Learning Numerical Representations of Biomedical Concepts from 28 Million Abstracts

Jesus E. Vazquez[1,2], Anna Yannakopoulos[3], Kayla Johnson[3,4], Christopher Mancuso[3], Arjun Krishnan[3,4]

[1]Dept. Mathematics and Statistics, [2]Dept. Economics, University of New Mexico
[3]Dept. Computational Mathematics, Science and Engineering, [4]Dept. Biochemistry and Molecular Biology, Michigan State University

Introduction

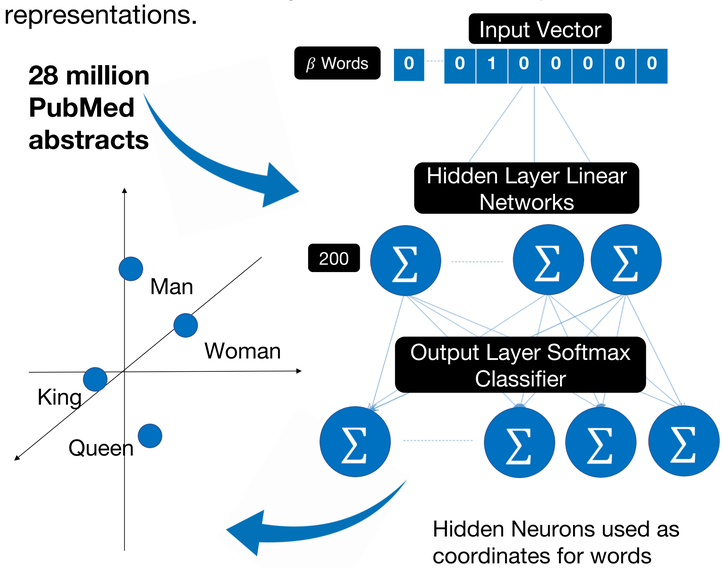
Natural Language Processing is a suite of analytical techniques for discerning meaning from vast text corpuses. Word2Vec is a neural network model that can learn numerical vector representations of words.

We apply this model to learn vector representations of biomedical concepts and explore their relationships based on analyzing abstracts of >28 million paper abstracts.

Approach

Using word2vec to learning vector representations

Titles and abstracts from 28 million biomedical papers from PubMed were processed using stemming, removing stop words, and named-entity-recognition. Two-layer neural networks were then used to train a word2vec model. The 200-dimensional hidden layer in this model provides concept representations.



Similarities between concepts

Similarities between pairs of concepts were calculated using Cosine Similarity and Euclidean Distance between their vectors.

Prior knowledge

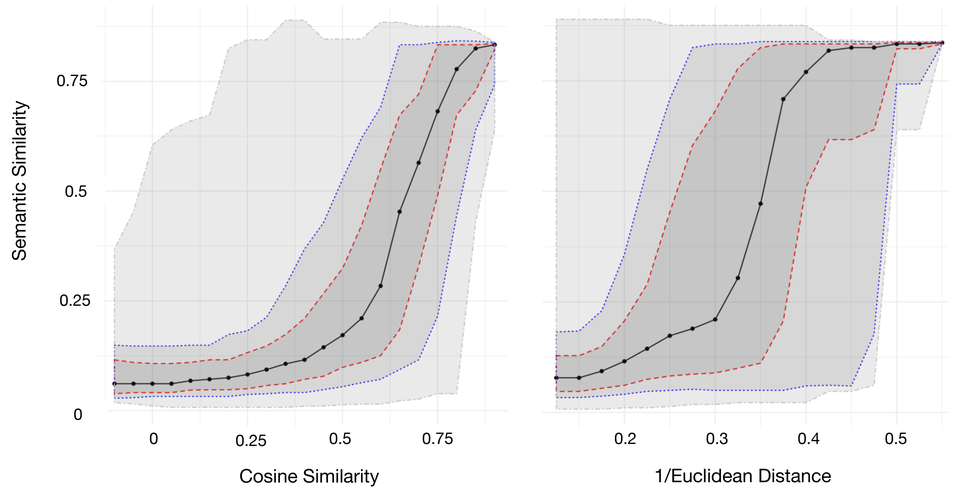
Prior knowledge about gene-gene, gene-function, function-function, gene-diseases, and disease-disease relationships were obtained from biomedical ontologies.

Ontologies to Semantic Similarities

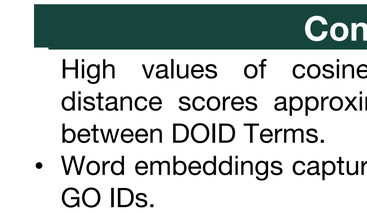
Gene Ontology (GO) and Disease Ontology (DO) represent our current knowledge about biological concepts (functions or diseases) and their relationships derived based on expert curation. Our word2vec models contain vectors representing these concepts. We examined if distances between these vectors capture semantic relationships between the concepts based on the underlying ontology.

