Intrinsic and Extrinsic Evaluations of Word Embeddings

Michael Zhai, Johnny Tan, Jinho D. Choi
Department of Mathematics and Computer Science
Emory University
Atlanta, GA 30322
{michael.zhai,johnny.tan,jinho.choi}@emory.edu

Abstract

In this paper, we first analyze the semantic composition of word embeddings by cross-referencing their clusters with the manual lexical database, WordNet. We then evaluate a variety of word embedding approaches by comparing their contributions to two NLP tasks. Our experiments show that the word embedding clusters give high correlations to the synonym and hyponym sets in WordNet, and give 0.88% and 0.17% absolute improvements in accuracy to named entity recognition and part-of-speech tagging, respectively.

Introduction

Distributional semantics, the field of finding semantic similarities between entities using large data, has recently gained lots of interest. Word clusters induced from distributional semantics have shown to be helpful for handling unseen words in several NLP tasks (Turian, Ratinov, and Bengio 2010). Furthermore, recent advances in embedding approaches have produced superior word representations for word similarity and analogy tasks (Mikolov et al. 2013). In this paper, we analyze the semantic composition of word embeddings by comparing their clusters to the manual lexical database, WordNet (Fellbaum 1998), and give extrinsic evaluations of different word embedding approaches through two NLP tasks, named entity recognition and part-of-speech tagging.

Intrinsic Evaluation

Word embeddings are continuous-valued vectors representing word semantics. In our experiments, they are generated by using bag-of-words (CBOW), skip-gram with negative sampling (SGNS), and GloVe (Pennington, Socher, and Manning 2014), and clustered by the *k*-means, *g*-means, hierarchical *g*-means, and agglomerative clustering algorithms using cosine similarity. Brown clusters are induced directly from the text.

For intrinsic evaluation, WordNet is used as the reference for our semantic analysis of word embedding. From WordNet, sets of synonyms and hyponyms of the 100 most frequent nouns and verbs in the New York Times corpus[1] are extracted and compared to the clusters generated from the word embeddings.

[1]https://catalog.ldc.upenn.edu/LDC2008T19

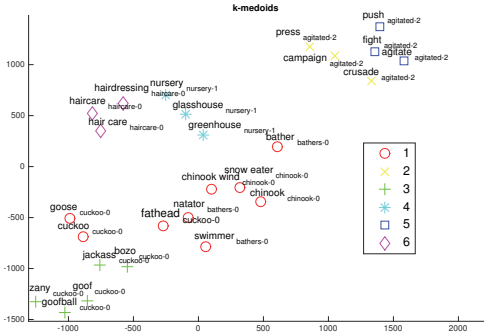

Figure 1: The t-SNE projection of word embeddings with respect to the synonym sets in WordNet.

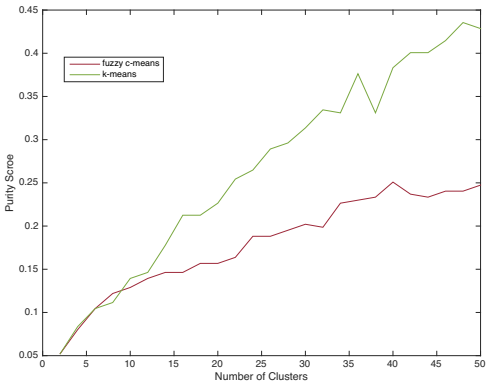

Figure 2: The purity scores achieved by *k*-means and *c*-means clustering with respect to the number of clusters.

Figure 1 shows that our *k*-means clustering (colored shapes) display high degree of agreement with the WordNet synonym sets (subscripts).[2]

Figure 2 shows that hard-bound clustering such as *k*-means achieves much higher purity scores than fuzzy-bound clustering such as *c*-means.

[2]The other clusters show similar results as Figure 1.

Embedding	Cluster	F1-score
Baseline	-	85.31
-	Brown	86.15
SGNS	agglomerative	**86.19**
SGNS	*k*-means	85.72
SGNS	g-means	85.83
SGNS	g-means hier	85.68
SGNS (w+c)	agglomerative	86.14
SGNS (w+c)	*k*-means	85.65
SGNS (w+c)	g-means	85.70
SGNS (w+c)	g-means hier	85.71
CBOW	agglomerative	85.98
CBOW	*k*-means	85.81
CBOW	g-means	85.67
CBOW	g-means hier	85.70
GloVe	agglomerative	86.08
GloVe	*k*-means	85.72
GloVe	g-means	85.71
GloVe	g-means hier	86.10

Table 1: Named entity recognition results on the test set.

