DUTIR-BioNLP@BC8 Track 3: Genetic Phenotype Extraction and Normalization with Biomedical Pre-trained Language Models

Jiewei Qi, Ling Luo*, Zhihao Yang, Hongfei Lin

School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China

*Corresponding author: lingluo@dlut.edu.cn

Abstract

It is important to automatically extract and normalize key medical findings from the observation results written during the physical examination of teratology. The BioCreative VIII Track 3 endeavors to facilitate the advancement and assessment of systems designed to automatically extract and normalize the phenotype entities from electronic health records (EHRs). This paper describes our method used to create our submissions to the track. Our pipelined method for the phenotype concept extraction partitions the process into two subtasks: Named Entity Recognition and Named Entity Normalization. The cutting-edge biomedical pre-trained language models are used for both subtasks. Then the ensemble method is further used to improve the final performance. The official results on the test set show that our best submission achieves the F1-scores of 0.7632 on Subtask 3a and 0.7112 on Subtask 3b.

Keywords: Phenotype Normalization; Pre-trained Models; Biomedical Concept Recognition

Introduction

A physical examination to identify abnormal phenotypic characteristics is instrumental in determining the underlying genetic etiology of a pathological condition. Phenotypic findings have a direct impact on clinical diagnosis, the determination of appropriate genetic analyses, and the interpretation of testing outcomes, especially in cases where variants of uncertain clinical significance are detected. Moreover, phenotypic data proves valuable for researchers aimed at delineating novel genetic disorders and augmenting current comprehension of known conditions. However, the phenotypes are often stored as unstructured text in electronic health records (EHRs), which makes them difficult to use for further computational analysis. Extracting human phenotype concepts manually from text is labor-intensive and time-consuming. Recently, some natural language processing methods (such as PhenoTagger (1) and Doc2Hpo (2)) have been developed to automate the identification of phenotypic concepts.

To facilitate the automatic extraction of phenotype information from EHRs, BioCreative VIII organized Track 3 for the phenotype normalization task. This task included two subtasks. Subtask 3a required participants to submit the Human Phenotype Ontology (HPO) term identifiers (IDs) of all key phenotype findings mentioned in the observation. Subtask 3b required participants to submit the spans of the key findings and their corresponding HPO term IDs. We participated in

both two subtasks and developed a deep learning-based pipeline approach using the biomedical pre-trained language models. The official results on the test set show that our best submission achieves the F1-scores of 0.7632 and 0.7112 on the subtasks 3a and 3b, respectively.

Methods

In this track, the official corpus included 1,716 de-identified observations for training, 454 de-identified observations for development, and the testing phase comprises of 966 de-identified observations, supplemented with 2,427 decoy observations. To standardize the description of dysmorphic findings, we used the HPO in version 2022-10-05, which includes 17,061 mappings of HPO IDs and HPO Terms. According to the characteristic of the corpus, the training set contained about 12% of entities with negative polarity, which indicated that this kind of discovery was not present in the observation result, and there are about 14% of discontinuous entities. Therefore, the additional challenges of this task include the identification of negations and the extraction of discontinuous entities. To effectively utilize the training dataset and address these challenges comprehensively, we proposed a pipeline method based on our previous work PhenoTagger. Specifically, we partitioned the extraction process into two subtasks and addressed them incrementally: first, we used W^{2}NER (3) for named entity recognition (NER), and then we used the classification part of the PhenoTagger method for named entity normalization (NEN). Finally, the ensemble method and post-processing rules are further used to improve the final performance. The overview of our method is shown in Figure 1.

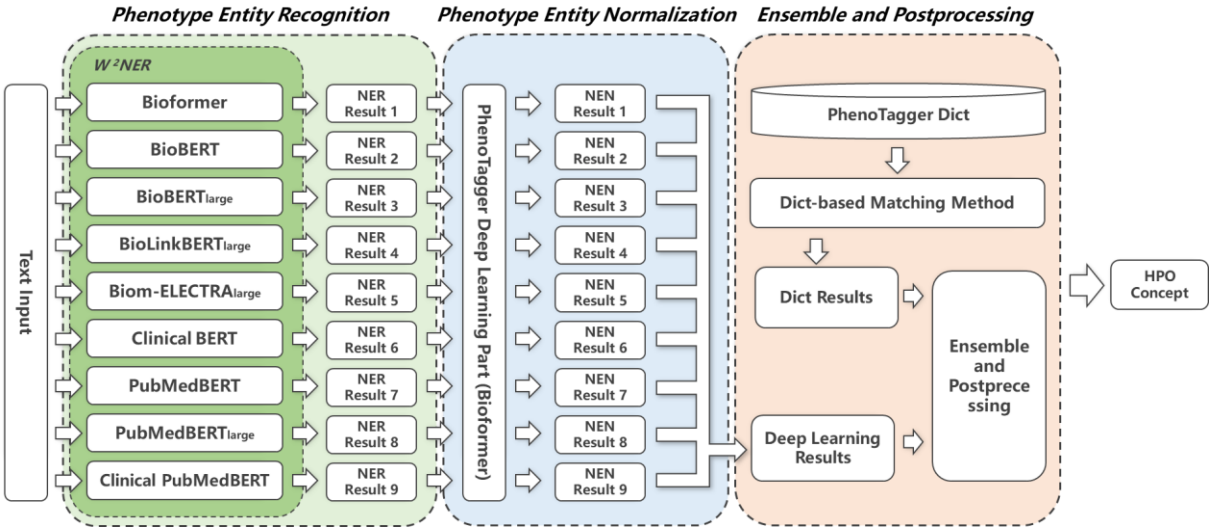

Figure 1 Overview of our method for genetic phenotype extraction and normalization.

