Word Sense Discrimination with four feature vectors

Ryuichiro Nakanishi (0910041)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 8, 2011

Keywords: Word Sense Discrimination, Corpus, Clustering.

The new senses of words grow day by day. When new senses are born, it is necessary to add them to a dictionary. It could contribute to the effective management of the dictionary if it is able to discover new senses automatically. In this thesis, I propose a new method of clustering of example sentences as an important fundamental technique to discover the new senses automatically from corpora. The detail procedures of clustering of example sentences are as follows. When a set of word instances(example sentences) for the target word is given, the system makes clusters where sentences with the same sense are collected. If we can judge whether the obtained cluster is the set of sentences with a new sense, we can discover new senses automatically from corpora. In this thesis, I aim at improvement of performance of clustering by using four feature vectors. Kuoka's work is related to clustering of example sentences. The flow of the processing in his method is as follows. First, he represented each instance as four feature vectors. Next, his system performed clustering with each feature vector. Finally, his system chose the best result from four results of clustering automatically. The basic idea of his method is that the points of view useful for clustering are different for target words. However, in general, effective points of view are different for senses of the same target word. In order to make clusters which have the same sense, it might better to use four feature vectors at the same time in process of clustering.

The overview of the proposed method is as follows. First, each word instance is represented with four feature vectors. Feature vectors in this

research are almost same with ones used in Kuoka's method. Kuoka's four feature vectors are $Collocation\ Vector$,$Context\ Vector$,$Assosiation\ Vector$ and $Topic\ Vector$. In this method, $Collocation\ Vector$ is modified so that window width of the vector is extended from one to two, because one width $Collocation\ Vector$ would give high similarity for pairs of example sentences with different senses. Features of $Collocation\ Vector$ are words and their parts-of-speech. Weight of word and parts-of-speech are equal in Kuoka's $Collocation\ Vector$, while a weight for parts-of-speech are half of one for words in this method.

The agglomerative clustering algorithm is used for clustering of example sentences. In the case of agglomerative clustering, it is easy to implement the idea to use multiple points of view. The similarity between clusters is defined as the highest similarity among four feature vectors. That is, a pair of example sentences could be regarded as the same word instances if the similarity of even one of four feature vectors is high enough. One of the problems of this approach is that the averages of similarity of feature vectors are different each other. In such a situation, only one type of feature vector is always chosen. Thus similarity between feature vectors are normalized in two ways. The first normalization method is a relative measure that is relative value in the range between minimum and maximum similarities. The second normalization method is to use standard derivation as normalized similarity.

On using four feature vectors, it would be better if one can interpret what the point of view is introduced for the constructed cluster. So the following constraint is introduced for clustering; example sentences which are similar for only one point of view are collected as a cluster.

When the system merges two clusters, the new cluster has the "cluster label $L(C)$",where the cluster label is one of types of feature vectors which is chosen for calculation of similarity between two clusters. Furthermore, two clusters are never merged when their cluster labels are different. This constraint makes a cluster a group of example sentences which are similar from only one point of view.

Two stopping conditions are set for agglomerative clustering: (1)when the number of clusters is less than the threshold Tc, (2)when the ratio of the size of the biggest cluster to the total number of example sentences is

more than the threshold Tr.

In the evaluation experiment, I use the training corpus of SemEval-2 Japanese Task. Forty target words of the task are used in the experiment. For each target word, 40 to 50 word instances are used. Each instance is represented as four feature vectors. Then clustering is performed for these feature vectors. In this experiment, stopping condition Tc is set to 10 and 15. Among 9 evaluation criteria of clustering, $Purity$,$Homogeneity$ and $Paired\ Precision$ are considered as the most important ones. Among the propose methods using four feature vectors, the method using normalized similarity of standard derivation is the best. And it is better than Kuoka's method which chooses the best one among four clustering results. However, the method using only $Collocation\ Vector$ achieved better performance than the proposed methods. The reason why is that the methods using one feature vector produce many small clusters consisting of only one example sentence, raising $Purity$ *etc.* unreasonably. However, such clusters are useless to discover new word senses. Then I investigated number of the clusters that have 2 or more than 1 example sentences, and the average precision, which is the ratio of the number of the most frequent sense within a single cluster, for these clusters. Comparing the results, the proposed method is better than the method using only $Collocation\ Vector$. For details, the average precision is 0.8565 for the proposed method with normalized similarity of relative measure,0.8279 for the method with normalized similarity of standard derivation, and 0.8190 for the method using $Collocation\ Vector$.The number of clusters consisting of two or more example sentences for these three method is 258, 347 and 211, respectively. These results show it is effective to use many feature vectors to improve the quality of the obtained clusters.