Deep Learning Approach to Automatically Extract Gene-Phenotype Relationships from Unstructured Literature Data

Tiffany Eulalio
Department of Biomedical Informatics
Stanford University
eulalio@stanford.edu

Bo Yoo
Department of Computer Science
Stanford University
byoo1@stanford.edu

Abstract

Mendelian disease is a set of genetic disorders that are caused by one broken gene. In order to identify the causative (i.e. broken) gene and offer a diagnosis, a doctor has to read through papers about each potentially malfunctioning genes in the patient. Often, this requires reviewing hundreds of papers, and for this reason, there is not enough bandwidth to diagnose all patients. The diagnosis process can be sped up if all symptoms (i.e. phenotypes) that are caused by each broken gene are organized in a structured database. Although some manual and automatic efforts have been made to build an accurate database of gene-phenotype relationships, there is still a lot of room for improvements. Here, we propose a deep learning based method to automatically extract gene-phenotype relationships from published literature to help make faster and more accurate diagnoses for patients with Mendelian diseases. We use the named entity recognition (NER) method to identify the gene and phenotype occurrences, convert these into recognizable identifiers, learn word vectors for these identifiers, and calculate cosine similarity between the vectors. From manual inspection on a small set, we see that our deep learning based method is able to capture some false negative and false positive mentions. Then we apply a gene prioritizing method, Phrank, with our automatically extracted gene-phenotype relationships and observe that our deep learning based method is able to rank the causative gene as the top 1 comparably to HPO-A and AMELIE.

1 Introduction and Related Work

Every year, approximately 7 million births worldwide are affected by severe genetic diseases (1). The simplest form of such disorders are Mendelian or monogenic diseases, which are caused by 1-2 rare variants in a single gene. Widespread use of Whole Exome Sequencing (**WES**) has revolutionized diagnosis (i.e. finding the causative variant in their DNA) of patients with Mendelian diseases. WES reveals the underlying genetic bases (i.e. which nucleobase – A, T, C, or G – is in each location of the person's genome), and clinicians can use this information to identify which gene contains a rare variant that is not generally found in the healthy population (e.g. a patient has a C in a loci when 99.99% of the population has A). However, diagnosing a patient is a very time-consuming task since after removing commonly observed variants there are 200-300 genes that contain a rare variant and are thus candidates for the causative gene (2). To find the correct *one* gene out of a few hundred candidate genes, clinicians compare known, published phenotypic information (i.e. symptoms and signs known to be displayed if a gene is perturbed) about each candidate gene to the patient's phenotypes (e.g. large head, difficult feeding, and missing 5th finger). Clinicians would consult existing databases that are manually curated by extracting gene to phenotype relationships from publications. However, each day there are new papers published, and novel gene-phenotype relationships take a long time for manual curators to update to the existing databases. To help clinicians make faster diagnoses, it is imperative that these databases are up to date and accurate.

Here, we propose a deep learning, natural language processing (NLP) based method to automate the curation of a structured gene-phenotype relationship database from unstructured publication data. Accurate automated extraction pipeline would allow clinicians to offer faster diagnosis by reducing their manual time required to read all novel publications. They will be able to compare the patient's symptoms with all known phenotypes caused by variations in the patient's candidate genes faster and offer diagnosis for more patients with suspected Mendelian disease. In the Bejerano Lab, AMELIE was built to automate the gene-phenotype extraction process but without using deep learning (3). Although AMELIE has been shown to help speed up the diagnosis process significantly through automatic extraction, it contains many false positive (i.e. a gene to phenotype relationship that has no causative relationship) examples. From our project, we hope to improve the accuracy of the extraction using deep learning and compare our results with the AMELIE curated database.

