NEUNLPLab Chinese Word Sense Induction System for SIGHAN Bakeoff 2010

Hao Zhang **Tong Xiao** **Jingbo Zhu**

1. Key Laboratory of Medical Image Computing (Northeastern University), Ministry of Education

2. Natural Language Processing Laboratory, Northeastern University

zhanghao1216@gmail.com

{xiaotong, zhujingbo}@mail.neu.edu.cn

Abstract

This paper describes a character-based Chinese word sense induction (WSI) system for the International Chinese Language Processing Bakeoff 2010. By computing the longest common substrings between any two contexts of the ambiguous word, our system extracts collocations as features and does not depend on any extra tools, such as Chinese word segmenters. We also design a constrained clustering algorithm for this task. Experiemental results show that our system could achieve 69.88 scores of *FScore* on the development data set of SIGHAN Bakeoff 2010.

1 Introduction

The goal of word sense induction (WSI) is to group occurrences containing a given ambiguous word into clusters with respect to sense. Most researchers take the problem of word sense induction as a clustering problem. Pantel & Lin (2002) clustered words on the basis of the distances of their co-occurrence vectors, and used global clustering as a solution. Neill (2002) used local clustering, and determined the senses of a given word by clustering its close associations.

In this paper, we propose a simple but effective method to extract collocations as features from texts without pre-segmentations, and design a constrained clustering algorithm to address the issue of Chinese word sense induction. By using our collocation extraction method, our Chinese WSI system is independent of any extra natural language processing tools, such as Chinese word segmenters. On the development set of SIGHAN 2010 WSI task, the experimental results show that our system could achieve 69.88 scores of *FScore*. In addition, the official results show that the performance of our system is 67.15 scores of *FScore* on the test set of SIGHAN Bakeoff 2010.

The rest of this paper is organized as follows. In Section 2, we present the task description of Chinese word sense induction. In Section 3, we first give an overview of our Chinese WSI system, and then propose our feature extraction method and constrained clustering algorithm. In Section 4, we describe the evaluation method and show the experimental results on the development and test data sets of the Bakeoff 2010. In Section 5, we conclude our work.

2 Task Description

Given the number of senses *S* and occurrences of the ambiguous word *w*, a word sense induction system is supposed to cluster the occurrences into *S* clusters, with each cluster representing a sense of the ambiguous word *w*. For example, suppose that there are some sentences containing the ambiguous word “暗淡” (gloomy), and the sense number *S* is 2, the job of WSI system is to cluster these sentences into 2 clusters, with each cluster representing a sense of “暗淡”. Based on this task description, it is obvious to regard the problem of WSI as a clustering problem.

Figures 1-2 shows example input and output of our WSI system , where there are 6 sentences and 2 resulting clusters. In Figure 1, the first column are the identifiers of sentences containing the word “暗淡”, and the second column are

part of the sentences. In Figure 2, the first column represents the identifiers of sentences, and the second column represents the identifiers of clusters generated by our Chinese WSI system.

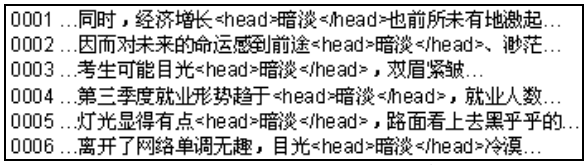

Figure 1 Part of input of word "暗淡" for our WSI system

0001	C1
0002	C0
0003	C1
0004	C0
0005	C0
0006	C1

Figure 2 Output of our WSI system for word "暗淡"

3 NEU Chinese WSI System

3.1 System overview

Our Chinese word sense induction system is built based on clustering work-frame. There are four major modules in the system, including *data pre-processing*, *feature extraction*, *clustering* and *data post-processing* modules. The architecture of our Chinese WSI system is illustrated in Figure 3.

