Supplemental – Semantic Word Clusters Using Signed Spectral Clustering

1 Gold Standard

One large issue that we had was choosing an appropriate gold standard for cluster evaluation. From our evaluation SimLex-999 is the dataset having the largest and most exact task of estimating semantic similarity between words and avoiding relatedness. Table 1 shows the difference between SimLex-999 and WordSim-353.

2 Other Word Embeddings

For GloVe we used pretrained 200 dimensional vector embeddings[1] trained using Wikipedia 2014 + Gigaword 5 (6B tokens). Eigenwords were trained on English Gigaword with no lowercasing or cleaning. Finally, we used 50 dimensional vector representations from Huang et al. (2012), which used the April 2010 snapshot of the Wikipedia corpus (Lin, 1998; Shaoul, 2010), with a total of about 2 million articles and 990 million tokens.

In table 2 we show a qualitative comparison between multiple word embedding. We show that many word embeddings contain antonyms, and also that thesauri include rare words and rare senses. It should be noted that signed clustering can easily be applied to word sense aware embeddings and thesauri.

3 Further Cluster Evaluation

Next we evaluated our clusters using an external gold standard. Cluster purity and entropy (Zhao and Karypis, 2001) is defined as,

$$Purity = \sum_{r=1}^{k} \frac{1}{n} max_i(n_r^i)$$

$$Entropy = \sum_{r=1}^{k} \frac{n_r}{n} \left(-\frac{1}{\log q} \sum_{i=1}^{q} \frac{n_r^i}{n_r} \log \frac{n_r^i}{n_r} \right)$$

where q is the number of classes, k the number of clusters, n_r is the size of cluster r, and n_r^i number of data points in class i clustered in cluster r. The purity and entropy measures improve (increased purity, decreased entropy) monotonically with the number of clusters.

The number of disconnected components (NDC) in the cluster where we only use synonym edges.

$$NDC = \sum_{r=1}^{k} \sum_{i=1}^{C} (n_r^i)$$

4 Hyperparameter Optimization

We show the optimization results using 10-fold cross validation on our 5108 word dataset. The optimal hyperparameters are chosen by minimizing error, as seen in table 3.

Table 3 shows out of sample results from the grid search of hyperparameter optimization. Here we show that Eigenword + MSW outperforms Eigenword + Roget, which is in contrast with the other word embeddings where the combination with Roget performs better. Another interesting result from the hyperparameter optimization is that Word2Vec with Roget has two very different optima.

5 Explanded Results

When compared with the MS Word thesaurus, Word2Vec, Eigenword, GloCon, and GloVe word

[1] http://nlp.stanford.edu/projects/GloVe/

Pair	Simlex-999 rating	WordSim-353 rating
coast - shore	9.00	9.10
clothes - closet	1.96	8.00

Table 1: Comparison between SimLex-999 and WordSim-353. This is from `http://www.cl.cam.ac.uk/~fh295/simlex.html`

.

Ref word	Roget	WordNet	MS Word	W2V	GloDoc	EW	Glove
accept	adopt accept your fate be fooled by acquiesce	agree get fancy hold	take swallow consent assume	accepts reject agree accepting	seek consider know ask	approve declare endorse reconsider	agree reject willin refuse
negative	not advantageous pejorative pessimistic no	unfavorable denial resisting pessimistic	severe hard wasteful charged	positive adverse Negative negatively	reverse obvious calculation cumulative	unfavorable positive dire worrisome	positive impact suggesting result
unlike	**no synonyms**	incongruous unequal separate hostile	different dissimilar	Unlike Like even But	whereas true though bit	Unlike Like Whereas whereas	instance though whereas likewise
absurd	discord dissension nonsense	appalling awful cruel insane irrational terrible	bizarre mysterious odd rare strange unusual	OOV	crazy foolish funny irrational silly loony rich	bizzarre irrational mad silly strange	foolish insane mad

Table 2: Qualitative comparison of clusters.

