Deep Semi-supervised Ensemble Method for Classifying Co-mentions of Human Proteins and Phenotypes

Morteza Pourreza Shahri and Indika Kahanda

July 13, 2020

Introduction: Identifying protein-phenotype relations is of paramount importance for applications such as uncovering rare and complex diseases. Human Phenotype Ontology (HPO) is a recently introduced standard vocabulary for describing disease-related phenotypic abnormalities in humans. Since the experimental determination of HPO categories for human proteins is a highly resource-consuming task, developing automated tools that can accurately predict HPO categories has gained interest recently [1].

The primary resource that capture protein-phenotype relationships is the biomedical literature. Two key steps involved in an automated curation pipeline can be identified as the following: (1) extracting all protein-phenotype co-mentions from biomedical literature, and (2) classifying extracted co-mentions for identifying valid relations. In our previous study [2], we developed a supervised SVM-based model called PPPred (Protein-Phenotype Predictor), which, to the best of our knowledge, is the first machine learning classifier that can classify a given sentence-level co-mention of a human protein and a phenotype. Using a gold standard dataset composed of manually curated sentence co-mentions, we demonstrated that PPPred significantly outperformed several baseline methods while allowing for noticeable room for improvement.

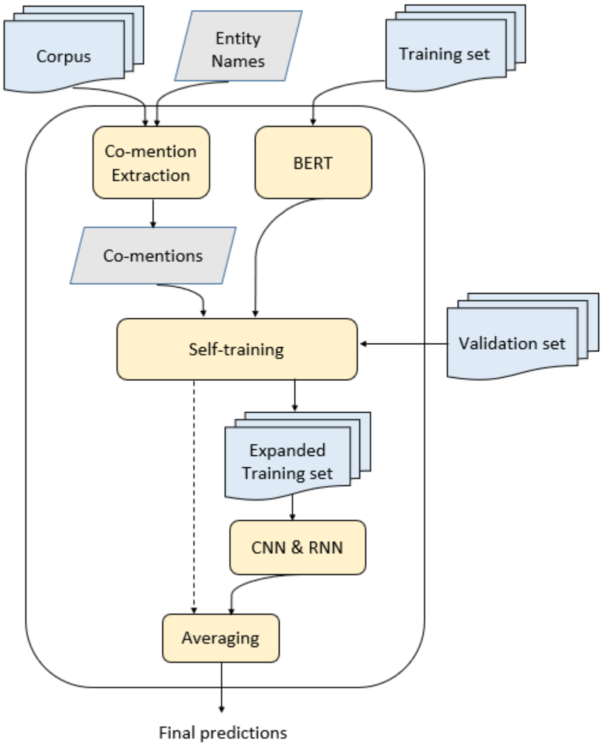

Figure 1: The proposed model

Methodology: In this work, we develop a deep semi-supervised ensemble method for the task of protein-phenotype co-mention classification. This novel model is motivated by the fact that while the manual annotation of co-mentions is extremely prohibitive, technically we have access to millions of unlabeled protein-phenotype co-mentions. Hence, in this study, we explore the feasibility of incorporating such unlabeled data for improving the overall predictive accuracy by developing a model that is depicted in Figure 1.

Our model is composed of several components. The *Co-mention Extraction* component extracts all co-mentions of protein-phenotype sentence-level co-mentions using the *corpus* and *entity names* as input. The corpus is a combination of 27 million abstracts and 1.7 million full-text articles retrieved from ProPheno online database developed in a previous study [3]. The entity names are protein names taken from UniProt, and phenotype names taken from Human Phenotype Ontology. At the same time, a supervised BERT (Bidirectional Encoder Representations from Transformers) model is trained using the labeled *Training set* data. The *Self-training* component uses this trained BERT model to make predictions on the extracted co-mentions. Using the labeled *Validation set* data, it selects the top $5,000$ co-mentions with the highest predicted confidence scores to generate the *Expanded Training Set*. This set is then used to train a convolutional neural network (CNN) & recurrent neural network (RNN) model which are averaged to get the final predictions. These two models employ Word2Vec word embeddings and the shortest dependency path between protein and phenotype entities as input.

Experimental Results: We use PyTorch for implementing neural networks. The length of sequences is set to 80 and 100-dimension Word2Vec word embeddings are trained on the entire corpus. Both CNN and RNN models are trained for 20 epochs. We also fine-tune the BERT model in four epochs and employ binary cross-entropy loss and Adam optimizer. We compare our ensemble model with PPPred [2], and S3VM [4], which is a well-known method using semi-supervised Support Vector Machines model. We report precision, recall, F1, and AUROC scores using the labeled *Test set*. The training, validation, and test sets have sizes of $1,010$, 337, and 337, respectively. The labeled data were annotated by two biologists (Cohen's kappa = 0.64).

Our results indicate that the proposed deep semi-supervised ensemble model significantly improves performance in comparison with PPPred and S3VM (see Table 1). We are in the process of setting up a web service that utilizes the proposed model, which returns a list of the most relevant sentences for an input pair of human protein and phenotype names. These sentences will be ranked by their corresponding confidence scores. The findings and the insight of this work have implications for biocurators, researchers, and bio text mining tool developers.

Table 1: Comparison of the proposed model with PPPred and S3VM

Method	**Precision**	**Recall**	**F1**	**AUROC**
PPPred	0.890	0.910	0.900	0.845
S3VM	0.854	0.785	0.818	0.761
Our model	0.898	0.972	0.936	0.876

References

[1] Morteza Pourreza Shahri and Indika Kahanda. Extracting co-mention features from biomedical literature for automated protein phenotype prediction using PHENOstruct. In *10th International Conference on Bioinformatics and Computational Biology*, pages 123–128. ISCA, 2018.

[2] Morteza Pourreza Shahri, Gillian Reynolds, Mandi Marie Roe, and Indika Kahanda. PPPred: Classifying protein-phenotype co-mentions extracted from biomedical literature. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 414–422. ACM, 2019.

[3] Morteza Pourreza Shahri and Indika Kahanda. ProPheno 1.0: An online dataset for accelerating the complete characterization of the human protein-phenotype landscape in biomedical literature. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pages 416–423. IEEE, 2020.

[4] Kristin P Bennett and Ayhan Demiriz. Semi-supervised support vector machines. In *Advances in Neural Information processing systems*, pages 368–374, 1999.