AUTOMATIC ACQUISITION OF LEXICAL KNOWLEDGE USING LATENT SEMANTIC MODELS

Masato Hagiwara, Yasuhiro Ogawa, Katsuhiko Toyama

Graduate School of Information Science, Nagoya University

ABSTRACT

For statistical acquisition of lexical knowledge from corpora, it is important to deal not only with such superficial information as the context of the words but also their latent semantics. The paper describes how to utilize latent semantic models to acquire synonyms from corpora, based on word-context co-occurrences. The experiments show PLSI achieves a better performance than conventional methods such as LSI, making it effective for automatic synonym acquisition.

1. INTRODUCTION

Lexical knowledge, especially synonyms, is one of the most important knowledge for natural language processing, having a broad range of applications such as query expansion in information retrieval and automatic thesaurus construction.

There have been many studies for automatic synonym acquisition from large corpora. They are usually based on the distributional hypothesis, i.e. semantically similar words share similar contexts, and calculate the word similarities using co-occurrences of words and their contexts such as dependencies. These methods, however, suffer from noises and sparseness, because corpora contain some amount of meaningless information, and co-occurrence data are often sparse and inappropriate for calculation. To tackle these issues, not only surface information but also latent semantics should be considered.

Several latent semantic models have been proposed, including Latent Semantic Indexing (LSI) [1]. LSI alleviates the noise and sparseness problems by dimensionality reduction and it has been widely used for document indexing. However, it lacks firm theoretical basis and it often requires some ad-hoc weighting such as idf. On the contrary, PLSI [2] is a probabilistic version of LSI, where it is formalized that documents and terms co-occur through a latent variable, which represents their latent meanings. PLSI is experimentally shown to outperform the former model.

This study shows that PLSI is also effective for the automatic synonym acquisition by estimating each word's latent meanings. Firstly, pairs of nouns and their contexts are collected from large corpora, assuming the distributional hypothesis. Secondly, these co-occurrences are fit into the PLSI model, and the probability distribution of the latent classes is calculated for each noun. Finally, similarity for noun pairs is calculated by measuring the distance between two probability distributions using an appropriate measure.

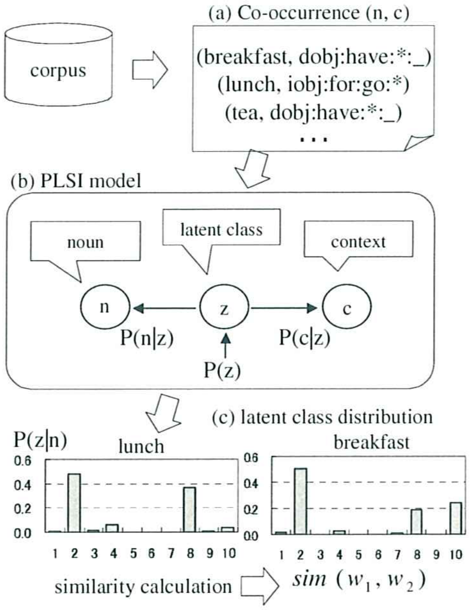

Fig. 1. Approach

2. APPROACH

We used dependency relationships as context of nouns. Using RASP Toolkit [3], grammatical relationships are extracted from corpora, which are then converted into co-occurrences of nouns and their contexts, as shown in Fig. 1 (a).

These co-occurrences are then fit into PLSI model:

$$P(n, c) = \sum_{z} P(z)P(n|z)P(c|z), \qquad (1)$$

whose graphical representation is shown in Fig. 1 (b). The original PLSI model deals with documents and terms, but it is also applicable to other kinds of co-occurrences. This model assumes that noun n and context c co-occur through latent class z, and the parameters $P(z), P(n|z), P(c|z)$ are deter-

