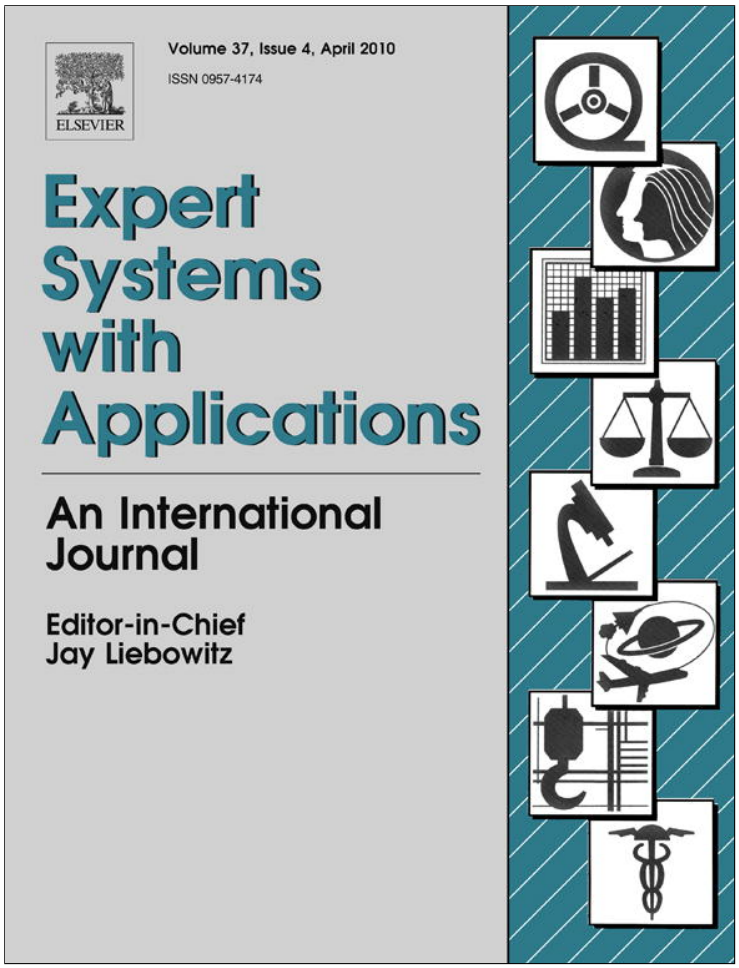
Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/copyright

Author's personal copy

Expert Systems with Applications 37 (2010) 3219–3228

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

An autonomous assessment system based on combined latent semantic kernels

Young-Bum Kim, Yu-Seop Kim *

Department of Computer Engineering, Hallym University, 39 Hallymdaehak-gil, Chuncheon, Gangwon-do 200-702, Republic of Korea

ARTICLE INFO

Keywords:
Autonomous assessment system
Latent semantic kernel
WordNet
Singular value decomposition
Corpus
Combined kernel
BLEU

ABSTRACT

In this paper, we develop an autonomous assessment system based on the kernel combinations which are mixed by two kernel matrices from the WordNet and corpus. Many researchers have tried to integrate these two resources in many applications, to utilize diverse information extracted from each resource. However, since two resources have been represented in quite different ways, one resource has been secondary to another. To fully integrate two resources at the same level, we first transform the WordNet, which has a hierarchical structure, into a matrix structure. Concurrently, another matrix, which represents a co-occurrence of words in the collection of text documents, is constructed. We then build two initial latent semantic kernels from both matrices and merge them into a new single kernel matrix. When we merge two matrices, we split each initial matrix into independent columns and mix the columns with various methods. We acquire a few combined kernel matrices which show various performances in experiments. Compared to the basic vector space model, original kernel matrices, and the BLEU based method, the combined matrices improve the accuracy of assessment.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

1.1. Background of this research

With the explosive growth of information technology, remote education (or e-learning) has widely spread into many areas. E-learning requires achievement evaluations of students after completing their learning. For the evaluation, both a multiple-choice exam and an essay exam have been used most widely in a conventional education community (Geiger, 1996; Martinez, 1999; Steele, 1997).

In spite of its objectiveness and ease of implementation, the multiple-choice exam has limits on evaluating the various learning effects on students. To the contrary, the essay-style exam can supplement the limitations of multiple-choice exams, but human instructors have to be involved in the scoring process because of the insufficient natural language processing technology (Bang, Hur, Kim, & Lee, 2001).

1.2. Related studies

In the area of long essay evaluation, a number of researches have been done. A hybrid feature identification method has been used, including syntactic structure analysis, rhetorical structure analysis, and topical analysis (Burstein et al., 1998; Page & Peterson, 1995).

The Intelligent Essay Assessor (IEA) based on the Latent Semantic Analysis (LSA) (Landauer, Foltz, & Laham, 1998) is built (Landauer, Laham, & Foltz, 2000), which is both a computational model of human knowledge representation and a method for extracting the semantic similarity of words and passages from a text. The research shows about a 0.9 correlation coefficient with two expert human graders.

Research related to the long essay grading has been improved and has also become more practical. However, that research is basically designed for long essays composed of many sentences and terms, which can provide sufficient information about the contents and structure of the essay.

An intelligent grading system based on case-based reasoning (Aamodt & Plaza, 1994) is proposed to recover the insufficiency of information contained in the answer (Briscoe-Smith & Evangelopoulos, 2002). The measurement of similarity among a submitted answer and multiple stored answers is one of the primary issues in this literature. However, detailed descriptions of the similarity measurement are not shown in the paper.

Computational linguistics technologies in marking short free text responses automatically are investigated (Pulman & Sukkarieh, 2005). They initially use information extraction technologies, such as an HMM POS tagger, and an NP and VG finite state machine, and write patterns by hand. Later, they use machine learning technologies to avoid the very high cost of building the patterns. Although they utilize automated learning methods, their system still depends much on human experts when building patterns.

* Corresponding author. Tel.: +82 33 248 2324; fax: +82 33 242 2524.
E-mail addresses: stylemove@hallym.ac.kr, stylebbum@gmail.com (Y.-B. Kim), yskim01@hallym.ac.kr (Y.-S. Kim).
URL: http://www.hallym.ac.kr/~yskim01 (Y.-S. Kim).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.09.058

Author's personal copy

The BLEU (BiLingual Evaluation Understudy) score (Papineni, Roukos, Ward, & Zhu, 2002) is the best known machine translation evaluation method. The BLEU estimates the similarity between the translated sentence and the reference sentences. This process is very similar to the autonomous assessment process in terms of estimating the similarity among sentences. The BLEU is used for assessing answers in English and Spanish (P'erez, Alfonseca, & Aodr, 2004). More detailed explanation on the BLEU is given in later section.

