

# 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 beginning 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 evaluate 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. Introduction

Word embeddings have proven extremely useful in many areas of NLP and machine learning.

## 2. Distributed Representations for Words and Concepts

The idea of a vectorized or distribution representation of a word has its roots in the neural language model of Bengio,<sup>1</sup> 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<sup>2</sup> underpinning the wildly successful *word2vec* software package which demonstrated that collapsing the neural language model of Bengio<sup>1</sup> 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 distributed representations are not an instance of deep learning, but are actually a specific kind of linear model, with explicit connections to many well known forms of matrix factorization.<sup>3</sup>

### 2.1. *Word2Vec and Glove*

### 2.2. *Medical Concept Embeddings*

## 3. Benchmarks

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

## 4. 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.