Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

PheneBank: a literature-based database of phenotypes

Mohammad Taher Pilehvar [1,*], Adam Bernard [2], Damian Smedley[2] and Nigel Collier [1,*]

[1]Language Technology Lab, Department of Theoretical and Applied Linguistics, University of Cambridge, Cambridge, UK
[2]The William Harvey Research Institute, Queen Mary University of London, London, UK

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

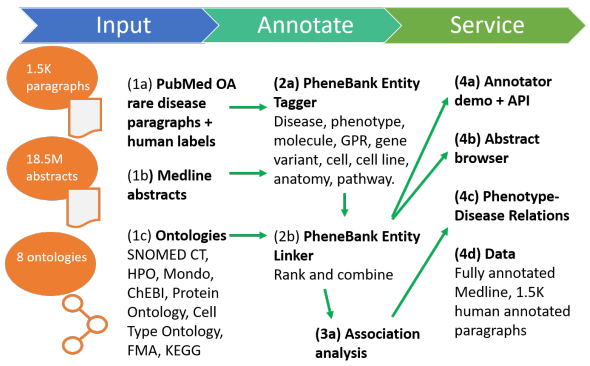
Motivation: Significant effort has been spent by curators to create coding systems for phenotypes such as the Human Phenotype Ontology (HPO), as well as disease-phenotype annotations. We aim to support the discovery of literature-based phenotypes and integrate them into the knowledge discovery process.
Results: PheneBank is a Web-portal for retrieving human phenotype-disease associations that have been text-mined from the whole of Medline. Our approach exploits state-of-the-art machine learning for concept identification by utilising an expert annotated rare disease corpus from the PMC Text Mining subset. Evaluation of the system for entities is conducted on a gold-standard corpus of rare disease sentences and for associations against the Monarch initiative data.
Availability: The PheneBank Web-portal freely available at http://www.phenebank.org. Annotated Medline data is available from Zenodo at DOI: 10.5281/zenodo.1408800. Semantic annotation software is freely available for non-commercial use at GitHub: https://github.com/pilehvar/phenebank.
Contact: nhc30@cam.ac.uk
Supplementary information: Supplementary data is available at *Bioinformatics* online.

1 Introduction

We contribute to the goal of understanding and curating human diseases by developing a high throughput Natural Language Processing (NLP) system that identifies phenotype and disease mentions in the scientific literature and links them to concept unique identifiers (CUIs) in biomedical ontologies. Integration of meaning between literature and concepts is an important task that has traditionally been accomplished using manually designed rules. In this work we apply a BiLSTM-CRF neural network in order to go beyond straightforward lexico-orthographic variations such as *carotid arteries* and *Carotid artery*. Additional challenges of matching text strings to concept labels include (i) minimal lexical overlap between synonyms such as *reduced serum calcium concentration* and *hypocalcaemia*; (ii) polysemous relations such as between *digit* and *proximal phalanges* , (iii) partial matches such as *normal hearing sensitivity* and *hearing test normal* and (iv) complex compositionality relations such as *right-sided colorectal cancer* matching to a relation between *right* and *colorectal cancer* in SNOMED CT. Neural network approaches have recently been used in concept identification systems such as PubTator Central, although to the best of our knowledge PheneBank is the first to perform concept identification of phenotypic abnormalities directly to 13K Human Phenotype Ontology terms (Köhler *et al.*, 2016). PheneBank brings together (i) API access to a state-of-the-art neural network model trained on complex sentences from full text articles for identifying concepts. The model exploits latent semantic representations (embeddings) to infer text-to-concept mappings in 8 ontologies that would often not be apparent to conventional string matching approaches; (ii) text-level recognition of phenotype-disease associations calibrated against known biological relations provided by the Monarch Initiative using HPO-Mondo mappings; (iii) text search of all Medline abstracts incorporating PheneBank concepts; (iv) fully annotated 18.5M Medline abstracts accessed and the Europe PMC Annotations API (`https://europepmc.org/AnnotationsApi`).