1.3. Our approach

For a decade, the WordNet (Fellbaum, 1998), a broad coverage lexical network of English words (Budanitsky & Hirst, 2006), has been one of the most widely used linguistic resources. For its unique and informative hierarchical structure, the WordNet has been widely used not only to measure the semantic distance between concepts (Budanitsky & Hirst, 2001) but also to disambiguate word senses (Voorhees, 1993).

In spite of its abundant and fine-tuned semantic information, the resource has been known to cause additional operations to be applied to many semantics-related applications. It is pointed out that the WordNet must be incorporated with the means for determining an appropriate sense during sense identification (Miller, 1995). The importance of achieving accurate automatic word sense disambiguation in the WordNet-based text retrieval is mentioned (Gonzalo, Verdejo, Chugur, & Cigarran, 1998). In addition to the polysemy, a domain specific semantic knowledge must be taken into account for being integrated with the WordNet. The WordNet is compared to domain specific linguistic resources, and shows that it must be supplemented with domain specific knowledge (Burgun & Bodenreider, 2001).

Considering the limitations noted above, we propose a new approach based on the latent semantic kernel (Cristianini, Shawe-taylor, & Lodhi, 2002) in this paper. The latent semantic kernel has shown as a simplified method to measure the similarity between documents and also has been applied to many applications such as an autonomous assessment system proposed by Kim, Cho, Lee, and Oh (2005). Unlike the WordNet, the kernel method has no need to consider the polysemy problem during the measurement. Also, a domain specific kernel could be built by using the corpus data. To integrate the WordNet with the domain specific kernel, we try to transform the WordNet structure into the kernel form, that is, a matrix.

To begin, we build a term–synset matrix (Kim & Kim, 2008) with a given Korean WordNet which is called KorLEX (Lee & Lim, 2004) and, concurrently, also build a term-document matrix of traditional IR society from domain specific corpus data. When building the WordNet-based matrix, we define each cell of the matrix as a distance between two synsets on the WordNet. At the same time, we also make a term-document matrix of a given document collection. Then we construct new semantic kernels from two matrices using the singular value decomposition (SVD) algorithm (Landauer et al., 1998). We, then, combine two kernel matrices into one to synthesize domain specific features and the WordNet inherent features. When synthesizing two matrices, first, we decompose matrices each with their column. Then we recombine the columns from the two matrices with three principles.

We build an autonomous assessment system by using the combined kernels which are used for the basic scoring schemes. The whole structure of the system will be shown later. Basically, the system already has multiple model answers for each question. After receiving the student answer, the system calculates the degree of similarity between the student answer and each model answer for the same question, by using the combined kernel. After a similarity estimation, the system decides the highest similarity value as the score of the given student answer.

In the experiment, we try to find out the most profitable kernel combination method for this system. To verify our approach, we compare our method to three existing methods: the basic vector model, which does not try to reduce the answer vectors, as the baseline, the Latent Semantic Kernel (Cristianini et al., 2002) based on the Latent Semantic Analysis (Landauer et al., 2000) and the BLEU score (Papineni et al., 2002).

The remainder of this paper is organized as follows: Section 2 explains three existing methods related to this research, like Basic Vector Space Model, Latent Semantic Kernel, and the BLEU. Section 3 describes the initial matrices constructed from the WordNet and corpus. Section 4 shows the proposed approach, the combined semantic kernel. Section 5 describes the whole structure and flow of our autonomous assessment system. The implementation and its experimental results are provided in Section 6, and concluding remarks are discussed in Section 7.

2. Three existing approaches

In this section, we introduce three existing approaches related to our research. First, we describe a little about the basic vector space model, which has been a base of many 'bag-of-words' approaches in the information retrieval society. Second, the latent semantic kernel method is explained with its background theory, latent semantic analysis. Finally, we describe the basics of the BLEU.

2.1. Vector space model

The vector space model (VSM) is a model for representing the retrieval process of information retrieval systems. Documents in the VSM are represented as points in an n-dimensional space, where n means the number of indexed terms extracted from the constructed document set (Cristianini et al., 2002). The VSM assumes an idea that the meaning of a document is to be represented by a combination of indexed terms appearing in the document (Wong, Ziarko, & Wong, 1985; Salton, Wong, & Yang, 1975; Salton, 1989). If it is possible, the relevance between the document and a given query is possibly measured by calculating the similarity between two vectors representing the document and the query respectively. In this model, the appropriate documents are selected by sorting the relevance degree of retrieved results by descending order of similarity values (Cheong, 2001). The VSM extracts indexed terms from a query and a document, gives weights to the indexed terms, and represents a document and a query as a vector form with n given weight values as follows:

$$d = \{(t_1, wd_1), (t_2, wd_2), \ldots (t_n, wd_n)\}, \tag{1}$$

$$q = \{(t_1, wq_1), (t_2, wq_2), \ldots (t_n, wq_n)\}, \tag{2}$$

where t_i means an indexed term extracted by any indexing scheme. And, wd_i and wq_i are weight values reflecting the degree of importance of an indexed term t_i in document d and a query q, respectively. The weight values are computed by just counting the frequencies of t or specially computing its $tf \cdot idf$ as will be described in Section 3.1.

Given a query q, the relevance degree, $RD(q, d)$, with a document d is simply computed as below:

$$RD(q,d) = cos(q,d) = \frac{q \cdot d}{\|q\|\|d\|} = \frac{\sum_{i=1}^{n}(wq_i \times wd_i)}{\sqrt{\sum_{i=1}^{n} wq_i^2}\sqrt{\sum_{i=1}^{n} wd_i^2}}. \tag{3}$$

However, in basic VSM, since every attribute in a vector has orthogonal relationships to each other, the model could not consider the

Author's personal copy

co-relationship of each indexed term. To solve this problem, a generalized vector space model (Wong et al., 1985), which uses co-occurrence information among the indexed terms appearing in the documents set, is proposed.

2.2. Latent semantic kernel

In this section, we explain the basic concept of the latent semantic kernel (Cristianini et al., 2002). By using the kernel, a similarity between documents, d_1 and d_2, is estimated as follows:

$$sim(d_1, d_2) = cos(P^T d_1, P^T d_2) = \frac{d_1^T P P d_2}{|P^T d_1||P^T d_2|}, \tag{4}$$

where P is a matrix, which maps a document vector into a feature vector of a semantic feature space. In basic VSM, $P = I_M$. (M means the size of vocabulary in the documents set and I_M is an identical matrix.) It is pointed out that a kernel function $k(d_1, d_2) = \langle \phi(d_1), \phi(d_2) \rangle$ uses the matrix P to replace $\phi(d_1)$ with $P^T d_1$ (Cristianini et al., 2002).

