A Silver Standard Corpus of Human Phenotype-Gene Relations

Diana Sousa[*], Andre Lamurias and Francisco M. Couto

LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal

Abstract

Human phenotype-gene relations are fundamental to fully understand the origin of some phenotypic abnormalities and their associated diseases. Biomedical literature is the most comprehensive source of these relations, however, we need Relation Extraction tools to automatically recognize them. Most of these tools require an annotated corpus and to the best of our knowledge, there is no corpus available annotated with human phenotype-gene relations. This paper presents the Phenotype-Gene Relations (PGR) corpus, a silver standard corpus of human phenotype and gene annotations and their relations. The corpus consists of 1712 abstracts, 5676 human phenotype annotations, 13835 gene annotations, and 4283 relations[1]. We generated this corpus using Named-Entity Recognition tools, whose results were partially evaluated by eight curators, obtaining a precision of 87.01%. By using the corpus we were able to obtain promising results with two state-of-the-art deep learning tools, namely 78.05% of precision. The PGR corpus was made publicly available to the research community.[2]

1 Introduction

Automatic extraction of relations between entities mentioned in literature is essential to obtain knowledge that was already available but required considerable manual effort and time to retrieve. Recently, biomedical relation extraction has gained momentum in several text-mining applications, such as event extraction and slot-filling (Lamurias and Couto, 2019b). Some of the commonly extracted biomedical relations are protein-protein interactions (Papanikolaou et al., 2015),

drug-drug interactions (Lamurias et al., 2019) and disease-gene relationships (Kim et al., 2017).

There are a few worth mention systems regarding biomedical Relation Extraction (RE) (Verga et al., 2018), and that specifically focus on the extraction of phenotype-gene relations regarding different species types like plants (Xing et al., 2018) and humans (Collier et al., 2015). The main problem that these systems face is a lack of specific high quality annotated corpora (gold standard corpus), mostly because this task requires not only a considerable amount of manual effort but also specific expertise that is not widely available. A solution to these limitations is to generate the corpus in a fully automated manner (silver standard corpus).

Connecting human phenotypes to genes helps us to understand the origin of some phenotypic abnormalities and their associated diseases. To extract human phenotype-gene relations, both entities, human phenotypes and genes have to be recognized. With genes, as a result of lexical features being relatively regular, many systems can successfully identify them in text (Leaman and Gonzalez, 2008). Even though Named-Entity Recognition (NER) research has significantly improved in the last years, human phenotype identification is still a complex task, only tackled by a handful of systems (Lobo et al., 2017).

To generate a silver standard for phenotype-gene relation extraction, we used a pipeline that performs: i) NER to recognize genes and human phenotype entities; ii) RE to classify a relation between human phenotype and gene entities. First, we gathered abstracts using the PubMed API with manually defined keywords, namely each gene name, *homo sapiens*, and *disease*. Then we used the Minimal Named-Entity Recognizer (MER) tool (Couto and Lamurias, 2018) to extract gene mentions in the abstracts and the Identifying Human Phenotypes (IHP) tool (Lobo et al.,

---
[*]dfsousa@lasige.di.fc.ul.pt

[1]Query 1, corresponds to the *10/12/2018* release of PGR
[2]https://github.com/lasigeBioTM/PGR

2017) to extract human phenotype mentions. At last, we used a gold standard relations file, provided by the Human Phenotype Ontology (HPO), to classify the relations obtained by co-occurrence in the same sentence as *Known* or *Unknown*.

To the best of our knowledge, there is no corpus available specific to human phenotype-gene relations. This work, overcame this issue by creating a large and versatile silver standard corpus. To assess the quality of the Phenotype-Gene Relations (PGR) corpus, eight curators manually evaluated a subset of PGR. We obtained highly promising results, for example 87.18% in precision. Finally, we evaluated the impact of using the corpus on two deep learning RE systems, obtaining 69.23% (BO-LSTM) and 78.05% (BioBERT) in precision.

