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

Abstract: Machine-learning (ML) has gained momentum as a critical component of Natural Language Processing (NLP), a suite of analytical techniques for discerning meaning from vast text corpuses. Specifically, learning word embeddings – numerical vector representations of words in high-dimensional spaces – has gained enormous popularity as a tool for deriving semantic relationships and similarities between words. However, the application of word embeddings and their subsequent interpretation is underexplored in the biomedical domain. In this research project, we explore the use of word embeddings to glean similarity and semantic relationships between biomedical entities – e.g. genes, cellular functions, diseases, and drugs– from PubMed, a corpus of 28 million biomedical abstracts produced over the past 52 years. We are specifically interested in testing the effect of Name Entity Recognition (NER) on the efficacy of the word embeddings in capturing previously-known relationships. We are also comparing different similarity scores and developing methods to assess how well these learned embeddings recapitulate various aspects of our prior biomedical knowledge.

Key Words: Natural Language Processing, Machine Learning, Computational Linguistics, Gene Ontology, Disease Ontology, Name Entity Recognition, Biomedical

1. Introduction

With the increasing number of biomedical terms, e.g. genes, there has been a need to construct classification systems to facilitate the description of these terms, and specify the functionality and relationships between them. The Gene Ontology (DO) was founded in 1998 with the purpose of creating a tree-structure system that would classify genes based on their type of functionality: Biological Processes (BP), Cellular Components (CC), Molecular Functions (MF). The relative position of genes in this tree structure classification system also relates the terms in terms of their functionality and relationship. Although the relative distance between GO terms should indicate a good measure of how similar their functionality is, the semantic meaning of their descriptions should also be considered when determining the extent to which the terms are alike. Little research has been done on determining to what extent is the semantic similarity on the definition of two GO terms associated with the relative position of them in the classification structures.

Just as the GO, the Disease Ontology (GO) is also an open source of ontological description for human diseases. The GO and DO pave the way to better understand complex relationships between terms and break big concepts into more granular terms, much like other classification systems such as the International Classification of Diseases from the World Health Organization does. There has been numerous studies that have aimed to relate GO-GO, DO-DO, and GO-DO and these studies are research papers that can be found in PubMed. Identifying biomarkers to associate GO terms between GO terms and other biomedical entities is an important component toward understanding the complex relationship of gene interactions, but manually curating the interactions between these biomedical entities is extremely laborious, expensive and unfeasible. The biomedical community has experienced a large increase in the number of academic papers generated every year regarding new found interactions and biomarkers between biomedical entities. These millions of papers account for most of the expert curated knowledge we have about biomedical entities but other relationships and interaction between these entities not yet explicitly studied hide in the complex structure of the natural language of these research papers.

There has been numerous computational attempts to extract biomedical entities interactions from this unstructured repository of knowledge but many of these methods have not been successful but are urgently needed to help in the curation of new interations. The tools in Natural Language Processing (NLP) have become a key resource in inferring interactions between entities by helping us discern meaning from vast text corpuses. In particular Word2Vec, a two-hidden layer machine learning technique, has been widely used in NLP to compute numerical representations of terms of interest. These numerical representation of terms, whose relative position and distance between them explain the similarity of terms of interest, have been used to predict the interaction and similarity of biomedical entities , e.g. GO-GO terms.

One problem that has been found when using word2vec is that not all researchers refer to specific biomedical terms with the same name which introduces “noise” to these word embeddings. Some researchers might refer to colorectal cancer as colon cancer and other researchers might refer to it as rectal cancer. In this project we use a tool called EXTRACT 2.0 from Name Entity Recognition (NER), which is a tool that finds related biomedical entities in text corpuses and assigns a common ID to identify them.

The 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. Word2vec models contain vectors representing these concepts. In this project we examined if the distances between these vectors capture semantic relationships between the concepts based on the underlying ontology. We aimed to do this by creating the word embeddings of GO terms found in the text corpus of PubMed abstracts and titles, and use the Uniform Manifold Approximation and Projection (UMAP) to reduce the dimensionality of the vector space to project the GO terms into a two dimensional graph. Color coding based on the ancestry of the GO term was used to visually inspect if semantic relationships from the GO can be captured through the use of EXTRACT 2.0 and word2vec.