When constructing a concept identification model, a major bottleneck is the lack of an openly available gold standard for evaluation. To address this issue we make available the PheneBank Corpus consisting of 1.5K

Fig. 1. Overview of the PheneBank system



expert-annotated paragraphs selected from the PMC Text Mining (PMC-TM) subset[1]. We believe this corpus offers advantages over Medline sentences due to the more complex linguistic structures they represent.

2 Methods

Figure 1 shows a high-level view of the information flow through the PheneBank system. End user services are deployed on a Linux server using Apache: **Demo.** This service enables users to submit texts and receive NER annotations from PheneBank, e.g. Phenotype, Cell, Gene variant. We leverage a BiLSTM network for tagging entity mentions that exploits the desirable properties of the conventional Conditional Random Fields (CRF) approach, such as sensitiveness to neighbouring context. The model implementation is based on anaGo[2]. **Browser.** This service provides a search engine for the retrieval of automatically annotated content from 24M MEDLINE abstracts. Abstracts are annotated using the Named Entity Recognition (NER) module with all entities mapped (if possible) to a concept in one of the five major ontologies: SNOMED, HPO, MeSH, PRO, and FMA. Users enter a query and retrieve all the relevant articles. System confidence is shown by color intensity and concept details are shown by clicking on the corresponding entry. Mapping entities to concepts is carried out by unifying ontology and text entities based on lexical semantic spaces. **Relations.** This service provides an interface to view the pre-computed disease-phenotype associations. Users can enter a disease name and check for the associated phenotypes and vice versa.

3 Performance Evaluation

Entity tagging. Training and evaluating NER taggers relies greatly on the availability of human-annotated data. To our knowledge, the *Gold Standard Corpora* (Groza *et al.*, 2015, GSC) is the only phenotype-tagged dataset. A major contribution of our work is the release of a large high-quality data set tagged with 9 classes of entities, including phenotypes. Supplementary Tables S1 & S2 show the NER tagging performance of our model and four other standard NER taggers on both the GSC and PheneBank data sets. Thanks to its usage of recurrent sequence encoders, our model greatly outperforms other systems (F1 0.69 on GSC versus 0.65 and F1 0.58 on PheneBank versus 0.36). **Entity linking.** Supplmentary Table S3 shows the results for the quality of entity linking. The objective here is to map entities to corresponding concepts in HPO. For comparison we show results for the NCBO Annotator and a string-based baseline which considers edit distance. Thanks to its usage of semantic composition of representations, PhenBank is able to improve on both conventional approaches (F1 0.78 versus 0.61 and 0.55). **Phenotype-Disease associations.** We employ a co-occurrence model where we assume that if a Disease and a Phenotype co-occur in a Medline abstract then there is a manifestation relationship between them. Given the volume of potential tuples we compared a variety of statistical association measures to assess whether any are likely to represent a true biological relationship. We evaluated the Fisher Exact Test, the Dice coefficient, and pointwise mutual information (PMI) ranking tuples from most significant to least significant. Then we examined which metric best corresponded to the known tuples available from the curated associations available in the Monarch Initiative (`https://monarchinitiative.org/`). A good metric will tend to have high rankings for known tuples. Results[3] showed that the Fisher Exact Test is clearly the best-performing of the three methods. Given that the Fisher Exact Test yields p-values, we then applied the Benjamini-Hochberg procedure to our dataset to find the cutoff for a false discovery rate of 1%; this occurs at p=0.0025, after 1.8M tuples.

[1] http://demo.phenebank.org/static/phenebank-data.zip

[2] https://github.com/Hironsan/anago

4 Conclusion

PheneBank is a database of phenotypes and their associations mined from the literature using machine learning techniques. We anticipate that the database will be useful in supporting biocuration and exploring phenotype-based similarity between diseases and patients as well as downstream text mining applications.

Acknowledgements

The authors gratefully acknowledge the help of Sebastian Köhler for evaluating a subset of phenotype-disease associations and Aravind Venkatesan for facilitating upload of PheneBank data to Europe PMC.

