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

health and biomedical informatics

Motivation

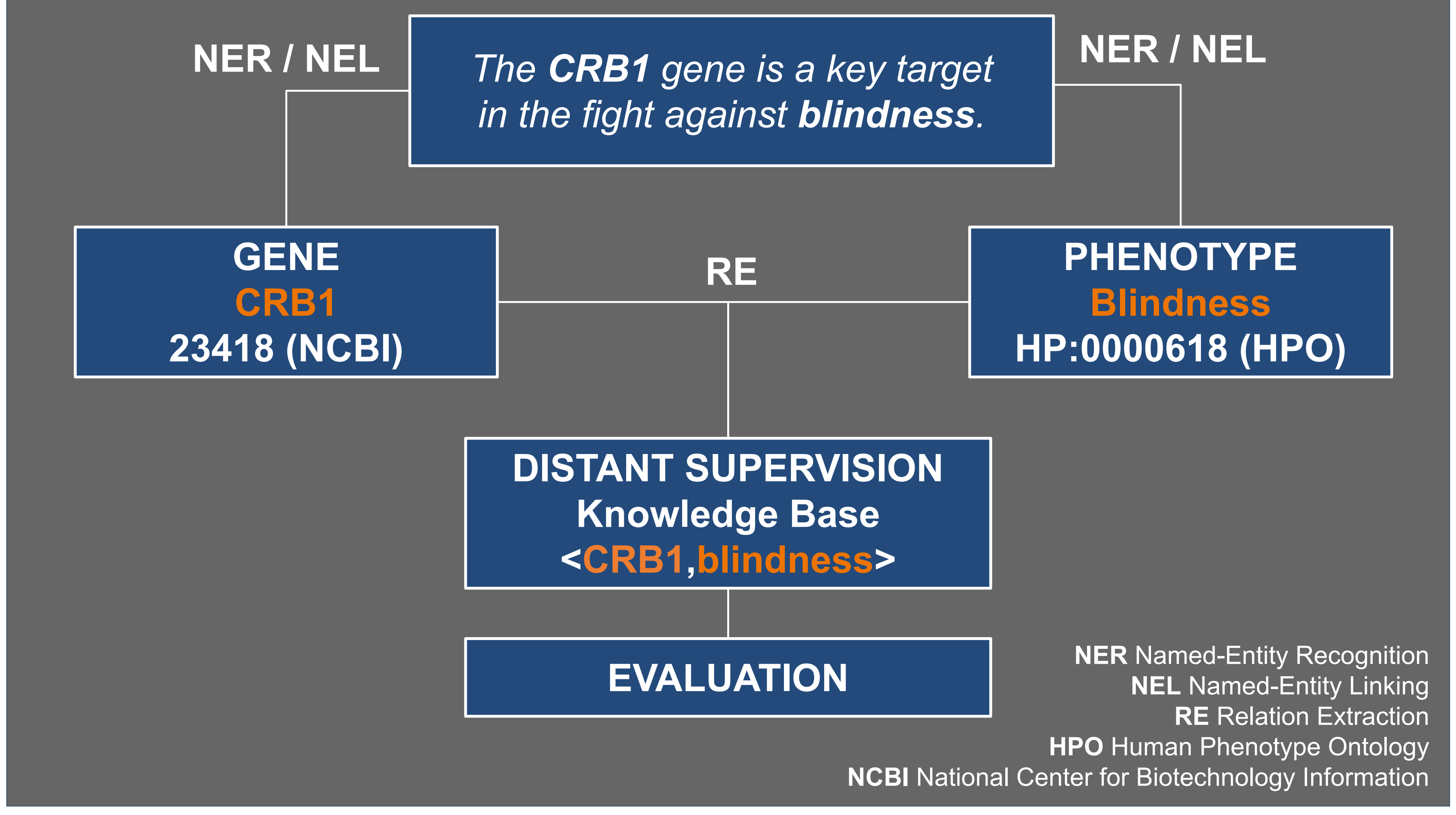
Detect the **origin of phenotypic abnormalities** and their **associated diseases** through relations expressed in biomedical literature, using Relation Extraction tools.

Relation Extraction tools require an annotated corpus and, to the best of our knowledge, there is **no corpus available** annotated with human phenotype-gene relations.

Goal

This paper presents the **Phenotype-Gene Relations (PGR)** corpus, a silver standard corpus of human phenotype and gene annotations and their relations.

Methodology



NER Named-Entity Recognition
NEL Named-Entity Linking
RE Relation Extraction
HPO Human Phenotype Ontology
NCBI National Center for Biotechnology Information

Results

Table 1. Corpus statistics. The *Known* relations are relations that are in the knowledge base and the *Unknown* relations are relations that are not yet identified or that do not exist.

Query	Abstracts	Annotations		Relations		
		Phenotype	Gene	Known	Unknown	Total
1 (10/12/2018)	1712	5676	13835	1510	2777	4283
2 (11/03/2019)	2657	9553	23786	2480	5483	7963

2 PubMed Queries
2 Named-Entity Recognition Tools:
- Minimal Named-Entity Recognizer (MER)
- Identifying Human Phenotypes (IHP)

1 HPO Knowledge Base of Gold Standard Relations
8 Curators

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

Relations		Marked Relations				Metrics		
Known	Unknown	True Positive	False Negative	False Positive	True Negative	Precision	Recall	F-Measure
77	143	67	86	10	57	87.01	43.79	58.26

87.58%
Inter-curator Agreement

Impact on Deep Learning

BO-LSTM System

A deep learning system that is used to extract and classify relations via long short-term memory networks along biomedical ontologies.

BioBERT Application

A pre-trained biomedical language representation model for biomedical text mining based on the **BERT** architecture.

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

Method	Precision	Recall	F-Measure
Co-occurrence	35.00	100.00	51.85
BO-LSTM	69.23	42.00	52.28
BioBERT	78.95	58.44	67.16

Adaptability for the creation of other RE silver standards.

github.com/lasigeBioTM/PGR

Acknowledgments

We acknowledge the help of Márcia Barros, Joana Matos, Rita Sousa, Ana Margarida Vasconcelos, Maria Teresa Cunha and Sofia Jesus in the curating phase. This work was supported by FCT through funding of DeST: Deep Semantic Tagger project, ref. PTDC/CCI-BIO/28685/2017 (http://dest.rd.ciencias.ulisboa.pt/), and LASIGE Research Unit, ref. UID/CEC/00408/2019.

U LISBOA Ciências ULisboa FCT Fundação para a Ciência e a Tecnologia LASIGE driven by excellence