In this research study we also predict the functionality of genes. We compare the performance of the Cosine Similarity and the Euclidean Distance when ranking the GO

terms that we know that are either related or not related with Ensembl Gene IDs. We then compare the distributions of auROC and auPRC under different levels of prior.

2. Approach

2.1 Preparing the Data and Building Word2Vec Model

Titles and abstracts from 28 million biomedical papers from PubMed were processed using stemming, removing stop words, and Extract 2.0 from NER. Two-layer neural networks were then used to train a word2vec model. For our Word2Vec model we decided to implement the recommendations from *"How to Train Good Word Embeddings for Biomedical NLP"* by Billy Chiu et al. We used a negative sample size (neg) of 10, a 200 word embedding dimension (dim), a learning rate (alpha) of 0.05, subsampling rate (samp) of 1e-4, window size of 30 (win) and a word minimum-count (min-count) of 5 (Billy Chiu 2016). Figure 1, explains the process of how Word2Vec works. After the Word2Vec model was trained, we used the word embeddings to associate terms of interest.

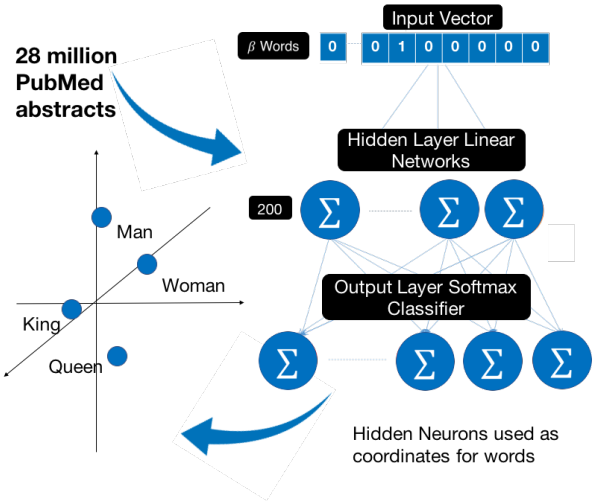

Figure 1: 28 million PubMed abstracts and titles are used to train the neural networks model. A one-hot input vector of size β-words is then used to calculate the 200-hidden-layer linear networks, which are then used as the coordinates in vector space. The location and placement of these word-embeddings have information regarding the relationship between terms.

In this example, the position and distance of the points say something about the relationship between King and Queen with Man and Woman.

2.2 GO Structure through Dimensionality Reduction

The coordinates of points given by the word-embeddings gives us relative distances and positions between biomedical terms. Through dimensionality reduction of these terms we hope that we can reconstruct the Gene Ontology structure. We use the Uniform Manifold Approximation and Projection (UMAP) to map the 200 dimensional vector of our word-embeddings into a two-dimensional platform with the goal of identifying clear patters that group Biological Processes (BP) GO terms of similar functionality together. The number of BP-GO terms are numerous, so terms that are either positive or negatively associated to other terms were used in this study. This measure was taken to be consistent with the prior knowledge we have of about the relationships and interactions between BP-GO terms. We also decided to color code BP-GO terms based on their ontological-father term, not the term itself. For some BP-GO terms they have multiple fathers, so we decided to only keep the father with the highest Cosine Similarity Score. The graphical illustrations of this method would then help us visually inspect if word-embeddings can recapitulate the structure of the BP-GO structure.

2.3 Calculating Similarity Scores

The Cosine Similarity has been used to associate points of interest in vector space for numerous research studies (Dat Duong 2017) (Fatima Zohra Smaili 2018). Although this measure is widely used there is no clear consensus of the legitimacy of this similarity score over other similarity scores when associating biomedical entities in vector space. Since the use of NLP is underexplored in the biomedical domain, in this research study we want to compare how the Cosine Similarity and Euclidean Distance similarity measures perform when ranking positively and negatively associated biomedical terms. We first identified all of the word embeddings that began with either “go” (Gene Ontology), “doid” (Disease Ontology), or “ensp” (Ensembl Peptide ID). We then calculated the Cosine Similarity and the Euclidean Distance for these and stored them in separate matrices depending on the terms that were compared, e.g. go-doid matrix.

