Word Sense Induction by Community Detection

David Jurgens[1,2]
[1]HRL Laboratories, LLC [2]Department of Computer Science
Malibu, California, USA University of California, Los Angeles
jurgens@cs.ucla.edu

Abstract

Word Sense Induction (WSI) is an unsupervised approach for learning the multiple senses of a word. Graph-based approaches to WSI frequently represent word co-occurrence as a graph and use the statistical properties of the graph to identify the senses. We reinterpret graph-based WSI as community detection, a well studied problem in network science. The relations in the co-occurrence graph give rise to word communities, which distinguish senses. Our results show competitive performance on the SemEval-2010 WSI Task.

1 Introduction

Many words have several distinct meanings. For example, "law" may refer to legislation, a rule, or police depending on the context. Word Sense Induction (WSI) discovers the different senses of a word, such as "law," by examining its contextual uses. By deriving the senses of a word directly from a corpus, WSI is able to identify specialized, topical meanings in domains such as medicine or law, which predefined sense inventories may not include.

aWe consider graph-based approaches to WSI, which typically construct a graph from word occurrences or collocations. The core problem is how to identify sense-specific information within the graph in order to perform sense induction. Current approaches have used clustering (Dorow and Widdows, 2003; Klapaftis and Manandhar, 2008) or statistical graph models (Klapaftis and Manandhar, 2010) to identify sense-specific subgraphs.

We reinterpret the challenge of identifying sense-specific information in a co-occurrence graph as one of *community detection*, where a community is defined as a group of connected nodes that are more connected to each other than to the rest of the graph (Fortunato, 2010). Within the co-occurrence graph, we hypothesize that communities identify sense-specific contexts for each of the terms. Community detection identifies groups of contextual cues that constrain each of the words in a community to a single sense.

To test our hypothesis, we require a community detection algorithm with two key properties: (1) a word may belong to multiple, overlapping communities, which is necessary for discovering multiple senses, and (2) the community detection may be hierarchically tuned, which corresponds to sense granularity. Therefore, we adapt a recent, state of the art approach, Link Clustering (Ahn et al., 2010). Our initial study suggests that community detection offers competitive performance and sense quality.

2 Word Sense Induction

A co-occurrence graph is fundamental to our approach; terms are represented as nodes and an edge between two nodes indicates the terms' co-occurrence, with a weight proportional to frequency. While prior work has focused on clustering the nodes to induce senses, using Link Clustering (Ahn et al., 2010), we cluster the *edges*, which is equivalent to grouping the word collocations to identify sense-specific contexts. We summarize our approach as four steps: (1) selecting the contextual cues, (2) building a co-occurrence graph, (3) performing community detection on the graph, and (4) sense labeling new contexts using the discovered communities.

Context Refinement Representing the co-occurrence graph for all terms in a context is

Proceedings of the TextGraphs-6 Workshop, pages 24–28,
Portland, Oregon, USA, 19-24 June 2011.

prohibitively expensive. Moreover, often only a subset of the terms in a context constrain the sense of an ambiguous word. Therefore, we refine a word's context to include only a subset of the terms present. Following previous work (Véronis, 2004), we select only nouns in the context.

Early experiments indicated that including infrequent terms in the co-occurrence graph yielded poor performance, which we attribute to having too few connecting edges to identify meaningful community structure. Therefore, we include only those nouns occurring in the most frequent 5000 tokens, which are likely to be representative the largest communities in which a term takes part. Last, we include all the nouns and verbs used in the SemEval 2010 WSI Task (Manandhar et al., 2010), which are used in our evaluation. The selected context terms are then stemmed using the Porter stemmer.

Building the Co-occurrence Graph The graph is iteratively constructed by adding edges between the terms from a context. For each pair-wise combination of terms, an edge is added and its weight is increased by 1. This step effectively embeds a clique if it did not exist before, connecting all of the context's words within the graph. Once all contexts have been seen, the graph is then pruned to remove all edges with weight below a threshold $\tau = 25$. This step removes edges form infrequent collocations, which may not contribute sufficient graph structure for community detection, and as a practical consideration, greatly speeds up the community detection process. However, we note that parameter was largely unoptimized and future work may see a benefit from accounting for edge weight.

