Dear Editor,

Thank you very much for providing us with the opportunity to revise our manuscript. We addressed all comments in a revised manuscript. We wish to thank the reviewers whose comments helped us to improve the quality of our manuscript substantially. Please find our detailed answers to the comments below.

Reviewer Comments:

Reviewer 1
The authors explore "the use of the different ontology components for disease/phenotype NER under the distant supervision scheme." The authors recaptured related works (e.g. ChemNER and PhenoTagger), i.e. weakly labelled sets created with names and synonyms from ontologies/vocabularies. The novelty of the work presented relies in using definitions and complex axioms from ontologies (e.g. DO, MEDIC and HPO) for the NER task under the distant supervision scheme. The study utilised different datasets (listed in Table 1) and different methods, including the language model BioBERT. The results for the various experiments are summarised in Table 3 and 4.

Some modifications (listed below) can enhance the readability and replicability of the experiments reported.

Suggested modifications:
1) Improve the abstract by recapturing all the different methods applied and reporting some concrete results, such as the highest value of the F-score from Table 3 and 4.
Response: We amended the abstract accordingly. We explained what components were used and which model performed achieved competitively.

2) Ontology components are described in subsection 1.3 and 1.4. However, the paper does not include some statistics about the ontology components used in the different experiments. For example, Table 3 and 4 do not have how many formal axioms or how many definitions were applied per corpus experiment.

Response: The statistics are in subsections 2.4.2, 2.4.3, and 2.4.4. We also added one more table to the manuscript and present the statistics them from Table 2.

3) Although the discussion section reports the number of false positives, the number of instances per experiment is not explicitly mentioned. Hence, it is unclear if only the test instances from Table 1 were used in the results of the experiments reported in Table 3 and 4.

Response:

Our results are reported based on the test sets provided in Table 1. To make this explicit, we modified the following sentence in the Results section and mention that we used the test sets (listed in Table 1) for performance evaluation.

"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)"

Other minor modifications:
4) Please note that Figure 1 is missing in the main manuscript.

Response: We replaced this figure with a more detailed one.

5) The additional excel file contains only partial results of the experiments reported and does not seem to include the results from the experiments with the language model BioBERT (results in Table 3 and 4). It will be highly beneficial the inclusion of results from all the different experiments (or at least a sample) in the additional excel file.

Response: We added all the results to the Additional file (excel)

6) The only 2022 reference is a python library. However, there is relevant work for the study published in 2022 and 2023, such as: "A Distant Supervision Corpus for Extracting Biomedical Relationships Between Chemicals, Diseases and Genes" (2022, arXiv:2204.06584); and "Not so weak PICO: leveraging weak supervision for participants, interventions, and outcomes recognition for systematic review automation" (2023, doi:10.1093/jamiaopen/ooac107)

Response: We added two recent references (published in 2022 and 2023) (9 and 10)

7) The reference [24] has some missing information (see '???') and TF-IDF needs a reference.

Response: We added a Reference for TFIDF and resolved the issueof ref 24.

Reviewer 2
This paper experimented the use of ontology components, such as definitions, synonyms, labels, axioms for a weakly supervised BERT NER method and concluded that this method achieves the state-of-the-art performance. In addition, zero-shot predictions were identified by methodology using ontology components.

Although this is an novel experiment, due to the significant problems with the paper writing, I have to regretfully reject the journal publication of this paper at the current stage. Instead, I think this is an excellent paper for a conference presentation after the revision.

Below are summaries or emphases of the issues need to be addressed. All issues are listed as comments in my reviewer PDF file.

1. There are significant issues with the writing, and reviewer's comments were uploaded in a redacted PDF file. Please address my comments point by point. A redline file and a clean version will be needed for a revision.

Response: We edited the text accordingly.

2. Figure 1 is too simple and is not useful. Please make more comprehensive figure according to my comments #20.

Response: We replaced it with a more detailed figure

3. All the results listed in the tables 3 and 4 needs to be narrated in the writing of result section. [see Comment #21]

Response: We modified the text to narrate the results.

