Exploring the Use of Ontology Components for Distantly-Supervised Disease and Phenotype Named Entity Recognition

Sumyyah Toonsi[1,2,†], Şenay Kafkas[1,2,†] and Robert Hoehndorf[1,2,*]

[1]*Computer, Electrical and Mathematical Sciences & Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955, Kingdom of Saudi Arabia*

[2]*Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955, Kingdom of Saudi Arabia*

Abstract

The lack of curated corpora is one of the major obstacles for Named Entity Recognition (NER). With the advancements in deep learning and development of robust language models, distant supervision utilizing weakly labelled data is often used to alleviate this problem. Previous approaches utilized weakly labeled corpora from Wikipedia or from the literature. However, to the best of our knowledge, none of them explored the use of the different ontology components for disease/phenotype NER under the distant supervision scheme. In this study, we explored whether different ontology components can be used to develop a distantly supervised disease/phenotype entity recognition model. We trained different models by considering ontology labels, synonyms, definitions, axioms and their combinations in addition to a model trained on literature. Results showed that content from the disease/phenotype ontologies can be exploited to develop a NER model performing at the state-of-the-art level. In particular, models that utilised both the ontology definitions and axioms showed competitive performance compared to the model trained on literature. This relieves the need of finding and annotating external corpora. Furthermore, models trained using ontology components made zero-shot predictions on the test datasets which were not observed by the models training on the literature based datasets.

Keywords

Named Entity Recognition, Text mining, ontologies

1. Introduction

Named Entity Recognition (NER) is a form of Natural Language processing (NLP) that aims to identify and classify named entities such as organisation, person, disease and genes in text. NER is a challenging task due to the nature of language which includes abbreviations, synonymous entities, and in general variable descriptions of entities.

Early methods for NER used dictionaries due to their applicability and time efficiency. Lexical approaches such as the NCBO (National Center for Biomedical Ontology) annotator [1], ZOOMA

ICBO'23: International Conference on Biomedical Ontology,

*Corresponding author.

†These authors contributed equally.

sumyyah.toonsi@kaust.edu.sa (S. Toonsi); senay.kafkas@kaust.edu.sa (Kafkas); robert.hoendorf@kaust.edu.sa (R. Hoehndorf)

[2], and the OBO (Open Biological and Biomedical Ontologies) annotator [3] are not able to recognise new concepts and cannot detect all variations of expressions. This is because once dictionaries are constructed with terms, they can only find exact matches to those terms. Hence, dictionary-based approaches suffer from low recall.

With the emergence of machine learning, better NER methods were developed. This was possible through exposing statistical models to curated text where mentions of entities are identified by human curators and provided to these models. Subsequently, these models were able to generalize to unseen entities better than previous methods. For instance, GNormPlus [4] was developed to find gene/protein mentions using a supervised model which demonstrated competitive results at the time. Although supervised methods showed remarkable improvements in performance, they require curated instances for the model to learn. That is, the model expects instances of text where mentions of entities are clearly provided to learn to distinguish concepts of interest. This becomes a serious problem when one wants to recognise a novel/unexplored concept. Moreover, supervised methods often fail to recognise concepts uncovered by the curated corpora.

To alleviate the need for curated corpora, distant-supervision was explored for NER. In particular, distantly supervised models are trained on a weakly labeled training set, i.e., obtained from an imprecise source. For instance, dictionaries could be used to annotate text with exact matches which can produce both false positives and false negatives. Methods like BOND[5], PatNER[6], ChemNER[7], PhenoTagger [8], Conf-MPU [9], and Dong et al. [10] demonstrated the potential of distant supervision for NER. The aforementioned methods created weakly labeled sets using labels and synonyms found in ontologies/vocabularies to extract training instances from unlabeled corpora. Later, these instances were used to train different models which in some cases outperformed state-of-the-art methods.

