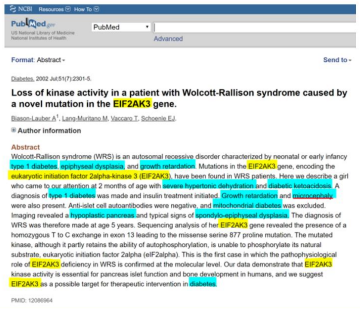
Deep Learning Approach to Automatically Extract Gene-Phenotype Relationships from Unstructured Literature Data

Tiffany Eulalio[1] and Bo Yoo[2]
Departments of 1.Biomedical Informatics, 2. Computer Science, Stanford University, Stanford, CA 94035

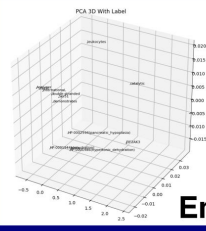
Motivation

Approximately 7 million births each year are affected by Mendelian diseases which are caused by one gene. To identify one gene responsible, doctors have to read thousands of papers describing phenotypic (symptoms and signs) abnormalities caused by each gene. Diagnosis is sped up if these relationships are in a structured format. Here, we use deep learning and NLP methods to extract gene-phenotype relationships from literature to help diagnosis of patients with Mendelian diseases and compare our result to AMELIE [1], a non deep learning based automatic gene phenotype extractor.



Data Description

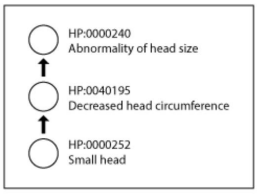
In this project, we extract gene phenotype relationships from **262,773** Mendelian disease related papers from Pubmed. These papers are identical to the set classified as relevant papers in AMELIE [1]. We train our network on **345,851** abstracts that are not part of the extraction set. Most of these abstracts are not Mendelian disease related but do contain phenotype and gene mentions. Phenotype descriptions are anchored on **13,559** Human Phenotype Ontology (HPO) terms. HPO terms have a DAG structure with parent-child relationships between related terms. This limited set helps us identify exactly which strings are phenotype mentions. Patient phenotypes are also annotated in HPO terms. Finally, there are **39,495** human genes we are interested in extracting from papers with synonyms and labeled with Ensembl identifiers.



3D PCA projection of word embedding vectors. A gene and phenotype mentions are much closer in space compared to random words

Word Embedding

Patient Diagnosis

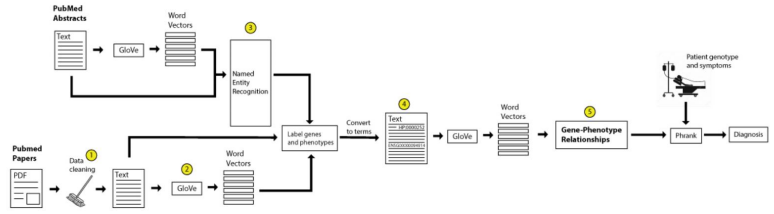


Phrank [4] is a gene prioritization tool that ranks genes with rare variants by their phenotypic Phenotype terms are organized in a hierarchical DAG and Phrank takes advantage of this structure to meaningfully compare similar terms. To measure our extraction's performance, we run Phrank on comparable structured gene phenotype relationships: HPOA and AMELIE [1]

Extracted Labels

Term	AMELIE	DL
Full	○	✗
Pancreas Insufficiency	✗	○
Hypertonic dehydration	✗	○
performed	✗	✗
dose	✗	✗
p30	○	○
DBH	○	○

Pipeline



1 Data Cleaning
- Select for papers mentioning Mendelian diseases only
- Convert papers from PDF format to text
- Identify and remove unsuccessful PDF conversions
- Normalize text

2 Glove
- Create global vectors for word representation to cover biological terms [2]
- Unsupervised learning algorithm
- 300-dimensional word vectors

3 Named Entity Recognition
- Implemented using Tensorflow
- Bi-directional
- Long- Short Term Memory (LSTM) model [3]

4 Convert to Terms
- Change all genes and phenotypes to single-word labels
- Replace genes with Ensembl identification
- Replace phenotypes with Human Phenotype Ontology identification

5 Gene-Phenotype Relationships
- From all word vectors calculate angles between the vector of a gene and a phenotype mentions
- Take varying perfect of gene-phenotype relationships with smallest angles

References

[1] Birgmeier et al. AMELIE accelerates Mendelian patient diagnosis directly from the primary literature., August 2017
[2] Pennington et al. GloVe: Global Vectors for Word Representation. 2014
[3] 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.
[4] Jagadeesh et al. Phrank measures phenotype sets similarity to greatly improve Mendelian diagnostic disease prioritization. July 2017