Understanding and Evaluating Medical Concept Embeddings

Andrew L. Beam*, Inbar Fried, Nathan P. Palmer, Isaac S. Kohane
Department of Biomedical Informatics, Harvard Medical School, Boston, MA, 02115, USA
**E-mail: Andrew_Beam@hms.harvard.edu*
www.university_name.edu

Benjamin Kompa
University of North Carolina, Chapel Hill, Chapel Hill, NC, 27514, USA
E-mail: kompa@live.unc.edu

Word embeddings, also known as distributed representations, have seen rapid adoption in natural language processing (NLP) and machine learning. Though they are now standard practice in many areas of NLP and machine learning, they are just now begining to attract interest in biomedical and clinical informatics. In this article, we present an overview of the existing word embedding methodology and investigate their use for biomedical concepts. In addition, we propose a set of benchmarks so that researchers can evaulate concept embeddings and understand what aspects of the source data they capture. We provide the benchmarks and a set of reference embeddings as an R package to the community to encourage quick, easy, and reproducible comparisons of new embeddings in the future.

Keywords: Machine Learning; Distributed Representations; Word Vectors; Concept Embeddings; Unsupervised Learning

1. Distributed Repsentations for Words and Concepts

The idea of a vectorized or distribution representation of a word has it roots in the neural language model of Bengio,[1] though this model is actually a formalization of the ideas first put forth in [paper from the 50s]. However, it wasn't until the paper[2] underpinning the wildly successful *word2vec* software package which demonstrated that collapsing the neural language model of Bengio[1] to a linear model enabled greater accuracy through training on much larger datasets that the idea of word embeddings finally came of age. Though they are often conflated, current distrubted representations are not an instance of deep learning, but are actually a speific kind of linear model, with explicit connections to many well known forms of matrix factorization.[3]

Word embeddings have ignited a furious amount of research after the the results of Mikolov[2] et. al demonstrated that they are capable of capturing a surprising amount of semantic information. The central idea of a word embedding is to represent a word as a dense, real-valued vector that projects the word into d-dimensional space. Words that are similar in this space encode certain semantic and linguistic regularities from the source text. While classic NLP tasks, such as sentiment analysis and text classification, have been shown to benefit from distributed representations, what caused this approach to gain considerable attention was the observation that analogies could be solved using arithmetic vector operations. The now famous example of $man : woman :: king :?$ can be sovled by the following operations on their

corresponding word vectors:

$$king - man + woman \approx queen$$

Where the vector for *queen* indicates a high cosine similarity to the vector resulting from $king - man + woman$. Thus, the analogy task is reduced to addition and subtraction on the word vectors.

1.1. *Word2Vec and Glove*

The two most popular algorithms for computing word vectors to emerge from the last several years of work are *word2vec*[2] and *Glove*.[4] The ideas in *word2vec* were originally presented in terms of two predictive models, the skip-gram and the CBOW. Though the objective function as originally presented appeared somewhat mysterious, later it was shown that the skip-gram model with negative sampling was equivalent to factorizing a shifted pointwise mutual information matrix[3] of word-context pairs.

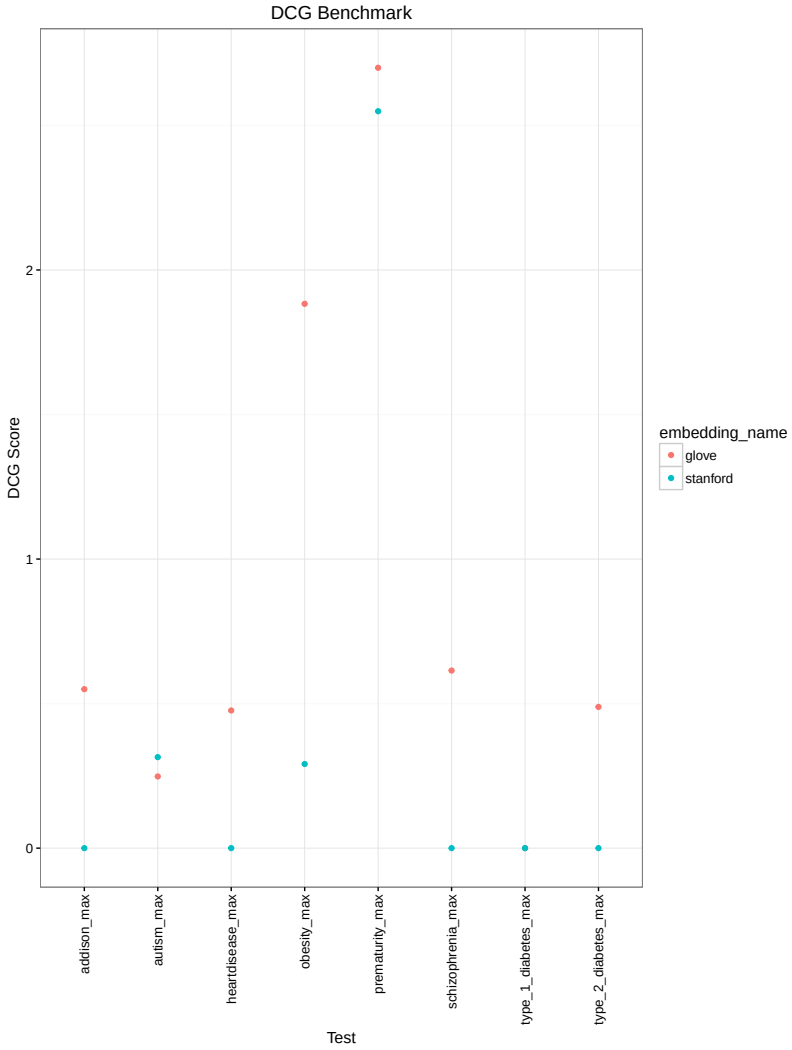
1.2. *Medical Concept Embeddings*

2. Benchmarks

Here is where we will put the description of all of the benchmarks, put in `\subsection{}` tags

3. Results

Here is where we will present the results for all of the different embeddings.



References

1. Y. Bengio, R. Ducharme, P. Vincent and C. Jauvin, *journal of machine learning research* **3**, 1137 (2003).
2. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean, Distributed representations of words and phrases and their compositionality, in *Advances in neural information processing systems*, 2013.
3. O. Levy and Y. Goldberg, Neural word embedding as implicit matrix factorization, in *Advances in neural information processing systems*, 2014.
4. J. Pennington, R. Socher and C. D. Manning, Glove: Global vectors for word representation., in *EMNLP*, 2014.