2 PGR Corpus

The HPO is responsible for providing a standardized vocabulary of phenotypic abnormalities encountered in human diseases (Köhler et al., 2017). The developers of the HPO also made available a file that links these phenotypic abnormalities to genes. These phenotype-gene relations are regularly extracted from texts in Online Mendelian Inheritance in Man (OMIM) and Orphanet (ORPHA) databases, where all phenotype terms associated with any disease that is related with a gene are assigned to that gene in the relations file. In this work, we used the relations file created by HPO as a gold standard for human phenotype-gene relations.

We started by retrieving abstracts from PubMed, using the genes involved in phenotype-gene relations and *homo sapiens* as keywords, and the Entrez Programming Utilities (E-utilities) web service (https://www.ncbi.nlm.nih.gov/books/NBK25501/), retrieving one abstract per gene (Query 1).

Later, we added the keyword *disease* and filter for abstracts in English (Query 2)[3]. Query 2 represents a more focused search of the type of abstracts to retrieve, such as abstracts regarding diseases, their associated phenotypes and genes.

For each gene, we opted for the most recent abstract (Query 1) and the two most recent abstracts (Query 2).

We opted by searching per gene and not human phenotype or the combination of both terms because this approach was the one that retrieved abstracts with the higher number of gene and human phenotype annotations, in the following NER and RE phases. We removed the abstracts that did not check the conditions of being written in English, with a correct XML format and content. The final number of abstracts was 1712 for Query 1 and 2657 for Query 2 as presented in Table 1. Then we proceeded to use the MER tool (Couto and Lamurias, 2018) for the annotation of the genes and the IHP framework (Lobo et al., 2017) for the annotation of human phenotype terms.

2.1 Gene Extraction

MER is a dictionary-based NER tool which given any lexicon or ontology (e.g., an OWL file) and an input text returns a list of recognized entities, their location, and links to their respective classes.

To annotate genes with MER we need to provide a file of gene names and their respective identifiers. To this goal, we used a list created by the HUGO Gene Nomenclature Committee (HGNC) at the European Bioinformatics Institute (http://www.genenames.org/). The HGNC is responsible for approving unique symbols and names for human loci, including protein-coding genes, ncRNA genes, and pseudogenes, with the goal of promoting clear scientific communication. Considering that we intended not only to map the genes to their names but also their Entrez Gene (www.ncbi.nlm.nih.gov/gene/) identifiers, we used the API from MyGene (http://mygene.info/) with the keyword *human* in species. The MyGene API provides several gene characteristics, including the confidence score for several possible genes that match the query. For this work, we chose the Entrez Gene identifier with a higher confidence score.

After corresponding all gene names to their respective identifiers, we were left with three genes that did not have identifiers (*CXorf36*, *OR4Q2*, and *SCYGR9*). For the first two genes (*CXorf36* and *OR4Q2*), a simple search in Entrez Gene allowed us to match them to their identifiers. For the last gene (*SCYGR9*) we were not able to find an Entrez Gene identifier, so we used the HGNC identifier for that gene instead. We opted to use the Entrez Gene identifiers because of their widespread use in the biomedical research field.

To the original gene list, we added gene synonyms using a synonyms list file provided by

---

[3]Query 2, corresponds to the *11/03/2019* release of PGR

Query	Abstracts	Annotations		Relations		
		Phenotype	Gene	Known	Unknown	Total
1 (10/12/2018)	1712	5676	13835	1510	2773	4283
2 (11/03/2019)	2657	9553	23786	2480	5483	7963

Table 1: The final number of abstracts retrieved, number of phenotype and gene annotations extracted and the number of known, unknown and total of relations extracted between phenotype and genes, for Query 1 and 2.

https://github.com/macarthur-lab/gene_lists (expanding the original list almost 3-fold). These synonyms were matched to their identifiers and filtered according to their length to exclude one character length synonyms and avoid a fair amount of false positives. The number of genes in the original gene list was 19194, and by including their synonyms that number increased to 56670, representing a total gain of 37476 genes.