Community Detection Within the co-occurrence graph, communities may have partial overlap. For example, Figure 1 illustrates a part of the local graph for "mouse." Two clear senses emerge from the neighbors: one for the input device and another for the animal. However, the terms that correspond to one sense also co-occur with terms corresponding to the other sense, e.g., "information," which hinders finding communities directly from disconnected components in the local neighborhood. Finding sense-specific communities requires recognizing that the co-occurring terms may be shared by multiple communities. Therefore, to identify communities we selected the approach of Ahn et al. (2010), summarized next, which performs well for overlapping community structure.

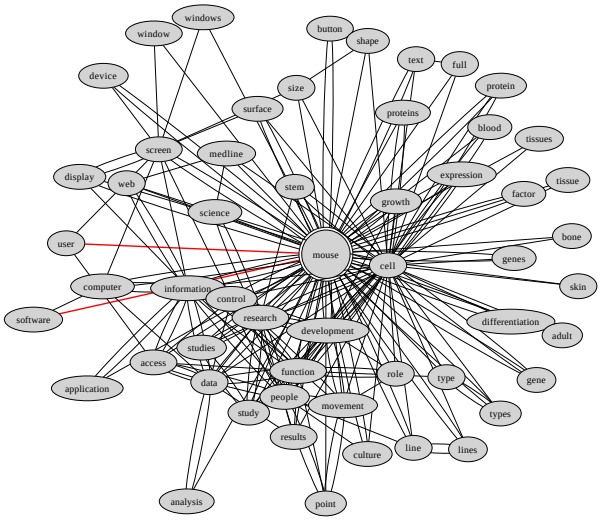

Figure 1: A portion of the local co-occurrence graph for "mouse" from the SemEval-2010 Task 14 corpus

First, the edges are clustered using an unweighted similarity function based on the neighbors of two edges, $e_{i,j}$ and $e_{i,k}$: $sim(e_{i,j}, e_{i,k}) = \frac{n_j \cap n_k}{n_j \cup n_k}$, where n_i denotes the node i and its neighbors. This similarity reflects the percentage of terms that co-occur in common with the term for nodes j and k, independent of the terms that co-occur with the shared term for i. For example, in Figure 1, the similarity for the edges connecting "mouse" with "user" and "software," $\frac{2}{5}$, measures the overlap in the neighbors of "user" and "software" independent of the neighbors for "mouse," such as "cell" and "size."

Using this similarity function, the edges are agglomeratively clustered into a dendrogram. We use the single-link criteria which iteratively merges the two clusters connected by the edge pair with the highest similarity. The dendrogram may then be cut at different levels to reveal different cluster granularities; cuts near the bottom of the dendrogram create a larger number of small groups of collocations, whereas cuts near the top create fewer, larger groups of collocations. To select the specific partitioning of the dendrogram into clusters, we select the solution that maximizes the partition density, which Ahn et al. (2010) define as $D = \frac{2}{M} \sum_c m_c \frac{m_c - (n_c - 1)}{(n_c - 2)(n_c - 1)}$, where M is the number of edges in the graph, c de-

notes a specific cluster, and n_c and m_c are the number of nodes and edges in cluster c, respectively.

The final set of communities is derived from these partitions: a node is a member of each community in which one of its edges occurs. Last, we remove all communities of size 3 and below, which we interpret as having too few semantic constraints to reliably disambiguate each of its terms.

Sense Induction from Communities Each term in a community is treated as having a specific sense, with one sense per community. To label a contextual usage, we identify the community that best maps to the context. For a given context, made of the set of words W, we score each community i, consisting of words C, using the Jaccard index weighted by community size: $score(C_i, W) = |C_i| \cdot \frac{|C_i \cap W|}{|C_i \cup W|}$. This similarity function favors mapping contexts to larger communities, which we interpret as having more semantic constraints. The final sense labeling consists of the scores for all overlapping communities.

3 Evaluation

We use the SemEval-2 Task 14 evaluation (Manandhar et al., 2010) to measure the quality of induced senses. We summarize the evaluation as follows. Systems are provided with an unlabeled training corpus consisting of 879,807 multi-sentence contexts for 100 polysemous words, comprised of 50 nouns and 50 verbs. Systems induce sense representations for target words from the training corpus and then use those representations to label the senses of the target words in unseen contexts from a test corpus. We use the entire multi-sentence context for building the co-occurrence graph.

