Evaluating distributed word representations for capturing semantics of biomedical concepts

Muneeb T H[1], **Sunil Kumar Sahu**[1] and **Ashish Anand**[1]

[1]Department of Computer Science and Engineering,
Indian Institute of Technology Guwahati, Assam - 781039, India
{muneeb, sunil.sahu, anand.ashish}@iitg.ac.in

Abstract

Recently there is a surge in interest in learning vector representations of words using huge corpus in unsupervised manner. Such word vector representations, also known as word embedding, have been shown to improve the performance of machine learning models in several NLP tasks. However efficiency of such representation has not been systematically evaluated in biomedical domain. In this work our aim is to compare the performance of two state-of-the-art word embedding methods, namely *word2vec* and *GloVe* on a basic task of reflecting semantic similarity and relatedness of biomedical concepts. For this, vector representations of all unique words in the corpus of more than 1 million full-length research articles in biomedical domain are obtained from the two methods. These word vectors are evaluated for their ability to reflect semantic similarity and semantic relatedness of word-pairs in a *benchmark data set* of manually curated semantic similar and related words available at `http://rxinformatics.umn.edu`. We observe that parameters of these models do affect their ability to capture lexico-semantic properties and word2vec with particular language modeling seems to perform better than others.

1 Introduction

One of the crucial step in machine learning (ML) based NLP models is how we represent word as an input to our model. Most of earlier works were treating word as atomic symbol and were assigning one hot vector to each word. Length of the vector in this representation was equal to the size of the vocabulary and the element at the word index is 1 while the other elements are 0s. Two major drawbacks with this representation are: first, length of the vector is huge and the second, there is no notion of similarity between words. The inability of one-hot vector representation to embody lexico-semantic properties prompted researchers to develop methods which are based on the notion that the "similar words appear in similar contexts". These methods can broadly be classified into two categories (Turian et al., 2010), namely, *distributional representation* and *distributed representation*. Both group of methods works in unsupervised manner with huge corpus. Distributional representations are mainly based on co-occurrence matrix O of words in the vocabulary and their contexts. Here, among other possibilities, contexts can be documents or words within a particular window. Each entry O_{ij} in the matrix may indicate either frequency of word i in the context j or simply whether the word i has appeared in the context j at least once. Co-occurrence matrix can be designed in variety of ways (Turney and Pantel, 2010). The major issue with such methods is size of the matrix O and reducing its size generally tends to be computationally very expensive. Nevertheless, the requirement of constructing and storing the matrix O are always there. The second group of methods are mainly based on language modeling (Bengio et al., 2003). We discuss more about these methods in the section 3.

Outside the biomedical domain, this kind of representation has shown significant improvement in the performance of many NLP tasks. For example, Turian et al. (2010) have improved the performance of chunking and named entity recognition by using word embedding also as one of the features in their CRF model. In one study, Collobert et al. (2011) have formulated the NLP tasks of parts of speech tagging, chunking, named entity recognition and semantic role labeling as multi-

Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015), pages 158–163,
Beijing, China, July 30, 2015.

task learning problem. They have shown improvement in the performance when word vectors are learned together with other NLP tasks. Socher et al. (2012) improved the performance of sentiment analysis task and semantic relation classification task using recursive neural network. One common step among these models is: learning of word embedding from huge unannotated corpus like Wikipedia, and later use them as features.

Motivated by the above results, we evaluate performance of the two word embedding models, word2vec (Mikolov et al., 2013a; Mikolov et al., 2013b) and GloVe (Pennington et al., 2014) for their ability to capture syntactic as well as semantic properties of words in biomedical domain. We have used full-length articles obtained from PubMed Central (PMC) open access subset[1] as our corpus for learning word embedding. For evaluation we have used publicly available validated reference dataset (Pakhomov et al., 2010; Pedersen et al., 2007) containing semantic similarity and relatedness scores of around 500 word-pairs. Our results indicate that the word2vec word embedding is capturing semantic similarity between words better than the GloVe word embedding in the biomedical domain, whereas for the task of semantic relatedness, there does not seem to be any statistical significant difference among different word-embeddings.

