Latent Semantic Kernels for WordNet: Transforming a Tree-like Structure into a Matrix

Young-Bum Kim
Department of Computer Engineering
Hallym University
Chuncheon, Gangwon, 200-702, Korea
stylemove@hallym.ac.kr

Yu-Seop Kim
Department of Computer Engineering
Hallym University
Chuncheon, Gangwon, 200-702, Korea
yskim01@hallym.ac.kr

Abstract

WordNet is one of the most widely used linguistic resources in the computational linguistics society. However, many applications using the WordNet hierarchical structure are suffering from the word sense disambiguation (WSD) caused by its polysemy. In order to solve the problem, we propose a matrix representing the WordNet hierarchical structure. Firstly, we transform a term as a vector with elements of each corresponding to a synset of WordNet. Then, with singular value decomposition (SVD), we reduce the dimension size of the vector to represent the latent semantic structure. For evaluation, we implement an automatic assessment system for short essays and acquire reliable accuracy. As a result, the scores which are assessed by the automatic assessment system are signi£cantly correlated with those of human assessors. The new WordNet is expected to be easily combined with other matrix-based approaches.

1 Introduction

For a decade, WordNet [4] has been one of the most widely used linguistic resources, and is a broad coverage lexical network of English words [2]. [1] made use of WorNet to measure the semantic distance between concepts and [13] used WordNet to disambiguate word senses. [9], [5], and [11] measured the relatedness of concepts, not words themselves, and the relatedness of words were estimated by using an assumption proposed by [11]. [11] measured the relatedness between words by calculating the relatedness between the most-related pair of concepts.

The assumption would be proper to measure the relatedness of words themselves. However, it is less proper to measure the similarity between sentences, paraphrases, and even documents, and a new complicated method is needed to be developed to disambiguate sense of each word with considering its context.

Latent Semantic Kernel, which was proposed by [3], has shown simpli£ed methods to measure the similarity between documents and also has been applied to many applications such as an automatic assessment system of [7]. Unlike the WordNet based methods, the kernel method fortunately has no need to consider the polysemy problem during measurement process. This is the reason we tried to transform the WordNet structure into the kernel form, that is, a matrix. This matrix from WordNet has another nice property that it can easily be integrated with other matrix-based information in other application later.

At £rst, we built a term-synset matrix, like a term-document matrix in traditional IR society, with a given Korean WordNet which is called KorLEX [10]. We initially gave a score, which was calculated based on a distance to the given synset, to each element of the matrix. Then we constructed a new semantic kernel from the matrix using the SVD algorithm. For evaluation of the new kernel matrix, we implemented an automatic assessment system for short essay questions. The correlation coef£cient between the automatic system and a human was 0.922.

Section 2 will describe the representation of the transformed WordNet. The latent semantic kernel method integrated with WordNet will be explained in section 3. The experimental results and the concluding remarks will be described in section 4 and 5, respectively.

2 Transformed WordNet Structure

Synsets (synonym sets) are representing speci£c underlying lexical concepts in the WordNet. Even though the WordNet has been utilized to solve many semantic problems in computational linguistics and information retrieval societies, its inherent polysemy problem, that is, one term could be shown in multiple synsets, has also caused another problems. In order to solve the problem, we adapted the

latent semantic kernel to the WordNet.

Firstly, we built a term-synset matrix by transforming the WordNet hierarchical structure. Each row vector of the matrix is associated with each term listed in WordNet and the terms are also appearing more than once in corpus data having about 38,000 documents. The vector can be represented like:

$$t_j =< s_1, s_2, \ldots, s_i, \ldots, s_N > \quad (1)$$

, where t_j means a row vector for the j-th term and N is the total number of synsets in the WordNet. And s_i, set to zero initially, is calculated by

$$s_i = \frac{\alpha}{2^k} \quad (2)$$

, where α is a constant value. The s_i is decreasing along with the number of edges, $0 \leq k \leq k_{max}$, on the path from the synset including the term t_j to the i-th synset. The k_{max} decides the range of synsets related with the term. If k_{max} is increasing, more synsets will be regarded as to be related with the synset of the term. In this paper, we decided that $k_{max} = 2$ and also $\alpha = 2$.

