Baroni, Bgoc-Quynh Do and Chung-chieh Shan studied the entailment of the adjective-noun structure and the quantifier-noun structure using distributional vector expression of the phrase [7]. The adjective-noun structure and quantifier are also expressed as semantic vectors, and entailment can be found using SVMs and classifiers as distributed vectors.

Marek Rei and Ted Briscoe performed a research for a hyponym using vectors [8]. They found pattern-based hyponym results in a very low recall because it depends only on the two words mentioned. We used a vector similarity method that can be applied to the other domains and other languages without supervised learning or pattern structure. We used dependence-based vector representations to achieve state-of-the-art performance using neural networks and windows-based models.

For extracting a hypernym and a hyponym in Korean, Chan-Seong Pang and Hae-Yun Lee proposed a method for extracting the hypernym and hyponym relation pattern using the corpus [9]. When displaying an enumeration of nouns for the purpose of pattern extraction, there is difficulty in capturing the fixed patterns because of various postpositions or punctuation marks, and when context-dependent vocabularies appear, it is difficult to differentiate between a hypernym and a hyponym alone. Choi Yu-mi and Sakong Chukl studied automatic hypernym extraction for automatic thesaurus construction [10]. The syntactic characteristics of the sentences described in the Glossary of Library and Information Science were examined. Ten algorithms were developed using the syntax information obtained from a sample survey and they showed accuracy of 89.4%.

3. The proposed method

We use synonyms and embedding vectors for words to assign semantic categories to words that do not appear in the Sejong electronic dictionary. First, we embed the words in the Sejong electronic dictionary using large external documents. We use Google's word2vec to create word embedding [11]. We assign semantic categories to new words using similarity between these values and new words that do not appear in the Sejong electronic dictionary. In addition, semantic categories are assigned to new words by comparing the synonyms in the Sejong electronic dictionary with those not appearing in the Sejong electronic dictionary. We assign the semantic category to a new word based on the semantic category of a word with the highest cosine similarity and the Pearson correlations of all experiments and the semantic category of a word with the lowest value for the Euclidean distance. We determined the weight after performing the adjustment experiments.

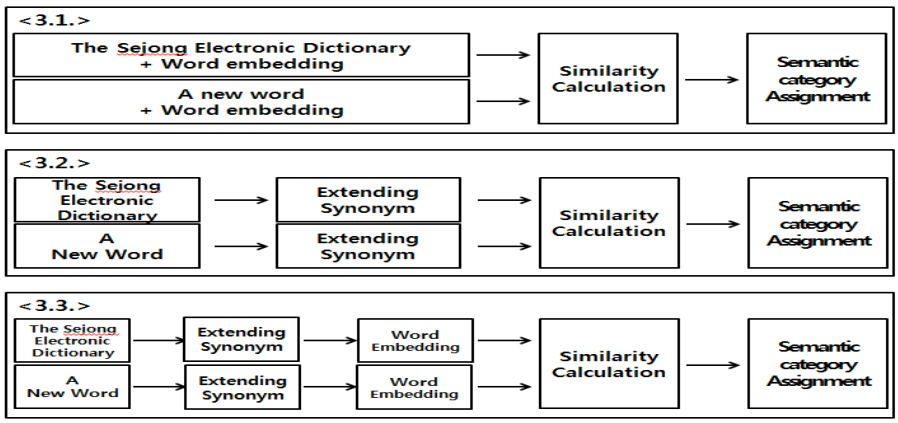

Figure 1. The proposed method

Figure 1 summarizes the proposed method. These three methods are described in detail in Sections 3.1, 3.2 and 3.3.

2.3 Word embeddings similarity

Our first attempt is to assign a semantic category using word-embedding vectors. The word-embedding values for words in the Sejong electronic dictionary are obtained using large external documents. In addition, word-embedding values are obtained in the same way for words that have to be assigned a semantic category. The semantic category is assigned by calculating the similarity of the embedding values between these words.

We used 280 million morpheme corpora collected from Internet newspapers and Korean Wikipedia and labeled automatically using the Expresso part-of-speech tagger [12] to obtain the embedding values for the words.

For similarity calculation, we used Cosine similarity (equation 1) [13,14,15], Euclidean distance (equation 2) [16] and Pearson correlations (equation 3) [17].