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

2 Dataset and Features

2.1 Gene Data

We are only interested in extracting gene-phenotype relationships for human genes since we would be using this database to diagnose human diseases. We have a full set of 39,495 human genes we can extract from papers. Most journals have recently been enforcing all publications to use the standardized HUGO Gene Nomenclature Committee (HGNC) (4) approved gene symbol for human genes, but in our extraction we also consider their synonyms as well as full names. To reduce conflicts among different methods of writing gene names, we standardize by converting gene names to Ensembl (a project to annotate vertebrate genes and other biological information) gene identifier (5). For example, a human protein coding gene KMT2A has a full name Lysine Methyltransferase 2A, a synonym MLL, and an Ensembl gene identifier ENSG00000118058. Generally, human genes are written in all capital letters and can include numbers and dashes but no other symbols. Since we have a finite set of gene names and they have distinctive standardized nomenclature, we expect identification of genes in publications to be a simpler task. However, some genes can have the same acronym as daily used words (e.g. SHE and FULL) which can be quite tricky.

2.2 Phenotype Data

Human Phenotype Ontology (**HPO**) is a standardized set of vocabularies that define phenotypic abnormalities shown in humans (6). Since phenotype terms naturally have a hierarchical structure, HPO is also organized in a directed acyclic graph (**DAG**) format. Each HPO term has a ontology value and set of definitions. For example, *HP:0001263* is a term for *Global developmental delay*, and it has many synonyms including Mental and motor retardation, Psychomotor developmental delay, Developmental retardation, Developmental delay, and Delayed cognitive development. It also has a parent *HP:0012758 (Neurodevelopmental delay)* and a child *HP:0011344 (Severe global developmental delay)*. HPO terms are particularly useful in our task since patient symptoms are also encoded in HPO terms by clinicians allowing us to easily compare gene-phenotype relationships with patient phenotypes. We have a total of 13,624 phenotypic terms, and we will use their descriptions as synonyms to identify the terms from publications. Since phenotypes are not enforced to be standardized in publications, identifying phenotype terms from papers is a much more difficult task.

2.3 Publication Data

PubMed is an online resource that contains more than 28 million biomedical literature from MEDLINE, life science journals, and online books. In this project we chose to use full publications rather than restricting ourselves to just abstracts. From about 28 million publications in PubMed, we subset our papers to 262,545 Mendelian disease relevant papers (i.e. papers that are expected to mention Mendelian disease, a gene, and the phenotypes caused by that gene) classified by the AMELIE pipeline so that we can fairly compare our results to AMELIE (3). We obtained these papers in PDF format which were filtered based on their relevance to our task. The PDF papers were converted into string text format using the Poppler library's pdftotext utility (7). Since old publications were stored as images, the converter was not able to successfully convert all publications. Therefore, we enforce that all converted publications to contain at least one sections of the paper we most suspect to include gene-phenotype relationships – abstract, introduction, results, and discussion – and remove any papers that didn't have these sections. After filtering out these unsuccessful conversions, we were left with 215,218 converted texts. The text files range in length, with the mean word count per paper of 3,508, minimum of 131, maximum of 35,294, and median of 3,339.