Figure 1 from [12] shows a part of WordNet extract for the term 'car'. The term appears in multiple synsets, {car, gondola}, {car, railway_car}, and {car, automobile}. Then the values of s_i for the synsets are determined to $\frac{(\alpha=2)}{(2^0=1)} = 2$ in this paper. The adjacent synsets to {car, automobile}, which are {motor_vehicle}, {coupe}, {sedan}, and {taxi} in £gure 1, are all given $\frac{(\alpha=2)}{(2^1=2)} = 1$s as their s_i values. This procedure is continued until the k_{max}th adjacent synsets are faced.

3 Latent Semantic Kernel for Similarity Measurement

With the initial term-synset matrix, A, created above, we build the latent semantic kernel [3]. Similarity between documents, d_1 and d_2, is estimated as follows.

$$sim(d_1, d_2) = cos(P^T d_1, P^T d_2) = \frac{d_1^T P P d_2}{|P^T d_1||P^T d_2|} \quad (3)$$

, where P is a matrix transforming documents from an input space to a feature space. A kernel function $k(d_1, d_2) =< \phi(d_1), \phi(d_2) >$ uses the matrix P to replace $\phi(d_1)$ with $P^T d_1$.

To £nd P, the term-synset matrix A is transformed by using SVD like

$$A = U \Sigma V^T \quad (4)$$

, where Σ is a diagonal matrix composed of nonzero eigenvalues of AA^T or $A^T A$, and U and V are the orthogonal eigenvectors associated with the r nonzero eigenvalues of AA^T and $A^T A$, respectively. The original term-synset matrix (A) has size of $m \times n$. One component matrix (U), with $m \times r$, describes the original row entities as vectors of derived orthogonal factor value, another (V), with $n \times r$, describes the original column entities in the same way, and the third (Σ) is a diagonal matrix, with $r \times r$, containing scaling values when the three components are matrix-multiplied, the original matrix is reconstructed. The singular vectors corresponding to the k ($k \leq r$) largest singular values are then used to de£ne k-dimensional synset space. Using these vectors, $m \times k$ and $n \times k$ matrices U_k and V_k may be rede£ned along with $k \times k$ singular value matrix Σ_k. It is known that $A_k = U_k \Sigma_k V_k^T$ is the closest matrix of rank k to the original matrix A. And U_k is replaced with P. [8] explains more details of above SVD-based methods, latent semantic analysis (LSA).

4 Evaluation

In this section, we will £rstly explain the automatic assessment procedure developed to evaluate the usefulness of the new kernel based on WordNet, and then show the usefulness by giving the assessment accuracy.

4.1 Automatic Assessment for Short Essay

Figure 2 shows the whole process of the automatic assessment system developed to evaluate the new kernel. The whole process is started with the sentence input written by a student. At £rst, Korean Morphological Analyzer (KMA) [6] extracts main indexed terms from the input sentence. With the term list, an initial vector is constituted with elements of the vocabulary generated from both a large document collection and WordNet. 16,000 words were listed in the vocabulary. Then, the dimension is reduced by computing $P^T d$, where P is the kernel from WordNet and d is the initial vector. With model sentences, which were created by instructors and were transformed on the same way as the input student sentence, the similarities are estimated by using the equation (3), where the student sentence is mapped into d_1 and the model sentences are mapped into d_2. In this research, we stored £ve model sentences for each question. Finally, the highest similarity value is determined as the £nal score of the student sentence.