2 Related Work

In a recent study, Miñarro-Giménez et al. (2015) have evaluated the efficiency of word2vec in finding clinical relationships such as "may treat", "has physiological effect" etc. For this, they have selected the manually curated information from the National Drug File - Reference Terminology (NDF-RT) ontology as reference data. They have used several corpora for learning word-vector representation and compared these different vectors. The word-vectors obtained from the largest corpus gave the best result for finding the "may treat" relationship with accuracy of 38.78%. The relatively poor result obtained for finding different clinical relationships indicates the need for more careful construction of corpus, design of experiment and finding better ways to include domain knowledge.

In another recent study, Nikfarjam et al. (2015) have described an automatic way to find adverse drug reaction mention in social media such as twitter. Authors have shown that including word embedding based features has improved the performance of their classifier.

Faruqui and Dyer (2014) have developed an online suit to analyze and compare different word vector representation models on a variety of tasks. These tasks include syntactic and semantic relations, sentence completion and sentiment analysis. In another recent work, Levy et al. (2015) have done extensive study on the effect of hyperparameters of word representation models and have shown their influence on the performance on word similarity and analogy tasks. However in both the studies (Faruqui and Dyer, 2014; Levy et al., 2015) the benchmark datasets available for NLP tasks are not suitable for analyzing vector representations of clinical and biomedical terms.

3 Word Embedding

As discussed earlier, word embedding or distributed representation is a technique of learning vector representation for all words present in the given corpus. The learned vector representation is generally dense, real-valued and of low-dimension. As contrast to one-hot vector representation each dimension of the word-vector is supposed to represent a latent feature of lexico-semantic properties of the word. In our work we considered two state of the art word embedding techniques, namely, *word2vec* and *GloVe*. Although in literature there exists several word-embedding techniques (Hinton et al., 1986; Bengio et al., 2003; Bengio, 2008; Mnih and Hinton, 2009; Collobert et al., 2011), the selected two word embedding techniques are very much computationally efficient and are considered as state-of-the art. We have summarized the basic principles of the two methods in subsequent sections.

3.1 word2vec Model

word2vec generates word vector by two different schemes of language modeling: continuous bag of words (CBOW) and skip-gram (Mikolov et al., 2013a; Mikolov et al., 2013b). In the CBOW method, the goal is to predict a word given the surrounding words, whereas in skip-gram, given a single word, window or context of words are predicted. We can say skip-gram model is opposite of CBOW model. Both models are neural network based language model and take huge corpus as an input and learn vector representation for

[1] http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/

each words in the corpus. We used freely available *word2vec*[2] tool for our purpose. Apart from the choice of architecture skip-gram or CBOW, word2vec has several parameters including *size of context window*, *dimension of vector*, which effect the speed and quality of training.

3.2 GloVe Model

GloVe (Pennington et al., 2014) stands for Global Vectors. In some sense, GloVe can be seen as a hybrid approach, where it considers global context (by considering co-occurrence matrix) as well as local context (such as skip-gram model) of words. *GloVe* try to learn vector for words w_x and w_y such that their dot product is proportional to their co-occurrence count. We used freely available *glove*[3] tool for all analysis.

4 Materials and Methods

4.1 Corpus Data and Preprocessing

PubMed Central® (PMC) is a repository of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM). We have downloaded the gzipped archived files of full length texts of all articles in the open access subset [4] on 19^{th} April, 2015. This corpus contains around 1.25 million articles having around 400 million tokens altogether.

In pre-processing step of the corpus, we mainly perform following two operations-

- we put all numbers in different groups based on number of digits in them. For example, all single digit numbers are replaced by the token "number1", all double digit numbers by the token "number2" and so on.
- each punctuation mark is considered as separate token.

