Journal of the Korean Society of Speech Sciences Vol. 7 No. 4 (2015.12.31) ksss.jams.or.kr
ISSN 2005-8063, pp. 49~55 http://dx.doi.org/10.13064/KSSS.2015.7.4.049

Korean Semantic Similarity Measures for the Vector Space Models

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

ABSTRACT

It is argued in this paper that, in determining semantic similarity, Korean words should be recategorized with a focus on the semantic relation to ontology in light of cross-linguistic morphological variations. It is proposed, in particular, that Korean semantic similarity should be measured on three tracks, human judgements track, relatedness track, and cross-part-of-speech relations track. As demonstrated in Yang et al. (2015), GloVe, the unsupervised learning machine on semantic similarity, is applicable to Korean with its performance being compared with human judgement results. Based on this compatability, it was further thought that the model's performance might most likely vary with different kinds of specific relations in different languages. An attempt was made to analyze them in terms of two major Korean-specific categories involved in their lexical and cross-POS-relations. It is concluded that languages must be analyzed by varying methods so that semantic components across languages may allow varying semantic distance in the vector space models.

Keywords: semantic similarity in Korean, semantic relatedness, lexical relation, cross-POS-relations

1. Introduction

As well known in the literature (Budanitsky & Hirst, 2006), semantic similarity has attracted a great deal of interest in natural language processing in recent years. A number of approaches have been implemented, involving word sense disambiguation, text structure, and automatic corrections of word errors in text use measures of relatedness and distance. One of the most widely used machine learning approaches is the semantic vector space model. Based on a huge size of corpora, the models represent the degree of semantic similarity or relatedness of two target words by using distance in the vector space as a measure for semantic similarity to evaluate. For example, the semantic difference between vector spaces of 'king' and 'queen' is regarded very close to the one of 'man' and 'woman'.

It has been observed that the Global Vector Models (henceforth, GloVe) show a great performance in multiple similarity tasks (Pennington et al., 2014). Recently, Yang et al. (2015) made the very first attempt to examine the GloVe's applicability to Korean. Yang et al. compared the model's performance to human judgement results and found that the model gave an insight into contributing to evaluating semantic similarity in Korean.

The measures utilized in Yang et al., on the other hand, mainly examined Korean lexical categories, and it gave us an incentive to enhancing them by employing other techniques tailored close to a set of Korean-specific properties. For example, as noted in other studies of various languages (Lopukhin et al., 2015), the degree of similarity between two words must depend on the lexical relations in association with synonyms ('mom' and 'mummy', for example), antonyms ('old' and 'young', for example), among others. Besides, Korean has hardly been used in testing the validity of GloVe model, and it is likely that Korean-specific features will provide a new insight into how word vector models may be built cross-linguistically.

1) Sogang University, youngin.lee721@gmail.com
2) Sogang University, indeed1122@gmail.com
3) Sogang University, mwkoo@sogang.ac.kr
4) Sogang University, swcho@sogang.ac.kr, corresponding author

This work was supported by the Sogang University Research Grant of 2012 (201210046.01).

Received: December 9, 2015
Revised: December 17, 2015
Accepted: December 17, 2015

This paper is concerned with constructing a set of categorization with the aim to build up a word dictionary on the basis of semantic similarity pertaining to Korean. For this purpose, Section 2 discusses the basic notion of semantic similarity and relatedness. Section 3 compares two major approaches to semantic similarity and briefly reviews previous research. Section 4 and 5 present our new dataset and lexical recategorization of Korean vocabulary. Section 5 concludes the paper.

2. Semantic Similarity

Semantic similarity or semantic relatedness has been adopted in the literature frequently with reference to the degree to which words are close to each other in contents (Lopukhin et al., 2015; Budanitsky and Hirst, 2006). For Budanitsky and Hirst, on the other hand, tho two terms are not exactly alike. In their view, similarity may involve radically different types of close meanings for a certain set of words and hence may not always apply in a straightforward way. Words are semantically ‘similar’ in cases where they are not only synonymous, or semantically close to each other, but also opposite in meaning or semantically relatively distant to each other. To take an example, from this perspective, ‘big’ and ‘large’ are semantically similar in that they both associate themselves with the concept of ‘size’ or ‘volume.' Words such as ‘happy’ and ‘unhappy,’ on the other hand, are opposite in meaning, yet can also be conceived as semantically related in light of the fact that they both denote an inner state of mind. Given this view, it would not be unreasonable to assume that semantic relatedness subsumes semantic similarity, and may alternately be used in the literature for that reason.

