Clustering Paraphrases by their Vector Representations

Edidiong Okon, Faculty Advisor: Dr. Callison-Burch, Mentor: Anne Cocos

Introduction and Motivation

Polysemous words are words with multiple meanings. The translation of a paraphrase from one language to another is not complete without knowing the sense in which a polysemous noun is being used. For a computer scientist, the word *bug* likely means *glitch* or *error*. However, a physician hearing *bug* more likely makes associations with *virus* or *microbe*. This project presents a method for clustering paraphrases by their vector representations.

Vector Representation 1: AlexNet

- A convolutional neural network (CNN) that notably competed in the ImageNet Large Scale Visual Recognition Challenge in 2012
- Used to obtain vector representations of images scraped from the web

Figure 1: 96 convolutional kernels learned by AlexNet [2].

Vector Representation 2: Word2vec

- Continuous vector representations of words from large datasets
- Pre-trained vectors trained on part of the Google News dataset (about 100 billion words)
- Model contains 300-dimensional vectors for 3 million words and phrases

Paris – *France* + *Italy* = *Rome*
Paris – *France* + *Japan* = *Tokyo*
Copper – *Cu* + *zinc* = *Zn*
Copper – *Cu* + *gold* = *Au*
Microsoft – *Windows* + *Google* = *Android*
Microsoft – *Windows* + *IBM* = *Linux*

Figure 2: Word-pair relationships learned by the word2vec model [3].

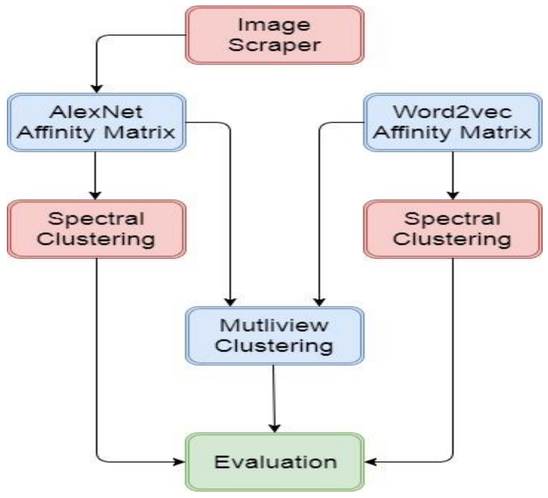
Spectral Clustering

- Performs a low dimensional embedding of the affinity matrix
- Then performs k-means in the lower dimensional space

Multiview Clustering

- Learns coefficient matrices from different affinity matrices, or views
- Regularizes the coefficient matrices towards a common consensus between the views
- The joint matrix factorization algorithm also takes into account inconsistencies between each view's coefficient matrix and the consensus

Pipeline



Sample Output Clusters

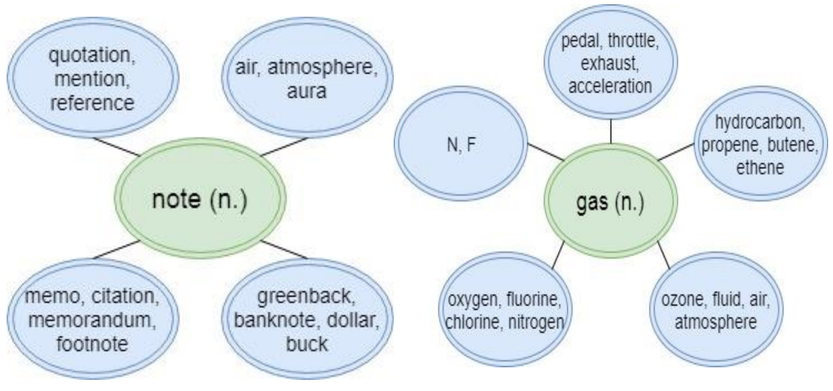
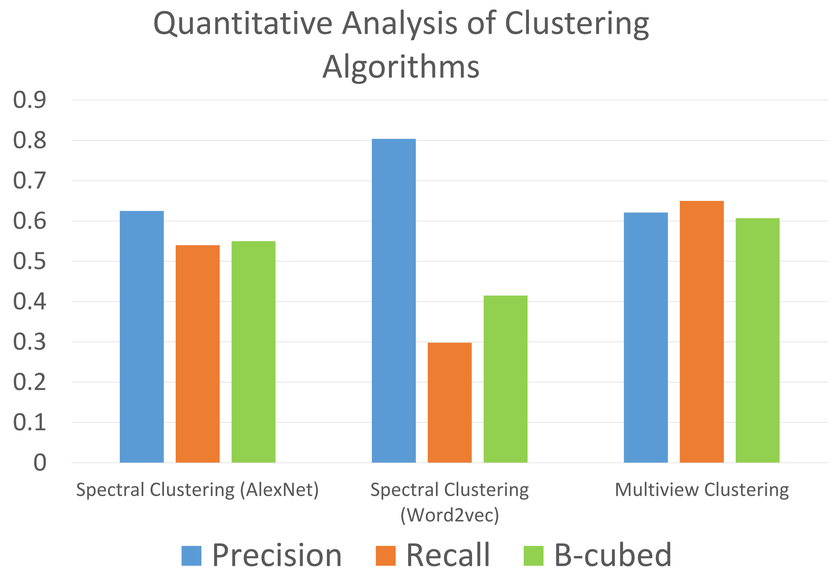


Figure 3: Sample multiview clustering output on the nouns *note* and *gas.*

Results



Graph 1: Experiments were run on a set of 110 polysemous nouns. Clusters were compared to those generated by WordNet [5]. While spectral clustering of the word2vec affinity matrix yields the highest precision, the multiview clustering produced the best recall and b-cubed score.

Future Work

- Run experiments that test different ways of finding the number of clusters
- Train a classifier than can learn weights for the different views for multiview clustering

References

[1] A. Cocos and C. Callison-Burch. Clustering paraphrases by word sense. In NAACL, 2016.
[2] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In NIPS, 2012.
[3] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. ICLR Workshop, 2013.
[4] J. Liu, C. Wang, J. Gao, and J. Han, "Multi-view clustering via joint nonnegative matrix factorization," in *Proc. SIAM Int. Conf. Data Min.*, 2013, pp. 252–260.
[5] G. A. Miller, "WordNet: a lexical database for English," vol. 38, p. 39-41, 1995.