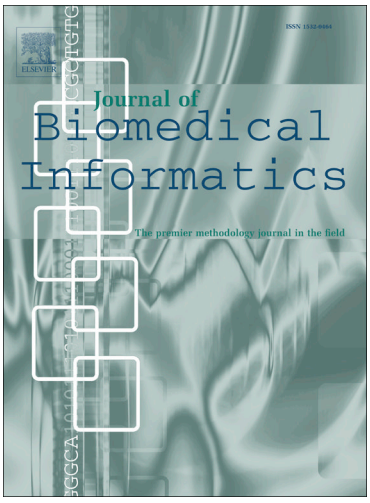

Vector representations of multi-word terms for semantic relatedness

Sam Henry, Clint Cuffy, Bridget T. McInnes

PII: S1532-0464(17)30276-9
DOI: https://doi.org/10.1016/j.jbi.2017.12.006
Reference: YJBIN 2901

To appear in: *Journal of Biomedical Informatics*

Received Date: 25 July 2017
Revised Date: 9 October 2017
Accepted Date: 12 December 2017

Please cite this article as: Henry, S., Cuffy, C., McInnes, B.T., Vector representations of multi-word terms for semantic relatedness, *Journal of Biomedical Informatics* (2017), doi: https://doi.org/10.1016/j.jbi.2017.12.006

Vector representations of multi-word terms for semantic relatedness

Sam Henry*, Clint Cuffy and Bridget T. McInnes

Department of Computer Science, Virginia Commonwealth University, 401 S. Main St., Richmond, VA 23284, USA

Abstract

This paper presents a comparison between several multi-word term aggregation methods of distributional context vectors applied to the task of semantic similarity and relatedness in the biomedical domain. We compare the multi-word term aggregation methods of summation of component word vectors, mean of component word vectors, direct construction of compound term vectors using the compoundify tool, and direct construction of concept vectors using the MetaMap tool. Dimensionality reduction is critical when constructing high quality distributional context vectors, so these baseline co-occurrence vectors are compared against dimensionality reduced vectors created using singular value decomposition (SVD), and word2vec word embeddings using continuous bag of words (CBOW), and skip-gram models. We also find optimal vector dimensionalities for the vectors produced by these techniques. Our results show that none of the tested multi-word term aggregation methods is statistically significantly better than any other. This allows flexibility when choosing a multi-word term aggregation method, and means expensive corpora preprocessing may be avoided. Results are shown with several standard evaluation datasets, and state of the results are achieved.

Keywords: Natural Language Processing, Semantic Similarity and Relatedness, Distributional Similarity

*Corresponding author

Email address: henryst@vcu.edu (Sam Henry*, Clint Cuffy and Bridget T. McInnes)