It is also important to recognize that the similarity scores derived from word-embeddings represent word-similarity scores and not semantic-similarity scores. Two words can have a similar use which would incur in a high word-similarity score but that does not mean that their semantic relationship is present, e.g. day and night. For this reason we use the DOSE R-package to calculate the functional-semantic relationship between DO terms (Yu G 2015). The Resnik similarity score implemented from this package was used to represent the functional-semantic of the DOID terms. The scores calculated were then used in following steps to either correlate word-embedding derived measures with semantic related measures, as well as how these measures rank positive and negative associated biomedical entities when building ROC and Precision-Recall Curves.

Comment [JV1]: Arjun, is this the package and measure you used?

Source: https://bioconductor.org/packages/devel/bioc/vignettes/GOSemSim/inst/doc/GOSemSim.html#bma

2.4 Validation

Gene to function prediction models are urgently needed to increase the speed at which gene-function interactions are curated. From previous knowledge, we had data that that specified either positive or negative relationships between ENTREZ Gene IDs with GO terms. We first converted ENTREZ Gene IDs to Ensembl Peptide IDs (ENSP), then predicted GO terms associated with ENSP terms. In our research study we calculated auROC and auPRC scores using the Cosine Similarity and the Euclidean Distance as our ranking scores of ENSP terms that had at least 10 positive/negative labels. We then divided the distribution of scores by plotting them according to their prior. Visual inspection of these scores was then carried to decide which ranking criterion had higher auROC and $\log_2(\frac{auROC}{Prior})$.

3. Results

3.1 Approximation of Semantic Similarity Between DO Terms

The Cosine and 1/Euclidean Distance similarity scores were divided into 20 equally size classes. From each equally spaced bin the minimum, maximum, median, 10th percentile

and 90th percentile of their respective semantic similarity score was plotted. From the plot we can see that the median follows a sigmoidal curve which signifies that there is some relationship between the word-embedding derived similarity scores with the semantic similarity of the DO 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 to some extent the semantic relationships between DO terms can be captured by word-embeddings. This gives promises that word-embeddings reflect the functionality and semantic descriptions of biomedical terms in the disease domain. Figure 2 illustrates this observation.

