Refining the Concept Network for Automatically Chinese Thesaurus Construction

Weidong Qu and Katsuhiko Shirai

Department of Information and Computer Science, Waseda University

Abstract

Most attempts to automatically construct thesaurus have relied on the Distributional Hypothesis that words that appear in similar contexts are semantically similar. Due to the random co-occurrence (words occur with each other in same context by chance) and polysemy (a word may has several meanings), the terms co-occurrence space is more noisy and complicated. Moreover, the clustering method or the nearest neighbor search in such high dimensional space is often unstable. In this work, we use a refined concept network for Chinese thesaurus construction. A context window is used to extract the corresponding features of a word and to build a concept network. To extract the semantic relationships of a word, we first search the concept network and obtain the corresponding feature subset. Then a refined subset was obtained by using dimension reduction. Finally, the similarity of words that share some or most features is calculated using Cosine Function or Jaccard Measure. The output of these comparisons is ranked words used in the most similar way to that word. We run our initial experiments in a Chinese corpus and give same experiment results.

1. Introduction

A thesaurus is a useful lexical resource for natural language processing. In information retrieval, people use thesauri to solve the word mismatch problems. Thesauri have been also used to solve various questions in many tasks such as language model smoothing, word sense disambiguation and machine translation. The widespread use of Wordnet and EDR thesauri showed the need for these kinds of thesaurus resources. However, most available thesauri are general thesauri and many tasks need some domain specific thesauri. Manually constructing a domain specific resource is mostly labor-intensive and time consuming. Moreover, it is impossible to manage and modify these resources in time with the language development and evolution. Automatic thesaurus construction is a feasible approach to solve this problem.

Most attempts to automatically construct thesaurus have relied on the Distributional Hypothesis that words that appear in similar contexts are semantically similar [1]. By extracting the context co-occurrence information over a large corpus, the semantic distance between words can be measured. Typically, these methods have been to construct a lexical co-occurrence matrix based on frequency counts of the words in a context window of a predetermined size. The co-occurrence matrix also can be constructed based on dependency relationships of a sentence such as the relation between heads and modifiers that generated by parsing a large corpus. In these methods, a word can be treated as a point in a high dimensional space and the semantic distance between two words can be calculated by Cosine or Jaccard measure. Due to the random co-occurrence (words occur with each other in same context by chance) and polysemy (a word may has several senses), the terms co-occurrence space is more noisy and complicated. Moreover, the clustering method or the nearest neighbor search in such high dimensional space is often unstable.

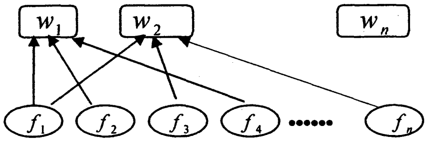
In this paper, we present a refined concept network (here we use concept network as a substitute for co-occurrence matrix) for thesaurus construction. Same as the standard method, a context window is used to extract the corresponding features of a word and to build a concept network [2]. Given a word to retrieval the most similar words, we first search the concept network and obtain the corresponding feature subset. Then a refined subset was obtained by using dimension reduction. Finally, the similarity of words that share some or most features is calculated using Cosine Function or Jaccard Measure. The output of these comparisons is ranked words used in the most similar way to that word. We run our experiments in a Chinese corpus and give the experiment results.

2. Method

We use the simplest approach taken in majority of the studies [2], a context window, to extract the co-occurrence words and build the concept network. The difference between our work and others is that we use a dimension reduction technique to refine the concept network before performing the similarity comparison.

2.1 Concept network

Given a word, the context window is used to extract the corresponding features. The nouns or adjective (features) that appear within ±*k* words of the target word ***wi*** are linked to the target word ***wi*** as features ***fi***. The co-occurrence frequency of target word ***wi*** with feature ***fi*** is defined as the number of times that ***wi*** and ***fi*** occurs in the window of *k* words surrounding ***wi***, summed when the window shift over all corpus. After obtaining the frequency of co-occurrence, the conditional probability was used to weight the strength of target word ***wi*** and feature ***fi***. The following diagram shows the structure of our concept network.



Concept network

To retrieval a given word's similar words, we first search the concept network and obtain the word's corresponding feature set. Then the words share some or most features are selected. Usually, similarity between two words is calculated using cosine function or jaccard measure. The output of these comparisons is ranked words used in the similar way to that word.

2.2 Refining the Concept Network

Since features may occur with a target word in the context window by chance, the word co-occurrence space is noisy. A word may have several meanings; this makes the word's feature space more complicated. Furthermore, due to the sparsity of data objects in the high dimensional space and all pairs of points are almost equidistant from one another for a wide range of data distributions and distance functions, the nearest neighbor search is often unstable. To solve this problem, we reduce the matrix's dimension by linear algebraic technique LAS/SVD (Latent Semantic Analysis/Singular Value Decomposition) [3]. The SVD is known for its capabilities of deriving the low dimensional refined feature space from a high dimensional raw feature space [3]. Given a matrix A and rank (A)=*r*, the SVD of A and the rank-k approximation matrix A_k are defined as follows:

$$A = U\Sigma V^T = \sum_{i=1}^{n} u_i \bullet \sigma_i \bullet v_i^T \qquad (1)$$

$$A_k = U_k \Sigma_k V_k^T = \sum_{i=1}^{k} u_i \bullet \sigma_i \bullet v_i^T \qquad (2)$$

Where the matrix A is decomposed into left and right singular vectors U, V and a diagonal matrix Σ. Truncated SVD matrix A_k can be constructed from the k-largest singular triples of A. A_k with rank-k can be seen the best approximation to A for any unitray invariant norm [4].

3. Experiments

We use a Chinese corpus to run our initial experiment. The corpus consists of six months newspaper of Chinese People's Daily in 1998. The corpus is segmented and tagged with part of speech information. On this corpus, we select ten nouns as target words to retrieval their similar words. In most cases, the refined concept network works better than the raw matrix of the top ten ranked words. The following table gives same results of target word "银行" (bank):

Top rank	Refined matrix	Raw matrix
1	银行	银行
2	监管	企业
3	比例	国家
4	存款	金融
5	金融	经济
6	机构	市场
7	风险	问题
8	资产	改革
9	业务	国有
10	贷款	资金

Table 1: The top 10 most similar words of "银行".

4. Conclusion and Future Work

In this paper, we present a dimension reduction method for Chinese thesaurus construction. This method solves the problems of popular used context window method. In future work, we plan to use other subset selection and noise removing techniques such as clustering or projection method to improve the automatic constructed thesaurus's quality. Since evaluating the performance of thesaurus construction is often subjective, we plan to evaluate the quality of constructed thesaurus in an information retrieval application.

5. References

[1] Dekang Lin and Patrick Pantel. Indeuction of semantic classes from natural language text. Proc. KDD01, 2001

[2] Scott McDonald and Michael Rtamscar. Testing the Distributional Hypothesis: The Influence of Contecxt on Judgements of Semantic Similarity. 2001. 23rd Annual Conference of the Cognitive Science Society.

[3] Scott Deerwester, Susan Dumais, Goerge Furnas, Thomas Landauer. 1990. Indexing by latent semantic analysis. Journal of the american scoiety for information science.

[4] Zlatko Drmac Michael W. Berry and Elizabeth R. Jessup. 1999. Matrices, vector spaces, and infromation retrieval. SIAM Review, 41(2):335-362.