To find P, we employ the SVD to factorize the initial matrix M which is represented by

$$M = U \Sigma V^T, \tag{5}$$

where Σ is a diagonal matrix composed of nonzero singular values of MM^T or $M^T M$, and U and V are the orthogonal eigenvectors corresponding to the r nonzero singular values of MM^T and $M^T M$, respectively. The original matrix has a size of $m \times n$. One component matrix (U), $m \times r$, describes the original row entities as vectors of derived orthogonal factor value, another (V^T), $r \times n$, describes the original column entities in the same way, and the third (Σ) is a diagonal $r \times r$ matrix, containing scaling values when the three components are matrix-multiplied, the original matrix is reconstructed. The eigenvectors corresponding to the k $(k \leqslant r)$ largest singular values are then used to define k-dimensional document and WordNet spaces in this paper. Using these vectors, $m \times k$ and $n \times k$ matrices U_k and V_k may be redefined along with $k \times k$ singular value matrix Σ_k. It is known that $M_k = U_k \Sigma_k V_k^T$ is the closest matrix of rank k to the original matrix M. And U_k is replaced with P in formula (4). More details of above SVD-based methods, latent semantic analysis (LSA), are explained in Landauer et al. (1998).

2.3. Bilingual evaluation understudy

BLEU (Papineni et al., 2002) is an IBM-developed metric and is probably the best known and used in the machine translation community.[1] The BLEU starts from the following idea, that is, "The closer a machine translation is to a professional human translation, the better it is." The BLEU focuses on the closeness between human translations and the given machine translation, just as the autonomous assessment system does between model answers and the given student answer.

The BLEU score is based on the following two factors. First, the number of shared n-gram between the candidate translation and its references affects the score in a positive way. To estimate this, the precision of a block of candidate sentences, p_n, is defined and also computed like

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count(n\text{-}gram)}, \tag{6}$$

where $Count_{clip}(n\text{-}gram)$ means the number of shared n-gram between the candidate and the references, whereas $Count(n\text{-}gram)$ means the number of all n-gram in the candidate.

[1] http://www.ics.mq.edu.au/.

Second, on the contrary, if the length of the candidate translation is less than those of references, then the BLEU is affected in a negative way. The brevity penalty, BP, is computed like

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r, \\ e^{(1-r/c)} & \text{if } c \leqslant r, \end{cases} \tag{7}$$

where c is the length of the candidate translation and r is the reference length. Then the BLEU score is to be computed as follows.

$$\mathrm{BLEU} = \mathrm{BP} \cdot exp\left(\sum_{n=1}^{N} w_n \log p_n \right) \tag{8}$$

where N is the maximum length of the n-gram and w_n is a weight value for each n-gram model.

3. Initial matrix construction

In this section, we describe the process of constructing the initial matrices representing the corpus and the WordNet spaces. In order to combine two matrices, we select vocabulary terms appearing in corpus and WordNet at the same time.

3.1. Corpus matrix

First, we extract n indexed terms from N documents. A document d_j has n dimensions and is represented by.

$$d_j = \langle w_{1,j}, w_{2,j}, \ldots, w_{i,j}, \ldots, w_{n,j} \rangle, \tag{9}$$

where $w_{i,j}$ denotes a weight of ith component of the document d_j. In consideration of the relative occurrence frequency and inverse document frequency of the term, the weights are computed as follows:

$$w_{i,j} = tf_{i,j} \cdot idf_i = tf_{i,j} \cdot log\left(\frac{N}{df_i} \right). \tag{10}$$

In formula (10), $tf_{i,j}$ stands for occurrence frequency of the ith component in the given document d_j, df_i, called as a document frequency, denotes the number of documents including the ith component, and idf_i is an inverse document frequency.

An initial corpus matrix M_C is constructed by considering every single d_j vector as a column vector of M_C. Then, the row vector of the M_C represents each indexed term vector,

$$t_i = \langle w_{i,1}, w_{i,2}, \ldots, w_{i,j}, \ldots, w_{i,N} \rangle. \tag{11}$$

3.2. WordNet matrix

Synsets (synonym sets) represent specific 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 other problems. In order to solve this problems, we adapt the latent semantic kernel to the WordNet.

First, the original hierarchical structure of the WordNet is transformed to a term–synset matrix. Each row vector of the matrix is associated with each term listed on WordNet and the terms are also appearing in corpus data of 38,000 documents. The term vector can be represented by

$$t_i = \langle s_1, s_2, \ldots, s_l, \ldots, s_T \rangle, \tag{12}$$

where t_i means a row vector for the ith term and T is the total number of synsets listed on the WordNet. The matrix M_W is composed by collecting t_is as its row vectors, and s_l, set to 0 initially, is calculated by

Author's personal copy

$$s_l = \frac{\alpha}{2^k}, \tag{13}$$

where α is a constant value. The s_l is decreasing along with the number of edges, $0 \leqslant k \leqslant k_{max}$, on the path from the synsets including the term t_i to the lth synset. The k_{max} decides the range of synsets related to the term. As k_{max} is increasing, more synsets will be regarded as to be related to the term. In this paper, we set $k_{max} = 2$ and also $\alpha = 2$. The WordNet matrix M_W is now built by collecting t_i vectors as its row vectors.

For example, Fig. 1 from Richardson et al. (1994) shows a part of WordNet extract for the term 'car'. The term appears in multiple synsets: {car,gondola}, {car,eleavator_car}, {car,railway_car}, and {car,automobile}. Then four s values of the above four synsets are calculated to $\frac{(\alpha=2)}{(2^0=1)} = 2$. The adjacent synsets to {car,automobile}, which are {motor_vehicle}, {coupe}, {sedan}, and {taxi} in Fig. 1, are all given $\frac{(\alpha=2)}{(2^1=2)} = 1$s as their s values. This procedure is continued until the k_{max}th adjacent synsets are faced. Table 1 shows all the s values of synsets in Fig. 1 and their distance to the synsets including 'car'.

More details about the WordNet matrix are provided in Kim and Kim (2008). We use only the is–a relationship since the KorLex defines only the relationship.

Table 1
This table shows the s values of all synsets which are shown in Fig. 1 and connected to the 'car' synset with is–a relationship. The second and the fifth columns show the shortest distance from the 'car' synset. Finally, the third and the sixth columns show the s values of the synsets.

Synset	Distance	s Value
{Entity}	7	0
{Artifact}	5	0
{Instrumentality}	4	0
{Conveyance}	3	0
{Vehicle}	2	0.5
{Wheeled_vehicle}	1	1
{Car,Gondola}	0	2
{Car,Railway_car}	0	2
{Caboose}	1	1
{Coupe}	1	1
{Taxi}	1	1
{Object}	6	0
{Structure}	4	0
{Area}	3	0
{Room}	2	0.5
{Compartment}	1	1
{Motor_vehicle}	1	1
{Car,Elevator_car}	0	2
{Car,Automobile}	0	2
{Freight_car}	1	1
{Sedan}	1	1

4. Combined latent semantic kernel

Given the initial matrices, M_C and M_W, from corpus and WordNet, respectively, we extract latent semantic kernels P_C and P_W. P_C is the result of conventional latent semantic kernel method mentioned in Section 2.2. With two kernel matrices, we construct six combined kernels. The combined kernels $K_i,\ i = 1,\ldots,6$, are developed based on three concepts: concatenation, selection in turn, and sorting by singular values.