3 Methods

3.1 Named-Entity Recognition

3.1.1 Generating Labels for Supervised Learning

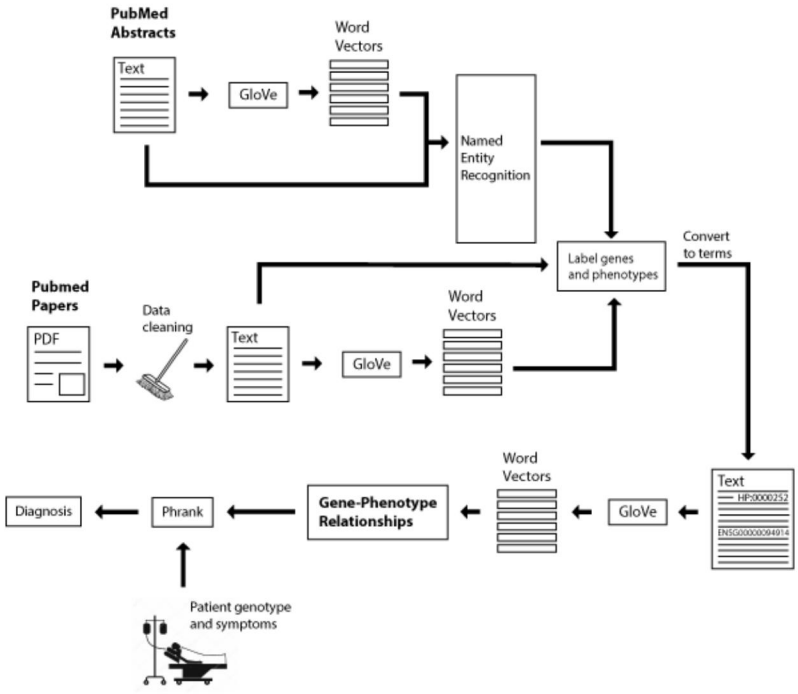
In order to draw out relationships between genes and phenotypes from the published literature, we first need to be able to identify the genes and phenotypes from the texts. To standardize these mentions, once an instance of gene or phenotype is found, we will replace those mentions with Ensembl gene identifiers for genes and HPO terms (e.g. HP:0001263 for global developmental delay) for phenotypes. We are using a deep learning approach to accomplish Named-Entity Recognition (**NER**).

This is a supervised approach, meaning we need labelled data. We create the labelled data by parsing the publication data and finding instances of gene and/or phenotype terms. For the gene extraction, we normalize both the text from papers as well as our full set of gene names. We scan the text using a window-size of eight to find references to genes in the paper. Indices of identified genes within the text are stored for later use. Similarly the phenotype extraction is performed by normalizing both the phenotype terms and the texts using common text-analysis techniques such as removal of stopwords, punctuation and converting case, making use of the Python Natural Language Toolkit (NLTK) library (8). We also perform matching using permutations of orders of phenotype description words that appear to catch more instances. To increase uniformity, some word that appear often are mapped to their synonyms. For instance, "decreased" is replaced with "reduced". We scan the texts in windows of eight words to find matches with phenotype terms. We record the matches along with their indices of the occurrence in the texts.

Once the matching had been performed, we use the generated indices to create the labelled data that will serve as our training, validation and test sets for NER. We create labels based on the IOBES tagging scheme (9) which provides more data than other commonly used schemes such as IOB. In the IOB format (Inside, Outside, Beginning), words are labelled according to whether they are at the beginning, inside, or outside of an entity, which in our case is a gene or phenotype. In the IOBES scheme, words are additionally labelled as being at the end of an entity, or as a single-word entity (9).

Since our goal for NER is to label genes and phenotypes for our 215,218 papers and these are unlabelled, these are our application set. We cannot manually label all 215,218 for accurate measurement of performance so after we train our model, we will run our model on these set of papers and then manually verify a small random subset. To train our model, we will use 345,851 PubMed paper abstract texts that contain gene and phenotype instances but are not part of our 215,218 papers of interests (e.g. are not Mendelian disease related papers).

Figure 1: Pipeline used to extract gene-phenotype relationships from PubMed Papers. Named Entity Recognition was trained using PubMed abstracts and used to label genes and phenotypes in relevant PubMed papers. Gene and phenotype mentions were replaced with identifiers and word vectors were created using GloVe. The resulting word vectors were analyzed to obtain gene-phenotype relationships. We use the tool Phrank to make comparisons on the extracted relationships.