4.2 Reference Dataset

Pakhomov et al. (2010) have constructed a reference dataset of semantically similar and related word-pairs. These words are clinical and biomedical terms obtained from control vocabularies maintained in the Unified Medical Language System(UMLS). This reference dataset contains 566 pairs of UMLS concepts which were manually rated for their semantic similarity and 587 pairs of UMLS concepts for semantic relatedness. We removed all pairs in which at least one word has less than 10 occurrences in the entire corpus as such words are removed while building vocabulary from the corpus. After removing less frequent words in both reference sets, we obtain 462 pairs for semantic similarity having 278 unique words, and 465 pairs for semantic relatedness having 285 unique words. In both cases, each concept pair is given a score in the range of $0 - 1600$, with higher score implies similar or more related judgments of manual annotators. The semantic relatedness score span the four relatedness categories: completely unrelated, somewhat unrelated, somewhat related, closely related.

4.3 Experiment Setup

We generate the word vectors using the two word embedding techniques under different settings of their parameters and compare their performance in semantic similarity and relatedness tasks. Dimension of word-vector is varied under the two different language models, CBOW and SKIP-GRAM, for word2vec word embedding. For GloVe, only dimension of word vector is changed. For each model, word vectors of 25, 50, 100, and 200 dimensions are generated. Due to limited computing power, we could not go for higher dimensions. For window size, we did not perform any experiment and simply considered 9 as window size for all models.

4.4 Evaluation

As discussed earlier, both reference data have provided a score for each word-pair in them. We calculate cosine similarity between the two words of each pair present in the reference data using learned word vectors. Now, each word pair has two scores: one given in the dataset and the other cosine similarity based on learned word vectors. We calculate Pearson's correlation between these two scores.

Further we visualize a limited number of manually selected words for qualitative evaluation. For this we use the t-SNE (van der Maaten and Hinton, 2008) tool to project our high dimensional word vectors into two-dimensional subspace. t-SNE is being widely used for this purpose.

[2]https://code.google.com/p/word2vec/
[3]http://nlp.stanford.edu/projects/glove/
[4]http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/

Dimension	Semantic Similarity			Semantic Relatedness		
	CBOW	**Skip**	**GloVe**	**CBOW**	**Skip**	**GloVe**
25	0.32	0.39	0.28	0.30	0.34	0.27
50	0.36	0.44	0.34	0.33	0.38	0.36
100	0.42	0.48	0.41	0.39	0.43	0.41
200	0.46	0.52	0.42	0.41	0.45	0.42

Table 1: Correlation between cosine similarity and the score provided in the benchmark dataset.

5 Results and Discussion

Table 1 shows the correlation values in all cases. We observe that increasing the dimension of word vectors improve their ability to capture semantic properties of words. The above results indicate that less than $d = 200$ dimension will likely to be a bad choice for any NLP tasks. Due to the limited computing power, we could not complete our experiments with 500 and 1000 dimensional vector representations. We have also calculated the Spearman and Kendall-Tau's correlation in each case and have observed similar trends in all cases.

Skip-gram model seems to be better than both CBOW and GloVe models in the semantic similarity task for all dimensions. However this does not seem to be the case with the relatedness task. So we perform the statistical significance test to check whether correlation corresponding to word2vec skip-gram model is significantly higher than correlation corresponding to other two models. In the statistical test, we evaluate the null-hypothesis " correlation corresponding to alternate model (CBOW or GloVe) is equal to that corresponding to the skip-gram model" at significance level $\alpha = 0.05$. We use cocor (Diedenhofen and Musch, 2015) package for statistical comparison of dependent correlations.

