Chinese Word Sense Induction with Basic Clustering Algorithms

Yuxiang Jia[1,2], Shiwen Yu[1], Zhengyan Chen[3]
[1]Key Laboratory of Computational Linguistics, Ministry of Education, China
[2]College of Information and Engineering, Zhengzhou University, Zhengzhou, China
[3]Department of Information Technology, Henan Institute of Education, Zhengzhou, China
{yxjia,yusw}@pku.edu.cn chenzhengyan1981@163.com

Abstract

Word Sense Induction (WSI) is an important topic in natural langage processing area. For the bakeoff task Chinese Word Sense Induction (CWSI), this paper proposes two systems using basic clustering algorithms, k-means and agglomerative clustering. Experimental results show that k-means achieves a better performance. Based only on the data provided by the task organizers, the two systems get FScores of 0.7812 and 0.7651 respectively.

1 Introduction

Word Sense Induction (WSI) or Word Sense Discrimination is a task of automatically discovering word senses from un-annotated text. It is distinct from Word Sense Disambiguation (WSD) where the senses are assumed to be known and the aim is to decide the right meaning of the target word in context. WSD generally requires the use of large-scale manually annotated lexical resources, while WSI can overcome this limitation. Furthermore, automatically induced word senses can improve performance on many natural language processing tasks such as information retrieval (Uzuner et al., 1999), information extraction (Chai and Biermann, 1999) and machine translation (Vickrey et al., 2005).

WSI is typically treated as a clustering problem. The input is instances of the ambiguous word with their accompanying contexts and the output is a grouping of these instances into classes corresponding to the induced senses. In other words, contexts that are grouped together in the same class represent a specific word sense.

The task can be formally defined as a two stage process, feature selection and word clustering. The first stage determines which context features to consider when comparing similarity between words, while the second stage apply some process that clusters similar words using the selected features. So the simplest approaches to WSI involve the use of basic word co-occurrence features and application of classical clustering algorithms, more sophisticated techniques improve performance by introducing new context features, novel clustering algorithms, or both. (Denkowski, 2009) makes a comprehensive survey of techniques for unsupervised word sense induction.

Two tasks on English Word Sense Induction were held on SemEval2007 (Agirre and Soroa, 2007) and SemEval2010 (Manandhar and Klapaftis, 2010) respectively, which greatly promote the research of English WSI.

However, the study on Chinese Word Sense Induction (CWSI) is inadequate (Zhu, 2009), and Chinese word senses have their own characteristics. The methods that work well in English may not work well in Chinese. So, as an exploration, this paper proposes simple approaches utilizing basic features and basic clustering algorithms, such as partitional method k-means and hierarchical agglomerative method.

The rest of this paper is organized as follows. Section 2 briefly introduces the basic clustering algorithms. Section 3 describes the feature set. Section 4 gives experimental details and analysis. Conclusions and future work are given in Section 5.

2 Clustering Algorithms

Partitional clustering and hierarchical clustering are the two basic types of clustering algorithms.

Partitional clustering partitions a given dataset into a set of clusters without any explicit structure, while hierarchical clustering creates a hierarchy of clusters.

The k-means algorithm is the most notable partitional clustering method. It takes a simple two step iterative process, data assignment and relocation of means, to divide the dataset into a specified number of clusters, *k*.

Hierarchical clustering algorithms are either top-down or bottom-up. Bottom-up algorithms treat each instance as a singleton cluster at the beginning and then successively merge pairs of clusters until all clusters have been merged into a single cluster. Bottom-up clustering is also called hierarchical agglomerative clustering, which is more popular than top-down clustering.

We use k-means and agglomerative algorithms for the CWSI task, and compare the performances of the two algorithms.

Estimating the number of the induced clusters, *k*, is difficult for general clustering problems. But in CWSI, it is simplified because the sense number of the target word is given beforehand.

CLUTO (Karypis, 2003), a clustering toolkit, is used for implementation. The similarity between objects is computed using cosine function. The criterion functions for k-means and agglomerative algorithms are I2 and UPGMA respectively. Biased agglomerative approach is chosen in stead of the traditional agglomerative approach.

3 Feature Set

For each target word, instances are extracted from the XML data file. Then the encoding of the instance file is transformed from UTF-8 to GB2312. Word segmentation and part-of-speech tagging is finished with the tool ICTCLAS[1]. Then the following three types of features are extracted:

1. The part-of-speech of the target word

2. Words before and after the target word within window of size 3 with position information

3. Unordered single words in all the contextual sentences without the target word, punctuations and symbols of the part-of-speech "nx" (Each word is only counted once, which is different from the word frequency in the bag-of-words model)

The target word is not necessarily a segmented word. Their relations are as follows:

1. The target word is a segmented word.

E.g. 别/d 打/v 我/r 电话/n

Don't dial my phone.

