Evaluating distributed word representation for capturing semantics of biomedical concepts

Muneeb TH[1], **Sunil Kumar Sahu**[1], **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.

1. Introduction

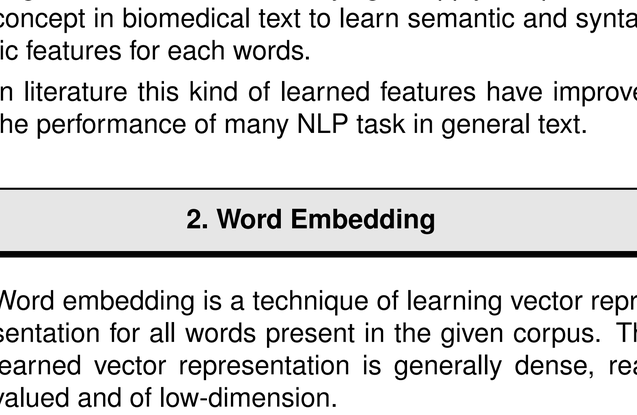
- Performance of any Machine Learning based system is highly dependent on the features of the model therefore, most of Machine Learning based NLP models uses very carefully hand-designed features and representation.
- Handcrafted features required domain knowledge, its time consuming, brittle and incomplete.
- In case of biomedical domain it will be difficult for none practitioners to design these features.
- Deep learning is one way of learning features. Here we can learn the features without any prior domain knowledge. In our work we are trying to apply deep learning concept in biomedical text to learn semantic and syntactic features for each words.
- In literature this kind of learned features have improved the performance of many NLP task in general text.

2. Word Embedding

- Word embedding 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.
- 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, the selected two word embedding techniques are very much computationally efficient and are considered as state-of-the art.

2.1 word2vec

- *word2vec* generates word vector by two different schemes of language modeling: continuous bag of words (CBOW) and skip-gram [1, 2]
- We used freely available *word2vec* 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.



Skipgram and CBOW model taken from [2]

2.2 GloVe

- *GloVe* [3] 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* tool for all analysis.

3. Corpus Data

- We downloaded 1.25 millions articles from PubMed Centeral® PMC.
- This corpus contains around 400 million tokens altogether.

4. Preprocessing

- 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.

5. Reference Datasets

- [4] 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).
- The table 1 is the sample of referenced dataset for semantic relatedness task. Same way they curated for semantic similarity task also.

Term1	Term2	Score
Carbatrol	Dilantin	797.5
Cardiomyopathy	Tylenol	417.25
Nausea	Vomiting	1456.75
Thrombocytopenia	Sciatica	443.5

Table 1: *relatedness score for medical concept pairs*

- 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.

6. Experiment Setup

- We generate the word vectors using the two word embedding techniques under different settings of their parameters and compare their performance.
- For each model, word vectors of 25, 50, 100, and 200 dimensions are generated.

7. Evaluation

- In reference dataset we have two score for each word-pair. 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.

8. Results and Discussion

8.1 Similarity and Relatedness Task

- Table 2 and 3 shows the correlation values in all cases for Semantic Similarity and Semantic Relatedness task respectively. We observe that increasing the dimension of word vectors improve their ability to capture semantic properties of words.
- SkipGram model seems to be better than both CBOW and GloVe models in the semantic similarity and relatedness task for all dimensions.

Dimenstion	CBOW	SkipGram	GloVe
25	0.32	0.39	0.28
50	0.36	0.44	0.34
100	0.42	0.48	0.41
200	0.46	0.52	0.42

Table 2: *Semantic Similarity*

Dimenstion	CBOW	SkipGram	GloVe
25	0.30	0.34	0.27
50	0.33	0.38	0.34
100	0.39	0.43	0.41
200	0.41	0.45	0.42

Table 3: *Semantic Relatedness*

8.2 Nearest Neighbors of Some of the seed-words

- 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. Table 4 shows the result of 10 nearest neighbor of some seed words for SkipGram models, similar result were obtain for CBOW and GloVe also.

Seed word	SkipGram
eye	eye, eyes, face, head, ocular, mouth, pupillary, fovea, angle, Eye
cough	cough, breathlessness, expectoration, coughing, wheezing, dyspnea, phlegm, shortness, haemoptysis, sore
surgery	surgery, surgical, operation, procedure, esophagectomy, surgeries, laparoscopic, elective, reintervention, postoperative
tumour	tumour, tumor, tumors, tumours, malignant, metastatic, metastasis, metastases, tumoral, melanoma

Table 4: *10 nearest neighbors of selected seed-words for SkipGram*

9. Conclusion and Future Work

- In this study, we have shown that *word2vec* with skip-gram model gave the best performance compared to other models in the semantic similarity and relatedness task.
- 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

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

[2] 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.

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

[4] 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.