The discovery of a new sense of a word based on clustering of sentences

Hiroki Tanaka (0710043)

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

February 5, 2009

Keywords: Corpus,New sense of a word,Dictionary,Clustering.

In this thesis, I describe techniques to discover a new sense of a word from a corpus automatically. Here, new sense indicates the meaning of a word that has not been defined in an existing dictionary. The flow of the processing in the proposed methods is as follows. First, I collect some instances (examples) of the word from the corpus,and apply the clustering technique to compile instances with the same meaning to a cluster. Next, both clusters of sentences and the sense of a word in a dictionary is converted to vectors,then similarity between these vectors are measured.Thus the sense corresponding to each example cluster is chosen. Finally,I calculate "existing sense likelihood",which is a measure how likely a cluster corresponds to one of existing senses in a dictionary,using the similarity of a cluster and a sense calculated before, and judge if a cluster corresponds to a new sense of a word based on that value.

I used the methods suggested by Kuoka when clustering of the example sentences of the word. Here,when I convert each sentence to a vector,I tested Kuoka's four feature vectors:Context Vector,Adjacency Vector with LDA,Association Vector and Topic Vector,and two methods to combine these vectors. Furthermore,I use three clustering algorithm;k-means method which Kuoka applied, k-means method selecting an initial cluster by KKZ method(k-means+KKZ) and Top-down division method. As a result of the experiment,for the purity,which indicates how much sentences

in the same cluster have the same meaning of the word, k-means+KKZ was the best.On the other hand ,for the discrimination rate of new sense of a word,which indicates how much new senses of words are collected in one cluster,top-down division method was the best.

Next,the methods to map sentence clusters to existing senses of a word in a dictionary is proposed. First,I convert a sentence cluster to a feature vector. Firstly, I make a co-occurrence matrix which shows strength of the co-occurrence between words,and I assume the row of the matrix is a co-occurrence vector of the word. Secondly,in a sentence cluster,I define the feature vector of cluster as co-occurrence vector of content words appearing in the context of the target word. Next, the method to make the feature vector of the sense of a word is proposed. Here,I use the sentences in the dictionary. In this research,I classify the sentences in the dictionary into four types;definition sentence,example sentence,reference entry and others. I use a definition sentence and an example sentence when I construct the feature vector of senses. More concretly,I define the feature vector of the sense as sum of co-occurrence vector define of the content words appearing in the definition sentence or the example sentence. However,the feature vectors of the senses made by this technique are different in terms of the degree of sparseness according to the length of the definition sentences and example sentences. As a result,accuracy of mapping a sentence cluster to a sense becomes low. Therefore,I suggest three methods :the first one is an improvement of constructing feature vector of senses,the second one is an improvement of the computation of similarity between asentense ckuster and senses , and the third one is the method to complete feature vectors. The method to complete feature vectors was the most effective among them. In the evaluation experiment,I use two kinds of sentence clusters sentence clusters:one is completely correct sentence clusters made by hand, ,the other is the example cluster autimaically constructed by the clustering technique that I suggested. The accurracy of mapping example clusters to senses was 61.9% when example clusters are correct,while 59.5% when clusters are automatically constracted.

Finally, I proposes the technique to judge whether the sentence cluster is a new meaning of a word. I define "existing sense likelihood" (K) as degree of new likely cluser is mapped to one of existing sense in a dictionary,and

I use this value for a judge. In this research,I compared three kinds of K:the first one is a variance of simiralities between sentense clusters and the existing senses of a word(K-Var),the second one is a difference between maximum and minimun of similarities (K-$Diff$),the third one is the maximum of similarities (K-Max). I try to judge that the sentense cluster is a new sense of a word when K is small. However,as a result of a preliminary experiment, I found that it is rather difficult to judge it a sentence duster is a new sense or not by simply setting a threshold for K. So,I suggest technique to discriminate a sentence clusters of senses from ones of new senses : first arrange sentence clusters in descendent order of K,then I find the point that where difference of K is the greatest and great enough.According to the of experiment,K-Max is the most effective approach among three kinds of K.Furthermore F-measure of the judgment of new senses is 0.615 when sentence cluster is correct.