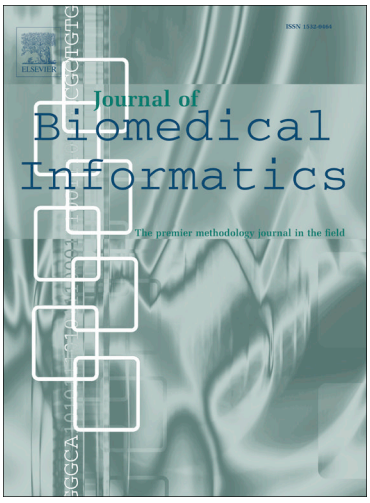

Word embeddings and recurrent neural networks based on Long-Short Term Memory nodes in supervised biomedical word sense disambiguation

Antonio Jimeno Yepes

PII: S1532-0464(17)30180-6
DOI: http://dx.doi.org/10.1016/j.jbi.2017.08.001
Reference: YJBIN 2837

To appear in: *Journal of Biomedical Informatics*

Received Date: 2 April 2016
Revised Date: 27 July 2017
Accepted Date: 2 August 2017

Please cite this article as: Yepes, A.J., Word embeddings and recurrent neural networks based on Long-Short Term Memory nodes in supervised biomedical word sense disambiguation, *Journal of Biomedical Informatics* (2017), doi: http://dx.doi.org/10.1016/j.jbi.2017.08.001

Word embeddings and recurrent neural networks based on Long-Short Term Memory nodes in supervised biomedical word sense disambiguation

Antonio Jimeno Yepes[a]

[a]*IBM Research Australia, Melbourne, VIC, Australia*

Abstract

Word sense disambiguation helps identifying the proper sense of ambiguous words in text. With large terminologies such as the UMLS Metathesaurus ambiguities appear and highly effective disambiguation methods are required. Supervised learning algorithm methods are used as one of the approaches to perform disambiguation. Features extracted from the context of an ambiguous word are used to identify the proper sense of such a word. The type of features have an impact on machine learning methods, thus affect disambiguation performance. In this work, we have evaluated several types of features derived from the context of the ambiguous word and we have explored as well more global features derived from MEDLINE using word embeddings. Results show that word embeddings improve the performance of more traditional features and allow as well using recurrent neural network classifiers based on Long-Short Term Memory (LSTM) nodes. The combination of unigrams and word embeddings with an SVM sets a new state of the art performance with a macro accuracy of 95.97 in the MSH WSD data set.

Keywords: Word sense disambiguation, Word embeddings, Recurrent neural networks, Biomedical domain