Parallel Clustering of English Verbs into Levin Classes

18.337/6.338 FINAL PROJECT PROGRESS REPORT

Mark Brigham Melanie Goetz Andrew Hogue
{brigham, melg, ahogue}@mit.edu

April 15, 2004

1 Introduction

We plan to implement an automated system for classifying verbs into Levin classes [4][1].

An example of a set of verbs and prepositions is shown in Figure 1. Each edge from a verb to a preposition represents the fact that the pair was found at least once in the sample corpus[2]. A possible clustering of the verbs based on the prepositions with which they appear is shown on the right.

2 Existing Work

2.1 Sentence Parsing

2.2 Bipartite Clustering

Many algorithms exist for clustering nodes in bipartite graphs. We will focus on methods that utilize the singular value decomposition (SVD) of the matrix representation of the graph, which allows us to utilize the parallel SVD algorithm developed by Jessup [3].

[1] This is a change from the original concept of our project proposal, which was to implement a parallel chart parser for assigning parts of speech to sentences. We believe that this project is significantly more interesting, while still dealing with the same field of natural language processing.

[2] These edges might also be assigned weights indicating the *number* of times the pair was encountered.

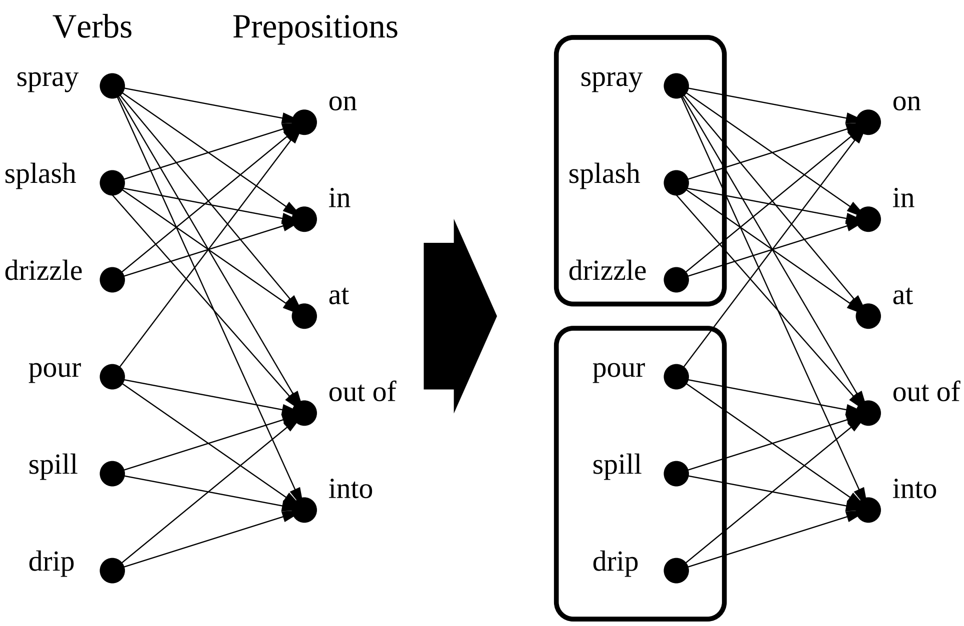

Figure 1: A simple example of a bipartite classification between verbs and prepositions.

Dhillon's algorithm [2] performs *co-clustering*, simultaneously clustering both documents and words by using the technique of spectral graph partitioning. This solution is equivalent to finding minimum cut vertex partitions on the graph. As this problem is NP-complete, a relaxation of it is solved using the second left and right singular vectors of the matrix representing the graph.

Other relevant bipartite clustering algorithms are those of Zha et. al. [5] and Chen et. al. [1].

While these algorithms have been published and tested, we were unable to locate any existing source code or libraries for the authors' implementations. From this we conclude that we will have to re-implement the algorithm we choose to integrate it with our verb classification system.

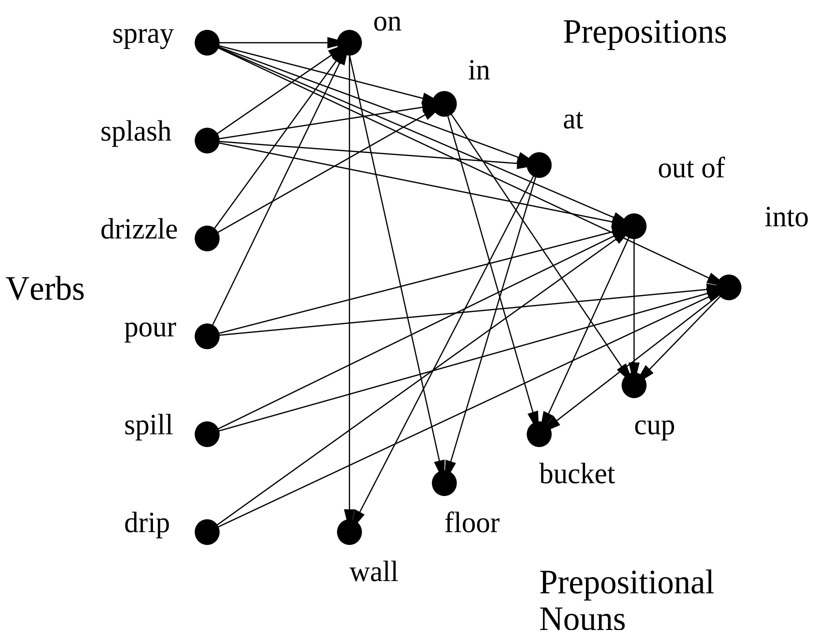

Figure 2: An example of a tripartite graph connecting verbs, prepositions, and prepositional nouns.

3 Expected Difficulties

4 Possible Extensions

One possible extension beyond bipartite clustering would be to include other features of the sentences, such as subject or object nouns or nouns taken by the prepositional phrases. This would split the bipartite graph into a tripartite graph (or more). An example of a tripartite graph for verbs, prepositions, and prepositional nouns is shown in Figure 2.

We were unable to locate any existing algorithms for clustering tripartite or n-partite graphs. There do, however, exist many algorithms for clustering graphs without partitioned structure. If we decide to extend in this direction, we will need to decide whether to utilize these existing algorithms or attempt to generalize the SVD bipartite clustering algorithms described above.

References

[1] Y. Chen, J. Wang, and R. Krovetz. Content-based image retrieval by clustering. In *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pages 193–200, 2003.

[2] I. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proc. of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*, pages 269–274, 2001.

[3] E. Jessup and D. Sorensen. A parallel algorithm for computing the singular value decomposition of a matrix. *SIAM Journal on Matrix Analysis and Applications*, 15(2):530–548, 1994.

[4] B. Levin. *English Verb Classes and Alternations.* University of Chicago Press, Chicago, 1993.

[5] H. Zha, X. He, C. Ding, H. Simon, and M. Gu. Bipartite graph partitioning and data clustering. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 25–32, 2001.