3.1.2 Create Word Vectors to Cover Biological Terms

GloVe (Global Vectors for Word Representation) is an unsupervised learning algorithm that creates word vectors from a corpus using aggregated global word-word co-occurrence statistics (10). Glove uses the cost function

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \hat{w}_j + b_i + \tilde{b_j} - \log(X_{ij}))^2 \tag{1}$$

where V is the size of the vocabulary, $f(X_{ij})$ is a weighting function, $w \in \mathbb{R}^d$ are the word vectors, $\tilde{w} \in \mathbb{R}^d$ are separate context word vectors, b and $\tilde{b}$ are the biases and X_{ij} denotes the number of times word j occurs in the context of the word i. The weighting function defined by GloVe is

$$f(x) = \begin{cases} (x/x_{max})^\alpha & \text{if } x < x_{max} \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

where x_{max} and α are parameters to be defined. This model combines advantages of global matrix factorization and local context window methods to improve on the performance of word vector applications for specific tasks, including our desired application NER (10).

Prior to performing NER, we create 300-dimensional vectors for all words in our texts. These vectors are a transformation of our words into numerical values based on the frequency with which words occur together and the distance between words, which is captured by the cost function above.

For all word vectors that we create, we stick to parameters that were found to deliver the best performance by the creators of the GloVe model in (10). They report best results using an $\alpha = 3/4$ and $x_{max} = 100$. We run 100 iterations to create word vectors of 300 dimensions and a symmetric window size of 15 (15 words to the left and to the right of the word of interest)

We use these vectors as input for our NER model. Although GloVe is a method to represent word frequencies in vector formats, this does not accomplish our final gene-phenotype relationship extraction since phenotypes are generally comprised of multiple words. For example, if two phenotypes mentioned are *small head* and *large foot* both *small* and *large* would be linked to the gene ignoring their relationships with *head* and *foot*.

3.1.3 Implementing NER

We build our NER pipeline from a Tensorflow (17) implementation of NER available on GitHub (11). We perform NER using a bidirectional Long Short Term Memory (**LSTM**) model which has previously been shown to produced state of the art results (12), (9). The bidirectional model uses passes in both the forward and backward directions allowing predictions to be made based on both past and future data (13), as opposed to only past data in the unidirectional model. The LSTM architecture allows us to to make connections between extensive time lags in the text while addressing the issue of exponentially vanishing and exploding gradients (14). We combine a conditional random field (**CRF**) network with our LSTM network since this has been shown to improve tagging accuracy (12). The CRF layer focuses on sentences level tags instead of individual positions to make use of the information from neighboring tags. We run 8,000 minimum steps with a dropout rate of 0.5.

3.2 GloVe for Extracting Relationships

Using NER we identify occurrences of genes and phenotypes in the publications. After standardizing these instances (i.e. replacing gene names with Ensembl gene identifiers and phenotype mentions with HPO identifier), we apply GloVe to obtain vector representations of gene and phenotype instances. By learning these word embeddings, we hope to minimize the distances between the positive gene-phenotype relationships while maximizing the distance among other words, which we evaluate by calculating cosine similarity between vectors.

4 Experiments/Results/Discussion

4.1 Performance Evaluation

Our ultimate goal of having accurate gene-phenotype relationships in structured data is to help clinicians diagnose patients faster. To do so, there are tools that prioritizes genes based on their phenotypic relevance to the patient's phenotypes. In the Bejerano lab, a tool called Phrank (16) has been developed to use information content to meaningfully compare (i.e. instead of a direct comparison use the HPO DAG to compare similar terms) the patients phenotypes with genes' phenotypes to rank the causative gene highest in the candidate gene list. The tool's performance is heavily dependent on the accuracy of gene-phenotype relationships. Therefore, we expect the best gene-phenotype database to produce the best result measured by the number of patients' causative genes ranked at the top.

We have access to 384 real Mendelian disease patient data with diagnosis. For each patient we have 200-300 candidate genes, clinician noted patient phenotypes in HPO terms, and their causative gene. We will run Phrank (16) on these patients with three different gene-phenotype databases: our deep learning method, AMELIE database (3), and Human Phenotype Ontology Annotations (HPO-A) (6), a manually curated gene-phenotype relationship database from literature.

