Applying Spectral Clustering for Chinese Word Sense Induction

Zhengyan He, Yang Song, Houfeng Wang
Key Laboratory of Computational Linguistics (Peking University)
Ministry of Education,China
{hezhengyan, ysong, wanghf}@pku.edu.cn

Abstract

Sense Induction is the process of identifying the word sense given its context, often treated as a clustering task. This paper explores the use of spectral cluster method which incorporates word features and n-gram features to determine which cluster the word belongs to, each cluster represents one sense in the given document set.

1 Introduction

Word Sense Induction(WSI) is defined as the process of identifying different senses of a target word in a given context in an unsupervised method. It's different from word sense disambiguation(WSD) in that senses in WSD are assumed to be known. The disadvantage of WSD is that it derives the senses of word from existing dictionaries or other corpus and the senses cannot be extended to other domains. WSI can overcome this problem as it can automatically derive word senses from the given document set, or a specific domain.

Many different approaches based on co-occurence have been proposed so far. Bordag (2006) proposes an approach that uses triplets of co-occurences. The most significant co-occurences of target word are used to build triplets that consist of the target word and its two co-occurences. Then intersection built from the co-occurence list of each word in the triplet is used as feature vector. After merging similar triplets that have more than 80% overlapping words, clustering is performed on the triplets. Triplets with fewer than 4 intersection words are removed in order to reduce noise.

LDA model has also been applied to WSI (Brody and Lapata, 2009). Brody proposes a method that treats document and topics in LDA as word context and senses respectively. The process of generating the context words is as follows: first generate sense from a multinomial distribution given context, then generate context words given sense. They also derive a layered model to incorporate different kind of features and use Gibbs sampling method to solve the problem.

Graph-based methods become popular recently. These methods use the co-occurence graph of context words to obtain sense clusters based on sub-graph density. Markov clustering(MCL) has been used to identify dense regions of graph (Agirre and Soroa, 2007).

Spectral clustering performs well on problems in which points cluster based on shape. The method is that first compute the Laplace matrix of the affinity matrix, then reform the data points by stacking the largest eigenvectors of the Laplace matrix in columns, finally cluster the new data points using a more simple clustering method like k-means (Ng et al., 2001).

2 Methodology

Our approach follows a common cluster model that represents the given context as a word vector and later uses a spectral clustering method to group each instance in its own cluster.

Different types of polysemy may arise and the most significant distinction may be the syntactic classes of the word and the conceptually different senses (Bordag, 2006). Thus we must extract the features able to distinguish these differences. They are:

Local tokens: the word occuring in the window -3 – +3;

Local bigram feature: bigram within -5 – +5 Chinese character range;

The above two features model the syntactic usage of a specific sense of a Chinese word.

Topical or conceptual feature: the content words (pos-tagged as noun, verb, adjective) within the given sentence. As the sentence in the training set seems generally short, a short window may not contains enough infomation.

We represent the words in a 0-1 vector according to their existence in a given sentence. Then the similarity measure between two given sentences is derived from their cosine similarity. We find that it is difficult to define the relative importance of different types of features in order to combine them in one vector space, and find that ignoring weight achieve better result. Brody (2009) achieves this in LDA model through a layered model with different probability of feature given sense.

Later we use a spectral clustering method from R kernlab package (Karatzoglou et al., 2004) which implements the algorithm described in (Ng et al., 2001). Instead of using the Gaussian kernel matrix as the similarity matrix we use the cosine similarity derived above.

One observation is that instances with the same target word sense often appear in the same context. However, for some verb in Chinese, it is often the case that one sense relates to a concrete object while the other relates to a more broad and abstract concept and the context varies considerably. Simple word co-occurence cannot define a good similarity measure to group these cases into one cluster. We must consider semantic relatedness measures between contexts.

3 Performance

Our system performs well on the training set. Two methods are used to evaluate the performance under different features.

method	precision	recall	F-score
Purity-based	81.11	83.19	81.99
B-cubed	74.41	76.51	75.33

Table 1: The performance of training set

Our system finally gets a F-score of 0.7598 on the test set.

4 Conclusion

Our experiment in the Chinese word sense induction task performs good with respect to the relative small corpus(only the training set). But only considering token co-occurence cannot achieve better result. Moreover, it is difficult to define a similarity measure solely based on lexicon infomation with no regard to semantic relatedness. Finally, combining different types of features seems to be another challenge in our model.

5 Acknowledgments

This research is supported by National Natural Science Foundation of Chinese (No.9092001).

References

Agirre, Eneko and Aitor Soroa. 2007. Ubc-as: a graph based unsupervised system for induction and classification. In *SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 346–349, Morristown, NJ, USA. Association for Computational Linguistics.

Bordag, Stefan. 2006. Word sense induction: Triplet-based clustering and automatic evaluation. In *EACL*. The Association for Computer Linguistics.

Brody, Samuel and Mirella Lapata. 2009. Bayesian word sense induction. In *EACL*, pages 103–111. The Association for Computer Linguistics.

Karatzoglou, Alexandros, Alex Smola, Kurt Hornik, and Achim Zeileis. 2004. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20.

Ng, Andrew Y., Michael I. Jordan, and Yair Weiss. 2001. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press.