At last, we identified some missed gene annotations that were caught using regular expressions. These missed gene annotations were next to forward/back slash and dashes characters (Example 1).

Example 1. Missed gene annotation because of forward slash.
- **Gene:** *BAX*
- **Gene Identifier:** 581
- **Abstract Identifier:** 30273005
- **Sentence:** According to the morphological observations and DNA fragmentation assay, the MPS compound induced apoptosis in both cell lines, and also cause a significant increase in the expression of **Bax**/Bcl-2.

2.2 Phenotype Extraction

IHP is a Machine Learning-based NER tool, specifically created to recognize HPO entities in unstructured text. It uses Stanford CoreNLP (Manning et al., 2014) for text processing and applies Conditional Random Fields trained with a rich feature set, combined with hand-crafted validation rules and a dictionary to improve the recognition of phenotypes.

To use the IHP system we chose to update the HPO ontology for the most recent version[4]. The annotations that originated from the IHP system were matched to their HPO identifier. There was a total of 7478 annotations for Query 1 and 10973 annotations for Query 2 that did not match any HPO identifier. We put aside these annotations to be confirmed or discarded manually as some of

them are incorrectly identified entities but others are parts of adjacent entities that can be combined for an accurate annotation.

We did not use the MER system for phenotype extraction mainly because a more efficient tool for this task was available without the limitations of a dictionary-based NER tool for complex terms as phenotypes are.

2.3 Relation Extraction

After filtering abstracts that did not have annotations of both types, gene and phenotype, we gathered a total of 1712 abstracts for Query 1 and 2656 abstracts for Query 2 as presented in Table 1. The abstracts retrieved by Query 1 were not specific enough for human phenotype-gene relations and therefore about half of them did not contained entities from both types, which we addressed in Query 2, increasing from about 2.5 relations per abstract to about 3.0 relations per abstract.

Using a distant supervision approach, with the HPO file that links phenotypic abnormalities to genes, we were able to classify a relation with *Known* or *Unknown*. For this end, we extract pairs of entities, of gene and human phenotype, by co-occurrence in the same sentence (Example 2). The final number of both *Known* and *Unknown* annotations is also presented in Table 1.

Example 2. Relation extraction.
- **Abstract Identifier:** 23669344
- **Sentence:** A homozygous mutation of **SERPINB6**, a gene encoding an intracellular protease inhibitor, has recently been associated with post-lingual, autosomal-recessive, nonsyndromic **hearing loss** in humans (DFNB91).
- **Gene:** *SERPINB6*
- **Gene Identifier:** 5269
- **Phenotype:** *hearing loss*
- **Phenotype Identifier:** HP_0000365
- **Relation: Known**

---
[4] *09/10/2018* release

1489

3 Evaluation

To evaluate the quality of the classifier, we randomly selected 260 relations from Query 1 to be reviewed by eight curators (50 relations each, with an overlap of 20 relations). All researchers work in the areas of Biology and Biochemistry. These curators had to evaluate the correctness of the classifier by attributing to each sentence one of the following options: *C* (correct), *I* (incorrect) or *U* (uncertain). The *U* option was given to identify cases of ambiguity and possible errors in the NER phase. We classified as a true positive (TP) a *Known* relation that was marked *C* by the curator, a false positive (FP) as a *Known* relation marked *I*, a false negative (FN) as a *Unknown* relation marked *I* and a true negative (TN) as a *Unknown* relation marked *C*.

3.1 State-of-the-art Applications

The PGR corpus was applied to two state-of-the-art systems that were compared against a co-occurrence (or all-true) baseline method.

3.1.1 BO-LSTM Application

The BO-LSTM system (Lamurias et al., 2019) is a deep learning system that is used to extract and classify relations via long short-term memory networks along biomedical ontologies. This system was initially created to detect and classify drug-drug interactions and later adapted to detect other types of relations between entities like human phenotype-gene relations. It takes advantage of domain-specific ontologies, like the HPO and the Gene Ontology (GO) (Ashburner et al., 2000). The BO-LSTM system represents each entity as the sequence of its ancestors in their respective ontology.