Embedding	Cluster	Accuracy
Baseline	-	97.34
-	Brown	**97.51**
SGNS	agglomerative	97.43
SGNS (w+c)	agglomerative	97.39
CBOW	agglomerative	97.42
GloVe	g-means	97.40

Table 2: Part-of-speech tagging results on the test set; only the best result is displayed for each approach.

Extrinsic Evaluation

For extrinsic evaluation, we use the word embedding clusters as features for two NLP tasks, named entity recognition and part-of-speech tagging. The English portion of OntoNotes 5 is used for experiments following the standard split suggested by Pradhan et al. (2013). AdaGrad is used for training statistical models. As recommended by Levy, Goldberg, and Dagan (2015), additional experiments are conducted by concatenating the word and contextual vectors (w+c).

For the NER experiments, the highest F1-score of 86.19 is achieved by the skip-gram with negative sampling embeddings (SGNS) using the agglomerative clustering. On the other hand, the highest accuracy of 97.51 is achieved by Brown clustering (using raw text instead of embeddings). These results outperform the previous work (Pradhan et al. 2013), showing the absolute improvements of 3.77% and 0.42% for the NER and POS tasks, respectively.

All of the above experiments are using the maximum cluster size of 1,500. We also tested on the max cluster size of 15,000, which showed very similar results. This implies that the increase of cluster size does not improve the quality of the clusters, at least for these two tasks. For the NER task, SGNS and Brown give constant additive increase in performance regardless of the size of the training data.

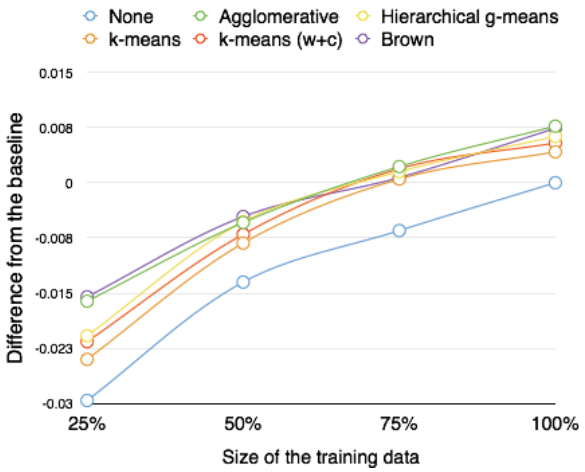

Figure 3: The F1-scores for named entity recognition with respect to different sizes of the training data using SGNS grouped by all clustering algorithms.

Conclusion

Word embeddings have shown to be useful for several NLP tasks. In this paper, we first analyze the nature of the vector spaces created by different word embedding approaches and compare their clusters to the ontologies in WordNet. From our experiments, we found that the embedding clusters show high correlations with the synonyms and hyponyms in WordNet although the correlation level decreases as the cluster size increases.

We also show the impact of different word embedding approaches couple with several clustering algorithms on two NLP tasks, named entity recognition and part-of-speech tagging. Our experiments show that hierarchical clustering algorithms such as Brown or agglomerative are more suitable for finding clustering features than partition-based clustering algorithms such as *k*-means and *g*-means for these tasks.[3]

References

Fellbaum, C., ed. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Levy, O.; Goldberg, Y.; and Dagan, I. 2015. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *TACL* 3:211–225.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*, 3111–3119.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

Pradhan, S.; Moschitti, A.; Xue, N.; Ng, H. T.; Björkelund, A.; Uryupina, O.; Zhang, Y.; and Zhong, Z. 2013. Towards robust linguistic analysis using ontonotes. In *CoNLL*, 143–152.

Turian, J.; Ratinov, L.; and Bengio, Y. 2010. Word Representations: A Simple and General Method for Semi-supervised Learning. In *ACL*, 384–394.

[3] All resources are available at http://github.com/emorynlp.