Inspired by the advances achieved by distant supervision, we explored the contribution of different components of ontologies (Labels and synonyms, definitions, and complex axioms) to the task of NER under the distant supervision scheme. In all of the previously mentioned distantly-supervised NER methods, only labels and synonyms of ontologies/vocabularies were used to create the weakly labeled corpora from literature. The use of different ontology components to develop NER models has not been comprehensively explored for diseases/phenotypes. In addition to the use of labels and synonyms, in this study, we go a step further to explore the use of definitions and axioms to develop a disease/phenotype NER model. We hypothesize that the dense and rich knowledge found in ontologies can be used to develop NER models without the need of external corpora such as literature abstracts. We conducted our experiments on disease and phenotype entity recognition because, the study of diseases and phenotypes is important for understanding disease diagnosis, treatment and epidemiology.

2. Materials and Methods

2.1. Ontologies, literature resource and benchmark corpora

2.1.1. Ontologies

We used the Disease Ontology (DO) [11] on 15/April/2022) (downloaded on 1/March/2022) and the MEDIC vocabulary [12] in our study. DO is an ontology from the Open Biomedical Ontologies (OBO) [11], whereas MEDIC is a vocabulary of disease terms represented in the Web Ontology Language (OWL) [12]. We used the Human Phenotype Ontology (HPO) [13] (downloaded on 5/Jan/2022) for the phenotype concepts.

2.1.2. Literature

We used Medline [14] as a literature resource to generate our abstract-based weakly labeled dataset. To select abstracts that cover ontology concepts, we used an in-house index covering 32,923,095 Medline records (downloaded on Dec-15-2022) generated using Elasticsearch [15].

2.1.3. Benchmark corpora

To evaluate the named entity recognition models, we used four benchmark corpus; the NCBI–Disease Corpus [16] and the MedMentions Corpus (disease and phenotype) [17] and GSC+ [18]. NCBI–Disease is a widely used corpus where disease mentions are annotated and reviewed by multiple annotators. MedMentions is a large corpus annotated by an extensive set of Unified Medical Language System (UMLS) concepts. We selected the abstracts with disease annotations from MedMentions and named this the MedMentions–disease Corpus. To form this corpus, we used UMLS-to-MESH mappings from UMLS to obtain the MESH codes and selected the disease concepts which exist in our disease dictionary (described in section 2.2). Similarly, we selected the abstracts with phenotype concepts where we found mappings from UMLS-to-HPO and named this dataset as MedMentions–phenotypes. GSC+ is a widely used benchmarking dataset covering phenotype concepts particularly from HPO. We used the test dataset version released by [8]. Table 1 shows the distribution of the abstracts and annotations in the four benchmark corpora.

Table 1
Statistics of benchmark corpora

Corpus	Abstracts	Annotations
NCBI–disease train	593	5146
NCBI–disease dev	100	788
NCBI–disease test	100	960
MedMentions–disease test	879	3726
MedMentions–phenotype train	1291	6772
MedMentions–phenotype dev	428	2287
MedMentions–phenotype test	405	2190
GSC+ test	228	1933

2.2. Dictionary generation

We generated and used two dictionaries to weakly label Medline abstracts for disease and phenotype concepts. To generate our dictionaries, first, we extracted the labels and synonyms of all concepts from MEDIC, DO and HPO. Second, we filtered out the possible ambiguous labels/synonyms which are often stop words, short labels/synonyms (1 or 2 character long) and labels/synonyms shared by two different concepts from the dictionary. For example, DO contains a synonym which is "go" for the "geroderma osteodysplasticum" concept (DOID:0111266). The synonym "go" is ambiguous with the verb "go". Filtering out ambiguous names is a common practice used in text mining workflows that rely on lexical matches. We used the Natural Language Toolkit (NLTK) stop words [19] and filtered out any exact match with the labels/synonyms in MEDIC and DO and HPO. In both sources, we did not find any match with the list of stop words. We also filtered out the labels/synonyms having less than 3 characters to avoid false positives. Additionally, for the generation of the dictionary for diseases, we filtered out all the disease labels/synonyms which exactly match with protein labels/synonyms from the HUGO Gene Nomenclature Committee (HGNC) Database [20] to avoid false positive matches with protein names. Third, we generated the plural form of each label/synonym by using the Inflect Python module [21]. For example, the module generates “tetanic cataracts” for the given multi-word term, “tetanic cataract” (DOID:13822). Our final disease dictionary covers 244,903 disease labels and synonyms of 29,374 distinct concepts from MEDIC and DO. The final phenotype dictionary covers 79,010 phenotype labels and synonyms of 14,631 distinct concepts from HPO.