Method	σ	thresh	# Clusters	Error $\downarrow$ $\frac{(NNE+NDC)}{\|V\|}$	Purity $\uparrow$	Entropy $\downarrow$
Word2Vec	0.2	0.04	750	0.716	0.88	0.14
Word2Vec + Roget	10.0	0	750	0.033	0.95	0.07
Word2Vec + Roget	0.7	0.04	750	0.033	0.94	0.09
Eigenword	2.0	0.07	200	0.655	0.84	0.25
Eigenword + MSW	1.0	0.08	200	0.042	0.95	0.01
GloCon	3.0	0.09	100	0.691	0.98	0.03
GloCon + Roget	0.9	0.06	750	0.048	0.94	0.02
Glove	9.0	0.09	200	0.657	0.72	0.33
Glove + Roget	11.0	0.01	1000	0.070	0.91	0.10

Table 3: Clustering evaluation after parameter optimization minimizing error using grid search.

embeddings had a total of 286, 235, 235, 220 negative edges, respectively. The results are similar with the other thesauri.

If we examined the number of disconnected components within the different word clusters, we observed that when K-means were used, the number of disconnected components were statistically significant from random labelling. This suggests that the word embeddings capture synonym relationships. By optimizing the hyperparameters we found roughly a 10 percent decrease in disconnected components using normalized cuts. When we added the signed antonym relationships using our signed clustering algorithm, on average we found a 39 percent decrease over the K-means clusters.

Model	Accuracy
NB (Socher et al., 2013)	0.818
VecAvg (W2V, GV, GC) (Faruqui et al., 2015)	0.812, 0.796, 0.678
RVecAvg (W2V, GV, GC) (Faruqui et al., 2015)	0.821, 0.822, 0.689
RNN, RNTN (Socher et al., 2013)	0.824, 0.854
CNN (Le and Zuidema, 2015)	0.881
LSTM-RNN GloVe (Le and Zuidema, 2015)	0.88
SC W2V	0.836
SC GV	0.819
SC GC	0.572
SC EW	0.820

Table 4: Sentiment analysis accuracy for binary predictions of signed clustering algorithm (SC) versus other models.

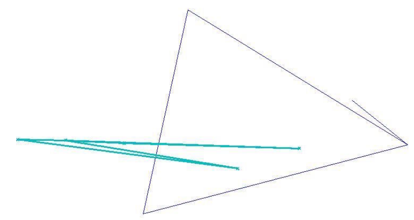

Figure 1.1: Cluster with two disconnected components. All edges represent synonymy relations. The edge colors are only meant to highlight the different components.

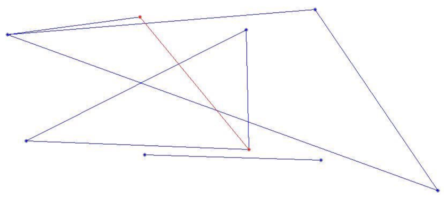

Figure 1.2: Cluster with one antonym relation. The red edge represents the antonym relation. Blue edges represent synonymy relations.

Figure 1: Disconnected component and number of antonym evaluations.

References

Manaal Faruqui, Jesse Dodge, Kumar Sujay Jauhar, Chris Dyer, Eduard Hovy, and A. Noah Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 1606–1615. https://doi.org/10.3115/v1/N15-1184.

Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 873–882. http://aclweb.org/anthology/P12-1092.

Phong Le and Willem Zuidema. 2015. Compositional distributional semantics with long short term memory. *arXiv preprint arXiv:1503.02510* .

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*. http://aclweb.org/anthology/P98-2127.

Cyrus Shaoul. 2010. The westbury lab wikipedia corpus. *Edmonton, AB: University of Alberta* .

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, D. Christopher Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1631–1642. http://aclweb.org/anthology/D13-1170.

Ying Zhao and George Karypis. 2001. Criterion functions for document clustering: Experiments and analysis. Technical report, Citeseer.