The target word is "打" (dial) and the segmented word is also "打" (dial). So they match.

2. The target word is inside of a segmented word.

E.g.同/p 媒体/n 打交道/v

deal with media

The target word is "打" (deal), but the segmented word is "打交道" (deal with). Then we split the segmented word and specify the part-of-speech of the target word as "1".

3. The target word is the combination of two segmented words.

E.g. 发/v 动/v "/w 文化大革命/nz "/w

launching the "Culture Revolution"

The target word is "发动" (launching), but it is split into two segmented words "发" (start) and "动" (move). Then we combine the two segmented words and specify the part-of-speech of the target word as "2".

4. The target word is split into two segmented words.

E.g. 刮/v 起/v 了/u 东/j 北风/n

blow up northeast wind

The target word is "东北", but it is segmented into two words "东" (east) and "北风" (north wind). In this case, we specify the postion of first segmented word as the position of the target word and the part-of-speech of the target word as "3".

If the target word occurs more than once in an instance, we consider the first occurrence.

4 Experiments

4.1 Data Sets

Two data sets are provided. The trial set contains 50 target words and 50 examples for each target word. The test set consists of 100 new target word and 50 examples for each target word. Both data sets are collected from the internet.

Table 1 shows the distribution of sense numbers of the target words in the two data sets. We can see that two sense words dominate and three

[1]http://ictclas.org/

sense words are the second majority. The word "打" (beat) in the trial set has 21 senses.

Table 1. Distribution of sense numbers

sense number	2	3	4	6	7	8	21
trial set	39	9	1	0	0	0	1
test set	77	10	7	4	1	1	0

Table 2. Distribution of relations between target words and segmented words

relation type	1	2	3	4	Total
trial set	2314	105	68	12	2499
test set	4031	710	212	47	5000

As is shown in table 2, the total instance number in the trial set is 2499 because there is a target word has only 49 instances. About 7.4% of the instances in the trial set and 19.38% of the instances in the test set have mismatched target words and segmented words (with relation types 2, 3 and 4).

4.2 Evaluation Metrics

The official performance metric for the CWSI task is *FScore* (Zhao and Karypis, 2005). Given a particular class C_i of size n_i and a cluster S_r of size n_r, suppose n_r^i examples in the class C_i belong to S_r. The F value of this class and cluster is defined to be:

$$F(C_i, S_r) = \frac{2 * P(C_i, S_r) * R(C_i, S_r)}{P(C_i, S_r) + R(C_i, S_r)},$$

where $P(C_i, S_r) = \frac{n_r^i}{n_r}$ is the precision value and $R(C_i, S_r) = \frac{n_r^i}{n_i}$ is the recall value defined for class C_i and cluster S_r. The *FScore* of class C_i is the maximum F value attained at any cluster, that is

$$FScore(C_i) = \max_{S_r} F(C_i, S_r)$$

and the *FScore* of the entire clustering solution is

$$FScore = \sum_{i=1}^{c} \frac{n_i}{n} FScore(C_i)$$

where c is the number of classes and n is the size of the clustering solution.

Another two metrics, *Entropy* and *Purity* (Zhao and Karypis, 2001), are also employed in this paper to measure our system performance. *Entropy* measures how the various classes of word senses are distributed within each cluster, while *Purity* measures the extent to which each cluster contained word senses from primarily one class. The entropy of cluster S_r is defined as

$$E(S_r) = -\frac{1}{\log c} \sum_{i=1}^{c} \frac{n_r^i}{n_r} \log \frac{n_r^i}{n_r}$$

The entropy of the entire clustering solution is then defined to be the sum of the individual cluster entropies weighted according to the cluster size. That is

$$Entropy = \sum_{r=1}^{k} \frac{n_r}{n} E(S_r)$$

The purity of a cluster is defined to be

$$P(S_r) = \frac{1}{n_r} \max_i (n_r^i),$$

which is the fraction of the overall cluster size that the largest class of examples assigned to that cluster represents. The overall purity of the clustering solution is obtained as a weighted sum of the individual cluster purities and is given by

$$Purity = \sum_{r=1}^{k} \frac{n_r}{n} P(S_r)$$

In general, the larger the values of *FScore* and *Purity*, the better the clustering solution is. The smaller the *Entropy* values, the better the clustering solution is.

The above three metrics are defined to evaluate the result of a single target word. Macro average metrics are used to evaluate the overall performance of all the target words.