The induced sense labeling is scored using two unsupervised and one supervised methods. The unsupervised scores consists of two contrasting measures: the paired FScore (Artiles et al., 2009) and the V-Measure (Rosenberg and Hirschberg, 2007). Briefly, the V-Measure rates the homogeneity and completeness of a clustering solution. Solutions that have word clusters formed from one gold-standard sense are homogeneous; completeness measures the degree to which a gold-standard sense's instances are assigned to a single cluster. The paired FScore reflects the overlap of the solution and the gold standard in cluster assignments for all pair-wise combination of instances. The supervised evaluation transforms the induced sense clusters of a portion of the corpus into a word sense classifier, which is then tested on the remaining corpus. An 80/20 train-test split, $S_{80/20}$, and 60/40 split, $S_{60/40}$, are both used.

	FScore	V-Meas.	$S_{80/20}$	$S_{60/40}$
S_{PD}	61.1 (3)	3.6 (18)	57.64 (18)	57.64 (16)
S_V	56.16 (9)	8.7 (6)	57.90 (18)	57.36 (17)
S_F	63.4 (1)	0 (26)	56.18 (21)	56.20 (21)
Best_F	63.3 (1)	0 (26)	58.69 (14)	58.24 (13)
Best_V	26.7 (25)	16.2 (1)	58.34 (16)	57.27 (17)
Best_S	49.8 (15)	15.7 (2)	62.44 (1)	61.96 (1)
MFS	63.4	0	58.67	58.95

Table 1: Performance results on the SemEval-2010 WSI Task, with rank shown in parentheses. Reference scores of the best submitted systems are shown in the bottom.

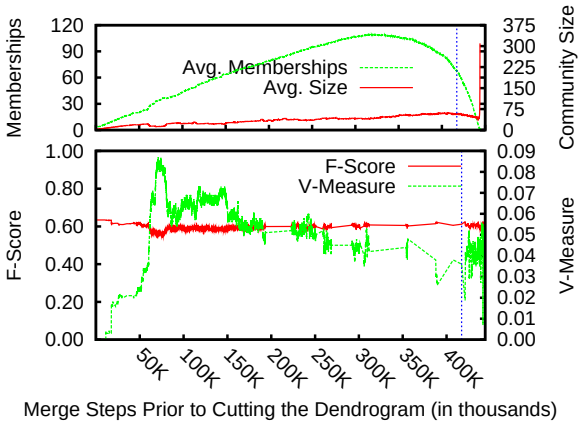

Figure 2: V-Measure and paired FScore results for different partitionings of the dendrogram. The dashed vertical line indicates S_{PD}

Results As a first measure of the quality of the induced senses, we evaluated both the solution that maximized the partition density, referred to as S_{PD}, and an additional 5,000 solutions, evenly distributed among the possible dendrogram partitionings. Figure 2 shows the score distribution for V-Measure and paired FScore. Table 1 lists the scores and rank for S_{PD} and the solutions that optimize the V-Measure, S_V, and FScore, S_F, among the 26 participating Task-14 systems. For comparison, we include the highest performing systems on each measure and the Most Frequent Sense (MFS) baseline.

Discussion Optimizing the partition density results in high performance only for the FScore; however, optimizing for the V-Measure yields competitive performance on both measures. The behavior is encouraging as most approaches submitted to Task 14 favor only one measure.

Figure 2 indicates a relationship between the V-Measure and community memberships. Therefore, using S_V, we calculated the Pearson correlation between a term's scores and the number of community memberships within a single solution. The correlation with the paired FScore, r = -0.167, was not statistically significant at $p < .05$, while correlation with the V-Measure, $r = 0.417$ is significant with $p <$ 1.6e-5. This suggests that at a specific community granularity, additional communities enable the WSI mapping process to make better sense distinctions between contexts. However, we note that V-Measure begins to drop as the average community membership increases in solutions after S_V, as shown in Figure 2. We suspect that as the agglomerative merge process continues, communities representing different senses become merged, leading to a loss of purity.