4.2 Results

4.2.1 Named-Entity Recognition Labelling

Upon completion of training our NER network on PubMed abstracts, we obtain an accuracy of 0.991, precision of 0.854, recall of 0.464, and F1 score of 0.602. We use Named-Entity Recognition to label genes and phenotypes in the PubMed papers and compare these labels with those of the non-deep learning method AMELIE (3). Examples of this comparison can be found in Figure 2A. We find that our deep learning (**DL**) method is correctly able to identify genes and non-genes as also identified by AMELIE. Additionally our DL method is able to identify phenotypes that are not found by AMELIE. Interestingly, our DL method is able to distinguish between genes that share their names with common English words such as "Full". This was one of the shortcomings of AMELIE and indicates an improvement in the labelling method.

4.2.2 Gene-Phenotype Relationships

We get the final gene-phenotype relationships from the cosine similarities of the word embedding vectors of the identifiers (e.g. ENSG00000118058 and HP:0001263). We take 50%, 30%, and 100% of gene-phenotype relationships with the smallest angle between two vectors, and compare these two sets against AMELIE and HPO-A. When doctors diagnose a patient they would run the candidate genes through a gene prioritization method and analyze them in the ranked order. If the causative gene (i.e. the right answer) is ranked near the top, diagnosis can be made faster. Therefore, we measure our performance on how many patient cases each gene-phenotype relationship set helps the gene prioritization method to rank the causative gene near the top (e.g. top 1, top 5 , etc.). Here, we use the prioritization method Phrank (16) which uses gene's phenotypic relevance to the patient's symptoms to rank potentially disease causing genes. Figure 2B shows the cumulative distribution of causative gene rankings, and from this plot we observe that AMELIE's performance is very close to HPO-A which is manually curated. Our deep learning based method is also able to rank the causative gene on top slightly more than HPO-A can especially when we take 30% of closest vectors. In some cases no causative gene's phenotypes are extracted by our method which then automatically ranks the gene at the bottom.

5 Conclusion/Future Work

Diagnosing a patient with Mendelian diseases is a very time consuming task for doctors. There are new papers written everyday about new gene-phenotype discoveries, and it is impossible for a doctor or manual curators to keep up with continuously growing information. To speed up the diagnosis process and build an accurate gene-phenotype relationship extractor, we propose an automatic NLP based

Figure 2: A) Word labels generated from our Deep Learning (DL) Named-Entity Recognition and non-Deep Learning AMELIE methods. Green circles represent positive labels (gene or phenotype) and red X's mark negative labels. Terms in purple boxes are genes and were labelled properly by both methods. Phenotype terms in yellow boxes were identified correctly by our DL method but not identified by the non-DL method. Terms in grey boxes are not genes or phenotypes and were all identified correctly by our DL method but AMELIE identifies one term incorrectly. B) Cumulative distribution of causative gene rankings. Each bar represents the frequency of patients where the causative gene was ranked less than or equal to top X when Phrank was ran with varying gene-phenotype relationship extractions.

