Discovering Word Senses from Corpora

Yusuke Kuoka (0610032)

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

February 7, 2008

Keywords: word sense, word sense discrimination, clustering.

Word Sense Disambiguation (WSD) is a fundamental technology for various NLP tasks. In Machine Translation, a sense of a polysemous word must be determined to get its proper translation. Moreover, in Information Retrieval, a sense of a polysemous keyword must be determined to extract a document in which the keyword has the same sense occurred.

The problem of WSD is that it rely heavily on manually compiled dictionaries. Common WSD methods select one of the meanings defined beforehand in a dictionary such as Iwanami Kokugo Jiten. However, a word might has a sense which is not defined in a dictionary, since senses of words always change day by day. When a word is used with an unknown sense, the past WSD methods are unable to select the correct senses of the word. To overcome this problem, dictionaries must be recompiled as quickly as possible when a new word sense has been occurred. However, compilations of dictionaries require a plenty of manual cost.

In this paper, we describe a new method of Word Sense Discrimination which discovers word senses by clustering instances of the target word such that instances with the same sense are grouped together. The method does not depend on dictionaries since it just discriminates a group of word instances with the same sense and does not label the group with a sense defined in an existing dictionary. The method allows us to identify instances of a word which has an unknown sense. Moreover, example sentences including the instances with the same sense can be automatically collected.

Our method can be applied as the way in assisting compilations of dictionaries and it also overcomes the problem of WSD that the unknown word sense cannot be handled.

Our word sense discrimination method consists of the following processes. At first, a corpus is prepared, and instances of a target word is extracted. Next, each instance is represented by feature vectors. Then, feature vectors are clustered so that similar vectors are grouped together. Finally, senses of the instances are distinguished by assuming that every instance in the same cluster have the same sense.

Our method has two characteristics. One is that every instance is represented by a variety of feature vectors utilizing co-occurrences of words, neighboring words or their parts of speech, or topics inferred by PLSI or LDA. To be precise, we propose the following feature vectors: Context Vector has words in the context of each instance as features, Adjacency Vector has collocations consisting a target word as features, Association Vector has topics or words related to words in the context as features, and Topic Vector has topics of the document including a target instance inferred by PLSI as features. The other characteristic is that each instance is represented by multiple feature vectors, combining them together to cluster instances, while many of previous methods utilize just one feature vector for clustering. We propose two combining methods. In the first method, similarity of clusters are calculated by weighted sum of similarities of different kinds of feature vectors. In the second method, evaluation measures for assessing the quality of clusters is introduced, and the feature vector producing the best clusters is chosen for each target word. The quality of clusters result is assessed in respect to high similarities between elements in each cluster, or low similarities between clusters. Evaluation measures involving a relative similarity between clusters or elements of a cluster is also introduced considering differences of similarity values between different feature vectors.

Our experiments has been held for 10 words extracted from Mainichi Shimbun, and 23 words extracted from Yahoo! Tiebukuro. For Mainichi Shimbun, 70 instances for each word are chosen. For Yahoo! Tiebukuro, 100 instances are chosen. The instances are clustered by Spherical k-means or Centroid Clustering. The result of each clustering is evaluated by Purity,

Entropy, Inverse Purity.

Adjacency Vector, Association Vector, Topic Vector, Context Vector performed well in descending order when instances has clustered with these feature vectors alone. Moreover, the best discriminating feature vector varied according to the target word. The combination of similarities between Context Vector, Adjacency Vector, Association Vector, Topic Vector performed better than each feature vector alone. Furthermore, The method choosing the best discriminative feature vector for each word by evaluation measures performed even more better than each feature vector or the combination of similarities. In addition, evaluation measures involving relative similarity performed better than absolute similarity. However, the evaluation measures did not always choose the best feature vector in respect to Purity or Inverse Purity. Thus, better evaluation measures should be investigated to improve a quality of clusters.