A. Phenotype Entity Recognition

In the phenotype entity recognition part, how to efficiently and accurately identify entities was our main focus. After analyzing the data set, we found that 80% of the training set and development set are flat entities, and about 20% of the entities are discontinuous entities. While our previous work with PhenoTagger demonstrates competent recognition of flat phenotype entities utilizing the HPO ontology for distant supervision, this approach faces limitations in effectively identifying discontinuous entities. In order to maximize the utility of the training dataset and tackle the challenge of discontinuous entities, we applied the W^{2}NER methodology for phenotype entity recognition. This method has exhibited robust performance in recognizing both flat and

discontinuous entities by transforming the NER task into predicting the relationship categories between word pairs. We tried Bioformer (4), BioBERT (5), BioLinkBERT$_{large}$ (6), Biom-ELECTRA$_{large}$ (7), Clinical BERT (8), PubMedBERT (9) and Clinical PubMedBERT (10) models for W^{2}NER. Both BioBERT and PubMedBERT also include large versions of pre-trained models. We used the training set to train the above models and then identified the entity mentions on the test set as the NER results.

B. Phenotype Entity Normalization

In the entity normalization part, we used the deep learning-based classification method in PhenoTagger to classify the candidate entities extracted by the W^{2}NER into a specific HPO ID. In this part, we also tried Bioformer, BioBERT and PubMedBERT models. According to the performance of these models on the development set, we chose Bioformer as the final classification model.

The candidate entities are classified by the softmax layer after passing through the Bioformer, and the output of the softmax layer is a probability score. We manually set a threshold and keep the results with a probability greater than this threshold, while discarding the results with a probability less than this threshold. We selected two thresholds to generate the NEN results: 0.8 and 0.95. When the threshold was set to 0.8, the model had a higher recall rate on the development set. When the threshold was set to 0.95, the model had a higher accuracy rate on the development set. Generally, when evaluating the performance of the development set across identical indicators, a threshold set at 0.8 yielded a better F1 score.

C. Ensemble and Postprocessing

As delineated in Section A, our methodology incorporated nine distinct models for entity recognition, inclusive of large-scale pre-trained models. Subsequent to the process of entity normalization, each model generated a unique recognition result. We then employed a voting ensemble method based on the recognition results derived from the nine models. Specifically, we set a threshold *m*. The entity that is predicted by more than *m* models is selected as the ensemble result. We conducted the experiment on the development set in which *m* was configured to 2, 3, 4, and 5 respectively. The results show that the best performance of the ensemble is achieved when *m* is set to 2.

For the ensemble result, we performed postprocessing rules for the overlapped and negative entities：

(1) Remove overlapping recognition results: When the recognized spans are overlapped, keep the one with the longest entity mention.
(2) Remove negative polarity results: When the recognition results contain negative words such as “no”, “not”, “without”, “abnormal”, etc., they are considered as negative results and removed from the results.

Finally, we added the entities recognized by the dictionary part of PhenoTagger as a supplement to the merged results, because the dictionary-based method has high accuracy and can be directly used as a supplement to the final result.

Results and Discussion

Table 1 shows the results (Subtask 3a: Standard Precision, Recall and F1-score; Subtask 3b: Strict Precision, Recall and F1-score) of the original PhenoTagger method with different pretrained-models on the validation set. The results show that the Bioformer model achieves better

performance than the other two models. Given that subtask 3a serves as the primary evaluation criterion for this track, our selection aligns with the adoption of Bioformer.

Table 1 Results of the original PhenoTagger on the validation set

Model	Subtask 3a Normalization			Subtask 3b exactExtNorm		
	P	R	F1	P	R	F1
PhenoTagger (Bioformer)	**80.20**	64.75	**71.65**	**78.35**	57.83	66.54
PhenoTagger (BioBERT)	77.71	66.06	71.42	75.89	**59.64**	**66.79**
PhenoTagger (PubMedBERT)	77.46	**66.23**	71.40	75.57	**59.64**	66.67

Table 2 shows the results of our pipeline method (i.e., W^{2}NER for NER then PhenoTagger for NEN) and their ensemble on the validation set. Here, we set 0.8 as the threshold for the Phenotype entity normalization part. From the results, we can see that the W2NER(PubMedBERTlarge) with PhenoTagger model achieves better performance than other models and it obtains improvements of 4.77% and 6.63% in F1-scores than the original PhenoTagger in Subtasks 3a and b, respectively. The final ensemble results achieve the highest F1-score.

