Neurocomputing 154 (2015) 127–138

Contents lists available at ScienceDirect

Neurocomputing

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

A link-based approach to semantic relation analysis

Xin Cheng [a,*], Duoqian Miao [a], Can Wang [b]

[a] Department of Computer Science, Tongji University, Shanghai, China
[b] Commonwealth Scientific and Industrial Research Organisation, Australia

ARTICLE INFO

Article history:
Received 27 May 2014
Received in revised form
29 September 2014
Accepted 1 December 2014
Communicated by Y. Chang
Available online 15 December 2014

Keywords:
Semantic relation analysis
Co-occurrence statistics
Neighbor information
Link-based relation

ABSTRACT

The semantic relation analysis is an interesting issue in natural language processing. To capture the semantic relation between terms (words or phrases), various approaches have been proposed by using the co-occurrence statistics within corpus. However, it is still a challenging task to build a robust relation measure due to the complexity of the natural language. In this paper, we present a novel approach for the semantic relation analysis, which takes account of both the pairwise relation and the link-based relation within terms. The pairwise relation captures the relation between terms from the local view, which conveys the co-occurrence pattern between terms to measure their relation. The link-based relation involves the global information into the relation measure, which derives the relation between terms from the similarity of their context information. The combination of these two relations creates a model for robust and accurate semantic relation analysis. Experimental evaluation indicates that our proposed approach leads to much improved result in document clustering over the existed methods.

1. Introduction

The study of semantic relation between terms is an important issue in natural language processing. It is a big challenge to capture the complete and precise semantic relation between terms due to the complexity of natural language. A number of semantic relation measures have been proposed in the previous literatures, and different approaches have been proven useful in the specific areas of document processing, such as document retrieval [1,2], document summarization [3,4] and word sense disambiguation [5–7].

It is worth noting that the precise analysis of the semantic relation between terms is an integral part of natural language processing. As with the accurate semantic relation between terms, we can capture the real semantic meaning of each document, and the original documents can be mapped into a novel feature space by integrating the semantic meaning into document representation. In the novel feature space, it is much easier to distinguish whether the documents are semantically similar or not. According to the enhanced document representation, we can further improve the performance of natural language processing, such as document retrieval, clustering and classification.

To capture the semantic relation between terms, a number of approaches have been proposed by using the co-occurrence information within corpus, and the resultant models have become known as word co-occurrence models. The basic idea of these models is simple that words with similar meanings will tend to occur in similar contexts, and hence word co-occurrence statistics can provide a natural basis for semantic representation. Generally, these semantic relation analysis models can be categorized into two groups: pairwise relation analysis model and similarity-based relation analysis model. The pairwise relation analysis means if two terms co-occur frequently, they are considered to be relational. Explicit simulations show that the pairwise relation analysis can be used to perform remarkably well on various performance criteria, but it only considers the relation between terms by themselves but overlooks their interaction with other terms (the context information) and fails to discover the underlying relation between terms. The similarity-based relation analysis is that if two terms have the similar distribution of co-occurrence with other terms, they are considered to be similar. Therefore, the similarity-based relation analysis takes the interaction with other terms into consideration for the relation measure, but it is insensitive to the strong relation which can be captured from the co-occurrence information in the pairwise relation analysis.

The motivation behind the work in this paper is that we believe that the semantic relation analysis should be based on not only the pairwise relation analysis, but also the similarity-based relation analysis. In this paper, we propose an approach for the semantic relation analysis from two points of view. The first is the pairwise relation analysis based on the co-occurrence information in the document collection. The second is the link-based relation analysis by considering their interaction with other terms. Then the pair and link-based relation analysis are combined to achieve a more

* Corresponding author. Tel.: +86 186 0027 1086; fax: +86 21 6958 9979.
E-mail address: cx1227@gmail.com (X. Cheng).

http://dx.doi.org/10.1016/j.neucom.2014.12.011
0925-2312/

accurate semantic relation analysis model. Besides, we propose an effective measure for the *pairwise relation* between terms, and a *link-based relation* measure to calculate relation based on the similarity of their context information. With the pairwise relation, which considers the local information, and the link-based relation, which considers the global information, the combination of them is further proposed to improve the accuracy of semantic relation analysis.

In our experiments, which are used for evaluating the performance of various semantic relation analysis strategies, our strategy achieves significant improvement on the experimental document collections with respect to clustering task. For Purity, which is used to measure the overall precision of the clusters, the average score is 0.791. For Rand Index, which is used to measure the quality by the percentage of the true positive and true negative decisions in all decisions during clustering, the average score is 0.794. For *F*-measure, which considers both the precision and recall for evaluation, the average score is 0.470. For Normalized Mutual Information, which is computed by dividing the Mutual Information between the clusters and the label of the dataset with the average of the clusters and the pre-exist classes entropy, the average score is 0.512. The document representation with the semantic relation captured by our strategy performs significantly better than the traditional bag-of-word approach on document clustering (the average scores of above four measures with the bag-of-word approach are 0.710, 0.748, 0.399 and 0.441, respectively).