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}, \qquad (1)$$

where, A and B are the two given vectors of attributes.

It is not affected by the distance of the vector. The range of values is from -1 to 1, and the closer they are to 1, the more similar they are.

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}, \qquad (2)$$

where, p, q are two points in Euclidean n-space.

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}, \qquad (3)$$

where, X is the value in the first set of data, Y is the value in the second set of data, and N is the total number of values.

The range of values is from -1 to 1, and 1 indicates that they are fully correlated, 0 indicates no association at all, and -1 indicates that they are completely reversed.

Table 1 shows the new word experimental results. The best 1 is the accuracy when the one with the highest similarity value is the correct answer and best 5 are the top 5 answers with the highest similarity value. In the test, we used 876 new words corresponding to a noun.

Table 1. Experimental results

Similarity	Best 1	Best 3	Best 5
Cosine similarity	25.57% (224/876)	42.57% (373/876)	49.77% (436/876)
Euclidean distance	25.00% (219/876)	42.57% (373/876)	49.77% (436/876)
Pearson correlations	25.45% (223/876)	42.35% (371/876)	50.68% (444/876)

We used word-embedding values for semantic similarity of words. Table 2 shows the words for which categories have to be assigned. As shown in Table 2, the similarity calculated by word-embedding cannot distinguish from the hypernym similarity and the hyponym similarity. This shows that word-embedding has its limitations in detailed semantics. In addition, it was found that incorrect values could be obtained for low frequency words.

Table 2. Words for which categories have to be assigned

Word in the Sejong electronic dictionary	바톤/NNG (Baton/NNG)	삐에로/NNG (Pierrot/NNG)	각각/NNG (Each/NNG)
Best 5 Words	콜/NNG (Call/NNG), 큐/NNG (Cue/NNG), 하프/NNG (Half/NNG Harp/NNG), 헤드/NNG (Head/NNG), 러브/NNG (Love/NNG)	고백/NNG (Confession/NNG), 제비족/NNG (Gigolo/NNG), 짝사랑/NNG (Crush/NNG), 진짜/NNG (Real/NNG), 무심/NNG (Indifference/NNG)	형태/NNG (Form/NNG), 방식/NNG (Way/NNG), 각자/NNG (Severalty/NNG), 순서/NNG (Sequence/NNG), 특징/NNG (Characteristic/NNG)

3.2 Word-embedding and synonym extension

In order to compensate for the disadvantages of word embedding value itself, we use synonym similarity. The synonyms used here are synonym information in the Korean dictionary, named entity dictionary information, and semantic similarity information using word embedding. We used the matching number of extended words as the second feature. Lexical similarity (α) and the matching number of extended synonym (β) were applied to the

linear combination. Table 3 shows the experimental results. Table 4 shows the weight applied to the linear combination.

Table 3. Experimental results

Similarity	Best 1	Best 3	Best 5
Cosine similarity	31.96% (280/876)	49.09% (430/876)	56.96% (499/876)
Euclidean distance	4.57% (40/876)	12.67% (111/876)	18.49% (162/876)
Pearson correlations	31.62% (277/876)	48.97% (429/876)	56.96% (499/876)

Table 4. Linear combination weight

similarity	α	β
cosine similarity	0.7	0.3
Euclidean distance	0.1	0.9
Pearson correlations	0.7	0.3

When using a synonym, it is impossible to evaluate the words that do not have word-embedding vector values when using similarity. Also, it can be seen that the correct answer, which was a relatively low value in the similarity measure, has a higher value through the synonym, and there is a rank change, and it is selected as the correct answer. However, lexical simple matching of the synonym alone lacks support for assigning semantic categories.

3.3 Word-embedding and extended synonym

The similarity of the synonym, which is additional information that can be obtained on synonyms, is calculated. We used the similarity of the extended words as the third feature. Lexical similarity (α), the matching number of extended synonym (β) and the similarity of the expanded lexical word and the synonym (γ) were applied to the linear combination. Table 5 shows the experimental results. Table 6 shows the weight applied to the linear combination.