3. Recategorization of Korean Words

Results from Pennington et al. (2014) indicate that GloVe has comprehended the concept of 'superordinate' and 'hyponym' (‘country’ and ‘capital city’, for example) and processed them successfully. A set of Korean data that we have so far observed, on the other hand, show an interesting contrast between the distribution of hyponymy and that of synonymy cases. As can be seen in Table 1, Spearman correlation coefficients were much higher for hyponymy (61.3%) than for synonymy (51.6%). This disparity in Korean may point to a possibility that the degree and types of similarity may vary across languages. In this light, it would be important to find out how GloVe and human judgments may differ in terms of synonymy relations.

Moreover, as also seen in Table 1, in addition to the two categories, hyponymy and synonymy, it is also observed that other subcategories do not behave alike, Spearman correlations coefficients varying from 39.9% to 75.9%. This finding has led us to take a closer examination of various types of word categories in Korean, and details are discussed in the sections below.

Table 1. Spearman correlation coefficient of word categorization

	Word Categories	Spearman correlations coefficient (%)*
Relations	Synonymy	51.6
	Antonymy	61.7
	Meronymy	67.8
	Kinship	75.9
	Hyponymy	61.3
Cross-POS -Relations	Adj - Noun	51.9
	Noun - Verb	58.8
	Noun - Noun	39.9
	No Categorization	65.1

*Note: in order to convert Spearman coefficient correlation into [0, 1] range, we used the ad hoc normalized Spearman coefficient correlation.

4. Word Synthesis

In the GloVe model, individual words are recognized in terms of spacing, and it was thought that it would be necessary to study what features or aspects of Korean differ themselves from those of English from the computational linguistic point of view. This is particularly because, when it processes linguistic input, the computer recognizes individual words in terms of spacing. It is hence possible for two elements to be processed as one unit in the absence of spacing between them. This is very important to make a note of for the Korean data. As an agglutinative language, Korean allows a root to have multiple particles after it without spacing. Thus, words consisting of the same root and different affixes are each recognized as a different individual word and processed in different ways in producing word pairs. It should be noted that there is a tendency for human beings to comprehend various words containing the same stem *kangaji-nun* ('puppy'-'Nominative') and *kangaji-to* ('puppy'-'also'). Hence, as this paper deals with semantic similarity, those words having the same root (or semantic notion) should be treated and grouped as the same unit.

5. Word Recategorization

This research has also noted that Korean has a larger kinship terms than English. Based on this observation, we have developed a separate category for kinship terms. While Pennington et al. (2014) did not categorize word pairs, we attempted to recategorize Korean input data by taking its own different grammatical features into consideration. Along this line, it was thought that we must (1) determine whether words belong to the same parts of speech (POS) (relations or cross-POS-relations, for example), (2) subclassify words in the same POS based on their lexical relations, and (3) divide other types of words of POS into the pairs of adjective-noun, noun-noun, and noun-verb. These categories are illustrated in an outline presented in Table 2. (See additional details in Appendix.)

Table 2. Word categorization

Categories	No. of pairs	Examples
Relations		
Synonymy	156	*emma* (mom) – *emeni* (mother), *appa (dad)* – *apeci* (father)
Antonymy	67	*kyelhon* (marriage) - *ihon* (divorce), *cohun* (good) – *nappun* (bad)
Meronymy	130	*nwuntongca* (eyeball) – *nwun (eye)*, *nwun* (eye) – *elkwul* (face)
Kinship	73	*emma* (mom) – *appa* (dad), *atul* (son) – *ttal* (daughter)
Hyponymy	46	*nala* (nation) - *hankwuk* (Korea), *tosi* (city) – *sewul* (Seoul)
Cross-POS-relations		
ADJECTIVES-NOUN	122	*chakhan* (generous) – *salam* (person), *masissnun* (delicious) – *pap* (meal)
NOUN-VERB	47	*pap* (meal) – *mekessta* (ate), *cam* (sleep; noun) – *cassta* (slept)
NOUN–NOUN	138	*siemeni* (mother-in-law) – *sitayk* (in-laws), *aki* (baby) – *pyengwen* (hospital)
Total	819	

5.1 Synonymy

It is assumed in this study that similarity is determined by semantic attributes of a pair of words to compare. For example, the synonyms *emma* (mom) and *emeni* (mother) share a number of attributes, and therefore their attributional similarity is high. Along this reasoning, it is thought that words are similar if their attributional similarity is high.