2.3. Ontology components used

An ontology O, as previously described in [22], has four main components:

- Classes and relations, where classes and relations are assigned unique identifiers.
- Domain vocabulary, where labels and synonyms are linked to ontology classes and relations.
- Textual definitions, where descriptions about classes and relations are provided, usually in natural language.
- Formal axioms, where relations between concepts are described in some formal language and possibly linked to other ontologies and sources.

We used labels and synonyms, textual definitions, and formal axioms components separately to create weakly labeled corpora and the statistics are reported in Table 2.

Table 2
Statistics of used ontology components

Component	DO and MEDIC	HPO
Labels/synonyms	35,333	16,307
Definitions	9,435 and 19,939 dummy	10,202 and 2,451 dummy
Axioms	30,834	37,062

2.4. Training dataset construction

2.4.1. Abstracts from literature

To generate the training set for distant supervision, first, we retrieved the relevant literature by searching the indexed Medline for the exact match of each label/synonym from the dictionaries. We retrieved the top [1-5] Medline abstracts/titles hits per concept that is identified based on the default Elastic Search Engine relevance scoring settings (TF-IDF [23] based scoring). Second, we used the dictionaries and annotated the downloaded abstracts lexically and converted the annotations to the I-O-B format (a common format for tagging tokens in a chunking task where B indicates the first token (Beginning) of an annotation, I subsequent (Inside) token of the same annotation and O representing a token that is not annotated (Outside)) [24] by using spaCy [25]. Finally, we obtained two sets of corpora; one for the disease concepts and the other for the phenotype concepts. We found 16,307 distinct phenotype labels/synonyms belonging to 6,962 classes from HPO in at least one Medline record by searching the indexed literature. These concepts are covered by 16096, 31372, 46032, 60098 and 74087 distinct Medline abstracts/titles at top 1, 2, 3, 4, 5 hits respectively, and we used them as our training sets for phenotypes. We found 35,333 distinct disease labels/synonyms linked to 8,400 distinct concepts from MEDIC and DO in at least one Medline records. These concepts are covered by 41698, 81007, 118295, 154060 and 187462 distinct Medline abstracts/titles at top 1, 2, 3, 4, 5 hits respectively and we used as our training sets for disease concepts.

Table 3
Example of using the class DOID:0040099 to create different weakly labeled sets. Text in bold refers to text annotated as B/I classes in the IOB format.

Component	Ontology representation	Dataset representation
Labels	name: Livedoid vasculitis	**Livedoid vasculitis**
Synonyms	synonym: "livedoid vasculopathy" EXACT	**Livedoid vasculopathy**
Axioms	DOID:0040099 SubClassOf DOID:865	**Livedoid vasculitis** is a **vasculitis**
Definitions	"A vasculitis with purpuric ulcers."	A **vasculitis** with purpuric **ulcers**.

2.4.2. Labels and synonyms

Using the direct labels and synonyms from ontologies, we created two sets for phenotypes and diseases. For phenotypes, the labels and synonyms extracted from HPO were directly considered as positives as shown in Table 3. We used the labels and synonyms from DO and added MEDIC as well. The labels and synonyms were retrieved from the dictionary described in 2.2.

2.4.3. Definitions

Definitions in DO are available in natural language. To associate the concept with its definition, we added the concept label/synonyms to the beginning of a definition as shown in Table 3. For concepts which lacked definitions, we simply included their labels/synonyms with a dummy sentence replicated for all. For instance, if a disease X does not have a definition, its dummy definition is "X is a disease". Since definitions can included other concepts (e.g. parent concepts)

in their description, mentions of such concepts can be troublesome. To partially resolve this issue, we annotated the definitions with the dictionaries described in 2.2 Matches against the dictionaries were treated as positive mentions of concepts. In total, we retrieved 9,435 definitions from DO and used dummy definitions for 19,939 concepts. For phenotypes, we included definitions for 10,202 concepts and used dummy definitions for 2,451 concept.