4.3 Results

The overall performance on the trial data is shown in table 3. From the Macro Average Entropy and Macro Average Purity, we can see that k-means works better than agglomerative method. The detailed results of the k-means system are shown in table 4.

Table 3. Result comparison on the trial data

	Entropy	Purity
k-means	0.4858	0.8288
agglomerative	0.5328	0.8020

Table 4. Detailed results of k-means system

TargetWord	SenseNum	Entropy	Purity
反射	2	0.855	0.72
翻身	2	0.692	0.78
发展	2	0.377	0.92
发动	3	0.207	0.94
扼杀	2	0.833	0.7
断气	2	0	1
断交	2	0.592	0.82
杜鹃	2	0.245	0.959
动力	2	0.116	0.98
东西	3	0.396	0.82
东方	2	0.201	0.96
东北	2	0.201	0.96
调动	3	0.181	0.9
导师	2	0.122	0.98
单纯	2	0.327	0.92
大人	2	0.653	0.82
大气	2	0	1
大陆	2	0.855	0.72
大军	2	0.5	0.8
打气	2	0.312	0.92
打破	2	0.519	0.86
打开	3	0.534	0.72
打断	2	0.846	0.7
打	21	0.264	0.48
戳穿	2	0.521	0.88
春秋	3	0	1
初二	2	0.76	0.78
出口	3	0.205	0.92
冲撞	2	0.854	0.72
冲洗	2	0.449	0.9
充电	2	0.467	0.9
吃饭	2	0.881	0.7
澄清	2	0.402	0.92
程序	2	0.39	0.92
草包	2	0.793	0.76
参加	2	0.904	0.68
采购	2	0.943	0.64
材料	3	0.548	0.74
哺育	2	0.583	0.86
补贴	2	0.999	0.52
病毒	2	0.242	0.96
标兵	2	0.75	0.74
便宜	3	0.464	0.84
比重	2	0.181	0.96
背离	2	0.672	0.78
报销	2	0.471	0.82
保管	3	0.543	0.7
保安	2	0.347	0.9
把握	4	0.508	0.66
暗淡	2	0.583	0.86

The official results on the test set are shown in table 5. Our k-means system and agglomerative system rank 5 and 8 respectively among all the 18 systems.

Table 5. System ranking

Rank	FScore	Rank	FScore
1	0.7933	6	0.7788
2	0.7895	7	0.7729
3	0.7855	8*	0.7651
4	0.7849	9	0.7598
5*	0.7812	18	0.5789

5 Conclusions and Future Work

This paper tries to build basic systems for Chinese Word Sense Induction (CWSI) task. Basic clustering algorithms including k-means and agglomerative methods are studied. No extra language resources are used except the data given by the task organizers.

To improve the performance of CWSI systems, we will introduce new features and study novel clustering algorithms. We will also investigate the bakeoff data sets to find some more characteristics of Chinese word senses.

Acknowledgements

The authors are grateful to the organizers of the Word Sense Induction task for their hard work to provide such a good research platform. The work in this paper is supported by grants from the National Natural Science Foundation of China (No.60773173, No.60970083).

References

D. Vickrey, L. Biewald, M. Teyssler, and D. Koller. 2005. Word sense disambiguation for machine

translation. In *Proceedings of HLT/EMNLP2005*, pp. 771-778.

E. Agirre and A. Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of SemEval2007*, pp. 7-12.

G. Karypis. 2002. CLUTO - a clustering toolkit. *Technical Report 02-017, Dept. of Computer Science, University of Minnesota*. Available at http://www.cs.umn.edu˜cluto.

H. Zhu. 2009. Research into Automatic Word Sense Discrimination on Chinese. *PhD Dissertation of Peking University*.

J. Y. Chai and A. W. Biermann. 1999. The use of word sense disambiguation in an information extraction system. *In Proceedings of AAAI/IAAI1999*, pp. 850-855.

M. Denkowski. 2009. A Survey of Techniques for Unsupervised Word Sense Induction. *Language & Statistics II Literature Review*.

O. Uzuner, B. Katz, and D. Yuret. 1999. Word sense disambiguation for information retrieval. In *Proceedings of AAAI/IAAI1999*, pp.985.

S. Manandhar and I. P. Klapaftis. 2010. SemEval-2010 Task 14: Evaluation Setting forWord Sense Induction &Disambiguation Systems. In *Proceedings of SemEval2010*, pp. 117-122.

Y. Zhao and G. Karypis. 2005. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10(2):141–168.

Y. Zhao and G. Karypis. 2001. Criterion functions for document clustering: Experiments and analysis. *Technical Report 01–40, Dept. of Computer Science, University of Minnesota.* Available at http://cs.umn.edu/˜karypis/publications.