It turns out that for the semantic similarity task, word2vec skip-gram model is significantly better (i.e., correlation is higher corresponding to skip-gram word vectors) than word2vec-CBOW (p-value: 0.01) and GloVe (p-value: 0.0007) models. On the other hand correlation in skip-gram model is not found significantly higher than the correlations in the other two models for the semantic relatedness task. The above observation is made for the 200 dimensional vectors. But we can not say the same for results obtained by lower dimensional vectors. For example, in case of 25-dimensional vectors, correlation obtained by skip-gram model is significantly higher than that obtained by GloVe model for both tasks. However similar observation was made in case of comparison between CBOW and skip-gram as in 200 dimensional case.

We further look at nearest neighbors of some manually selected words. If word-vectors truly represent latent features of lexical-semantic properties of words, then their nearest neighbors must be related words. We tested this hypothesis on a small set of manually selected seed-words and their nearest neighbors. We selected 8 seed-words representing *disease*, *disorder*, *organ* and *treatment*: *eye (organ)*, *liver (internal organ)*, *fever (disorder/symptom)*, *tumour (disease/disorder)*, *thyroid (gland)*, *cough (symptom)*, *surgery (procedure/treatment)*, *leg (external organ)*, *aids (disease)*. Table 2 shows the 10 nearest neighbors of some of the seed-words (similar results are observed for other seed-words) as picked by the three methods. As it can be seen from the table that the nearest neighbors are very much related to the seed-words. Not only words like "coughs", "coughing", but also words like "wheezing", "dyspnea" are within the top-10 nearest neighbors of "cough". The first set of examples indicates ability of the learned word-vectors to capture lexical properties of words, whereas the later set of words shows vectors' ability to capture semantic properties as well.

Next we visualize (Figure 1) the 4 seed-words (shown in Table 2) and their 25 nearest neighbors using t-SNE. Here we have shown the result obtained from the word2vec skip gram model (dimension = 200) only. Due to space constraints we have not shown the results of other methods but similar observation was made for the other methods. t-SNE projects high-dimensional vectors into $\mathbf{R}^2$ by preserving the local structure of high-dimensional space.

Figure 1 clearly shows the ability of the learned word-vectors to automatically group similar words together. This again provides another evidence of the vectors' ability to represent semantic properties.

seed word	CBOW	Skip	GloVe
eye	eye, eyes, eyeball, hemifield, hemibody, forelimb, eyebrow, midline, head, face	eye, eyes, face, head, ocular, mouth, pupillary, fovea, angle, Eye	eye, eyes, SEFsupplementary, ocular, visual, vision, cornea, optic, retina, ear
cough	cough, coughing, breathlessness, Cough, dyspnea, wheezing, wheeze, hemoptysis, coughs, haemoptysis	cough, breathlessness, expectoration, coughing, wheezing, dyspnea, phlegm, shortness, haemoptysis, sore	cough, coughing, shortness, breathlessness, TDITransition, dyspnea, wheezing, sore, bronchitis, expectoration
surgery	surgery, operation, decompression, dissection, resection, parathyroidectomy, stenting, surgeries, esophagectomy, resections	surgery, surgical, operation, procedure, esophagectomy, surgeries, laparoscopic, elective, reintervention, postoperative	surgery, surgical, BCSBreast-conserving, surgeries, operative, eBack, postoperative, PSMPositive, operation, resection
tumour	tumour, tumor, tumoral, tumoural, glioma, melanoma, PDAC, HNSCC, tumors, neoplastic	tumour, tumor, tumors, tumours, malignant, metastatic, metastasis, metastases, tumoral, melanoma	tumour, tumor, Tprimary, tumors, VHLVon-Hippel-Lindau, tumours, metastatic, metastasis, malignant, EHSEngelbreth-Holm-Swarm

Table 2: 10 Nearest neighbors of selected seed-words.

