Research on a Graph-based Method for Word sense Induction

YIN BO (1210211)

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

August, 7, 2014

Keywords: Word Sense Induction, Graph-based Method, Small World Graph, Key Player Problem, Syntactic Relation.

Word Sense Induction (WSI) is an open problem of natural language processing, which concerns automatic identification of senses of a word. The senses of words are usually defined by a dictionary, but it might not compile all senses, especially a new sense or a sense in a specific domain. Therefore, WSI is important to recognize all the senses for a given word and indispensable for various NLP. In general, WSI is a task to construct clusters of contexts in which the target word occurs or to build cluster of words related to the sense of the target word. Each obtained cluster corresponds to one sense of the target word. In order to solve WSI problem, two major approaches have been proposed, clustering based method and graph based method.

HyperLex is one of the graph-based word sense induction methods proposed by Véronis. HyperLex infers senses of the target word by the following procedures: (1) a co-occurrence graph where vertices are words appearing in the context of the target word is built, (2) hubs that are representative words of a sense are found from the graph, (3) the graph is subdivided into several sub-trees whose root node are the hubs, and (4) sub-trees are used for word sense disambiguation of unknown words. Although Véronis

reported promising results on HyperLex, there are some problems on it. The first problem concerns the way how to construct the co-occurrence graph. In HyperLex, there are several parameters to decide words to be added as vertices of the graph, or pairs of words to be added as edges. The structure of the co-occurrence graph highly depends on these parameters. However, parameters are determined in ad-hoc manner in HyperLex. Since the best parameter may be different for the target word or the corpus used for WSI, a scheme to determine the optimum parameters should be investigated. The second problem concerns how to choose the hub. In HyperLex, the frequent words in the corpus are simply chosen as the hubs. However, the structure of the graph should be considered in selection of the hubs, since the hub should be the center of the dense sub-graph. The third problem concerns how to determine weights of edges. In a graph-based WSI, edges in the co-occurrence graph are weighted so that the weight is small when two words correlate each other. In HyperLex, only the co-occurrence frequency of two words is considered to set weights of the edges. However, correlation between two words can be measured from other points of views. Syntactic relation is one of them. If two words are under a syntactic relation more frequently in the corpus, their relations might be strong and the weight of the edge connecting them should be set small. Although it is well-known that the syntactic relation is one of the useful features for word sense disambiguation, it is not considered in HyperLex. The goal of this research is to propose a method to tackle the above problems.

This thesis presents a graph based word sense induction method that is an extenstion of HyperLex. First, a co-occurence graph is constructed. Vertices correspond to words appearing in the context of the target word, while edges between vertices represent co-occurrence relations between words. Weights of the edges indicate how two words correlated. In this method, the weights of the edges are determined based on both co-occurrence and syntactic relations of words. When two words co-occur frequently and they frequently appear under any kinds of syntactic relations, the weight of the edge between them becomes low. Once the co-occurrence graph is built, a simple iterative algorithm is applied to obtain the hubs by considering the

connectivity of the graph. The hubs are detected by *Key Player Problem (KPP)* algorithm that measures a graph connectivity. Unlike HyperLex, a node that is connected from many other nodes is chosen as the hub. Next, an expanded graph is built. A dummy node representing the target word itself is added to the graph. Then each hub is linked to the target word with 0 weights. Finally, a minimum spanning tree (MST) is computed over the expanded graph by taking the target word as the starting node and making previously identified hubs as its first level children. The target word is removed from the MST, causing that the graph is subdivided into the sub-trees whose root nodes are hubs.

As already discussed, there are three parameters in HyperLex. Since the structure of the co-occurrence graph is highly dependent on them, this thesis proposes a method to optimize the parameters. We suppose that the co-occurrence graph should be a small world graph, which consists of several dense subsets of nodes (small world) that are loosely connected each other. A property of small world graph can be represented by $L \sim L_{rand}$ and $C >> C_{rand}$, where L and C respectively stand for the characteristic path length and clustering coefficient of the graph and the suffix 'rand' means a random graph. For various combination of parameters, we choose the parameters that fulfills the small world property best. In addition to the small world property, the number of the hubs in the graph is also considered in the optimization of the parameters. When the number of the hubs (the number of the induced senses in other words) is too large, it may be an incorrect result. In our method, for each parameter combination, two conditions, $|\frac{L}{L_{rand}} - 1| \leq T_1$ and (number of hubs) $\leq T_2$, are checked, where T_1 and T_2 are thresholds. If these conditions are not fulfilled, the combination of parameters is discarded. Then we choose the best parameter combination where $C_{sw} = |C_{weight}/C_{rand}|$ is maximum. Note that C_{weight} stands for a clustering coefficient in a weighted graph.

The performance of the proposed WSI method was evaluated on a sense tagged corpus for 57 target words. We compared Purity, Inverse Purity and the harmonic mean of them (P-IP hereafter), that are well-known criteria for clustering, of the original HyperLex and our extended model.

The average of $P\text{-}IP$ for 57 target words of HyperLex was 0.482. By optimization of the parameters, $P\text{-}IP$ was improved to 0.513. $P\text{-}IP$ of the model where the hubs are chosen by KPP was 0.609. Finally, considering syntactic relations in the weights of the edges, $P\text{-}IP$ was improved to 0.611. Furthermore, the number of induced senses in the proposed method was more similar to the genuine number of senses. These results proved the effectiveness of our proposed method.

Although this thesis proposed a new weighting scheme considering syntactic relations, the structure of the co-occurrence graph is independent of it. In future, we will examine a method to the build co-occurrence graph considering the weights between two words.