4.2 Comparison to Human Assessors

We gave 30 questions about Chinese proverbs to 100 students, requiring proper meaning of each proverbs. Firstly, a human assessor decided whether each answer is correct or not and gave one of two scores, 1, and 0, respectively. The

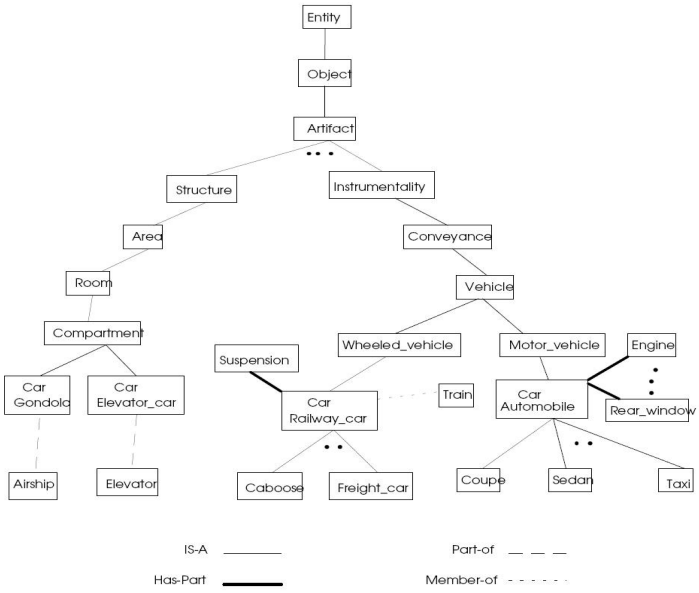

Figure 1. WordNet Extract for the term 'car'

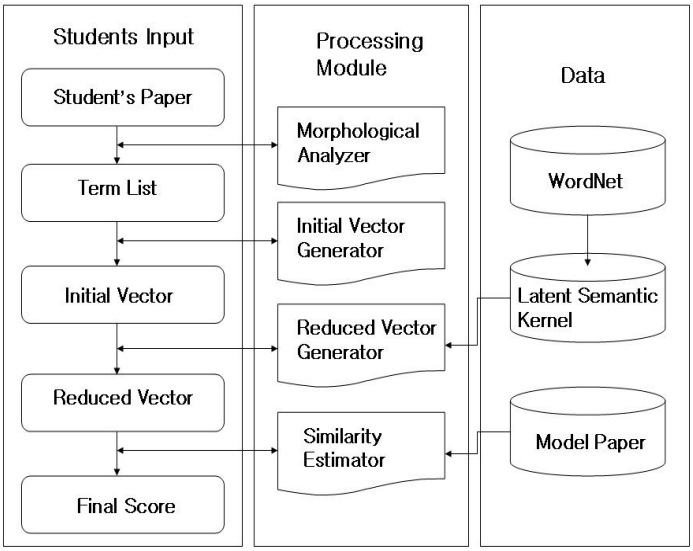

Figure 2. Automatic Assessment Process

score of each student are then computed. Likewise, our assessment system does the same thing except that the system gives a partial score to each answer ranged from 0 to 1.

Figure 3 shows the correlation between the scores from the human assessor and those of the automatic assessor. The correlation coef£cient value is 0.922. As shown in £gure 3, the assessment system tends to give a little lower score to the answer than the human. It is caused by the lack of information used in scoring. The less information has the assessor, the lower score it gives.

Table 1. Error rate for each threshold

threshold	error rate	threshold	error rate
0.1	0.184	0.2	0.229
0.4	0.228	0.6	0.248

We also evaluate the coincidence level of two assessors' decision whether the answer is correct or not. At £rst, we randomly decided the threshold value which roles as a boundary score of correct and wrong answer. If a similarity value is larger than the threshold value, for example, then the answer is decided as a correct answer. Then we can count the number of answers decided to be same by both assessors. Table 1 shows the comparison results, regarding to the threshold values. When we used threshold values lower than 1.0, the error rates were raised again.