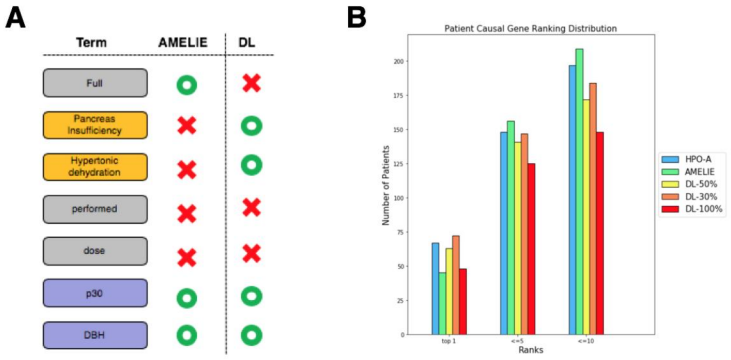
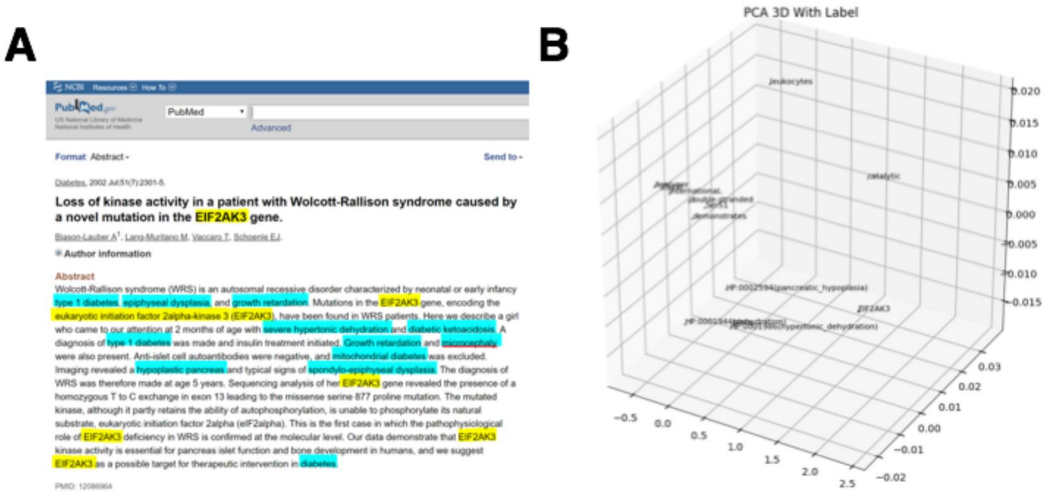


Figure 3: (A) A sample PubMed abstract with gene (yellow) and phenotype (blue) mentions highlighted. (B) A PCA 3D visualization of word embeddings generated through our deep-learning pipeline. Three phenotypes and one gene can be seen clustered towards the bottom of the graph.



method applying deep learning techniques. From our analysis comparing to exiting gene-phenotype extractions, we find that NER using deep learning is able to capture some false positive and false negative tagging of gene and phenotype mentions. However, there is much work that could be done to improve our results further. The biggest reason why our method does not outperform manually curated HPO-A and automatically curated AMELIE is because we are not able to extract phenotypes for all candidate genes. We expect improving identification of genes would make our performance much better. Our data was not perfect as our full paper data was converted from PDFs. We also only use texts in these papers but many important findings may appear in tables or figure format. Additional computer vision tasks on them are expected to improve the results. Also, which journal the paper is published usually matters on how significant the findings are (e.g. Nature papers are often more profound than other journals), therefore, in the future we can incorporate this information as a feature to weigh relationship origins differently.

6 Acknowledgements

In this project, we modified the NER implementation written by Guillaume Genthial and GloVe implementation written by the Stanford NLP lab. This project was inspired by Johannes Birgmeier's project AMELIE and advised by Gill Bejerano.

7 Contributions

Tiffany Eulalio: various pre-processing steps including converting pdfs to text, creating the phenotype-extractor, worked on main extraction code, working with repositories for NER and GloVe, labelling data, converting gene and phenotype mentions to identifiers, poster presentation

Bo Yoo: gathering of original data, various data pre-processing steps including parsing the original texts, filtering out unusable data, creating the gene-extractor and main extraction code, help running and modifying NER and GloVe code, write cosine angle code, implement Phrank on different databases, visualize word vectors, poster presentation

8 Code

Due to the use of private code from the Bejerano Lab, we have our code stored in a private Bitbucket repository. We have shared our code with our TA mentor Pedro Garzon for grading submission by sending as a zipped email attachment.

References

[1] Church, G. "Compelling Reasons for Repairing Human Germlines." Current Neurology and Neuroscience Reports.,U.S. National Library of Medicine, 16 Nov. 2017, www.ncbi.nlm.nih.gov/pubmed/29141159.

[2] Yang, Y, et al. "Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders." Current Neurology and Neuroscience Reports., U.S. National Library of Medicine, 17 Oct. 2013, www.ncbi.nlm.nih.gov/pubmed/24088041.