4.1. Concatenation

We construct two combined kernels, K_1 and K_2, based on concatenation of P_C and P_W. First of all, the row vectors, tc and tw from two kernels, P_c and P_w respectively, are defined as follows:

$$tc_i = \langle c_{i,1}, c_{i,2}, c_{i,3}, \ldots, c_{i,k-1}, c_{i,k} \rangle, \tag{14}$$
$$tw_i = \langle w_{i,1}, w_{i,2}, w_{i,3}, \ldots, w_{i,k-1}, w_{i,k} \rangle, \tag{15}$$

where $c_{i,j}$ means the jth element of the ith row vector tc_i. And $w_{i,j}$ means the same as $c_{i,j}$ except that it's for tw_i. The k is the number of the largest singular values described in Section 2.2.

A row vector for a term t_i of K_1 is defined as

$$t_i = \left\langle c_{i,1}, c_{i,2}, \ldots, c_{i,\frac{k}{2}}, w_{i,1}, w_{i,2}, \ldots, w_{i,\frac{k}{2}} \right\rangle \tag{16}$$

and a row vector in K_2 is also defined as

$$t_i = \left\langle w_{i,1}, w_{i,2}, \ldots, w_{i,\frac{k}{2}}, c_{i,1}, c_{i,2}, \ldots, c_{i,\frac{k}{2}} \right\rangle. \tag{17}$$

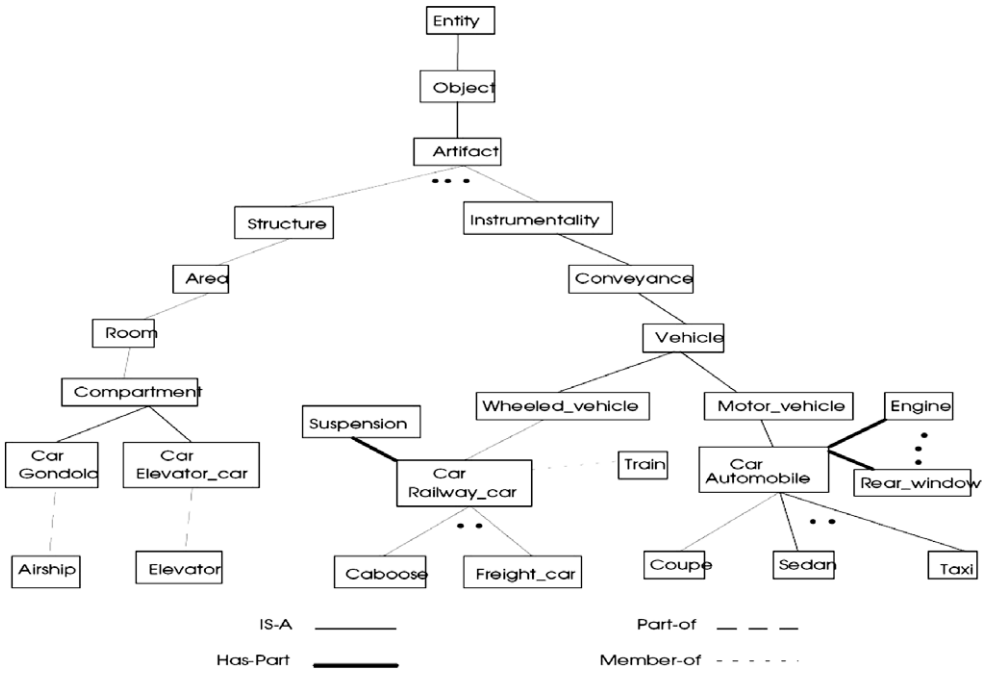


Fig. 1. WordNet Extract for the term 'car'.

Author's personal copy

From the original row vectors, shown in Eqs. (14) and (15), the first halves are extracted and concatenated like above. K_1 takes the first half of *tc* firstly and then the first half of *tw*. K_2 takes that of *tw* at first and then that of *tc*. The difference between K_1 and K_2 is the order of P_C and P_W.

4.2. Selection in turn

We also build a combined matrix by selecting column vectors from P_C and P_W in turn. If a column vector of P_C is inserted into the combined matrix, then a column vector of P_W is to be inserted in turn. A row vector for a term t_i of K_3 is:

$$t_i = \left\langle c_{i,1}, w_{i,1}, c_{i,2}, w_{i,2}, \ldots, c_{i,\frac{k}{2}}, w_{i,\frac{k}{2}} \right\rangle, \tag{18}$$

where $\alpha_{i,j}$ has the same meaning as that of Section 4.1. Fig. 2 shows the difference among three combined kernels, K_1, K_2, and K_3.

4.3. Sorting by singular values

On top of above kernels, we propose three more combined kernels, K_4, K_5, and K_6. Basically, the kernels are made by sorting the singular values of P_C and P_W in total. First, we extract $\frac{k}{2}$ largest singular values and their associated column vectors both from P_C and P_W like former kernels. Instead of simply concatenating two column vectors from two kernels, we mix and sort them in view of their own singular values. K_4 has an ascending order and, on the contrary, K_5 has a descending order.

Further, K_6 is constituted in nearly same way as K_5, except that we extract k largest singular values instead of $\frac{k}{2}$ largest ones. We then sort $2k$ singular values in descending order and only the k column vectors related to the largest k singular values are inserted to the K_6. Although K_4 and K_5 take the same number of column vectors from P_C and P_W, K_6 ordinarily has a different number of column vectors from the two matrices.

A row vector for a term t_i of K_4, K_5, and K_6 is

$$t_i = \langle e_1, e_2, \ldots, e_j, e_{j+1}, \ldots, e_k \rangle, \tag{19}$$

where e_js are from either P_C or P_W. The related singular value of jth column is always larger, smaller in case of K_4, than or equal to that of $(j+1)$th column.

5. Autonomous assessment system

In this paper, we develop an autonomous assessment system for short essay questions. We focus on the Korean questions for junior high school students.

Fig. 3 shows the whole process of the autonomous assessment system of our system. The whole system is divided into three major parts. The first part controls the flow of the student input. The second part consists of four processing modules. The final part includes data and knowledge used for the assessment process.

5.1. Control part and processing module part

The control part and the processing module part are integrated with close communication. The module part receives an answer from the control part and transforms the answer to transfer it to the control part. For this reason, we explain each module with its input and output. The module part is composed of four modules such as the morphological analyzer, the initial vector generator, the reduced vector generator, and the similarity estimator.