5 Concluding Remarks

We proposed a latent semantic kernel for WordNet. Firstly, we transformed the hierarchical structure into a matrix, representing a term as a vector. Like corpus-driven latent semantic kernel, then, we adapted the SVD algorithm to present the latent semantic structure of WordNet. We evaluated the usefulness of the new kernel by integrating it into an automatic assessment system which showed high correlation with a human assessor.

Terms in WordNet, in this paper, are represented as vectors similar to terms of other vector space models in many related applications. It will be possible for other researches to integrate the WordNet with other approaches based on 'bag of words' concept. We will try to £nd a new method to integrate the WordNet and many data-driven methods to integrate term similarity and term relatedness.

Acknowledgments

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Resrach Promotion Fund) (KRF-2006-331-D00534)

References

[1] A. Budanitsky and G. Hirst. Semantic distance in wordnet: An experimental, application-oriented evaluation of £ve measures. *Workshop on WordNet and Other Lexical Resources, Second Meeting of the North Americal Chapter of the Association for Computational Linguistics*, 2001.

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

[3] N. Cristianini, J. Shawe-Taylor, and H. Lodhi. Latent semantic kernel. *Journal of Intelligent Information Systems*, 18(2-3):127–152, 2002.

[4] C. Fellbaum. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA, 1998.

[5] J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of International Conference on Research in Computational Linguistics (ROCLING X)*, pages 19–33, 1997.

[6] S.-S. Kang. *Gerneral Purposed Morphological Analyzer HAM Ver. 6.0.0.* http://nlp.kookmin.ac.kr, 2004.

[7] Y.-S. Kim, W.-J. Cho, J.-Y. Lee, and Y.-J. Oh. An intelligent grading system using heterogeneous linguistic resources. *Proceedings of 6th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL-2005)*, pages 102–108, 2005.

[8] T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse Processes*, (25):259–284, 1998.

[9] C. Leacock and M. Chodorow. Combining local context and wordnet similarity for word sense identi£cation. *In Christiane Fellbaum, editor,* WordNet: An Electronic Lexical Database, pages chapter 11:265–283, 1998.

[10] E. R. Lee and S. S. Lim. *Korean WordNet ver.2.0.* Korean Language Processing Laboratory, Pusan National University, 2004.

[11] P. Resnik. Using information content to evaluate semantic similarity. *Proceedings of the 14th International Joint Conference on Arti£cial Intelligence*, 1995.

[12] R. Richardson, A. F. Smeaton, and J. Murphy. *Using WordNet as a Knowledge Base for Measuring Semantic Similarity between Words*. CA-1294, Dublin, Ireland, 1994.

[13] E. M. Voorhees. Using wordnet to disambiguate word senses for text retrieval. *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1993.

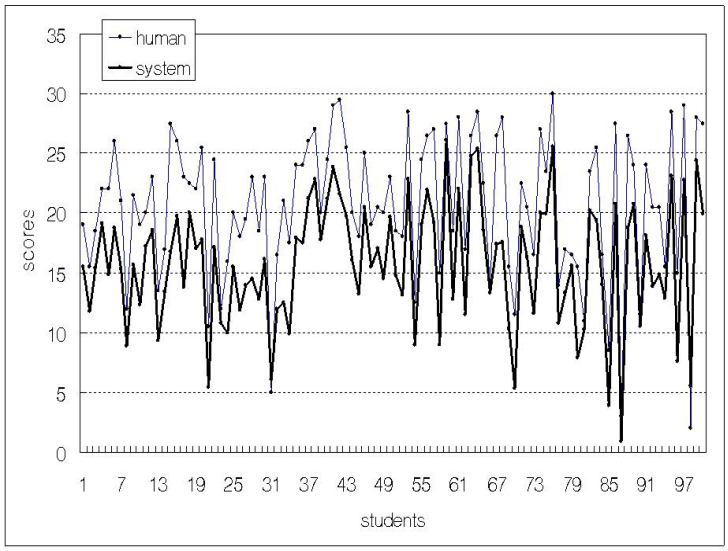

Figure 3. Scores graded by Human Assessor and Automatic Assessor