3.2 Feature extraction

Since there is no separators in Chinese like "space" in English to mark word boundaries, most Chinese natural language processing applications need to first apply a Chinese word segmenter to segment Chinese sentences. In our Chinese word sense induction system, we extract collocations from sentences containing the ambiguous word as features. To extract collocations, we might first segment the sentences into word sequences, and then conduct feature extraction on the word-segmented corpus. However, errors might be induced in the procedure due to unavoidable incorrect segmentation results. Addressing this issue, we propose a method to directly extract collocations from sentences without pre-segmentations.

In our method, we extract two kinds of collocations, namely "global collocation" and "local collocation". Here global collocations are defined to be the words (or character sequences) that frequently co-occur with the ambiguous word, and local collocations are defined to be the characters adjacent to the ambiguous word[1].

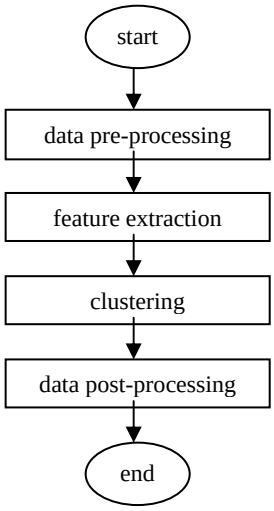

Figure 3 Architecture of our system

To extract global collocations, we first compute all the longest common substrings between any two of the sentences containing the ambiguous word to form the set of candidate global collocations. For each candidate global collocation, we count the number of sentences containing it. We then reduce the size of the candidate set by eliminating candidates which contain only one character or functional words. We also remove the candidate with other candidates as its substrings. Finally, we eliminate the candidates whose count of the number of sentences is below a certain threshold. The threshold equals to two in our experiments. We regard the candidates after the above processing as global collocations for WSI.

To extract local collocations, we simply extract one character on both left and right sides of the ambiguous word to form the set of candidate local collocations. We then refine the candidate set by eliminating candidates which are functional words or whose frequency is below a certain threshold. The threshold is set to two in our experiments.

After extracting global collocations and local collocations, we put them together to form the

[1] Definitions of global collocation and local collocation might be different from those in other papers.

final set of collocations and use them as features of our system. For each collocation (or feature), we compute the list of indices of sentences that containing the collocation. Thus, every element of the set of collocations has the data structure of pair of "key" and "value", where "key" is the collocation itself, and the "value" is the list of indices.

3.3 Clustering algorithm

We find that the high-confidence collocation is a very good indicator to distinguish the senses of an ambiguous word. However, the traditional clustering methods are based on the vector representations of features, which probably decreases the effect of dominant features (i.e. high-confidence collocations). To alleviate the problem, a nice way is to incorporate collocations into the clustering process as constraints. Motivated by this idea, we design a constrained clustering algorithm. In this algorithm, we could ensure that some occurrences of the ambiguous word *must* be in one cluster and some *must not* be in one cluster. The input for our constrained clustering algorithm is the set of collocations described in the previous section and the process of our clustering algorithm is shown in Table 1. Here the notation starting with character "C" represents a collocation, and the notations of "*Sin*" and "*Srlt*" represent the collocation set and the result set, respectively.

Every element in the result set *Srlt* is regarded as one cluster for a given ambigous word, and the list of the element records the indices of the sentences belonging to the cluster.

4 Evaluation of Our System

The evaluation method is *F-score* which is provided within the Bakeoff 2010 (Zhao and Karypis, 2005). Suppose *Cr* is a class of the gold standard, and *Si* is a cluster of our system generated. *FScore* is computed with the formulas below.

$$F-score(Cr,Si) = 2*P*R/(P+R) \quad (1)$$

$$FScore(Cr) = \max_{Si}(F-score(Cr,Si)) \quad (2)$$

$$FScore = \sum_{r=1}^{c} \frac{nr}{n} FScore(Cr) \quad (3)$$

We evaluate our Chinese word sense induction system on the development data set and the test data set of the Bakeoff 2010. The details of the development data set and the test data set are summarized in Table 2.