The whole process begins with an answer sentence, written by a student, for a given question, which asks the meaning of a Chinese idiom (in Korean, '*Sa-Ja-Seong-Eo*'). For example, when a student gets a question sentence, "What is the meaning of '*Goo-Sa-Il-Saeng*'?", the student writes an answer for the question like '*Jook-Eul Go-Bi-Reul A-Hop-Beon-I-Na Neom-Gim*'.

With this student answer paper, the morphological analyzer extracts indexed terms from the candidate words, called as '*Eo-Jeol*'s in Korean. The Korean Morphological Analyzer (KMA) (Kang, 2004) extracts the main indexed terms from the input Korean sentence. We extract only noun terms since the KMA used in this research is permitted only to extract noun terms. Due to the nice property of Korean that noun words which play roles of roots can be extracted from many of verbal and adjective words, relatively sufficient terms can be extracted from the sentence. From the above student answer sentence, the KMA extracts '*Jook*' from the first word, '*Jook-Eul*', '*Go-Bi*' from '*Go-Bi-Reul*', and '*A-Hop-Beon*' from '*A-Hop-Beon-I-Na*'. Table 2 shows other four examples of the extracted terms from given sentences by using the KMA.

Even though the first sentence and the last sentence in Table 2 are a little different from each other, the system considers them to

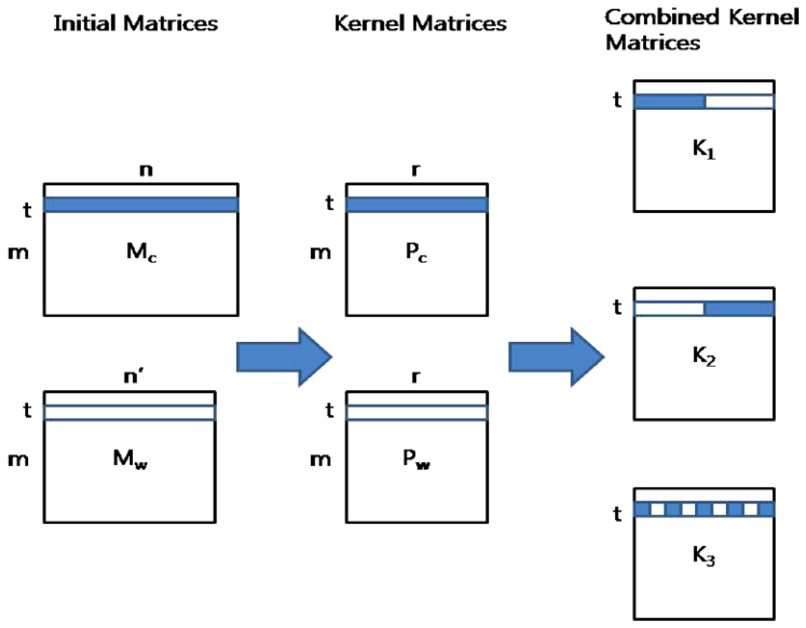


Fig. 2. The procedure of building K_1, K_2, and K_3 from the initial matrices.

Author's personal copy

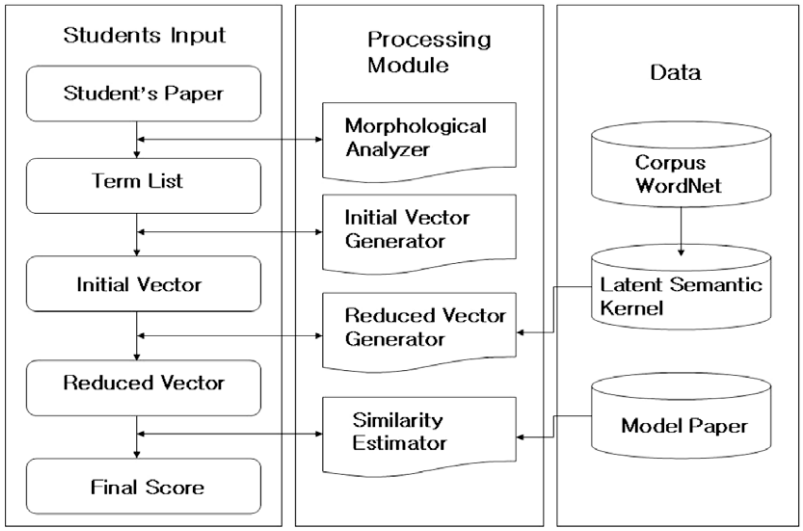


Fig. 3. Automatic assessment process.

be identical because the extracted terms from two sentences are exactly same.

With the term list, the initial vector generator makes an initial vector constituted with elements of the vocabulary generated from both a large document collection and WordNet at the same time. About 16,000 words are listed in the vocabulary in this research. The initial vector is filled with the occurrence frequency of the corresponding indexed terms in the input sentence.

Then, the reduced vector generator reduces the dimension of the initial vector by computing $P^T d$ of Section 2.2, where P is one of the kernels described above and d is the initial vector. We have eight kernels in total, two initial kernels and six combined kernels. Given 16,000 dimension vectors, we reduce to about six hundred dimensions which represent the semantic space. Also, the axis are orthogonal to each other, which makes it possible to utilize the vector space model more properly.

Finally, the similarity estimator calculates the score of the student answer by computing the similarity between the answer paper and the model answer papers. We build five model answers for each question. By using the formula (4), where the student answer is mapped into d_1 and one of the model answers is mapped into d_2 in Section 2.2, the similarities between the answer vector and five model vectors are estimated. The P can be replaced with one of the 8 kernels shown above. Finally, the highest similarity value is determined as the score of the student answer.

5.2. Data and knowledge part

For the autonomous assessment, we, basically, utilize two sorts of linguistic resources, WordNet and corpus. We compute the minimum distance between two synsets in advance, and make a distance matrix. The synset-to-synset distance matrix is extended to the term-to-synset matrix as described in Section 3.2. At this point, we collect tens of thousands of documents and also make a term-document matrix as Section 3.1.

The initial WordNet and corpus matrices are transformed to be the latent semantic kernel matrices as shown in Section 2.2. With the kernel matrices, the student answer and the model answer will be transformed to estimate their similarity, considering the semantic space.

Finally, the model answer papers for each question should be required before scoring the student answers. We prepared for five model answer papers per question. The model papers have the same process as student papers to be used for the assessment. Morphological analysis, vectorization, and the vector reduction process should be taken to the model papers. The only difference from the student answers is that the model answers are in advance stored as their reduced forms in the database.

Table 2
The result of term extraction by using Korean Morphological Analyzer (KMA). The first column shows the given sentences and the second column shows the extracted terms from the corresponding sentence. A word ('Eo-Jeol') can generate multiple terms with KMA as the third row.