Table 5. Experimental results

Similarity	Best 1	Best 3	Best 5
Cosine similarity	32.19% (282/876)	49.43% (433/876)	57.76% (506/876)

Euclidean distance	16.44% (144/876)	28.88% (253/876)	36.42% (319/876)
Pearson correlations	31.96% (280/876)	49.32% (432/876)	57.88% (507/876)

Table 6. Linear combination weight

Similarity	α	β	γ
Cosine similarity	0.4	0.3	0.3
Euclidean distance	0.1	0.4	0.5
Pearson correlations	0.4	0.3	0.3

We found experimental performance improvement using extended synonym and word-embedding. This is due to the synonyms, which are features that represent the meaning rather than the vector representing the similarity of words.

Table 7 shows examples of the target word and the semantic categories for the target word. The Sejong electronic dictionary words and new words in Table 7 are expected to have similar semantic categories or the same semantic category. However, the actual semantic category is not the same. The current experiment is correct only when the lexicon of the semantic category is the same. However, we did not consider only the semantic category of the current word as the correct answer.

Table 7. Examples of words and semantic categories

Word in the Sejong electronic dictionary		New words	
Lexicon	Semantic category	Lexicon	Semantic category
의료기관 (Medical institution)	기관 (Institution)	금융기관 (Financial institution)	금융기관 (Financial institution)
재혼 (Remarriage)	만남 (Meeting)	결혼 (Marriage)	대칭적행위 (Symmetrical acting)
개발도상국 (Developing country)	상황값 (Situation value)	선진국 (Developed country)	국가 (Country)

3.4 The method using the parent semantic category

Based on the above analysis, further experiments were conducted. The results of the experiment in Section 3.3 were applied to the parent semantic category. Table 8 shows the experimental results. Table 9 shows the weight applied to the linear combination.

Table 8. Experimental results

Similarity	Best 1	Best 3	Best 5
cosine similarity	51.14% (448/876)	71.58% (627/876)	78.20% (685/876)
Euclidean distance	30.37% (266/876)	49.66% (435/876)	57.88% (507/876)
Pearson correlations	51.26% (449/876)	70.55% (618/876)	78.31% (686/876)

Table 9. Linear combination weight

Similarity	α	β	γ
Cosine similarity	0.2	0.5	0.3
Euclidean distance	0.1	0.4	0.5
Pearson correlations	0.1	0.5	0.4

In the cognition experiment, we showed the best performance in the method proposed in Section 3.4, and we tried to conduct the experiment for the predicate. The predicate consists of adjectives and verbs, and it uses 355 new adjective words and 827 new verb words. The experimental method is the same as in Section 3.4.

Table 10 shows the experimental results applied to adjectives. Table 11 shows the weights applied to linear combinations applied to adjectives.

Table 10. Experimental results

Similarity	Best 1	Best 3	Best 5
Cosine similarity	18.03% (64/355)	34.93% (124/355)	44.79% (159/355)
Euclidean distance	27.89% (99/355)	44.79% (159/355)	59.15% (210/355)
Pearson correlations	18.59% (66/355)	33.52% (119.355)	43.94% (156/355)

Table 11. Linear combination weight

Similarity	α	β	γ
Cosine similarity	0.5	0.1	0.4
Euclidean distance	0.5	0.2	0.3
Pearson correlations	0.2	0.3	0.5

Table 12 shows the experimental results applied to verbs. Table 13 shows the weights applied to linear combinations applied to verbs.

Table 12. Experimental results

Similarity	Best 1	Best 3	Best 5
Cosine similarity	16.32% (135/827)	29.87% (247/827)	36.88% (305/827)
Euclidean distance	29.87% (247/827)	44.86% (371/827)	55.02% (455/827)
Pearson correlations	17.05% (141/827)	29.75% (246/827)	36.64% (303/827)

Table 13. Linear combination weight

Similarity	α	β	γ
Cosine similarity	0.3	0.4	0.3
Euclidean distance	0.4	0.2	0.4
Pearson correlations	0.2	0.3	0.5

We could compare the performance of the noun and the performance of the predicate. We could also determine that the performance of the verb is lower than the performance of the noun. Lexical similarity itself is lower than that of the noun by 10% or more, and the performance is lower because the synonym information of the verb is lesser than the synonym information of the noun.