5.2 Antonymy

As noted in the literature, antonymy is distinct, as well, among many other relations in that it displays both a sense of closeness and that of distance (Cruse, 1986). Antonyms convey a contrast as they co-occur in the same sentence (Murphy and Andrew, 1993). The contrast is considered a major property to represent the word meaning, evidenced by all major taxonomies across the languages. Antonymous concepts are not semantically similar, but semantically related.

5.3 Meronymy

Some lexical pairs hold part-whole relationships, and are called meronymy. That is, meronymy is a term to refer to the relationship of a smaller part with the whole. This part-whole relationship tends to be hierarchical since each part of the word is inherited from its superordinate: a finger is part of a hand, which is part of an arm. It is important, on the other hand, to notice that parts are not inherited upward because they may be characteristic only of specific kinds of things rather than the class as a whole. For example, an arm is a part of the body and has fingers, but not all kinds of body have fingers.

5.4 Kinship

Kinship systems have largely developed into terminology in many languages to refer to the persons whom an individual is related to through kinship. Every language has different expressions to describe family members and relatives. Korean has an especially rich system of kinship terms developed relatively well due to the socio-cultural tradition of taking a serious view on the various types of human relationships (H-K Kim 1967, 1983; H-S Wang 1988, 1990, 1992). In this section, we propose that kinship word pairs should be one of the categories to examine in the Korean corpus.

5.5 Hyponymy

Hyponymy is an important relation between two words in that the meaning of one word is included in that of the other word. Hyponymy relations are mostly related to the concept of entailment. That is, the denotations of the hyponym is part of

the denotation of the superordinate. In addition, a pair of hyponymy is a useful set for the automatic extraction for tasks such as document indexing and question answering (Mititelu, 2008). Our observation that hyponymy would be a useful category is supported by Oakes (2005) in which Hearst's patterns were employed to capture hyponym-hypernym pairs in a pharmaceutical text and proved their high effectiveness.

5.6 Adjective – Noun

According to the Montague grammar, each syntactic form involves a uniform semantic type. Hence, nouns and adjectives are the same in terms of properties of the entities that are encoded and required for an expression to be made. In addition, Lapata and Lascarides (2003) have also presented a supporting piece of evidence in their experiment where participants rated the degree of similarity of adjective-noun combinations as being high. These studies indicate that the adjective-noun pairs were observed to demonstrate semantic similarity.

5.7. Noun – Verb

Recent studies have reported on experimental data in which processing is activated and facilitated by the non-accidental noun-verb pairs, such as 'watch TV,' 'ride a bike,' and 'bake a cake,' as opposed to 'watch a soup,' 'ride a mountain,' and 'bake water,' amont others (McRae et al., 1998; McRae et al., 2005). This finding points to a possiblity that nouns and verbs are structured along a set of semantic features that may be compatiable each other, i.e., TV is something that is typically watched, but not cooked, a bike normally expected to be ridden, but not watched, and so on. These semantic compatibility between and among words is likely to determine the relation and occurrences of nouns and verbs as a set, ultimately being integrated into a net in the mental lexicon.

This noun-verb relation is thought in syntax to involve transitivity of the verb that requires a subject and an object argument. The syntactic mechanism, however, would require semantic properties of the nouns and verbs involved for a full account of their cooccurrences to be provided. For example, why 'ride a bike,' but not 'ride a cake, is grammatical cannot be explained away in the absence of the semantic relatedness between the two words, namely that 'bike,' but not 'cake,' is semantically something to be allowed to be ridden or ride-able.