Given sentence	Extracted terms
'A-Hop-Beon Jook-Go Han-Beon Sal-A-Nam'	'A-Hop-Beon', 'Han-Beon'
'Yeo-Reo-Beon-Eu Wi-Heom-Eul Moo-Sa-Hi Gyeon-Dyeo-Naem'	'Yeo-Reo-Beon', 'Yeo-Reo-Beon-Eu', 'Yeo-Reo', 'Beon-Eu'
'Eo-Ryeo-Woon Sang-Hwang-E-Seo Sal-A-Nam'	'Sang-Hwang'
'A-Hop-Beon Jook-Go Han-Beon San-Da'	'A-Hop-Beon', 'Han-Beon'

6. Implementation and experimental results

This section includes the implementation of the autonomous assessment system and experimental results. The first subsection will explain the whole structure of its client–server model. Also, the user interface of the system and its execution snapshot will be shown. In the next subsection, we show the experimental results. The experiments try to measure the contribution of each kernel to the assessment process.

6.1. Implementation of the autonomous assessment system

6.1.1. The client–server model

Fig. 4 shows the client–server model which is adapted to the autonomous assessment system. The centralized server executes the assessment process with input from clients, and many clients communicate with the server on TCP/IP. The linguistic resources like the WordNet and semantic kernels are stored in the Database. The server uses the Linux operating system and the Database is running on the MySQL.

6.1.2. Implementation

Students trying to use the system must download the client program from the server. Fig. 5 is showing the user interface of the client program right after executing the program.

Author's personal copy

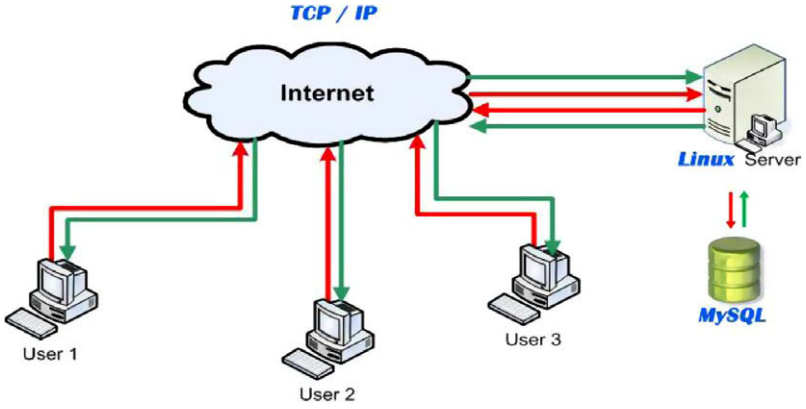


Fig. 4. The client–server model hired in this system.

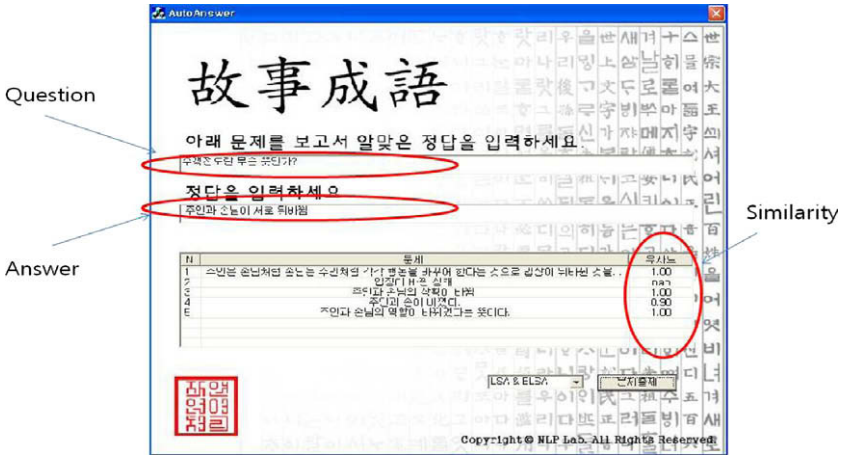


Fig. 5. The client program user interface.

Given a question shown in the first dialog box, students write their answers in the second dialog box. After running the program, they can get the assessment results in the third box having the similarity scores with five model answers. Among the similarity scores, the highest one is determined as the score of the answer for the question.

At this point, the server generates a question, accepts the answer, runs the Korean morphological analyzer, makes a vector, and finally computes the similarity scores with five model answers. The server processing can be checked in Fig. 6.

6.2. Experimental results

In this subsection, we describe two experimental results. The first is about the evaluation of the singular values used to transform the initial matrices. The second is about the assessment accuracy, compared to three existing methods.

6.2.1. Singular value evaluation

The two initial matrices, the corpus matrix, M_C, and the WordNet matrix, M_W, reduce their dimensions by using the singular value decomposition algorithm, and the reduced matrices, P_C and P_W, play a very important role in creating the new combined kernels. Table 3 shows the selected singular values of two reduced matrices and Fig. 7 shows their 600 largest singular values. In this experiment, we put a limitation on the largest k value to be 600, because we are faced with the critical time and space problem to extract singular values when the number of eigenvalues exceeded 600.

Compared to the singular values from the corpus matrix, those of the WordNet matrix show a little smaller values at the same k level. This theoretically means that WordNet space can be represented with a fewer number of axis than corpus. However, it will be clearly changed if we change the method of transforming the WordNet to a matrix.

Until now, a method to compute the optimal number of the largest eigenvalues has not been developed. Ordinarily, the number is determined empirically. In this paper, we do not try to find the optimal number empirically because we consider the problem to be beyond our research scope. Even though 200 can be treated as a turning point in Fig. 7, it cannot be said to be a optimal point.

6.2.2. Assessment accuracy evaluation

For the accuracy evaluation of the proposed system, we collect 30 questions about the Chinese proverbs and also gather 100 students to take this exam. We have a total of 3000 student answers for this experiment.

We perform two types of accuracy evaluations. First, we evaluate the correlation coefficient (Bain & Engelhardt, 1987) between human instructors and our system. We calculate a total score for each student. Every student gets a score from 0 to 30. Table 4 shows the correlation coefficient values from each model including the basic model, which does not transform the input vector, and the BLEU model. We use the bi-gram model for the BLEU test

