Chinese Word Sense Induction based on Hierarchical Clustering Algorithm

Ke Cai, Xiaodong Shi, Yidong Chen，Zhehuang Huang, Yan Gao
Cognitive Science Department, Xiamen University, Xiamen, 361005, China

Abstract

Sense induction seeks to automatically identify word senses of polysemous words encountered in a corpus. Unsupervised word sense induction can be viewed as a clustering problem. In this paper, we used the Hierarchical Clustering Algorithm as the classifier for word sense induction. Experiments show the system can achieve 72% F-score about train-corpus and 65% F-score about test-corpus.

1. Introduction

Word sense induction is a central problem in many natural language processing tasks such as information extraction, information retrieval, and machine translation [Vickrey et al., 2005].

Clp 2010 launches totally 4 tasks for evaluation exercise, these are: Chinese word segmentation, Chinese parsing, Chinese Personal Name disambiguation and Chinese Word Sense Induction. We participated in task 4, which is Chinese Word Sense Induction..

Because the contents surround an ambiguous word is related to its meaning, we solve the sense problem by grouping the instances of the target word into the supposed number of clusters according to the similarity of contexts of the instance. In this paper we used the hierarchical clustering algorithm to accomplish the problem.

The task can be defined as two stage process: Feature selection and word clustering. Researchers have proposed much approach to the sense induction task which involved the use of basic word co-occurrence features and application of classical clustering algorithms.

Because the meanings of unknown words can be inferred from the contexts in which they appear, Pantel and Lin (2002) map the senses to WordNet. More recently, the mapping has been used to test the system on publicly available benchmarks (Purandare and Pedersen, 2004; Niu et al., 2005).

However, this approach does not generalize to multiple-sense words. Each sense of a polysemous word can appear in a different context, there have been many attempts in recent years to apply classical clustering algorithms to this problem.

Clustering algorithms have been employed ranging from k-means (Purandare and Pedersen, 2004), to agglomerative clustering (Sch¨utze, 1998), and the Information Bottleneck (Niu et al., 2007). Senses are induced by identifying highly dense subgraphs (hubs) in the co-occurrence graph (V´eronis, 2004).The sIB algorithm was used to estimate cluster structure, which measures the similarity of contexts of instances according to the similarity of their feature conditional distribution(Slonim, et al.,2002). Each algorithm treats words as feature vectors, using the same similarity function based on context information.

The remainder of this paper is organized as follows. In section 2 the Featured set and word similarity definition is introduced. The hierarchical clustering algorithm is presented in section 3. Section 4 provides the experimental results and conclusion is drawn in section 5.

2. Feature Selection and Word Similarity Definition

2.1 Feature Selection

A feature set is used designed to capture both immediate local context in our experiment, wider context and syntactic context. Specifically, we experimented with several feature categories: ±5-word window (5w), ±3-word window (3w), part-of speech n-grams and dependency relations. These features have been widely adopted in various word sense induction algorithms. The overall best scores are achieved with local (5 words) context windows.

. 2.2 Similarity Definition

We treat the context words as feature vectors, using the same similarity function. Suppose $C_i(w_{i1}, w_{i2}\cdots w_{in})$ is the contexts set of sentence S_i, and $C_j(w_{j1}, w_{j2}\cdots w_{jn})$ is the contexts set of sentence S_j.

Then we defined $sim(S_i, S_j) = \sum_{\substack{W_{ik}\in C_i \\ W_{jl}\in C_j}} w_{kl} sim(w_{ik}, w_{jl})$, here w_{kl} is variable weight,

Where $sim(w_{ik}, w_{jl}) = \dfrac{\beta}{dis(w_{ik}, w_{jl}) + \beta}$, β is an adjustable parameter with a value of 1.2, and

$Dis(w_{ik}, w_{jl})$ is the path length between w_{ik} and w_{jl} based on the semantic tree structure used for TongYiCi CiLin (同义词词林).

3. The Hierarchical Clustering Algorithm Used In Word Sense Induction

Sense induction is viewed as an unsupervised clustering problem where to group a word's contexts into different classes, each representing a word sense. In this paper, we use the bottom-up clumping approach, which begin with n singleton clusters and successively merge clusters to produce the other ones.