5.8. Noun (Agent) – Noun (Location)

Nouns also activate expectations about other nouns occurring as co-arguments in the same sentence that denotes the same event (key – door) (Hare et al., 2009). Pado and Lapata (2007) conducted an experiment by collecting human judgments of similarity for noun-noun combinations using a rating scale. By adoopting correlation analyses, they examined the high semantic relationship between human ratings and corresponding vector values.[5)]

6. Conclusion

In conclusion, it is proposed in this paper that word pairs in Korean should be represented in terms of the new perspective of the categories closely tailored to Korean-specific morphological properties. As mentioned at the outset of this paper, this proposal was initially triggered by the results in Table 1. Based on the empirical data together with the Korean-specific lexical and cross-POS-relations in Table 2, it is concluded that, provided with the GloVe Models, languages must be analyzed by varying methods so that semantic components across languages may allow varying semantic distance in the vector spaces. It is hoped that the recategorization proposed above will provide an insight into the GloVe Models in constructing a Korean corpus.

References

[1] Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics,* 32(1), 13-47.

[2] 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.

[3] Yang et al. (2015). A Study on Word Vector Models for Representing Korean Semantic Information. *Journal of the Korean Society of Speech Sciences, 7(4),* 165-166.

[4] Lopukhin, A. S. (2015). The origin of life is the prerogative of primordial planets of novas. *Herald of the Russian Academy of Sciences, 85(5)*, 453-458.

[5] Cruse, D. A. (1986). *Lexical semantics*. Cambridge University

5) As seen in Figure 1, the Spearman correlation coefficient for the noun-noun pair was 39.9%. In our study we only looked into one of many semantic relations expressed in the pair, agent-locative, while we observed that the pair involves a variety of semantic relations, e.g., possessor-possessed, result-cause, etc., which will be examined in our future studies.

Press.

[6] Murphy, G. L., & Andrew, J. M. (1993). The conceptual basis of antonymy and synonymy in adjectives. *Journal of memory and language, 32(3)*, 301-319.

[7] Kim, H. K. (1967). Korean kinship terminology: A semantic analysis. *Language Research, 3(1)*, 70-81.

[8] Mititelu, V. B. (2008). Hyponymy patterns. In *Text, Speech and Dialogue* (pp. 37-44). Springer Berlin Heidelberg.

[9] Oakes, M. P. (2005). Using Hearst's Rules for the Automatic Acquisition of Hyponyms for Mining a Pharmaceutical Corpus. In *RANLP Text Mining Workshop,* 5, 63-67.

[10] Lapata, M., & Lascarides, A. (2003). A probabilistic account of logical metonymy. *Computational Linguistics, 29(2)*, 261-315.

[11] McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language, 38(3),* 283-312.

[12] McRae et al. (2005) A basis for generating expectancies for verbs from nouns. *Memory & Cognition*, *33(7)*, 1174-1184.

[13] Hare, M., Elman, J. L., Tabaczynski, T., & McRae, K. (2009). The wind chilled the spectators, but the wine just chilled: Sense, structure, and sentence comprehension. *Cognitive Science, 33(4)*, 610-628.

[14] Padó, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics, 33(2),* 161-199.

• **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

• **Koo, Myoung-Wan**
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

• **Cho, Sook Whan, corresponding author**
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

Appendix

5.1 Synonymy

high
emma (엄마) - emeni (어머니), appa (아빠) - apeci (아버지), manphyen (남편) - sinlang (신랑)

medium
pangpep (방법) - swutan (수단), salam (사람) - inkan (인간), mwun (문) - hyenkwan (현관)

low
nala (나라) - kwukka (국가)

4.2 Antonymy

high
kyelhon (결혼) - *ihon* (이혼), *napputa* (나쁘다) - *cohta* (좋다), *moluta* (모르다) - *alta* (알다)

medium
kantanhan (간단한) - pokcaphan (복잡한), celpun (젊은) - nulkun (늙은), tatta (닫다) - yelta (열다), cwungyohan (중요한) - sasohan (사소한), manhun (많은) - cakun (작은)

low
kamta (감다) - *ttuta* (뜨다)

4.3 Meronymy

high
son (손) - *phal* (팔), *nwun* (눈) - *elkwul* (얼굴), *meli* (머리) - *mom* (몸)

medium
ip (입) - *elkwul* (얼굴), *ipswul* (입술) - *hye* (혀), *son* (손) - *sonkalak* (손가락)

low
elkwul (얼굴) - *kho* (코), *tali* (다리) - *mom* (몸), *nwun* (눈)- *nwundongca* (눈동자)