Author's personal copy

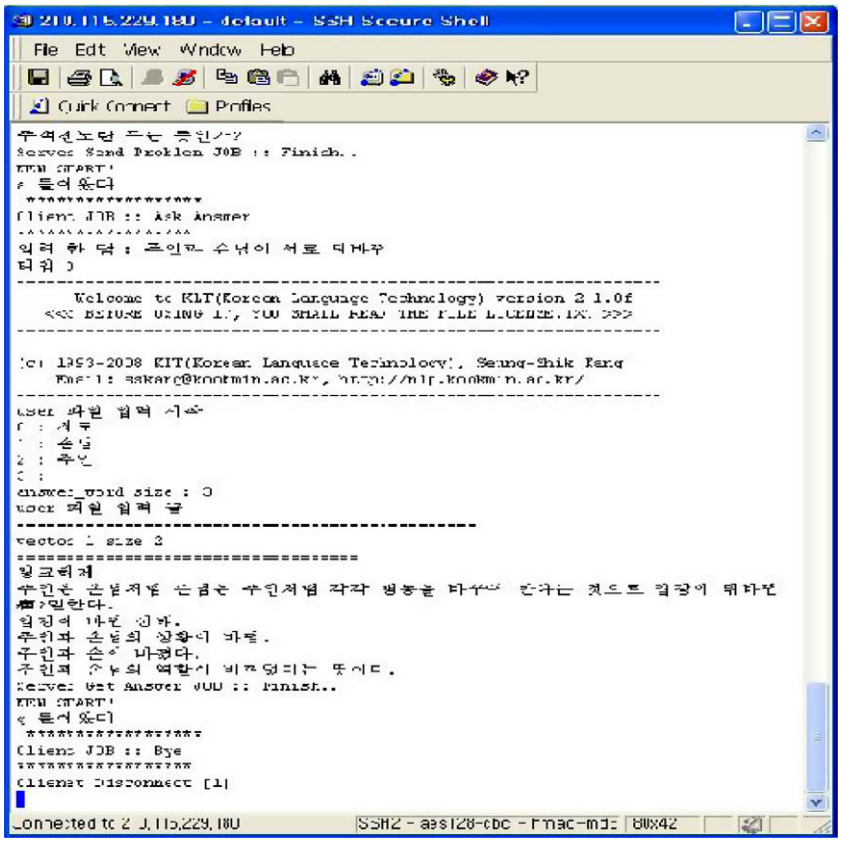


Fig. 6. A snapshot of the server processing.

Table 3
Singular values of P_C and P_W. We select fourteen kth largest singular values. The first row shows the list of selected ks. The second and the third rows denote the kth largest singular values of P_C and P_W, respectively.

k	1	25	50	100	150	200	250
P_C	98.164	31.548	26.109	22.007	19.757	18.333	17.321
P_W	99.787	17.923	13.178	10.040	8.774	7.590	7.067
k	300	350	400	450	500	550	600
P_C	16.492	15.805	15.250	14.761	14.301	13.913	13.565
P_W	6.562	6.148	5.783	5.494	5.261	5.054	4.856

because the uni-gram is considered to be a similar method to the basic model. The K_C model is identical to the latent semantic kernel (LSK) model.

In view of the correlation coefficient values, we cannot discover a distinguishable performance difference among those models. Except for the bi-gram BLEU model, every model shows its value higher than 0.9. Among all of the models, only K_1, K_2, and K_3 show a little distinguishable higher value than the basic model. The other models show a very little higher or even lower value than the basic one. From these very high correlation coefficient values, we can infer that the short answers which include a few words less than 10 can be assessed very similar to human instructors especially when being graded for each student, not for each answer. On the contrary, the very few words also cause a problem. The fewer keywords the answers involve, the less the difference among the models is.

Next, we compare the grading results of each answer, one from the human and the other from the models. A human assessor decides whether each answer is correct or not and gives it one of two scores, 1 or 0, respectively. Our models also give the similarity values as the score of each answer, ranged from 0 to 1. The threshold similarity score to decide whether the answer is correct or not has to be determined for each model. We divide all the answer papers into two parts, 25%, 750 answers, for threshold decision and the other, 2250 answers, for the later accuracy evaluation. We increase the threshold value from 0 to 1 by 0.01 and try to find the optimal threshold value making the accordance ratio of 750 papers the highest. Table 5 shows the threshold value of each model. We use the 4-fold cross validation method to simulate the real population. The scores shown in the table are the averages of four validation processes.

Compared to three existing models, our combined kernel models ordinarily show better accordance ratios except K_6. Especially, K_1, K_2, and K_3 show higher accordance ratios than the original kernels from corpus and the WordNet. It is apparently said that the combined kernels can improve the performance when assessing the answers. Also, merging by concatenation and selection in turn show higher possibility to improve the accuracy than merging by sorting by singular values. Merging by sorting causes irregularity of columns from two different matrices. The irregularity seems to be one of the reasons to decrease the accuracy.

In addition, we measure the changes of accordance ratios by decreasing the number of the model answers from 5 to 1. This experiment is performed on only the BLEU model and a combined kernel model, K_1, which shows the highest accordance ratio in the above experiment. The kernel based models selects the highest similarity value from a given number of similarity values of a give number of model answers, while the BLEU merges the given number of model answers into a single answer and calculated the BLEU score with the merged answer. Table 6 shows the changes of their accordance ratios.

Author's personal copy

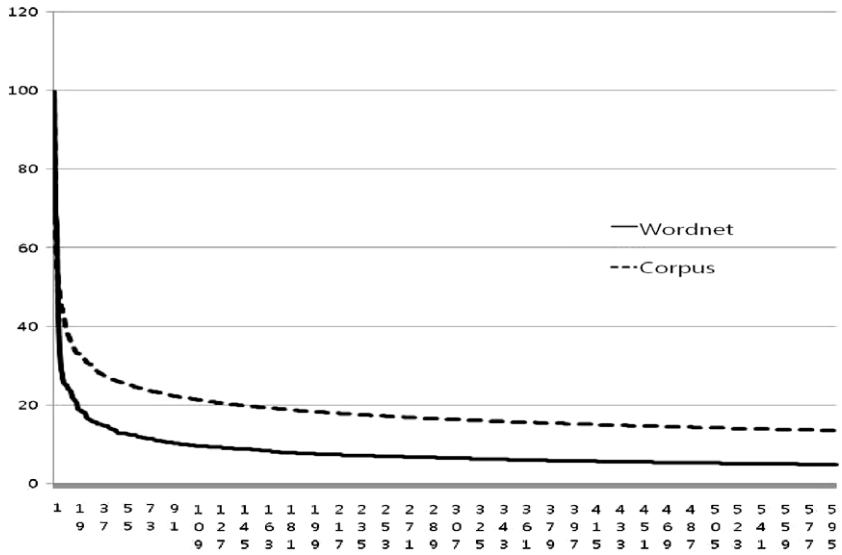


Fig. 7. Singular values of P_C and P_W. The x-axis denotes the k, from 0 to 600, and the y-axis denotes the corresponding singular values. The dashed line shows the singular values of P_C and the plain line shows the singular values of P_W.

Table 4
This table shows the correlation coefficient values between human instructors and models suggested in this paper. The first row shows the name of each model. The second row shows the coefficient values for corresponding models.

Models	Basic	P_C (=LSK)	BLEU	P_W	K_1
Values	0.919632	0.920784	0.850988	0.920922	0.934333
Models	K_2	K_3	K_4	K_5	K_6
Values	0.931380	0.921686	0.907796	0.912037	0.903492

Table 5
This table shows the accordance of each answer from all models. The first row shows the name of models used in this experiment. The second row is for threshold value calculated from the training data and the third row is for the accordance ratio of the rest data. The numbers shown in the second and the third rows mean the average numbers of four validations.