4. Conclusion

In this paper, we propose a method to extend the Sejong electronic dictionary by using word embeddings and synonyms. We can assign the lexical category to a new word using word embedding similarity of words in the Sejong electronic dictionary and new words. In addition,

semantic categories are also assigned by comparing the synonym of the word not found in the Sejong electronic dictionary with the synonym of the new word. We extended the semantic category by adding the correct word to the corresponding semantic category in the best 5 result of the experiment using the parent semantic category.

Experimental results show that lexical similarity, synonym extension, and word-embedding features have helped to find words that need to be assigned semantic categories in the Sejong electronic dictionary. However, the lexical similarity feature is helpful in finding word similarity, but it is difficult to distinguish between lexical meanings. Synonym feature helped to assign semantic categories.

A morpheme expressed as a vector using word embedding generates a vector containing various meanings. There is a problem that it is not possible to know a word indication a meaning of various meanings even if the semantic category of the vector is assigned. Therefore, there is a need for a method to detect the ambiguity in a vector of words with various meanings. In order to solve the problem, we plan to develop a word vector clustering method that can solve the ambiguity of words.

In addition, we will apply various word embedding models, and we plan to implement the model through machine learning.

5. Acknowledgments

The research was supported by 'Software Convergence Technology Development Program', through the Ministry of Science, ICT and Future Planning (S0177-16-1056).

References

[1] Daniel Jurafsky and James H. Martin, Speech and Language Processing, 2 (2007), 546.

[2] National Institute of Korean Language, Final Achievement of the 21st Century Sejong Plan, Ministry of Culture, Sports and Tourism, (2010), https://www.korean.go.kr/ .

[3] Hearst, M.A., Automatic Acquisition of Hyponyms from LargeText Corpora, Association for Computational Linguistics, (1992), 539-545.

[4] Cederberg, S. and Widdows, D., Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction, Proc.of the Conference on Natural Language Learning, (2003), 111-118.

[5] Verginica Barbu Mititelu, Automatic Extraction of Patterns Displaying Hyponym-Hypernym Co-Occurrence from Corpora, Proceedings of First Central European Student Conference in Linguistics, (2003).

[6] Sang, Erik Tjong Kim, Katja Hofmann, and Maarten De Rijke, Extraction of Hypernymy

Information from Text, Interactive Multi-modal Question-Answering, (2011), 223-245.

[7] Baroni, M., Bernardi, R., Do, N. Q., and Shan, C. C., Entailment above the word level in distributional semantics, Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, (2012), 23-32.

[8] Rei, Marek, and Ted Briscoe., Looking for Hyponyms in Vector Space, the Conference on Natural Language Learning, (2014), 68-77.

[9] Pang. Chan-Seong and Lee Hae-Yun, A Study of the Automatic Extraction of Hypernyms arid Hyponyms from the Corpus, The Korean Society for Cognitive Science, 19 (2008).

[10] Choi Yu-Mi and Sakong Chul, Development of Algorithm for the Automatic Extraction of Broad Term, The 5th Preceddings of the Korean Society for information Management Conference, (1998), 227-230.

[11] Mikolov and Tomas, et al., Efficient estimation of word representations in vector space, CoRR, (2013).

[12] URL: https://air.changwon.ac.kr/~airdemo/kg_tagger/, 2016-09-10.

[13] Gerard Salton, A. Wong, and C. S. Yang, A vector space model for information retrieval, Communications of the ACM, (1975), 613–620.

[14] Singhal and Amit., Modern information retrieval: A brief overview, IEEE, (2001), 35-43.

[15] Manwar, A. B., Mahalle, H. S., Chinchkhede, K. D., and Chavan, V, A Vector space model for information retrieval: A MATLAB approach, Indian Journal of Computer Science and Engineering (IJCSE), (2012), 222-229.

[16] Danielsson and Per-Erik, Euclidean distance mapping, Computer Graphics and image processing, (1980), 227-248.

[17] Pearson K., Notes on the history of correlation, Biometrika, 13 (1920), 25–45.

*Corresponding author: Prof. CHA Jeong-Won, Ph.D.

Department of Computer Engineering,

Changwon National University,

9 Sarim-dong, Changwon Gyoungnam, Republic of Korea 641-773, South Korea

E-mail: jcha@changwon.ac.kr