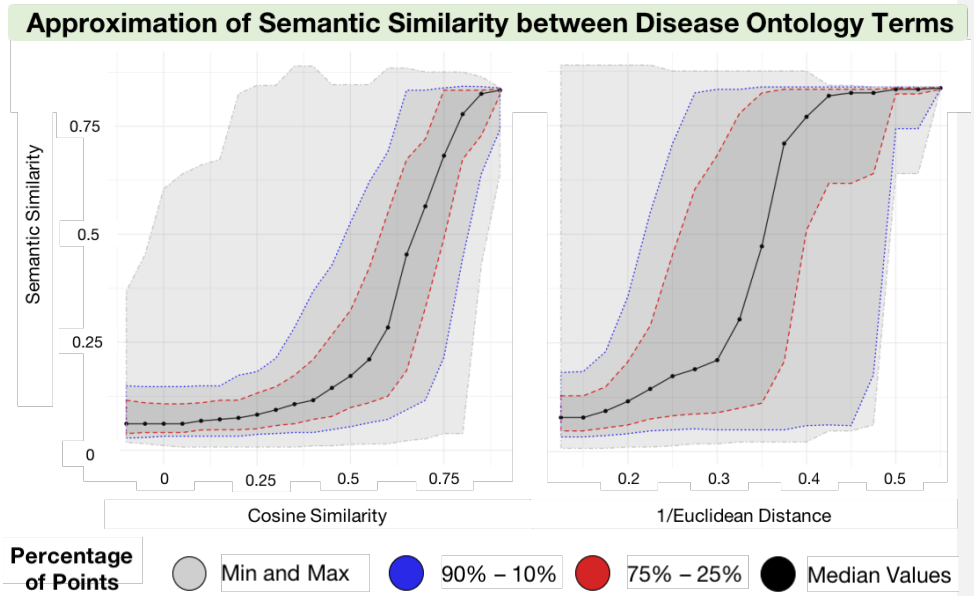

Figure 2: On this figure, the x-axis represents the similarity scores between Disease Ontology term using the word-embeddings using the Cosine Similarity (left) and the 1/Euclidean Distance (right). The y-axis represents the respective functionality-semantic similarity scores of the same terms as in the x-axis. The minimum, maximum, 10th percentiles, 90th percentiles and medians of the equal sized classes are plotted above. We can see that functionality-semantic similarity of DO terms are approximated by high values of similarity scores derived from word embeddings

3.2 Replicating the GO Structure

Dimensionality reduction of the word-embeddings using UMAP was implemented to access if the vector-representations of GO terms can capture the semantic functionality relationship between them. Results demonstrate that when projecting the word-embeddings into two dimensions and color coding GO terms based on their father-ancestor we find that terms with similar biological process are grouped together. We found 43 unique ancestor GO terms which were assigned different colors. Figure 3 shows the color definitions and how they are grouped in when projected into a two dimensional space.

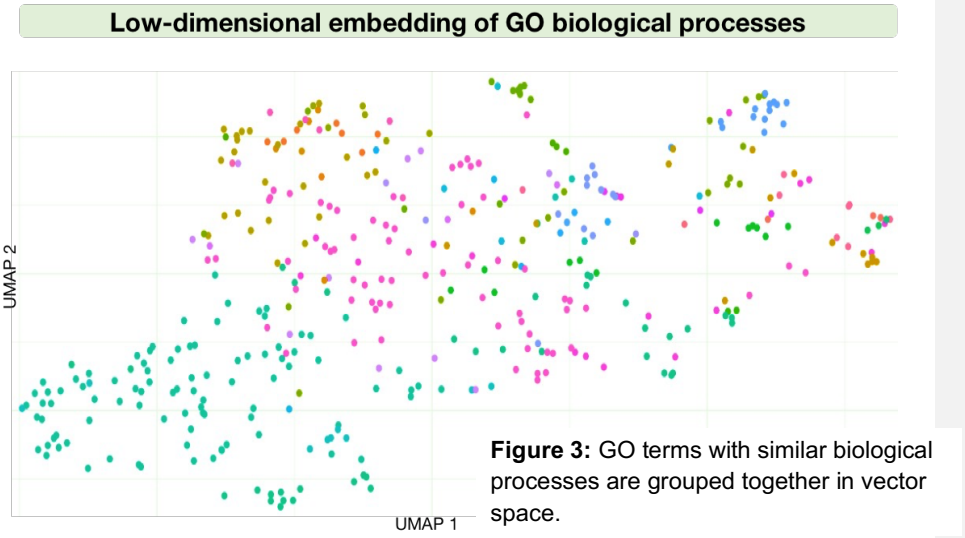

Figure 3: GO terms with similar biological processes are grouped together in vector space.

3.3 Predicting ENSP-GO Associations

To validate the predicting-power of our two word-embedding derived ranking scores we predicted GO terms that are associated with ENSP terms. In this research study we decided to only consider those ENSP terms that had at least 10 positive or negative labels (associated GO terms). The distributions of auROC and $\log_2(\frac{auPRC}{Prior})$ were also divided into prior levels. After calculating all of the auROC and $\log_2(\frac{auPRC}{Prior})$ for valid ENSP terms we found that the distribution of auROC per quartile of Prior was higher in the Cosine Similarity than in the Euclidean Distance when predicting Gene-function relationship for higher levels of prior. We see that this observation is present but not as strong when using $\log_2(\frac{auPRC}{Prior})$ as our metric of prediction-power. We can conclude that in the biomedical domain and when predicting Gene-Function relationships, the Cosine Similarity outperforms the Euclidean Distance as the better ranking criterion. Future studies should use the Cosine Similarity as the ranking criterion between word-embeddings. Distributions of the auROC and $\log_2(\frac{auPRC}{Prior})$ can be found in Figure 4.