2.4.4. Axioms

Axioms are not readily available for natural language tasks since they are expressed in formal language. To tackle this issue, we first processed axioms as previously described in [26]. Next, we replaced ontology identifiers with their labels/synonyms. We also included axioms which reference external ontologies and replaced their identifiers with names as shown in Table 3.

For diseases, we used 30,834 axioms from DO. For phenotypes, we included 37,062 axioms from HPO. Axioms of both concepts included references to external ontologies which we downloaded and processed to map their identifiers to their names. The external ontologies that were included are: the Basic Formal Ontology (BFO) [27], the Chemical Entities of Biological Interest (ChEBI) [28], the Cell Ontology (CL) [29], the Gene Ontology (GO), the Relation Ontology (RO) [30], and the Uber-anatomy Ontology (UBERON) [31].

2.5. Named entity recognition using distant supervision

NER refers to identifying boundaries of entity mentions in text (disease and phenotype mentions in our case). We used distant supervision to train our models by using BioBERT to recognise disease and phenotype mentions in text. Figure 1 depicts the system overview.

BioBERT is a BERT (Bidirectional Encoder Representations from Transformers) [32] pre-trained language model based on large biomedical corpora. BERT is a contextualized word representation model trained using masked language modeling. It provides self-supervised deep bidirectional representations from unlabeled text by jointly conditioning on both left and right contexts. The pre-trained BERT model can be fine-tuned with an additional output layer to generate models for various desired NLP tasks. We used *simpletransformers* [33] which provides a wrapper model to distantly supervise an entity recognition model. More specifically, the wrapped model is used to fine-tune BERT models by adding a token-level classifier on top that classifies tokens into one of the output classes which are I-O-B (Inside-Outside-Beginning). In the training phase, our models are initialised with weights from BioBERT-Base v1.1 [34] and then fine-tuned on the disease and phenotype entity recognition task using our training corpora.

3. Results

We set up our experiments on four separate benchmarking corpora covering phenotype and disease concepts; NCBI–disease, MedMentions–disease, MedMentions–phenotype and GCS+. We reported our NER results using the Precision, Recall and F-score metrics. We used a relaxed scheme to calculate the metrics where we considered any partial overlap between the prediction and the curated annotations to be a true positive. That is, predictions are considered to be

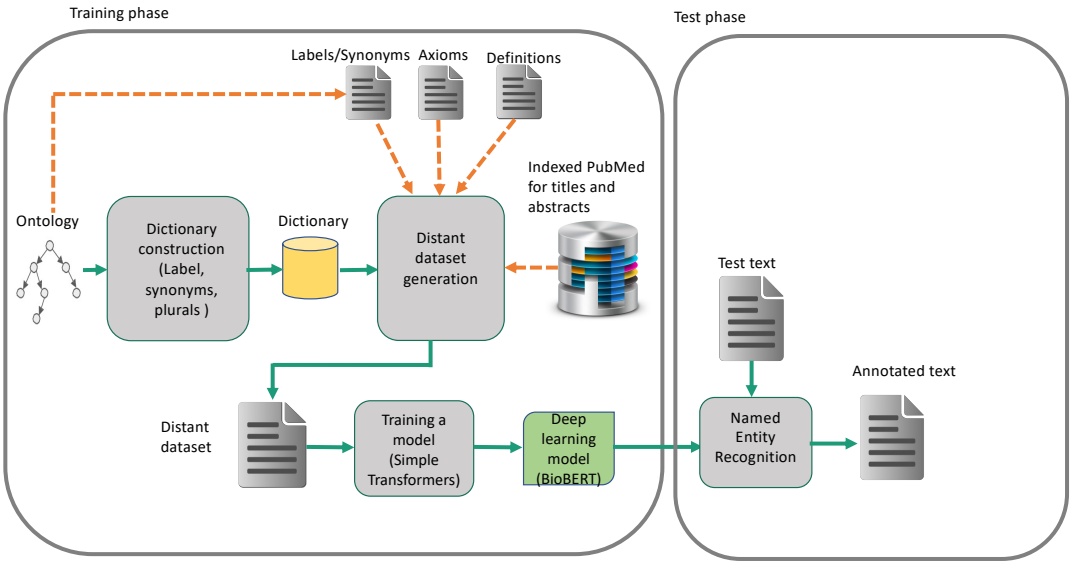

Figure 1: System Overview
This figure depicts the training and test phases in our system. In the training phase, we used ontologies to create a dictionary from the labels, synonyms and their plural forms. We used this dictionary to create distant datasets from Medline abstracts and different ontology parts (labels/synonyms, axioms and definitions). Later, this distant dataset is used for training a BioBERT NER model by using the SimpleTranformers wrapper. In the test phase, the trained model is tested on different benchmarking corpora.