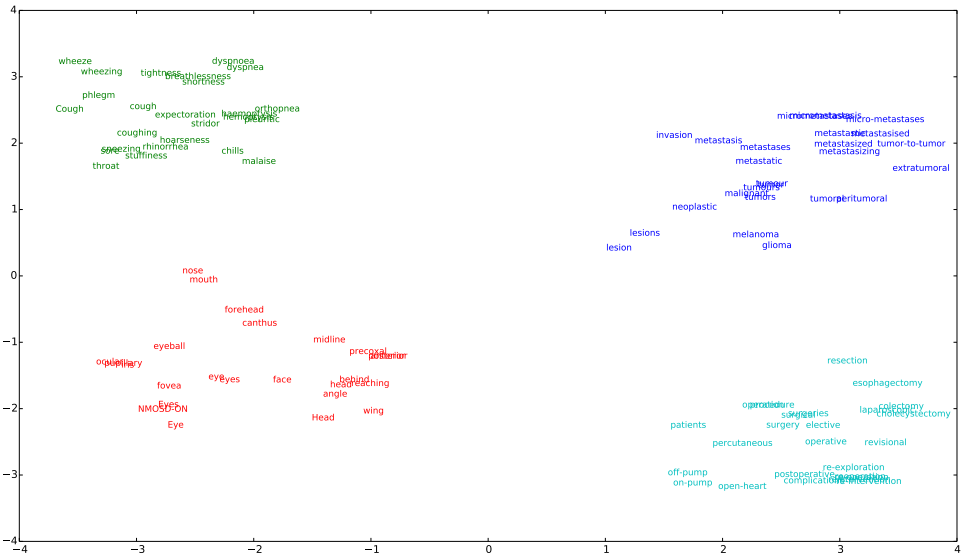

Figure 1: t-SNE projection of 100 biomedical words after applying word2vec skip-gram model. These words are nearest neighbors of the 4 seed-words 'eye', 'cough', 'surgery', and 'tumour'. All nearest neighbors of a particular seed-word are in closer proximity of each other than the nearest-neighbors of other seed-words.

6 Conclusion and Future Work

In this study, we have shown that while *word2vec* with skip-gram model gave the best performance compared to other models in the semantic similarity task, none of the model significantly outperformed others in the semantic relatedness task. Our results indicate that word-vectors should be at least of dimension 200, irrespective of the embedding model. However, further systematic evaluation of all models on more complex NLP tasks, such as medical concept and relation extraction, is required to find out which model will work best.

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March.

Yoshua Bengio. 2008. Neural net language models. *Scholarpedia*, 3(1):3881.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, November.

Birk Diedenhofen and Jochen Musch. 2015. cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE*, 10(4):e0121945, 04.

Manaal Faruqui and Chris Dyer. 2014. Community evaluation and exchange of word vectors at wordvectors.org. In *Proceedings of ACL: System Demonstrations*.

Geoffrey E Hinton, James L McClelland, and David E Rumelhart. 1986. Distributed representations. In *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*, pages 77–109. MIT Press.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

José Antonio Miñarro-Giménez, Oscar Marín-Alonso, and Matthias Samwald. 2015. Applying deep learning techniques on medical corpora from the world wide web: a prototypical system and evaluation. *CoRR*, abs/1502.03682.

Andriy Mnih and Geoffrey E Hinton. 2009. A scalable hierarchical distributed language model. In *Advances in neural information processing systems*, pages 1081–1088.

Azadeh Nikfarjam, Abeed Sarker, Karen OConnor, Rachel Ginn, and Graciela Gonzalez. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, page ocu041.

Serguei Pakhomov, Bridget McInnes, Terrence Adam, Ying Liu, Ted Pedersen, and Genevieve B Melton. 2010. Semantic similarity and relatedness between clinical terms: an experimental study. *AMIA annual symposium proceedings*, 2010:572.

Ted Pedersen, Serguei V.S. Pakhomov, Siddharth Patwardhan, and Christopher G. Chute. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288 – 299.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic Compositionality Through Recursive Matrix-Vector Spaces. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 384–394, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188, January.

L.J.P. van der Maaten and G.E. Hinton. 2008. Visualizing high-dimensional data using t-sne.