The lower performance of S_{PD} and the impact of community memberships raises the important question of how to best select the communities. While co-occurrence graphs have been shown to exhibit small-world network patterns (Véronis, 2004), optimizing for the general criterion of partition density that has performed well on such networks does not result in communities that map well to sense-specific contexts. We believe that this behavior is due to impact of the sense inventory; selecting a community solution purely based on the graph's structure may not capture the correct sense distinctions, either having communities with too few members to distinguish between senses or too many members, which conflates senses. However, a promising future direction is to examine whether the there exist features of the graph structure that would allow for recognizing the specific community solutions that correspond directly to different sense granularities without the need for an external evaluation metric.

4 Related Work

We highlight those related works with connections to community detection. Véronis (2004) demonstrated that word co-occurrence graphs follow a small-world network pattern. In his scheme, word senses are discovered by iteratively deleting the more connected portions of the subgraph to reveal the different senses' network structure. Our work capitalizes on this intuition of discovering sense-related subgraphs, but leverages formalized methods for community detection to identify them.

Dorow and Widdows (2003) identify sense-related subgraphs in a similar method to community detection for local region of the co-occurrence graph. They use a random walk approach to identify regions of the graph that are sense-specific. Though not identical, we note that the random walk model has been successfully applied to community detection (Rosvall et al., 2009). Furthermore, Dorow and Widdows (2003) performs graph clustering on a per-word basis; in contrast, the proposed approach identifies communities for the entire graph, effectively performing an all-word WSI.

Klapaftis and Manandhar (2010) capture hierarchical relations between collocations using a Hierarchical Random Graph model where nodes are collocations and edges indicate their co-occurrence, which improved performance over non-hierarchical models. Our community detection approach also captures the hierarchical structure of the collocation graph, but uses a much simpler graphical representation that for n terms requires $O(n)$ nodes and $O(n^2)$ edges, compared to $O(n^2)$ nodes and $O(n^3)$ edges for the above approach, which allows it to build the collocation graph from a larger set of terms.

5 Conclusion

We have proposed a new graph-based method for WSI based on finding sense-specific word communities within a co-occurrence graph, which are then identify distinguish senses in new contexts. An initial analysis using the SemEval-2010 WSI task demonstrates competitive performance. Future research will address two potential avenues: (1) the impact of word frequency on community size and memberships and (2) identifying both graph properties and semantic relations within hierarchical communities that distinguish between sense granularities. Software for the WSI model and for Link Clustering is available as a part of the S-Space Package (Jurgens and Stevens, 2010).

References

Yong-Yeol Ahn, James P. Bagrow, and Sune Lehmann. 2010. Link communities reveal multiscale complexity in networks. *Nature*, (466):761–764, August.

Javier Artiles, Enrique Amigó, and Julio Gonzalo. 2009. The role of named entities in web people search. In *Proceedings of EMNLP*, pages 534–542. ACL.

Beate Dorow and Dominic Widdows. 2003. Discovering corpus-specific word senses. In *Proceedings of the 10th EACL*, pages 79–82.

Santo Fortunato. 2010. Community detection in graphs. *Physics Reports*, 486(3-5):75–174.

David Jurgens and Keith Stevens. 2010. The S-Space Package: An Open Source Package for Word Space Models. In *Proceedings of the ACL 2010 System Demonstrations*.

Ioannis P. Klapaftis and Suresh Manandhar. 2008. Word sense induction using graphs of collocations. In *Proceeding of ECAI 2008*, pages 298–302.

Ioannis P. Klapaftis and Suresh Manandhar. 2010. Word sense induction & disambiguation using hierarchical random graphs. In *Proceedings of EMNLP*, pages 745–755. ACL.

Suresh Manandhar, Ioannis P. Klapaftis, Dmitriy Dligach, and Sameer S. Pradhan. 2010. SemEval-2010 task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68. Association for Computational Linguistics.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the Joint Conference of EMNLP-CoNLL*. ACL, June.

M. Rosvall, D. Axelsson, and C.T. Bergstrom. 2009. The map equation. *The European Physical Journal-Special Topics*, 178(1):13–23.

J. Véronis. 2004. HyperLex: lexical cartography for information retrieval. *Computer Speech & Language*, 18(3):223–252.