Funding

All authors were supported by the Medical Research Council (grant MR/M025160/1). NC was supported also by the Engineering and Physical Sciences Research Council (grant EP/M005089/1). We gratefully acknowledge the donation of a GPU card from the NVIDIA Grant Program.

References

Groza, T., Köhler, S., Doelken, S. C., Collier, N., Oellrich, A., Smedley, D., Couto, F. M., Baynam, G., Zankl, A., and Robinson, P. N. (2015). Automatic concept recognition using the human phenotype ontology reference and test suite corpora. volume 2015. bav005.

Köhler, S., Vasilevsky, N. A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S. M., Boerkoel, C. F., Boycott, K. M., *et al.* (2016). The human phenotype ontology in 2017. *Nucleic acids research*, **45**(D1), D865–D876.

Lobo, M., Lamurias, A., and Couto, F. M. (2017). Identifying human phenotype terms by combining machine learning and validation rules. *BioMed Research International*, **2017**.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *ACL System Demonstrations*, pages 55–60.

Shah, N. H., Bhatia, N., Jonquet, C., Rubin, D., Chiang, A. P., and Musen, M. A. (2009). Comparison of concept recognizers for building the open biomedical annotator. *BMC Bioinformatics*, **10**(Suppl 9), S14.

Taboada, M., RodrÃguez, H., MartÃnez, D., Pardo, M., and Sobrido, M. J. (2014). Automated semantic annotation of rare disease cases: a case study. *Database*, **2014**.

[3] Data available at DOI:10.5281/zenodo.142283

5 Supplementary Information

Table S1. **Phenotype tagging performance on the GSC dataset.**

System	Precision	Recall	F1
PheneBank BiLSTM-CRF	0.69	0.69	**0.69**
IHP (Lobo *et al.*, 2017)	0.56	0.79	0.65
OBO Annotator (Taboada *et al.*, 2014)	0.69	0.44	0.54
Bio-LarK CR (Groza *et al.*, 2015)	0.65	0.49	0.56
NCBO Annotator (Shah *et al.*, 2009)	0.54	0.39	0.45

The NCBO Annotator is based on the ontologies available in BioPortal[4], the largest repository of biomedical ontologies. The OBO Annotator is a semantic Natural Language Processing tool capable of combining any number of OBO ontologies from the OBO foundry to identify their terms in a given text. Bio-LarK CR (Groza *et al.*, 2015) is an HPO concept recognition tool which defines a set of manually crafted pattern matching rules that enable capturing conjunctive terms. IHP (Lobo *et al.*, 2017) is an NER system tuned for recognizing phenotypic entities in unstructured texts. The system is based on Stanford CoreNLP (Manning *et al.*, 2014) for text preprocessing and Conditional Random Fields (CRF) for named entity recognition (NER). The CRF model leverages a rich set of features including linguistic, lexical, morphologic, orthographic, lexical, and context features. The system also benefits from a validation step that can filter incorrect annotations based on a set of manually crafted rules, such as the negative connotation analysis. We report results provided by Lobo *et al.* (2017): NCBO API[5] targeted towards the HPO, the HPO-specific version of OBO Annotator available. linked to lexicons such as HPO.

Table S2. **Phenotype tagging performance on the PheneBank dataset.** We experimented with two settings: (1) phrases are regarded as whole units; partially tagging the phrases would not count towards correct results; (2) phrases are regarded as multiple disjoint entities; tagging any of the words counts toward overall performance.

	System	Precision	Recall	F1
Setting 1	BiLSTM-CRF	0.59	0.57	**0.58**
	IHP (Lobo *et al.*, 2017)	0.27	0.55	0.36
Setting 2	BiLSTM-CRF	0.78	0.79	**0.79**
	IHP (Lobo *et al.*, 2017)	0.49	0.58	0.53

Table S3. **Results for grounding to HPO entities.**

System	Accuracy
PheneBank - Semantic grounding	**0.78**
NCBO Annotator Shah *et al.* (2009)	0.61
Exact match baseline	0.55

[4] https://bioportal.bioontology.org/

[5] http://data.bioontology.org/ documentation