Abstract

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.

1 Introduction

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

2 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 [**?**], 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[**?**] underpinning the wildly successful *word2vec* software package which demonstrated that collapsing the neural language model of Bengio[**?**] 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[**?**].

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.