Table 2 Results outcomes of all nine models used in W^{2}NER on the validation set

Model	Subtask 3a Normalization			Subtask 3b exactExtNorm		
	P	R	F1	P	R	F1
Original PhenoTagger	80.20	64.75	71.65	78.35	57.83	66.54
W^{2}NER(Bioformer)+PhenoTagger	85.19	68.20	75.76	84.25	63.43	72.37
W^{2}NER(BioBERT)+PhenoTagger	84.38	66.72	74.52	83.55	62.77	71.68
W^{2}NER(BioBERT$_{large}$)+PhenoTagger	84.65	69.03	76.04	83.73	64.42	72.81
W^{2}NER(BioLinkBERT$_{large}$)+PhenoTagger	82.69	70.84	76.31	81.67	66.06	73.04
W^{2}NER(Biom-ELECTRA$_{large}$)+PhenoTagger	84.71	69.36	76.27	83.76	64.58	72.93
W^{2}NER(Clinical BERT)+PhenoTagger	85.07	68.53	75.91	84.20	64.09	72.78
W^{2}NER(PubMedBERT)+PhenoTagger	85.48	67.88	75.67	84.65	63.59	72.63
W^{2}NER(PubMedBERT$_{large}$)+PhenoTagger	**85.83**	68.86	76.42	**84.97**	64.25	73.17
W^{2}NER(Clinical PubMedBERT)+PhenoTagger	83.70	68.53	75.36	82.84	64.42	72.48
Ensemble Result	85.63	**74.63**	**79.75**	84.36	**67.55**	**75.02**

During this task, we submitted three runs as our final submissions. Our submitted three runs in the main task are based on the following configurations.

- Run 1: The nine NER models are trained on the training set and the number of training epochs is chosen by early stopping strategy by the performance on the development set (50 epochs at most). Then the models with the NEN threshold of 0.8 and 0.95 are evaluated on the development set, respectively. The better threshold for each model is selected to generate the final results.
- Run 2: We augmented the training set with the development set, and subsequently retrained the nine NER models with 30 epochs. Then the nine models with the NEN threshold of 0.95 are used to generate final results.
- Run 3: All nine NER models in Run 1 and nine NER models in Run2 with the NEN threshold of 0.8 are used to generate final results.

Table 3 shows the overall results of our runs on the official test set. Run 2 achieves the highest overall F1-scores on both subtasks. And 'MEAN' represents the official mean of this task. Our results are all above the average, thus show that our method is effective.

Table 3 Overall results on the test set and the average result

	Subtask 3a Normalization			Subtask 3b exactExtNorm		
	P	R	F1	P	R	F1
Run1	79.18	72.26	75.56	77.19	**64.31**	70.16
Run2	**83.07**	70.59	**76.32**	**81.44**	63.12	**71.12**
Run3	78.02	**72.81**	75.33	75.57	63.43	68.97
MEAN	-	**-**	72.59	-	-	66.84

Conclusion

In this paper, we present a pipeline approach based on pre-trained language models for the phenotype concept recognition task. We leveraged both the W^2NER and PhenoTagger methods and further enhanced overall performance through the implementation of ensemble techniques. The experimental results show that our approach successfully extracts phenotypic concepts from free text.

Funding

This work was supported by the Fundamental Research Funds for the Central Universities [DUT23RC(3)014].

References

1. Luo L, Yan S, Lai P T, et al. (2021) PhenoTagger: a hybrid method for phenotype concept recognition using human phenotype ontology. *Bioinformatics*, 37(13), 1884-1890.

2. Liu C, Peres Kury F S, Li Z, et al. (2019) Doc2Hpo: a web application for efficient and accurate HPO concept curation. *Nucleic acids research*, *47*(W1), W566-W570.

3. Li J, Fei H, Liu J, et al. (2022, June) Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 10, pp. 10965-10973).

4. Fang L, Chen Q, Wei C H, et al. (2023). Bioformer: an efficient transformer language model for biomedical text mining. arXiv preprint arXiv:2302.01588.

5. Lee J, Yoon W, Kim S, et al. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234-1240.

6. Yasunaga M, Leskovec J, Liang P. (2022, June). LinkBERT: Pretraining Language Models with Document Links. In *ICML 2022 2nd AI for Science Workshop*.

7. Alrowili S, Vijay-Shanker K. (2021, June). BioM-transformers: building large biomedical language models with BERT, ALBERT and ELECTRA. In *Proceedings of the 20th workshop on biomedical language processing* (pp. 221-227).

8. Alsentzer E, Murphy J R, Boag W, et al. (2019). Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

9. Gu Y, Tinn R, Cheng H, et al. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, *3*(1), 1-23.

10. Taylor N, Zhang Y, Joyce D W, et al. (2023). Clinical prompt learning with frozen language models. *IEEE Transactions on Neural Networks and Learning Systems*.