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, North Carolina ZIP/Zone, USA
E-mail: an_author@laboratory.com

Word embeddings, also known as distributed representations, have seen rapid adoption in natural langauge 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 its applicability to biomedical informatics, as well as proposing a set of benchmarks for medical concept embedding evaluation. We provide these benchmarks as an R package to the community to encourage quick, easy, and reproducible comparisons for new embeddings in the future.

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

1. Introduction

Here is where we will motivate the paper and introduce the key ideas

2. Overview of Word Embeddings

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.

2.1. *Word2Vec*

2.2. *GLOVE*

2.3. *Embeddings as random walks*

2.4. *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.

5. References

Leave this here for now, I will compile a bibtex file References are to be listed in the order cited in the text in Arabic numerals. BIBTEX users, please use our bibliography style file `ws-procs11x85.bst` for references. Non BIBTEX users can list down their references in the following pattern.

```
\begin{thebibliography}{9}

\bibitem{jarl88} C. Jarlskog, in {\it CP Violation} (World Scientific,
   Singapore, 1988).

\bibitem{lamp94} L. Lamport, {\it \LaTeX, A Document Preparation System},
   2nd edition (Addison-Wesley, Reading, Massachusetts, 1994).

\bibitem{ams04} \AmS-\LaTeX{} Version 2 User's Guide (American Mathematical
   Society, Providence, 2004).

\bibitem{best03} B.~W. Bestbury, {\em J. Phys. A} {\bf 36}, 1947 (2003).

\end{thebibliography}
```

6. BIBTEXing

If you use the BIBTEX program to maintain your bibliography, you do not use the `thebibliography` environment. Instead, you should include

```
\bibliographystyle{ws-procs11x85}
\bibliography{ws-pro-sample}
```

where `ws-procs11x85` refers to a file `ws-procs11x85.bst`, which defines how your references will look. The argument to `\bibliography` refers to the file `ws-pro-sample.bib`, which should contain your database in BIBTEX format. Only the entries referred to via `\cite` will be listed in the bibliography.

Sample output using `ws-procs11x85` bibliography style file:

BIBTEX database entry type	Sample citation
article	... text.[3–5]
proceedings	... text.[6]
inproceedings	... text.[7]
book	... text.[8,9]
edition	... text.[10]
editor	... text.[11]
series	... text.[12]
tech report	See Refs. 13 and 14 for more details
unpublished	... text.[15]
phd thesis	... text.[16]
masters thesis	... text.[17]
incollection	... text.[18]
misc	... text.[19]

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. B. W. Bestbury, *J. Phys. A* **36**, 1947 (2003).
4. P. X. Deligne and B. H. Gross, *C. R. Math. Acad. Sci. Paris* **335**, 877 (2002).
5. J. M. Landsberg and L. Manivel, *Adv. Math.* **171**, 59 (2002), `http://www.url.com/triality.html`.
6. G. H. Weiss (ed.), *Contemporary Problems in Statistical Physics* (SIAM, Philadelphia, 1994).
7. R. K. Gupta and S. D. Senturia, Pull-in time dynamics as a measure of absolute pressure, in *Proc. IEEE Int. Workshop on Microelectromechanical Systems (MEMS'97)*, (Nagoya, Japan, 1997).
8. C. Jarlskog, *CP Violation* (World Scientific, Singapore, 1988).
9. L. F. Richardson, *Arms and Insecurity* (Boxwood, Pittsburg, 1960).
10. R. V. Churchill and J. W. Brown, *Complex Variables and Applications*, 5th edn. (McGraw-Hill, 1990).
11. F. Benhamou and A. Colmerauer (eds.), *Constraint Logic Programming, Selected Research* (MIT Press, 1993).
12. D. W. Baker and N. L. Carter, *Seismic Velocity Anisotropy Calculated for Ultramafic Minerals and Aggregates*, in *Flow and Fracture of Rocks*, eds. H. C. Heard, I. V. Borg, N. L. Carter and C. B. Raleigh, Geophys. Mono., Vol. 16 (Am. Geophys. Union, 1972), pp. 157–166.
13. J. D. Hobby, *A User's Manual for MetaPost*, Tech. Rep. 162, AT&T Bell Laboratories (Murray Hill, New Jersey, 1992).
14. B. W. Kernighan, *PIC—A Graphics Language for Typesetting*, Computing Science Technical Report 116, AT&T Bell Laboratories (Murray Hill, New Jersey, 1984).
15. H. C. Heard, I. V. Borg, N. L. Carter and C. B. Raleigh, VoQS: Voice Quality Symbols, Revised to 1994, (1994).

16. M. E. Brown, An interactive environment for literate programming, PhD thesis, Texas A&M University, (TX, USA, 1988), pp. ix + 102.
17. G. S. Lodha, Quantitative interpretation of ariborne electromagnetic response for a spherical model, Master's thesis, University of Toronto (1974).
18. D. Jones, The term 'phoneme', in *Phonetics in Linguistics: A Book of Reading*, eds. W. E. Jones and J. Laver (Longman, London, 1973) pp. 187–204.
19. B. Davidsen, Netpbm (1993), `ftp://ftp.wustl.edu/graphics/graphics/packages/NetPBM`.