An Empirical Study on Dimensionality Optimization in Text Mining for Linguistic Knowledge Acquisition

Yu-Seop Kim[1], Jeong-Ho Chang[2], and Byoung-Tak Zhang[2]

[1] Division of Information and Telecommunication Engineering, Hallym University
Kang-Won, Korea 200-702
yskim01@hallym.ac.kr

[2] School of Computer Science and Engineering, Seoul National University
Seoul, Korea 151-744
{jhchang, btzhang}@bi.snu.ac.kr***

Abstract. In this paper, we try to find empirically the optimal dimensionality in data-driven models, Latent Semantic Analysis (LSA) model and Probabilistic Latent Semantic Analysis (PLSA) model. These models are used for building linguistic semantic knowledge which could be used in estimating contextual semantic similarity for the target word selection in English-Korean machine translation. We also facilitate k-Nearest Neighbor learning algorithm. We diversify our experiments by analyzing the covariance between the value of k in k-NN learning and accuracy of selection, in addition to that between the dimensionality and the accuracy. While we could not find regular tendency of relationship between the dimensionality and the accuracy, however, we could find the optimal dimensionality having the most sound distribution of data during experiments.

Keywords: Knowledge Acquisition, Text Mining, Latent Semantic Analysis, Probabilistic Latent Semantic Analysis, Target Word Selection

1 Introduction

Data-driven models in this paper are much beneficial in natural language processing application because the cost for building new linguistic knowledge is very expensive. But only raw text data, called untagged corpora, are needed in data-driven models of this paper. LSA is construed as a practical expedient for obtaining approximate estimates of meaning similarity among words and text segments and is applied to various application. LSA also assumes that the choice of dimensionality can be of great importance[Landauer 98]. The PLSA model is based on a statistical model which has been called aspect model[Hofmann 99c]. In this paper, we ultimately have tried to find out regular tendency of covariance

*** This work was supported by the Korea Ministry of Science and Technology under the BrainTech Project

K.-Y. Whang, J. Jeon, K. Shim, J. Srivatava (Eds.): PAKDD 2003, LNAI 2637, pp. 111–116, 2003.

between dimensionality and soundness of acquired knowledge in natural language processing application, especially in English-Korean machine translation. We utilized k-Nearest neighbor learning algorithm to resolve the meaning of unseen word or data sparseness problem. Semantic similarity among words should be estimated and the semantic knowledge for similarity could be built from the data-driven models to be shown in this paper[Kim 02b]. We also extended our experiment by computing covariance values between k in k-NN learning and selectional accuracy.

2 Knowledge Acquisition Models

Next two subsections will explain what LSA and PLSA are and how these two models can build the linguistic knowledge for measuring semantic similarity among words.

2.1 Latent Semantic Analysis Model

LSA can extract and infer relations of expected contextual usage of words in passages of discourse[Landauer 98]. LSA applies singular value decomposition (SVD) to the matrix explained in [Landauer 98]. SVD is defined as

$$A = U\Sigma V^T \tag{1}$$

,where Σ is a diagonal matrix composed of nonzero eigen values of AA^T or A^TA, and U and V are the orthogonal eigenvectors associated with the r nonzero eigenvalues of AA^T and A^TA, respectively. The details of above formula are described in [Landauer 98]. There is a mathematical proof that any matrix can be so decomposed perfectly, using no more factors than the smallest dimension of original matrix. The singular vectors corresponding to the $k(k \leq r)$ largest singular values are then used to define k-dimensional document space. Using these vectors, $m \times k$ and $n \times k$ matrices where m and n is the number of rows, U_k and V_k may be redefined along with $k \times k$ singular value matrix $\sum_k$. It is known that $A_k = U_k\Sigma_k V_k^T$ is the closest matrix of rank k to the original matrix. The term-to-term similarity is based on the inner products between two row vectors of A, $AA^T = U\Sigma^2U^T$. One might think of the rows of $U\Sigma$ as defining coordinates for terms in the latent space. To calculate the similarity of two words, $\mathbf{V_1}$ and $\mathbf{V_2}$, represented in the reduced space, cosine computation is used:

$$\cos\phi = \frac{\mathbf{V_1} \cdot \mathbf{V_2}}{\parallel \mathbf{V_1} \parallel \cdot \parallel \mathbf{V_2} \parallel} \tag{2}$$

Knowledge acquired from above procedure has a characteristics like that of people before and after reading a particular text and that conveyed by that text.

