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](image)

### 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](image)

### 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

![Diagram of pipeline](image)

### Results

<table>
<thead>
<tr>
<th>Quantitative Analysis of Clustering Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectral Clustering (AlexNet)</td>
</tr>
<tr>
<td>Precision</td>
</tr>
<tr>
<td>0.1</td>
</tr>
<tr>
<td>0.4</td>
</tr>
<tr>
<td>0.7</td>
</tr>
</tbody>
</table>

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

### Sample Output Clusters

*Note (n.)*

<table>
<thead>
<tr>
<th>quotation, mention, reference</th>
<th>air, atmosphere, aura</th>
<th>petal, throttle, exhaust, acceleration</th>
</tr>
</thead>
<tbody>
<tr>
<td>memo, citation, memorandum, footnote</td>
<td>greenback, banknote, dollar, buck</td>
<td>hydrocarbon, propene, butene, ethene</td>
</tr>
<tr>
<td>note (n.)</td>
<td>gas (n.)</td>
<td>ozone, fuel, air, atmosphere</td>
</tr>
</tbody>
</table>

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