Table1: Hierarchical Clustering Algorithm:

1. initialize number of senses n 、 number of clusters m

   and clusters $C_i(w_{i1}, w_{i2}\cdots), i = 1, 2\cdots m$

2. Set $k = n$
3. Set $k = k - 1$

4. Find the nearest clusters C_i and C_j , Merge C_i and C_j

5. If $k > m$, go to step 3, otherwise go to step 6;
6. return m clusters

The merging of the two clusters in step 4 simply corresponds to adding an edge between the nearest pair of nodes in C_i and C_j .To find the nearest clusters, the following clustering similarity function is used:

$$sim(S_i, S_j) = \sum_{\substack{W_{jk} \in C_i \\ W_{jl} \in C_j}} w_{kl} sim(w_{ik}, w_{jl}) .$$

Our model incorporates features based on lexical information and parts of speech. So we propose a improved hierarchical clustering algorithm based on parts of speech.

Table2: improved algorithm based on parts of speech.

---

1. initialize number of senses n 、 number of clusters m

   and clusters $C_i(w_{i1}, w_{i2}\cdots), i = 1,2\cdots m$

2. Part of Speech Tagging on the corpus

3. Divided n senses into nn classes base on the information of parts of speech.

4. If $nn = m$, return m clusters

5. If $nn < m$, invoke hierarchical clustering algorithm in different classes，merge clusters into m cluster.

6. if $nn > m$, invoke hierarchical clustering algorithm in different tagging, merge clusters into m cluster.

7. return m clusters

---

4. Experimental Results

The test data includes totally 100 ambiguous Chinese words, every word have 50

untagged instances. Table3 show the best/worst/average F-Score of our system about train-corpus and test-corpus.

	Best word	Worst word	All words
Train-corpus	0.98	0.5	0.73
Test-corpus	------	-----	0.65

Table 3 Model performance with deferent corpus

Table 4 shows the performance of our model about train-corpus when using 3w and 5w word windows, which represent more immediate, local context.

	Best word	Worst word	All words
3w(±3-word window)	0.98	0.5	0.73
5w(±5-word window)	0.92	0.52	0.72

Table 4 Model performance with deferent windows

Table 5 summarizes the F-score in our system about train-corpus when using deferent similarity definition.

	Best word	Worst word	All words
This article	0.98	0.5	0.73
Qun LIU	0.99	0.59	0.78

Table 5 Model performance with deferent similarity definition

Experimental results show that the Hierarchical Clustering Algorithm can be applied to sense induction. Considering words to be feature vectors and applying clustering algorithm can improve accuracy of the task. A significant gap still exists between the results of these techniques and the gold standard of manually compiled word sense dictionaries.

5. Conclusions

Sense induction is treated as an unsupervised clustering problem. In this paper we adopt hierarchical clustering algorithm to accomplish the problem. Generate context words according to this distribution of key words and formalize the induction problem in a generative mode. Experiments show the system can achieved 72% F-score about train-corpus and 65% F-score about test-corpus. The basic cluster algorithm can sorts the word sense into clusters corresponding to the context.

References

Boyd-Graber, Jordan, David Blei, and Xiaojin Zhu. 2007.A topic model for word sense disambiguation. In Proceedings of the EMNLP-CoNLL. Prague, Czech Republic,pages 1024–1033.

David Vickrey, Luke Biewald, Marc Teyssler, and Daphne Koller. Word-sense disambiguation for machine translation. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, page

771-778, 2005.

Qun LIU , Sujian LI. Word Similarity Computing Based on How-net. Computational Linguistics and Chinese Language Processing

Niu, Zheng-Yu, Dong-Hong Ji, and Chew-Lim Tan. 2007. I2r: Three systems for word sense discrimination, chineseword sense disambiguation, and english word sense disambiguation. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007). Association for Computational Linguistics, Prague, Czech Republic, pages 177–182.

Niu, Z.Y., Ji, D.H., & Tan, C.L. 2005. Word Sense Disambiguation Using Label Propagation Based Semi-Supervised Learning. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics.

Pantel, Patrick and Dekang Lin. 2002. Discovering word senses from text. In Proceedings of the 8th KDD. New York, NY, pages 613–619.

Pedersen, Ted. 2007. Umnd2 : Senseclusters applied to the sense induction task of senseval-4. In Proceedings of SemEval-2007. Prague, Czech Republic, pages 394–397.

Purandare, Amruta and Ted Pedersen. 2004. Word sense discrimination by clustering contexts in vector and similarity spaces. In Proceedings of the CoNLL. Boston, MA, pages 41–48

V'eronis, Jean. 2004. Hyperlex: lexical cartography for information retrieval. Computer Speech & Language. 18(3):223–252.