2.2 Probabilistic Latent Semantic Analysis

PLSA is based on *aspect model* where each observation of the co-occurrence data is associated with a latent class variable $z \in Z = \{z_1, z_2, \ldots, z_K\}$ [Hofmann 99c]. A word-document co-occurrence event, (d, w), d for documents and w for words, is modelled in a probabilistic way where it is parameterized as in

$$\begin{aligned} P(d,w) &= \sum_z P(z)P(d,w|z) \\ &= \sum_z P(z)P(w|z)P(d|z), \end{aligned} \tag{3}$$

$P(w|z)$ and $P(d|z)$ are topic-specific word distribution and document distribution, respectively. The objective function of PLSA, unlike that of LSA, is the likelihood function of multinomial sampling. The parameters $P(z), P(w|z)$, and $P(d|z)$ are estimated by maximizing the log-likelihood function

$$L = \sum_{d \in D} \sum_{w \in W} n(d,w) \log P(d,w), \tag{4}$$

and this maximization is performed using the EM algorithm. Details on the parameter estimation are referred to [Hofmann 99c]. To estimate the similarity of two words, w_1 and w_2, we compute the similarity of w_1 and w_2 in the constructed latent space. This is done by

$$sim = \frac{\sum_k P(z_k|w_1)P(z_k|w_2)}{\sum_k P(z_k|w_1)P(z_k|w_1) \sum_k P(z_k|w_2)P(z_k|w_2)} \tag{5}$$

$$P(z_k|w) = \frac{P(z_k)P(w|z_k)}{\sum_l P(z_l)P(w|z_l)} \tag{6}$$

3 Target Word Selection Process

We used grammatical relations stored in the form of a dictionary for target word selection. The structure of the dictionary is as follows [Kim 01]:

$$T(S_i) = \begin{cases} T_1 \text{ if } Cooc(S_i,\ S_1) \\ T_2 \text{ if } Cooc(S_i,\ S_2) \\ \quad \ldots \\ T_n \text{ otherwise,} \end{cases} \tag{7}$$

where $Cooc(S_i, S_j)$ denotes grammatical co-occurrence of source words S_i and S_j, which one means an input word to be translated and the other means an argument word to be used in translation, and T_j is the translation result of the source word. $T(\cdot)$ denotes the translation process. One of the fundamental difficulties is the problem of data sparseness or unseen words. We used k-nearest neighbor method that resolves this problem. The linguistic knowledge acquired from latent semantic spaces is required when performing k-NN search to select the target word. The nearest instance of a given word is decided by selecting a word having the highest semantic similarity, which is extracted from the knowledge. The k-nearest neighbor algorithm for approximating a discrete-valued target function is given in [Cover 67].

4 Experimental Result

For constructing latent space, we indexed 79,919 documents in 1988 AP news text from TREC-7 data to 21,945,292 words. From this, 19,286 unique words with above 20 occurrence was collected and the resulting corpus size is 17,071,211. Kim *et. al*[Kim 02b] built a 200 dimension space in SVD of LSA and 128 latent dimension of PLSA. We, however, made the dimensionality of LSA and PLSA ranging from 50 to 300 individually because we tried to discover the relationship between the dimensionality and the adequateness of semantic knowledge acquired. We utilized a single vector Lanczos method of SVDPACK[Berry 93] when constructing LSA space. The similarity of any two words could be estimated by performing cosine computation between two vectors representing coordinates of the words in the spaces. We extracted 3,443 example sentences containing predefined grammatical relations, like *verb-object*, *subject-verb* and *adjective-noun*, from Wall Street Journal corpus and other newspapers text of totally 261,797 sentences. 2,437,188, and 818 examples were utilized for *verb-object*, *subject-verb*, and *adjective-noun*, respectively.

In the first place, we selected an appropriate target word using a default meaning, which is selected when there is no target word found in a built-in dictionary. About 76.81% of accuracy was acquired from this method. Secondly, we also evaluated the selection from non-reductive dimensionality which is same as the number of documents used in knowledge acquisition phase, 79,919. Up to 87.09% of selection was successful. Finally, the reduction was performed on the original vector using LSA and PLSA. Up to 87.18% of successful selection was acquired by using LSA and PLSA. However, time consumed in selection by non-reductive dimensionality was much longer about from 6 to 9 times than those of LSA and PLSA reduction methods.

And we have also tried to find out regular tendency of covariance between dimensionality and distributional soundness of acquired knowledge. For this, we have computed the co-relationship between the dimensionality and the selection accuracy and between k in the k-NN learning method and the accuracy. Figure 1 shows the relationship between the dimensionality and selectional accuracy on each k value. In figure 2, the relationship between k in k-NN learning and accuracy of target word selection over the various dimensionality was shown.

From this experiment, we could calculate the covariance values among several factors [Bain 87]. Table 1 shows the covariance between dimensionality and accuracy and between k and accuracy. As shown in the table, the covariance value between k and accuracy is much higher than that between the dimension-

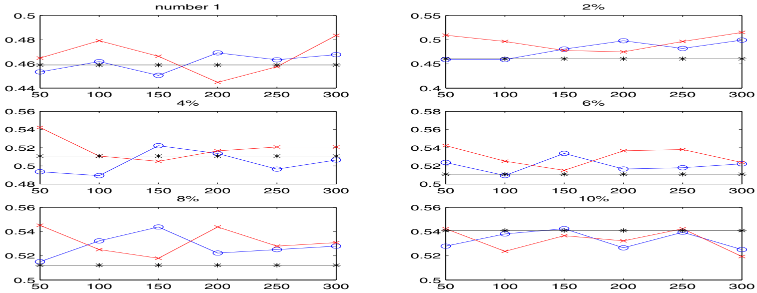

Fig. 1. This figure shows the tendency of selectional accuracy with various dimensionality over each k value of k-NN learning. 6 figures are shown for each k value. The line with 'X' (-X-) represents the result from SVD latent space, the line with 'O' (-O-) represents the result from PLSA space, and the other line (-*-) represents the result from non-reductive space. X-axis represents the dimensionality from 50 to 300 and y-axis represents selectional accuracy.