For comparison, we develop a baseline system that also uses the collocations as features and clustering based on the vector representations of features. On the development data set, we test our system and compare it with the baseline system. The performance of our Chinese WSI system and the baseline system are shown in Table 3. From Table 3, we see that using our constrained clustering algorithm is better than using the traditional hierarchical clustering methods by 7.06 scores of *FScore* for our Chinese WSI system. It indicates that our constrained clustering algorithm could avoid reducing the effect of

Input: collocation set *Sin*
while there is available collocation *Ci* in the input set *Sin*
  for each collocation *Ct* in the set *Sin*
    if *Ct* not equals to *Ci*, and *Ct* is available
      if list of *Ct* has intersection with that of *Ci*, or *Ct* and *Ci* have a meaningful substring (word or character), compose list of *Ct* into list of *Ci*, and mark *Ct* to be unavailable
      end if
    end if
  end for
  store *Ci* and its list into result set *Srlt*, and mark *Ci* to be unavailable
end while
if there are available collocations in the input set *Sin*
  if the size of result set *Srlt* does not satisfy the given cluster number, devide the rest collocations in *Sin* evenly into the rest clusters, and append their lists to their own clusters' lists respectively
  else add the rest collocations into the last cluster, and append their list to the list of the last cluster
  end if
end if
return the result set *Srlt*
Output: result set *Srlt*

Table 1 Constrained clustering algorithm

high-confidence features (i.e. high-confidence collocations) and lead to better clustering results. This conclusion is also ensured by the comparison between our constrained clustering algorithm and the traditional K-means clustering algorithm.

In addition, our system achieves 67.15 scores of *FScore* on the test data set reported by the SIGHAN Bakeoff 2010.

data	descriptions
Dev set	containing 50 ambiguous words, about 50 sentences for each ambiguous word
Test set	containing 100 ambiguous words, about 50 sentences for each ambiguous word

Table 2 Data sets of SIGHAN Bakeoff 2010

clustering methods	*FScore* of our system (%)
traditional hierarchical clustering	62.82
traditional K-means clustering	62.48
our constrained clustering	69.88

Table 3 System performance on dev set of Bakeoff 2010 using different clustering methods

5 Conclusions

In this paper, we propose a collocation extraction method and a constrained clustering algorithm for Chinese WSI task. By using the collocation extraction method and the clustering algorithm, our Chinese word sense induction system is independent of any extra tools. When tested on the test data set of the Bakeoff 2010, our system achieves 67.15 scores of *FScore*.

References

Vickrey, David, Luke Biewald, Marc Teyssler, and Daphne Koller. 2005. *Word-sense disambiguation for machine translation*. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Morristown, NJ, USA, pages 771-778.

Yarowsky, David. 1995. *Unsupervised word sense disambiguation rivaling supervised methods*. In Proceedings of 33rd Meeting of the Association for Computational Linguistics, Cambridge, MA, 189-196.

Schutze, Hinrich. 1998. *Automatic word sense discrimination*. Computational Linguistics, Montreal, Canada, 24(1):97–123.

Ng, Hwee Tou, Hian Beng Lee. 1996. *Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach*. In Proceedings of the 34th Meeting of the Association for Computational Linguistics, California, USA, pages 40-47.

Daniel, Neill. 2002. *Fully Automatic Word Sense Induction by Semantic Clustering*. In Computer Speech, Cambridge University, Master's Thesis.

Pantel, Patrick, Dekang Lin. 2002. *Discovering word senses from text*. In Proceedings of ACM SIGKDD, Edmonton, 613-619.

Rapp, Reinhard. 2004. *A Practical Solution to the Problem of Automatic Word Sense Induction*. In Proceedings of the 42nd Meeting of the Association for Computational Linguistics, Barcelona, Spain.

Zhao, Ying, George Karypis. 2005. *Hierarchical Clustering Algorithms for Document Datasets.* Data Mining and Knowledge Discovery, 10:141-168.