positives whenever the indices (locations in text) of the prediction and the curated annotations overlap.

Table 4
Disease NER results

Corpus	Precision	Recall	F1
NCBI-disease			
Labels and synonyms	0.64	0.36	0.46
Axioms	0.68	0.59	0.63
Definitions	0.87	0.80	0.83
Definitions and axioms	0.91	0.76	0.83
Literature abstracts	0.92	0.81	0.86
Curated NCBI train	0.91	0.96	0.94
MedMentions-disease			
Labels and synonyms	0.41	0.26	0.32
Axioms	0.43	0.42	0.43
Definitions	0.48	0.82	0.61
Definitions and axioms	0.58	0.79	0.67
Literature abstracts	0.60	0.78	0.68
Curated NCBI train	0.58	0.77	0.66

Table 4 shows the performance of the disease NER models which are distantly supervised on different ontology components or on abstracts (best F1-score is achieved at top 1, see Additional File 1) on the disease test sets (see Table 1). For the sake of comparison, we also included a supervised BioBERT model that is trained on the NCBI-disease training set. Our results showed that supervised BioBERT trained on the curated set performed the best on NCBI–disease (0.94 F1-score) because concepts are highly conserved in this dataset. To fairly compare the performance of the methods, we further evaluated the models on the MedMentions–disease dataset. Results showed that the distantly supervised models (trained on abstracts and definitions plus axioms) achieved higher F1 scores (0.68 for abstracts and 0.67 for definitions and axioms) compared to the model trained on the curated set (0.66 F1-score) which is actually biased towards the NCBI–disease dataset (we found out there is 80% overlap in concept IDs between NCBI training and test sets). The models trained on solely labels and synonyms, axioms, definitions showed lower F1-score compared to the model trained on abstracts. On the other hand, the model trained on definitions plus axioms achieved a competitive F1-score compared to the model trained on abstracts. This result is more evident on the MedMentions-disease test set.

Table 5
Phenotype NER Results

Corpus	Precision	Recall	F1
MedMentions-phenotype			
Labels and synonyms	0.33	0.75	0.46
Axioms	0.31	0.58	0.40
Definitions	0.47	0.80	0.59
Definitions and axioms	0.55	0.77	0.64
Literature abstracts	0.60	0.82	0.69
Curated MedMentions train	0.61	0.79	0.69
GSC+			
Labels and synonyms	0.32	0.71	0.44
Axioms	0.40	0.60	0.48
Definitions	0.61	0.77	0.68
Definitions and axioms	0.65	0.74	0.69
Literature abstracts	0.73	0.78	0.75
Curated MedMentions train	0.61	0.53	0.57

Table 5 presents the performance of the models in phenotype NER on the GSC+ and MedMentions-phenotype test datasets. We included the MedMentions-phenotype dataset to thoroughly test our models and to train the supervised model on sufficient data. With the inclusion of context at a large scale, the model trained on the weakly labelled abstracts achieved the highest F1-score (0.69 F1-score on MedMentions-phenotype and 0.75 on GSC+) compared to other models. On the other hand, the model trained on the curated set was not robust to the change of dataset as it performed poorly on GSC+ (0.57 F1-score). We observed 6% discrepancy between the model trained on abstracts and the model trained on weakly labelled definitions plus axioms. We discuss the reasons of this discrepancy in detail in the "Discussion" section.

