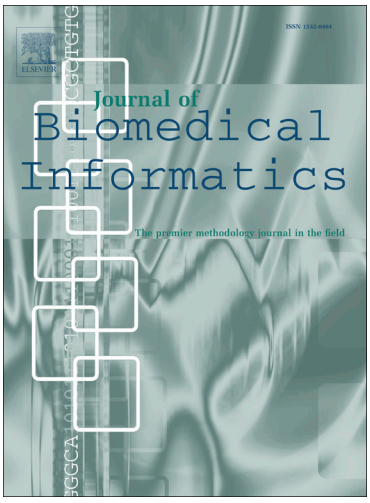

Bridging the gap: incorporating a semantic similarity measure for effectively mapping PubMed queries to documents

Sun Kim, Nicolas Fiorini, W. John Wilbur, Zhiyong Lu

PII: S1532-0464(17)30218-6
DOI: https://doi.org/10.1016/j.jbi.2017.09.014
Reference: YJBIN 2864

To appear in: *Journal of Biomedical Informatics*

Received Date: 12 May 2017
Revised Date: 1 September 2017
Accepted Date: 30 September 2017

Please cite this article as: Kim, S., Fiorini, N., John Wilbur, W., Lu, Z., Bridging the gap: incorporating a semantic similarity measure for effectively mapping PubMed queries to documents, *Journal of Biomedical Informatics* (2017), doi: https://doi.org/10.1016/j.jbi.2017.09.014

Bridging the gap: incorporating a semantic similarity measure for effectively mapping PubMed queries to documents

Sun Kim, Nicolas Fiorini, W. John Wilbur, Zhiyong Lu*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Abstract

The main approach of traditional information retrieval (IR) is to examine how many words from a query appear in a document. A drawback of this approach, however, is that it may fail to detect relevant documents where no or only few words from a query are found. The semantic analysis methods such as LSA (latent semantic analysis) and LDA (latent Dirichlet allocation) have been proposed to address the issue, but their performance is not superior compared to common IR approaches. Here we present a query-document similarity measure motivated by the Word Mover's Distance. Unlike other similarity measures, the proposed method relies on neural word embeddings to compute the distance between words. This process helps identify related words when no direct matches are found between a query and a document. Our method is efficient and straightforward to implement. The experimental results on TREC Genomics data show that our approach outperforms the BM25 ranking function by an average of 12% in mean average precision. Furthermore, for a real-world dataset collected from the PubMed® search logs, we combine the semantic measure with BM25 using a learning to rank method, which leads to improved ranking scores by up to 25%. This experiment demonstrates that the proposed approach and BM25 nicely complement each other and together produce superior performance.

*Corresponding author

Email addresses: sun.kim@nih.gov (Sun Kim), nicolas.fiorini@nih.gov (Nicolas Fiorini), wilbur@ncbi.nlm.nih.gov (W. John Wilbur), zhiyong.lu@nih.gov (Zhiyong Lu)