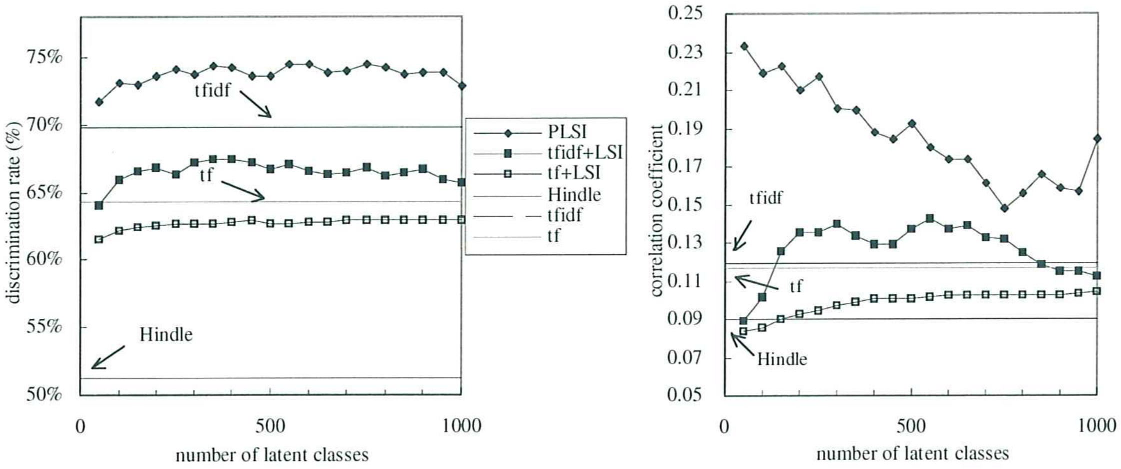

Fig. 2. Comparison of PLSI and otehr conventional methods

mined via EM algorithm by iteratively maximizing the likelihood of the co-occurrences.

We then obtain the latent class distribution $P(z|n)$ for each noun. The similarity between nouns n_1 and n_2 can be calculated by measuring the distance using KL divergence between the two corresponding distribution, $P(z|n_1)$ and $P(z|n_2)$.

3. EXPERIMENT

We conducted an experiment to compare the synonym acquisition performances of PLSI and these conventional acquisition methods: tf, tf.idf, tf+LSI, tf.idf+LSI, and Hindle's method [6]. For the first four vector-based models, the similarity is calculated using cosine of two vectors. For PLSI, tempered EM algorithm was employed to avoid over-fitting, where the inverse temperature parameter β was set to 0.86. Because the number of latent classes K must be given beforehand for PLSI and LSI, the performances for these models are measured varying K from 50 to 1,000 with a step of 50.

The performance was automatically evaluated using the existing thesaurus WordNet, as employed by Hagiwara et al. [5]. This evaluation scheme calculates two evaluation measures: discrimination rate (DR) and correlation coefficient (CC). The higher these values are, the better the results reflect WordNet, which means higher performance.

The result for the corpus WordBank (approx. 190,000 sentences, 3.5 million words) [4] is shown in Fig. 2. The performance of PLSI stays on top for all the values of K, strongly suggesting the superiority of PLSI over the conventional methods, especially when K is small, which is practically preferable. On the other hand, the performances of tf and tf+LSI, which are not weighted by idf, and Hindle's method are consistently low regardless of the value of K.

4. CONCLUSION

In this study, automatic synonym acquisition was performed using a latent semantic model PLSI by estimating the latent class distribution for each noun. For this purpose, co-occurrences of words and their contexts extracted from large corpora were utilized, and it was found that PLSI outperformed such conventional methods as tf·idf and LSI. These results make PLSI applicable enough for automatic thesaurus construction.

Note, however, that the performance of PLSI may greatly depend on such factors as the latent class number K. Although the fine tuning of the parameters enables PLSI to achieve the best performance, it might be a practical choice to adopt the stably performing tf·idf under certain circumstances where the available resource is limited.

Although synonymous nouns were extracted this time, the same framework can be applied to the other categories of words such as verbs and adjectives, where the co-occurrences of these categories of words and contexts are considered.

5. REFERENCES

[1] Scott Deerwester, et al., "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[2] Thomas Hofmann, "Probabilistic latent semantic indexing," in *Proc. of the 22nd SIGIR*, 1999, pp. 50–57.

[3] Ted Briscoe and John Carroll, "Robust accurate statistical annotation of general text," in *Proc. of th 3rd LREC*, 2002, pp. 1499–1504.

[4] Collins, *Collins Cobuild Major New Edition CD-ROM*, HarperCollins Publishers, 2002.

[5] Masato Hagiwara, et. al. "Selection of effective contextual information for automatic synonym acquisition," in *Proc. of the 21st COLING/ 44th ACL*, 2006, pp. 353–360.

[6] Donald Hindle, "Noun classification from predicate-argument structures," in *Proc. of the 28th ACL*, 1990, pp. 268–275.