4. Discussion

Our main goal was to explore whether ontology components can help to develop distantly supervised disease/phenotype entity recognition models which are competitive to the state-of-the-art. To that end, we exploited ontological components to create textual context using the labels/synonyms, axioms and definitions. We observed that utilising the context in ontologies via distant supervision aids in developing a NER model at the state-of-the-art level. While the models trained solely on labels and synonyms achieves lowest simply due to lack of context; the models incorporating context such as axioms and definitions improved the performance upon the models that lack context.

The disease NER model trained on the axioms and definitions achieved competitive F1-score compared to the model trained on the abstracts only. However, we observed 6% discrepancy between the phenotype NER models trained on the abstracts (best F1-score is achieved at top 2) and axioms and definitions together. To investigate the reason for this discrepancy, we focused on the False Positive (FP) predictions that we achieved on the GSC+ test corpus. The model trained on the weakly labeled abstracts produced 440 FPs while the model trained on the phenotype definitions and axioms produced 608 FPs. We found that 184 out of 608 FPs are produced distinctly by the model trained on definitions and axioms and not by the one trained on the abstracts. We randomly sampled 20 FPs from these 184 FPs for further manual analysis. Our manual analysis on these 20 FPs showed that all of them were actually True Positives but have been missed by the GSC+ dataset. For example, we found "Uniparental disomy" (HP:0032382) in PMID:8103288 was captured correctly by the model but was missed by GSC+ annotations. More importantly, we observed that the majority of the FPs were not introduced in the definitions and axioms training corpus but were rather predicted as zero-shot instances (i.e. instances that were not seen by the model during training). For example, "Angelman syndrome" in PMID:8786067 which does not correspond to any label/synonyms in HPO and does not exist in the corpus was annotated by the model trained on definitions and axioms. Furthermore, the model trained on literature abstracts did not have these FPs since they were specifically included as O classes in the training set. Details on our manual analysis can be found in the Additional Files 1.

We conducted our study on DO and HPO. These ontologies are widely used and therefore contain dense content which can help to generate sufficiently large weakly label datasets. Although the approach is generic and its utility can be explored for any given ontology; the performance would depend on the density of the content of the ontology of choice. That is, if the ontology does not sufficiently describe a concept, it is not possible to obtain a well-performing model.

5. Conclusion

In conclusion, our analysis showed that the ontology components can provide a suitable corpus to build a NER model that is competitive to state-of-the-art. This alleviates the need for annotating a large number of abstracts and facilitates the creation of weakly labeled training corpora. Easily obtained corpora are desirable since they reduce both the computational and

time overheads. To our best knowledge, this is the first work that uses ontology axioms to build disease/phenotypes NER models.

Additionally, the models trained on ontology components were capable of zero-shot learning on the test datasets. This was not the cases for the models trained on curated sets and the models trained on the large weakly labeled literature abstracts. Our approach is generic and its utility can be explored with any other given ontology which has sufficient content that describes the concept of interest.

Acknowledgments

We thank Dr. Mahmut Uludağ for his technical assistance in processing MEDLINE data.
This work has been supported by funding from King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. URF/1/4355-01-01, URF/1/4675-01-01, URF/1/4697-01-01, URF/1/5041-01-01, REI/1/5334-01-01, FCC/1/1976-46-01 and FCC/1/1976-34-01.

References

[1] C. Jonquet, N. H. Shah, M. A. Musen, The open biomedical annotator, in: American Medical Informatics Association Symposium on Translational BioInformatics, AMIA-TBI'09, San Francisco, CA, USA, 2009, pp. 56–60.

[2] M. Kapushesky, et al., Gene expression atlas update–a value-added database of microarray and sequencing-based functional genomics experiments, Nucleic Acids Research 40 (2011) D1077–D1081. URL: https://doi.org/10.1093/nar/gkr913. doi:10.1093/nar/gkr913.

[3] M. Taboada, H. Rodriguez, D. Martinez, M. Pardo, M. J. Sobrido, Automated semantic annotation of rare disease cases: a case study, Database 2014 (2014) bau045–bau045. URL: https://doi.org/10.1093/database/bau045. doi:10.1093/database/bau045.

[4] C.-H. Wei, H.-Y. Kao, Z. Lu, GNormPlus: An integrative approach for tagging genes, gene families, and protein domains, BioMed Research International 2015 (2015) 1–7. URL: https://doi.org/10.1155/2015/918710. doi:10.1155/2015/918710.