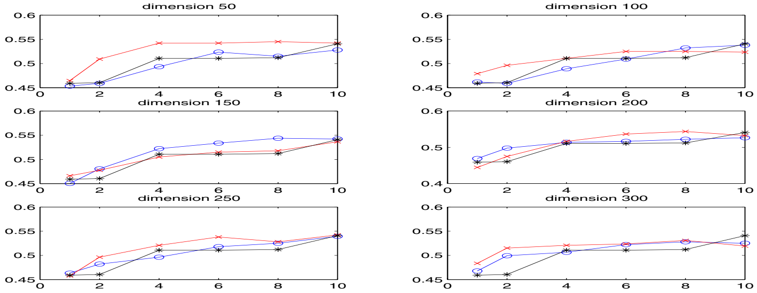

Fig. 2. This figure shows the tendency of selectional accuracy with various k over each dimensionality. Six figures are shown for each dimensionality value. The line with 'X' (-X-) represents the result from SVD latent space, the line with 'O' (-O-) represents the result from PLSA space, and the other line (-*-) represents the result from non-reductive space. X-dimension represents the k from 1 and from 2% to 10% of samples and y-dimension represents selectional accuracy.

ality and accuracy. It can be said that the value of k can affect the selectional accuracy much more than the dimensionality of a reduced vector. In PLSA, the soundness of space distribution is in the highest position when the dimensionality is 150, which could be inferred from the fact that the covariance between k and the accuracy is located at the top in the case of 150 dimensionality. It is known that the larger the k value is, the more robust to noisy data the sample data space is. From this, we can have an analogy that higher covariance value with k could make more sound distribution of latent space. In contrast, LSA has the most sound space in its 200 dimensionality. On average, PLSA has a little higher covariance then LSA. As a matter of fact, PLSA with 150 dimensionality selected accurate target words the most and LSA with 200 dimensionality se-

Table 1. The 3 upper rows of the table shows covariance values among k of k-NN learning and selectional accuracy in accordance to the dimensionality. And 3 lower rows show the covariance between dimensionality and accuracy for each k.

dim	50	100	150	200	250	300	avg
PLSA	52.89	58.35	59.83	32.45	49.02	34.84	47.90
SVD	43.54	29.20	44.52	60.09	47.90	19.75	40.83
k	1	2	4	6	8	10	
PLSA	7.89	23.91	6.46	0.12	1.79	-2.15	6.34
SVD	0.60	2.15	-5.50	-2.75	-3.11	-5.26	-2.31

lected the most accurately. Consequently, the distributional soundness and the selection accuracy could be said to have a strongly shared characteristics to each other.

5 Conclusion

LSA and PLSA were used for constructing the semantic knowledge in this paper. The dimensionality does not have a specific linkage to the semantic knowledge construction. However, the value of k could be the essential element. And, PLSA could build the semantic knowledge robust to the noisy data than LSA.

References

[Bain 87] Bain, L., M. Engelhardt, “Introduction to Probability and Mathematical Statistics,” *PWS publishers*, pp. 179,190, 1987.

[Berry 93] Berry., M., T. Do, G. O’Brien, V. Krishna, and S. Varadhan, “SVDPACKC: Version 1.0 User’s Guide,” *University of Tennessee Technical Report*, **CS–93–194**, 1993.

[Cover 67] Cover, T., and P. Hart, “Nearest Neighbor Pattern Classification,” *IEEE Transactions on Information Theory*, **13**, pp. 21–27, 1967.

[Hofmann 99c] Hofmann, T., “Probabilistic latent semantic indexing,” *Proceedings of the 22th Annual International ACM SIGIR conference on Research and Developement in Information Retrieval (SIGIR99)*, pp. 50–57, 1999.

[Kim 01] Kim, Y., B. Zhang and Y. Kim, “Collocation Dictionary Optimization using WordNet and k-Nearest Neighbor Learning,” *Machine Translation* **16**(2), pp. 89–108, 2001.

[Kim 02b] Kim, Y., J. Chang, and B. Zhang, “A comparative evaluation of data-driven models in translation selection of machine translation,” *Proceedings of the 19th Internation Conference on Computational Linguistics (COLING-2002)*, Taipei, Taiwan, pp. 453–459, 2002.

[Landauer 98] Landauer, T. K., P. W. Foltz, and D. Laham, “An Introduction to Latent Semantic Analysis,” *Discourse Processes*, **25**, pp. 259–284, 1998.