[3] Birgmeier, Johannes, et al. "AMELIE Accelerates Mendelian Patient Diagnosis Directly from the Primary Literature." BioRxiv, Cold Spring Harbor Laboratory, 1 Jan. 2017, www.biorxiv.org/content/early/2017/08/02/171322.

[4] Yates B, Braschi B, Gray K, Seal R, Tweedie S, Bruford E. Genenames.org: the HGNC and VGNC resources in 2017. Nucleic Acids Res. 2017 Jan 4; 45(D1):D619-625. PMID:27799471 PMCID: PMC5210531

[5] Daniel R. Zerbino, Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Giron, Laurent Gil, Leo Gordon, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G. Izuogu, Sophie H. Janacek, Thomas Juettemann, Jimmy Kiang To, Matthew R. Laird, Ilias Lavidas, Zhicheng Liu, Jane E. Loveland, Thomas Maurel, William McLaren, Benjamin Moore, Jonathan Mudge, Daniel N. Murphy, Victoria Newman, Michael Nuhn, Denye Ogeh, Chuang Kee Ong, Anne Parker, Mateus Patricio, Harpreet Singh Riat, Helen Schuilenburg, Dan Sheppard, Helen Sparrow, Kieron Taylor, Anja Thormann, Alessandro Vullo, Brandon Walts, Amonida Zadissa, Adam Frankish, Sarah E. Hunt, Myrto Kostadima, Nicholas Langridge, Fergal J. Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Dan M. Staines, Stephen J. Trevanion, Bronwen L. Aken, Fiona Cunningham, Andrew Yates, Paul Flicek Ensembl 2018. PubMed PMID: 29155950. doi:10.1093/nar/gkx1098

[6] Sebastian Köhler, Nicole Vasilevsky, Mark Engelstad, Erin Foster, et al. The Human Phenotype Ontology in 2017 Nucl. Acids Res. (2017) doi: 10.1093/nar/gkw1039

[7] Poppler library. (2018) [Online]. Available: https://poppler.freedesktop.org/

[8] Loper, E., Bird, S. "NLTK: The Natural Language Toolkit." CoRR., cs.CL/0205028.

[9] Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. "Neural Architectures for Named Entity Recognition." ArXiv:1603.01360 [Cs], March 4, 2016. http://arxiv.org/abs/1603.01360.

[10] Pennington, J., Socher, R., Manning, C.D. (2014) "GloVe: Global Vectors for Word Representation." [Online]. Available: https://nlp.stanford.edu/pubs/glove.pdf

[11] Genthial, Guillaume. Simple and Efficient Tensorflow Implementations of NER Models with Tf.Estimator and Tf.Data: Guillaumegenthial/Tf_ner. Python, 2018. https://github.com/guillaumegenthial/tf_ner.

[12] Huang, Zhiheng, Wei Xu, and Kai Yu. "Bidirectional LSTM-CRF Models for Sequence Tagging," August 9, 2015. https://arxiv.org/abs/1508.01991.

[13] Schuster, M., and K. K. Paliwal. "Bidirectional Recurrent Neural Networks." IEEE Transactions on Signal Processing 45, no. 11 (November 1997): 2673–81. https://doi.org/10.1109/78.650093.

[14] Graves, Alex, and Jürgen Schmidhuber. "Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures." Neural Networks: The Official Journal of the International Neural Network Society 18, no. 5–6 (July 2005): 602–10. https://doi.org/10.1016/j.neunet.2005.06.042.

[15] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13), C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.), Vol. 2. Curran Associates Inc., USA, 3111-3119.

[16] Jagadeesh, K A, et al. “Phrank Measures Phenotype Sets Similarity to Greatly Improve Mendelian Diagnostic Disease Prioritization.” Current Neurology and Neuroscience Reports., U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/pubmed/29997393

[17] Abadi, M, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org

[18] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.