[5] C. Liang, Y. Yu, H. Jiang, S. Er, R. Wang, T. Zhao, C. Zhang, Bond: Bert-assisted open-domain named entity recognition with distant supervision, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 1054–1064. URL: https://doi.org/10.1145/3394486.3403149. doi:10.1145/3394486.3403149.

[6] X. Wang, Y. Guan, Y. Zhang, Q. Li, J. Han, Pattern-enhanced named entity recognition with distant supervision, in: 2020 IEEE International Conference on Big Data (Big Data), 2020, pp. 818–827. doi:10.1109/BigData50022.2020.9378052.

[7] X. Wang, V. Hu, X. Song, S. Garg, J. Xiao, J. Han, ChemNER: Fine-grained chemistry named entity recognition with ontology-guided distant supervision, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp.

5227–5240. URL: https://aclanthology.org/2021.emnlp-main.424. doi:10.18653/v1/2021.emnlp-main.424.

[8] L. Luo, S. Yan, P.-T. Lai, D. Veltri, A. Oler, S. Xirasagar, R. Ghosh, M. Similuk, P. N. Robinson, Z. Lu, PhenoTagger: a hybrid method for phenotype concept recognition using human phenotype ontology, Bioinformatics 37 (2021) 1884–1890. URL: https://doi.org/10.1093/bioinformatics/btab019. doi:10.1093/bioinformatics/btab019.

[9] K. Zhou, Y. Li, Q. Li, Distantly supervised named entity recognition via confidence-based multi-class positive and unlabeled learning, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 7198–7211. URL: https://aclanthology.org/2022.acl-long.498. doi:10.18653/v1/2022.acl-long.498.

[10] H. Dong, V. Suárez-Paniagua, H. Zhang, M. Wang, A. Casey, E. Davidson, J. Chen, B. Alex, W. Whiteley, H. Wu, Ontology-driven and weakly supervised rare disease identification from clinical notes, BMC Medical Informatics and Decision Making 23 (2023). URL: https://doi.org/10.1186/s12911-023-02181-9. doi:10.1186/s12911-023-02181-9.

[11] L. M. Schriml, et al., Human Disease Ontology 2018 update: classification, content and workflow expansion, Nucleic Acids Research 47 (2018) D955–D962. URL: https://doi.org/10.1093/nar/gky1032. doi:10.1093/nar/gky1032.

[12] A. P. Davis, T. C. Wiegers, M. C. Rosenstein, C. J. Mattingly, MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database, Database 2012 (2012). URL: https://doi.org/10.1093/database/bar065. doi:10.1093/database/bar065, bar065.

[13] S. Köhler, et al., Expansion of the human phenotype ontology (HPO) knowledge base and resources, Nucleic Acids Research 47 (2018) D1018–D1027. URL: https://doi.org/10.1093/nar/gky1105. doi:10.1093/nar/gky1105.

[14] NCBI, Pubmed, 1996. https://pubmed.ncbi.nlm.nih.gov/, Last accessed on 2022-04-18.

[15] N. Elastic, Swiftype, Elastic search, 2010. https://www.elastic.co/, Last accessed on 2022-04-18.

[16] R. I. Doğan, R. Leaman, Z. Lu, NCBI disease corpus: A resource for disease name recognition and concept normalization, Journal of Biomedical Informatics 47 (2014) 1–10. URL: https://doi.org/10.1016/j.jbi.2013.12.006. doi:10.1016/j.jbi.2013.12.006.

[17] S. Mohan, D. Li, Medmentions: A large biomedical corpus annotated with umls concepts, 2019. URL: https://arxiv.org/abs/1902.09476. doi:10.48550/ARXIV.1902.09476.

[18] M. Lobo, A. Lamurias, F. M. Couto, Identifying human phenotype terms by combining machine learning and validation rules, BioMed Research International 2017 (2017) 1–8. URL: https://doi.org/10.1155/2017/8565739. doi:10.1155/2017/8565739.