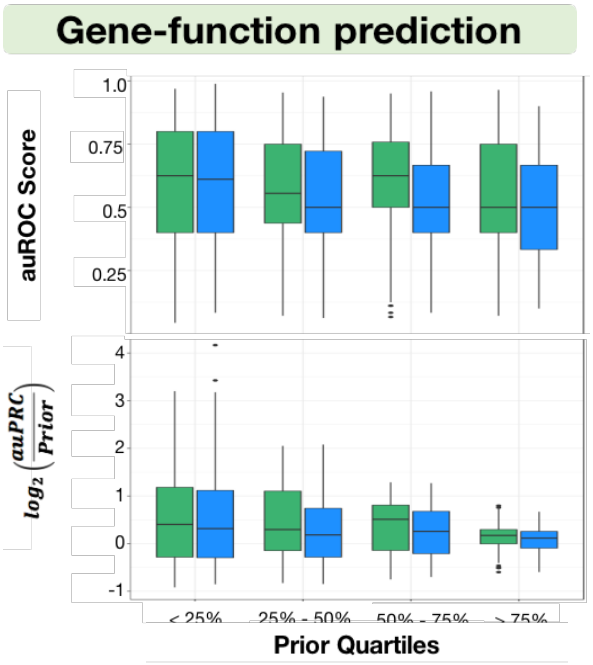

Figure 4: This plot illustrates the distributions of auROC and $\log_2(\frac{auPRC}{Prior})$ when using the Cosine Similarity and the Euclidean Distance as ranking criterions. This plot also divides the distributions by level of ENSP Prior.

4. Conclusion and Discussion

In this research project, we explore the use of word embeddings to glean similarity and semantic relationships between biomedical entities – e.g. genes, cellular functions, diseases, and drugs– from PubMed, a corpus of 28 million biomedical abstracts produced over the past 52 years. We find that high values of the Cosine Similarity and the 1/Euclidean Distance scores approximate the semantic relationship between DOID Terms. We also find that word-embeddings when reducing the dimensionality of word-embeddings and projecting it into a two dimensional space we can capture the GO structure for biological processes terms. We can also conclude that when predicting GO Biological Processes related terms the Cosine similarity is a better ranking criterion that the Euclidean Distance. These findings demonstrate that in the biomedical domain, we can extract relationships between biomedical entities from PubMed using word-embeddings.

For future studies we would recommend comparing other ranking criterions when validating the efficacy of word-embeddings at capturing relationships between biomedical terms. The GO and the DO are only two of the structures that we use to classify genes and diseases but there are also other classification systems in the biomedical community such as the International Classification of Diseases, 10^{th} Revision (ICD-10). Future Research studies should focus on determining if the structure of a classification system affects the way terms are semantically related. Findings from this analysis could shine light in to the way we structure our ontologies and the way we computationally extract data. Developing and enhancing current methods is extremely important as it can provide a way to reduce the amount of time and resources used to make expert curated biomedical entities interaction claims. Computational methods to link and develop the data mining and integration of large multi'omic data collections such as PubMed can help pave the way to a more rapid advancing medicine.

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. We thank our lab members Remy L., Nate D., Mark M., Jake C., Essenam B., Jainil S., Chinaza N., Janani R., for their support during the completion of this research project.

Bibliography

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. "How to Train Good Word Embeddings for Biomedical NLP." Language Technology Lab DTAL, University of Cambridge.

Dat Duong, Eleazar Eskin, and Jingyi Jessica Li. 2017. "A novel Word2vec based tool to estimate semantic similarity of genes by using Gene Ontology terms." *bioRxiv.*

Fatima Zohra Smaili, Xin Gao, and Robert Hoehndorf. 2018. "Onto2Vec: joint vector-based representation of biological entities and their ontology-based annotations." *Bioinformatics* .

ImmPort. n.d. *ImmPort Private Data.* Accessed 09 05, 2018. https://immport.niaid.nih.gov/home.

Yu G, Wang L, Yan G, He Q. 2015. "DOSE: an R/Bioconductor package for Disease Ontology Semantic and Enrichment analysis." *Bioinformatics* 31 (4): 608-609.