The main contributions associated with this work are as follows:

- A novel pairwise relation analysis approach, assigns levels of significance to the semantic relation between terms according to their co-occurrence information.
- A novel link-based relation analysis approach that captures semantic relation between terms from the similarity of their context information, which is based on the theory that the context information can be seen as a semantic description of each term, and the similar context information indicates terms are semantically related.
- An integration of the pairwise and link-based relation to capture the precise and reliable semantic relation between terms from corpus.
- Detailed analysis of the impact of pairwise and link-based relation with respect to the clustering task on the real document collections.

The rest of this paper is organized as follows. Section 2 provides an overview of the related work. Section 3 presents our proposed relation measure, which starts with the detailed description of our pairwise and link-based relation measures, and then introduces the optimal combination of these two measures to capture the semantic relation between terms. Section 4 illustrates the experimental results to show the improvement of our strategy with the comparison of the traditional bag-of-word approach, GVSM, HAL and LSA strategies. Then further analysis and discussion are provided based on the experimental results. Finally, the conclusion and future work is described in Section 5.

2. Related work

The solution to the problem that captures the semantic relation between terms using their co-occurrence information has attracted much research interest. The two most prevalent works are Latent Semantic Analysis (LSA) [8,9] and the Hyperspace Analogue to Language (HAL) [10]. Recently, some new approaches have also been proposed, including the context vector model for information retrieval (CVM-VSM) [2], compound features for document classification [11] and the term-term similarity model based on covariance matrix [12], that seek to improve on the previous works.

LSA learns the relation between terms by examining the patterns of term co-occurrence across the document collection. It firstly constructs the document-term matrix, and then compresses the matrix using singular value decomposition (SVD), which extracts the most important underlying dimensions from document-term matrix. Some variants of LSI have also been proposed. The Probabilistic Latent Semantic Analysis (PLSA) proposed by Hofmann [13] has a similar approach with LSA. In PLSA, the documents are mapped to a reduced vector space too, the latent semantic space. As opposed to LSA, the model has a solid statistical foundation. It is based on a latent variable model for general co-occurrence data, which associates an unobserved latent class variable with each observation. The number of latent factors will be much smaller than the number of words and the factors act as prediction variables for words. The factors are obtained using a generalization of the Expectation Maximization algorithm. Given the fact that the LSA focuses on the co-occurrence information between terms within a general context (e.g., a document), LSA is particularly well suited to capture the general association and relation which exist between terms. However, LSA overlooks the importance of the context information in deriving the term relation, and thus fails to capture the potential relation that exists between terms.

The HAL model is the other prevalent work, which has been successfully used in various areas, including word sense disambiguation [14], verb morphology modeling [15], abstract words and concrete words representing [16], emotional connotations [17] and word meaning learning [18]. Representation vectors in the HAL model are built from information about the proximal co-occurrence of terms within a large body of document. Using the Euclidean distance between co-occurrence vectors building with weighted 10 term windows, they are able to predict the degree of priming of one term by another in a lexical decision task. Given that these word vectors represent the positional context within which each word occurs, HAL is well suited to capturing the positional relation between words. However, in capturing positional information, the HAL model is largely insensitive to the types of the information to which LSA is sensitive.

The other approaches, which measure the semantic relation between terms by using the similarity of their context information, are introduced in recent literatures [19–21]. For instance, the semantic relation between terms can be measured by the ratio of the number of common words divided by the larger of the absolute number of words in their context information [19]. The performance of three different functions: the Jensen–Shannon divergence (total divergence to the average), the L1 norm, and the confusion probability, to measure the context similarity, are compared in [20]. Recently, deep learning methods have been successfully applied in language processing [21–23], in which words are represented as dense real-valued vectors. Such representation is referred as distributed word representation, which is designed to capture the semantic relation between words.

The generalized vector space model (GVSM), which was proposed by Wong et al. [24], captured the relation between terms by their co-occurrence information across the entire document set. It simply utilizes the document-term matrix W^T as S, and then each document vector is projected as $\vec{d}' = \vec{d}\,W^T$. The corresponding kernel between two document vectors is expressed as

$$k'(d_i, d_j) = \vec{d_i}\,W^T W\,\vec{d_j}^T \tag{1}$$

The entry in matrix $W^T W$ reflects the similarity between terms which is measured by their frequency of co-occurrence across the