[19] I. Brigadir, Nltk stop words, 2019. https://github.com/igorbrigadir/stopwords/blob/master/en/nltk.txt, Last accessed on 2022-09-14.

[20] S. Tweedie, B. Braschi, K. Gray, T. E. M. Jones, R. L. Seal, B. Yates, E. A. Bruford, Genenames.org: the HGNC and VGNC resources in 2021, Nucleic Acids Research 49 (2020) D939–D946. URL: https://doi.org/10.1093/nar/gkaa980. doi:10.1093/nar/gkaa980.

[21] P. Dyson, Inflect python module, 2022. https://pypi.org/project/inflect/, Last accessed on 2022-09-14.

[22] R. Hoehndorf, P. N. Schofield, G. V. Gkoutos, The role of ontologies in biological and biomedical research: a functional perspective, Briefings in bioinformatics 16 (2015) 1069–

1080.
[23] C. Sammut, G. I. Webb (Eds.), TF–IDF, Springer US, Boston, MA, 2010, pp. 986–987. URL: https://doi.org/10.1007/978-0-387-30164-8_832. doi:10.1007/978-0-387-30164-8_832.
[24] L. A. Ramshaw, M. P. Marcus, Text chunking using transformation-based learning, in: ACL Third Workshop on Very Large Corpora, 1995, pp. 82–94. doi:https://doi.org/10.48550/arXiv.cmp-lg/9505040.
[25] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017. To appear.
[26] F. Z. Smaili, X. Gao, R. Hoehndorf, Onto2Vec: joint vector-based representation of biological entities and their ontology-based annotations, Bioinformatics 34 (2018) i52–i60. URL: https://doi.org/10.1093/bioinformatics/bty259. doi:10.1093/bioinformatics/bty259.
[27] R. Arp, B. Smith, A. D. Spear, Building ontologies with Basic Formal Ontology, The MIT Press, Cambridge, Massachusetts;London, England;, 2015;2016;.
[28] J. Hastings, et al., Chebi in 2016: Improved services and an expanding collection of metabolites, Nucleic acids research 44 (2016) D1214—9. URL: https://europepmc.org/articles/PMC4702775. doi:10.1093/nar/gkv1031.
[29] T. Bakken, L. Cowell, B. D. Aevermann, M. Novotny, R. Hodge, J. A. Miller, A. Lee, I. Chang, J. McCorrison, B. Pulendran, et al., Cell type discovery and representation in the era of high-content single cell phenotyping, BMC bioinformatics 18 (2017) 7–16.
[30] R. P. Huntley, M. A. Harris, Y. Alam-Faruque, J. A. Blake, S. Carbon, H. Dietze, E. C. Dimmer, R. E. Foulger, D. P. Hill, V. K. Khodiyar, et al., A method for increasing expressivity of gene ontology annotations using a compositional approach, BMC bioinformatics 15 (2014) 1–11.
[31] C. J. Mungall, C. Torniai, G. V. Gkoutos, S. E. Lewis, M. A. Haendel, Uberon, an integrative multi-species anatomy ontology, Genome biology 13 (2012) 1–20.
[32] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, in: Proceedings of the 2019 Conference of the North, Association for Computational Linguistics, 2019. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.
[33] T. C. Rajapakse, Simple transformers, https://github.com/ThilinaRajapakse/simpletransformers, 2019.
[34] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert github respository, 2019. (https://github.com/dmis-lab/biobert).

A. Appendix

- Additional file 1 — AdditionalFile1.xls First sheet name as “performance_on_abstracts” contains the performances of the models trained on the weakly labeled abstract datasets selected based on top [1-5] hits from the ElasticSearch Index. Second sheet named as “manual_error_analysis” contains our manual analysis results on the False Positives from the GSC+ dataset. The file is available from github: https://github.com/bio-ontology-research-group/OntoNER

- This manuscript was first submitted to JBMS and it was advised to submit to ICBO as a full paper after the review process. ICBO_comments.pdf is available from https://github.com/bio-ontology-research-group/OntoNER and it contains our point-by-point responses to the reviewers' comments. In the same repository you can also find another version of the manuscript which shows our modifications (ToonsiKafkasICBO2023_redline.pdf) based on the reviewers' comments.