3.1.2 BioBERT Application

The BioBERT system (Lee et al., 2019) is a pre-trained biomedical language representation model for biomedical text mining based on the BERT (Devlin et al., 2018) architecture. Trained on large-scale biomedical corpora, this system is able to perform diverse biomedical text mining tasks, namely NER, RE and Question Answering (QA). The novelty of the architecture is that these systems (BioBERT and BERT) are designed to pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers. These feature allows easy adaption to several tasks without loss in performance.

4 Results and Discussion

The final results are presented in Table 2. The inter-curator agreement score, calculated from a total of 20 relations classified by eight curators, was 87.58%. Besides the fact that there were a few incorrectly extracted relations due to errors in the NER phase, that were discarded, the inter-curator agreement is not higher due to the complexity of the sentences where the relations between entities were identified. Even with highly knowledgeable curators in the fields of Biology and Biochemistry, most of them expressed difficulties in deciding which mark to choose on complex sentences that did not necessarily imply a relation between the identified entities (Example 3).

Example 3. Relation marked with *U* (Uncertain).

- **Abstract Identifier:** 27666346
- **Sentence: FRMD4A** antibodies were used to probe 78 paraffin-embedded specimens of **tongue squamous cell carcinoma** and 15 normal tongue tissues, which served as controls.
- **Mark:** *U*

The precision obtained from the test-set (about 6% of the total of relations), was 87.01%. Although we cannot state that this test-set is representative of the overall data-set, this is a solid evidence of the effectiveness of our pipeline to automate RE corpus creation, especially between human phenotype and genes, and other domains if a gold standard relations file is provided. Our lower recall is mostly due to incorrectly retrieved human phenotype annotations by IHP, that can be manually confirmed in a future optimized version of the PGR corpus, as some of them are parts of adjacent entities that can be combined for an accurate annotation.

4.1 Impact on Deep Learning

For BioBERT we used the available pre-trained weights for training and testing of RE model on our corpus. The results of BO-LSTM and BioBERT in the test-set are presented in Table 3. We also measured the performance of the co-occurrence (i.e. assuming all-true) baseline method. This baseline method assumes that all relations in the test-set are *Known* and therefore the recall is 100%. These results are comparable to

Relations		Marked Relations				Metrics		
Known	Unknown	True Positive	False Negative	False Positive	True Negative	Precision	Recall	F-Measure
77	143	67	86	10	57	0.8701	0.4379	0.5826

Table 2: The *Known* and *Unknown* number of relations selected, the number of true positives, false negatives, false positives and true negatives, and the evaluation metrics for the *Known* relations.

Method	Precision	Recall	F-Measure
Co-occurrence	0.3500	1.0000	0.5185
BO-LSTM	0.6923	0.4200	0.5228
BioBERT	0.7895	0.5844	0.6716

Table 3: Precision, recall and F-measure of the co-occurrence baseline, BO-LSTM and BioBERT.

the ones obtained from the evaluation stage by the curators, and show the applicability of our corpus.

BioBERT significantly outperforms BO-LSTM in all metrics proving that is indeed a viable language representation model for biomedical text mining. Even though the recall for both systems is relatively low, the purpose of this work was mainly to extract correct relations between entities to facilitate Machine Learning (ML), which was achieved by obtaining the precision of 69.23% (BO-LSTM) and 78.95% (BioBERT).

The most relevant metric for a silver standard corpus, directed towards ML tools, is precision. ML tools depend on correct examples to create effective models that can detect new cases, afterwards, being able to deal with small amounts of noise in the assigned labels.