Model Validation & Results

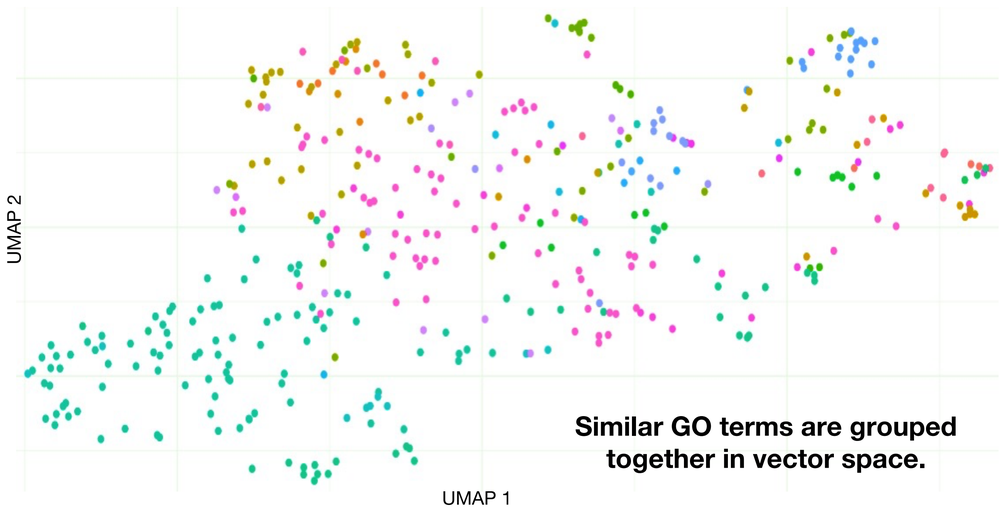
Approximation of Semantic Similarity between Disease Ontology Terms



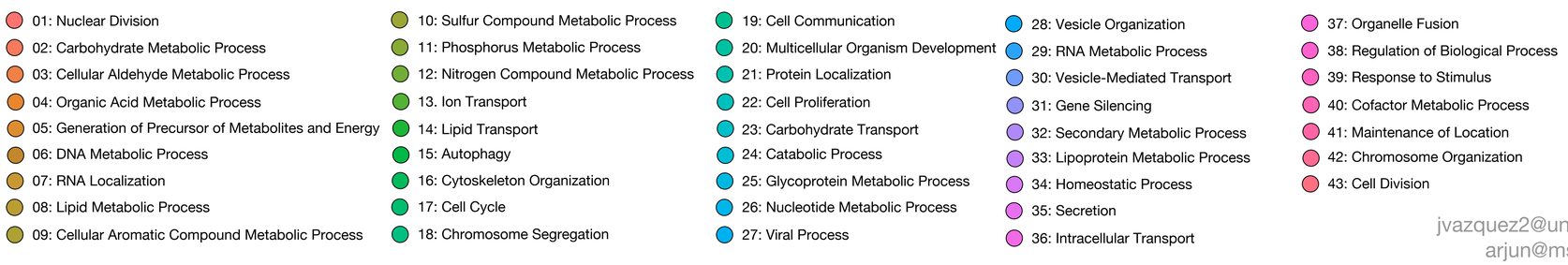
Results show that high values of cosine similarity approximate the semantic similarity between Diseases Ontology terms. This relationship also occurs for the 1/Euclidean Distance measure. **These findings state that semantic relationships can be captured by word-embeddings.**



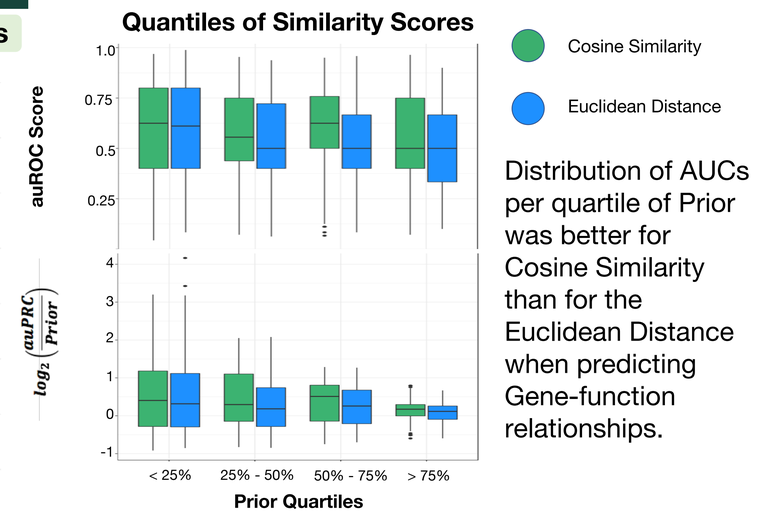
Low-dimensional embeddings of GO biological processes



Similar GO terms are grouped together in vector space.



Gene-function prediction



Distribution of AUCs per quartile of Prior was better for Cosine Similarity than for the Euclidean Distance when predicting Gene-function relationships.

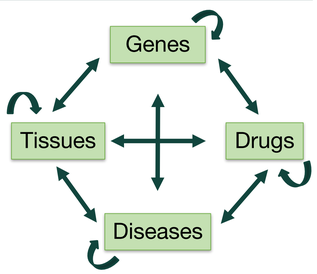
Conclusion

High values of cosine similarities and 1/euclidean distance scores approximate the semantic relationship between DOID Terms.

- Word embeddings capture the GO structure for particular GO IDs.
- The cosine similarity performs better as the ranking method when predicting gene-function relationships.

Discussion

Increasing the precision of word embeddings to associate terms can help us identify not only relationships between functions but relationships between genes, drugs, diseases, and tissues. Further exploration of the subject is needed to validate results.



Acknowledgements

This research was supported by the MSU ACRES REU program, which is supported by the National Science Foundation through grant ACI-1560168. I would like to thank Remy L., Nate D., Mark M., Jake R., Essenam B., Jainil S., Chinaza N. and Janani R. for their support this summer.

Email:
jvazquez2@unm.edu
arjun@msu.edu