Models	Basic	P_C (=LSK)	BLEU	P_W	K_1
Threshold values	0	0.0075	0.08875	0.025	0.0075
Accordance ratio (%)	74.875	78.6	75.87	78.325	81.8
Models	K_2	K_3	K_4	K_5	K_6
Threshold values	0.0075	0.0075	0.025	0.0025	0
Accordance ratio (%)	81.8	81.6	79.15	78.35	74.375

Table 6
This table shows the changes of accordance ratio for two models, the BLEU model and the K_1 model. The changes are due to the number of model answers used for the assessment. The first row shows the number of model answers. The second row and the third row are for the accordance ratios of the BLEU model and the K_1 model, respectively.

Models	5	4	3	2	1
BLEU model (%)	75.87	71.98	70.83	64.28	54.38
K_1 model (%)	81.80	81.88	80.63	79.58	75.85

Table 6 shows that the combined model is more robust to the shortage of model answers. The BLEU model shows the decrease of the accordance ratio as much as 5.13%, 6.64%, 15.28%, and 28.32% when the number of the model answers decreased to 4, 3, 2, and 1, respectively. However, the K_1 model shows only the decrease of −0.1%, 1.43%, 2.71%, and 7.27%. From these results, the kernel based model can be considered not to require many model answers. One or two model answers in kernel based models are enough to be used in the assessment process.

7. Conclusion

We propose an autonomous assessment system based on a combined latent semantic kernel to integrate various semantic information. First, we collect a large volume of documents to make a term-document matrix. Simultaneously, a lexical database, WordNet, is also transformed into a term–synset matrix. Secondly, we reduce two matrices by using the SVD algorithm, and two latent semantic kernels are created. With two kernels, we generate six other combined kernel matrices with three principles. For the evaluation, we compare the grading results of each student and of each answer to those of human instructors. The experimental results shows improved performance by using the combined kernel, compared to the other various methods. Also, the combined kernel requirs fewer model answers, compared to the BLEU model.

For future work, we will first have to make a plan to collect various domain specific documents and their related question–answer sets. This research will show that the WordNet kernel could help the corpus kernel supplement with its lack of semantic information. Secondly, we will develop more various approaches for combining multiple kernels. Also the new approaches must be proved by mathematic verifications. Finally, a syntactic analysis must be integrated with the semantics based approach presented in this paper. The 'bag-of-words' method cannot consider the syntactic role of each word. By integrating the syntactic features into the existing semantic features, the assessment process will have more robust rationale within itself.

Acknowledgement

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

References

Aamodt, A., & Plaza, E. (1994). Case-based reasoning; foundational issues, methodological variations, and system approaches. *AI Communications, 7*(1), 39–59.

Author's personal copy

Bain, L., & Engelhardt, M. (1987). Introduction to probability and mathematical statistics. *Thomson Learning*.

Bang, H., Hur, S., Kim, W., & Lee, J. (2001). A system to supplement subjectivity test marking on the web-based. In *Proceedings of KIPS* (in Korean).

Briscoe-Smith, A., & Evangelopoulos, N. (2002). Case-based grading: A conceptual introduction. In *Proceedings of ISECON 2002*, 19.

Budanitsky, A., & Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources, Second Meeting of the North Americal Chapter of the Association for Computational Linguistics, Pittsburgh, PA*.

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

Burgun, A., & Bodenreider, O. (2001). Comparing terms, concepts and semantic classes in WordNet and the unified medical language system. In *Workshop on WordNet and other lexical resources, second meeting of the North American chapter of the association for computational linguistics, Pittsburgh, PA*.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998). Automated scoring using a hybrid feature identification technique. In *Proceedings of annual meeting association of computational linguistics* (pp. 206–210).

Cheong, D. (2001). Evaluation of short and long essay questions by using vector similarity and thesaurus. MS Dissertation of Dongguk University (in Korean).

Cristianini, N., Shawe-taylor, J., & Lodhi, H. (2002). Latent semantic kernel. *Journal of Intelligent Information Systems, 18*(2–3), 127–152.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: The MIT Press.

Geiger, M. (1996). On the benefit of changing multiple-choice answers: Student perception and performance. *Education*(117), 1–108.

Gonzalo, J., Verdejo, F., Chugur, I., & Cigarran, J. (1998). Indexing with WordNet synsets can improve text retrieval. In *Proceedings of the 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics, Montreal, Canada* (pp. 38–44).

Kang, S. S. (2004). Gerneral purposed morphological analyzer HAM Version 6.0.0. <http://nlp.kookmin.ac.kr>.

Kim, Y. S., Cho, W. J., Lee, J. Y., & Oh, Y. J. (2005). An intelligent grading system using heterogeneous linguistic resources. In *Proceedings of 6th international conference on intelligent data engineering and automated learning (IDEAL-2005), Brisbane, Australia* (pp. 102–108).

Kim, Y. B., & Kim, Y. S. (2008). Latent semantic kernels for WordNet: Transforming a tree-like structure into a matrix. In *Proceedings on the 7th international conference on advanced language processing and web information technology (ALPIT-2008), Dalian, China*.

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

Landauer, T. K., Laham, D., & Foltz, P. W. (2000). The intelligent essay assessor. *IEEE Intelligent Systems, 15*(5), 27–31.

Lee, E. R., & Lim, S. S. (2004). Korean WordNet version 2.0. Korean Language Processing Laboratory, Pusan National University.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 98*(11), 39–41.

Page, E. B., & Peterson, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan, 76*, 561–565.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics (ACL-2002), Philadelphia* (pp. 311–318).

P'erez, D., Alfonseca, E., & Aodr, P. (2004). Application of the BLEU method for evaluating Free-text answers in an e-learning environment. In *Proceedings of the language resources and evaluation conference (LREC-2004)*.

Pulman, S. G., & Sukkarieh, J. Z. (2005). Automatic short answer marking. In *Proceedings of the second workshop on building educational applications using NLP* (pp. 9–16).

Richardson, R. Smeaton, A. F., & Murphy, J. (1994). Using WordNet as a knowledge base for measuring semantic similarity between words. CA-1294, Dublin, Ireland.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communication of the ACM, 19*(11), 613–620.

Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Addison Wesley.

Steele, C. W. (1997). Essays – Well worth the effort. *College Teaching, 45*(4), 371–416.

Voorhees, E. M. (1993). Using WordNet to disambiguate word senses for text retrieval. In *Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval, Pittsburgh, PA*.

Wong, S. K. M., Ziarko, W., & Wong, P. C. N. (1985). Generalized vector space model in information retrieval. In *Proceedings of ACM SIGIR conference on research and development in information retrieval* (pp. 18–25).