5 Conclusions

This paper showed that our pipeline is a feasible way of generating a silver standard human phenotype-gene relation corpus. The pipeline required the application of two NER tools, and the availability of a list of known relations. We manually evaluated the corpus using eight curators obtaining a 87.01% precision with an inter-agreement of 87.58%. We also measured the impact of using the corpus in state-of-the-art deep learning RE systems, namely BO-LSTM and BioBERT. The results were promising with 69.23%, and 78.95% in precision, respectively. We believe that our pipeline and silver standard corpus will be a highly useful contribution to overcome the lack of gold standards.

Future work includes manually correcting the human phenotype annotations that did not match any HPO identifier, with the potential of expanding the number of human phenotype annotations almost 2-fold and increasing the overall recall. Also, we intend to expand the corpus by identifying more missed gene annotations using pattern matching, which is possible due to our approach being fully automated. Another possibility is the expansion of the test-set for a more accurate capture of the variance in the corpus. For example, we can select a subset of annotated documents in which two curators could work to grasp the complexity of manually annotating some of these abstracts. Further, we intend to use semantic similarity to validate the human phenotype-gene relations. Semantic similarity has been used to compare different types of biomedical entities (Lamurias and Couto, 2019a), and is a measure of closeness based on their biological role. For example, if the *BRCA1* gene is semantically similar to the *BRAF* gene and the *BRCA1* has an established relation with the *tumor* phenotype, it could be possible to infer that *BRAF* gene also has a relation with the *tumor* phenotype, even if that is not evident by the training data. Finally, the effect of different NER systems applied to the pipeline should be studied.

References

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan Peter Davis, Kara Dolinski, Selina S. Dwight,

Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29.

Nigel Collier, Tudor Groza, Damian Smedley, Peter N. Robinson, Anika Oellrich, and Dietrich Rebholz-Schuhmann. 2015. Phenominer: from text to a database of phenotypes associated with OMIM diseases. *Database*, 2015:bav104.

Francisco M Couto and Andre Lamurias. 2018. MER: a shell script and annotation server for minimal named entity recognition and linking. *Journal of Cheminformatics*, 10(58):1–10.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jeongkyun Kim, Jung-jae Kim, and Hyunju Lee. 2017. An analysis of disease-gene relationship from Medline abstracts by DigSee. *Scientific Reports*, 7:40154.

Sebastian Köhler, Nicole Vasilevsky, Mark Engelstad, Erin Foster, et al. 2017. The human phenotype ontology. *2017 Nucleic Acids Research*.

Andre Lamurias and Francisco M Couto. 2019a. Semantic similarity definition. In *Encyclopedia of Bioinformatics and Computtational Biology*, volume 1, pages 870–876. Oxford: Elsevier.

Andre Lamurias and Francisco M Couto. 2019b. Text mining for bioinformatics using biomedical literature. In *Encyclopedia of Bioinformatics and Computtational Biology*, volume 1, pages 602–611. Oxford: Elsevier.

Andre Lamurias, Diana Sousa, Luka A. Clarke, and Francisco M. Couto. 2019. BO-LSTM: classifying relations via long short-term memory networks along biomedical ontologies. *BMC Bioinformatics*, 20(1):10.

Robert Leaman and Graciela Gonzalez. 2008. BANNER: an executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing*, pages 652–663.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *arXiv e-prints*, page arXiv:1901.08746.

Manuel Lobo, Andre Lamurias, and Francisco M Couto. 2017. Identifying human phenotype terms by combining machine learning and validation rules. *BioMed Research International*, 7.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford coreNLP natural language processing toolkit. In *ACL*.

Nikolas Papanikolaou, Georgios A. Pavlopoulos, Theodosios Theodosiou, and Ioannis Iliopoulos. 2015. Protein-protein interaction predictions using text mining methods. *Methods*, 74:47–53.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 872–884. Association for Computational Linguistics.

Wenhui Xing, Junsheng Qi, Xiaohui Yuan, Lin Li, Xiaoyu Zhang, Yuhua Fu, Shengwu Xiong, Lun Hu, and Jing Peng. 2018. A gene-phenotype relationship extraction pipeline from the biomedical literature using a representation learning approach. *Bioinformatics*, 34(13):i386–i394.

1492