4.4 Kinship

High
emma (엄마) - *appa* (아빠), *oppa* (오빠) - *enni* (언니), *oppa* (오빠) - *nwuna* (누나), *emma* (엄마) - *halmeni* (할머니), *emma* (엄마) - *ttal* (딸), *ttal* (딸) - *atul* (아들)

medium
siemeni (시어머니) - *siapeci* (시아버지), *oppa* (오빠) - *hyeng* (형), *namphyen* (남편) - *siemeni* (시어머니), *oppan* (오빠) - *yetongsayng* (여동생)

low
pwumonim (부모님) - *casik* (자식)

4.5 Hyponymy

high
ton (돈) - *saynghwalpi* (생활비), *ton* (돈) - *welkup* (월급), *pap* (밥) - *cemsim* (점심), *nala* (나라) - *hankwuk* (한국)

medium
pwuomonim (부모님) - *emma* (엄마), *salam* (사람) - *yeca* (여자), *ton* (돈) - *yongton* (용돈), *salam* (사람)- *namca* (남자), *nala* (나라) - *mikyuk* (미국), *pyengwen* (병원) - *chikwa* (치과), *pyengwen* (병원) - *soakwa* (소아과)

low
hakkyo (학교) - *tayhakkyo* (대학교), *hakkyo* (학교) - *kotunghakkyo* (고등학교), *pyengwen* (병원) - *ungkupsil* (응급실), *kamceng* (감정) - *cilthwu* (질투)

5.6. Adjective - Noun

high
chinhan (친한) - *chinkwu* (친구), *nappun* (나쁜) - *salam* (사람), *hwaksilhan* (확실한) - *pangpep* (방법), *massissnun* (맛있는) - *pap* (밥), *alamtawun* (아름다운) - *seysang* (세상), *calmostonyn* (잘못된) - *hayngtong* (행동)

medium
cohun (좋은) - *salam* (사람), *simkakhan* (심각한) - *mwuncey* (문제), *wuwulhan* (우울한) - *kipwun* (기분), *hayan* (하얀) - *nwun* (눈), *ikicekin* (이기적인) - *hayngtong* (행동), *celmun* (젊은) - *nai* (나이), *silin* (시린) - *kasum* (가슴), *palkun* (밝은) - *mosup* (모습)

low
sasohan (사소한) - *kamceng* (감정), *kantanhan* (간단한) - *pangpep* (방법), *mimyohan* (미묘한) - *kamceng* (감정), *yeyppun* (예쁜) - *elkwul* (얼굴), *kanunghan* (가능한) - *kyengwu* (경우)

5.7 Noun - Verb

high
mwun (문) - *yelta* (열다), *pap* (밥)- *mekta* (먹다), *cam* (잠) - *cata* (자다)

medium
nwun (눈) - *ttuta* (뜨다), *son* (손) - *capta* (잡다), *ai* (아이)

- *khiwuta* (키우다), *hakkyo* (학교) - *tanita* (다니다), *suthuleysu* (스트레스) - *patta* (받다), *pyengwen* (병원) - *kata* (가다)

low
kwansim (관심) - *kacita* (가지다), *hakkyo* (학교) - *colephata* (졸업하다), *cali* (자리) - *chacihata* (차지하다), *kitay* (기대) - *kelta* (걸다)

5.8 Noun – Noun

high
iyaki (이야기) - *tayhwa* (대화), *samwusil* (사무실) - *hoysa* (회사), *sitayk* (시댁) - *siemeni* (시어머니), *sacang* (사장) - *hoysa* (회사)

medium
yenlak (연락) - *tapcang* (답장), *yenlak* (연락) - *mwunca* (문자), *kwelhon* (결혼) - *yenay* (연애), *ihay* (이해) - *yongse* (용서), *hoysa* (회사) - *cikwen* (직원)

low
yenin (연인) - *kwankey* (관계), *kamceng* (감정) - *phyohyen* (표현), *pyengwen* (병원) - *chilyo* (치료), *nampwuk* (남북) - *kwankey* (관계)