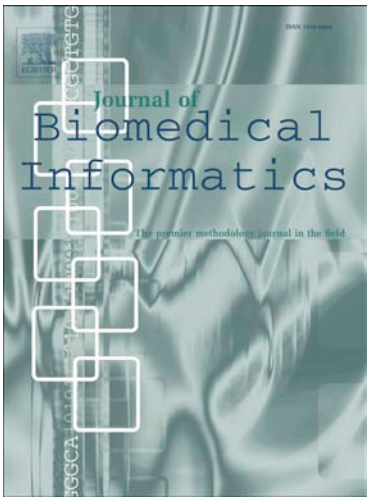

A Comparison of Word Embeddings for the Biomedical Natural Language Processing

Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, Hongfang Liu

PII: S1532-0464(18)30182-5
DOI: https://doi.org/10.1016/j.jbi.2018.09.008
Reference: YJBIN 3053

To appear in: *Journal of Biomedical Informatics*

Received Date: 4 April 2018
Accepted Date: 10 September 2018

Please cite this article as: Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., Kingsbury, P., Liu, H., A Comparison of Word Embeddings for the Biomedical Natural Language Processing, *Journal of Biomedical Informatics* (2018), doi: https://doi.org/10.1016/j.jbi.2018.09.008

A Comparison of Word Embeddings for the Biomedical Natural Language Processing

Yanshan Wang*, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, Hongfang Liu*

Department of Health Sciences Research

Mayo Clinic

Rochester, USA

Email: {Wang.Yanshan, Liu.Sijia, Afzal.Naveed, Mojarad.Majid, Wang.Liwei, Shen.Feichen, Kingsbury.Paul1, Liu.Hongfang}@mayo.edu

Abstract

Background Word embeddings have been prevalently used in biomedical Natural Language Processing (NLP) applications due to the vector representations of words capturing useful semantic properties and linguistic relationships between words. Different textual resources (e.g., Wikipedia and biomedical literature corpus) have been utilized in biomedical NLP to train word embeddings and these word embeddings have been commonly leveraged as feature input to downstream machine learning models. However, there has been little work on evaluating the word embeddings trained from different textual resources.

Methods In this study, we empirically evaluated word embeddings trained from four different corpora, namely clinical notes, biomedical publications, Wikipedia, and news. For the former two resources, we trained word embeddings using unstructured electronic health record (EHR) data available at Mayo Clinic and articles (MedLit) from PubMed Central, respectively. For the latter two resources, we used publicly available pre-trained word embeddings, GloVe and Google News. The evaluation was done qualitatively and quantitatively. For the qualitative evaluation, we arbitrarily selected medical terms from three medical categories (i.e., disorder, symptom, and drug), and manually inspected the five most similar words computed by word embeddings for each of them. We also analyzed the word embeddings through a 2-dimensional visualization plot of 377 medical terms. For the quantitative evaluation, we conducted both intrinsic and extrinsic evaluation. For the intrinsic evaluation, we evaluated the medical semantics of word embeddings using four published datasets for measuring semantic similarity between medical terms, i.e., Pedersen's dataset, Hliaoutakis's dataset, MayoSRS, and UMNSRS. For the extrinsic evaluation, we applied word embeddings to multiple downstream biomedical NLP applications, including clinical information extraction (IE), biomedical information retrieval (IR), and relation extraction (RE), with data from shared tasks.

Results The qualitative evaluation shows that the word embeddings trained from EHR and MedLit can find more relevant similar medical terms than those from GloVe and Google News. The intrinsic quantitative evaluation verifies that the semantic similarity captured by the word embeddings trained from EHR is closer to human experts' judgments on all four tested datasets. The extrinsic quantitative evaluation shows that the word embeddings trained

*Corresponding authors.