4. The results have to match the claim. [see Comment #24]
Response: We clarified our claim (the models incorporating context such as axioms and definitions improved the pertormance upon the models that lack context). See section 3.

5. Some technical terms need to be explained for the readers of this paper, such as I-O-B and zero shot. [See Comment #19, #26]
Response: Both concepts are now explained.

6. The 184 FPs does not match the numbers in attached table in AdditionalFile1 Sheet "manual_error_analysis". [See Comment #25]

Response:
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 those 184 FPs for further manual analysis. We modified section 4 accordingly.

Attachments:

- https://reviewer-feedback.springernature.com/download/attachment/d69bc72d-f7f5-4c02-ad5d-f2ba52b1fd6c

1. Rephrase this sentence: “Early methods for NER used dictionaries due to their simplicity and speed.”

Response: We modified the text as “Early methods for NER used dictionaries due to their applicability and time efficiency.”

2. Should this be curated corpora? “Moreover supervised methods often fail to recognise concepts uncovered by the corpora”

Response: Yes. We edited the text accordingly.

3. “i.e., obtained from a imprecise source (e.g., annotations generated by using rules or vocabularies). t is unclear why annotations generated by using rules or vocab an example of source. Is it because it is generated using machines but not the human curation? Please add some text to make sentence clearer.

Response: We modified “Introduction”, by adding the following sentence:

> “For instance, dictionaries could be used to annotate text with exact matches which can produce both false positives and false negatives”

4. “Later these corpora were used to train different models which in some cases outpertormed state-or-the-art methods” this sentence is not clear.
5. In the above sentence, What does these corpora mean refer to? refer t o ?From the sentence arranged, it seems that the corpora means the unlabeled corpora which is obviously not the case

Response for 4 and 5 : We modified “Introduction”, by adding the following sentence:

> “Later, these instances were used to train different models which in some cases outperformed state-of-the-art methods.”

6. “...NER models has not been comprehensively explored for diseases phenotypes. ” A bridge sentence is needed here to explicitly saying that ontology components, such as .... in addition to labels and synonyms.

Response: We modified “Introduction” accordingly. Please see the last paragraph.

7. “We conducted our experiments on diseases and phenotype concepts” There is a need to add the rational of testing for disease and phenotypes. Why is this use case selected?

Response: We modified “Introduction” accordingly. Please see the last sentence in the last paragraph.

8. “To select abstracts, we used an in-house index …..”Why there is a need to select the abstracts ? Can the search strategy be shared ?

Response: In section 2.1.2, we describe the literature source that we used. We need to select the abstracts that contain the disease and phenotype concepts to form our weakly labellled training dataset. We modified the last sentence of this section accordingly.

9. It seems there are only 3 corpora,unless the disease and phenotype are separate

Response: We used two parts from MedMentions, diseases and phenotypes, therefore in total we used 4 different datasets/corpora.

10. What follows,the MedMentions and GSC+ corpora were mentioned, but NCBI-Disease Corpus was missing

Response: We edited the text accordingly. (See section 2.1.3)

11. In section 2.1.3, What does the disease dictionary refert to ?It appears that the disease dictionay is in the next step in 2.2. So it is very confusing to use a dictinary before it is generated.

Response: We refer to the disease dictionary that we describe in section 2.2 . We made explicit reference to it from section 2.1.3 (See section 2.1.3)

12. section 2.2. : This section is producing two dictionaries for disease and phenotype. The goal and process neds to be cleared stated

Response: We edited the text accordingly.. See section 2.2, first sentence.

13. “we extracted thelabels and synonyms of all concepts” Extracted from where? The ontology, the abstracts? If ontology, what ontologies ?

Response: We edited the text accordingly.. See section 2.2, second sentence.

14. “... short names..” :Here the names may refer to "labels or should explicitly declare it.You could say " we extracted the Labels angsur mosmereamer.onal concepts from MEDIC, DO ,and HPO.

Response: We edited the text accordingly. See section 2.2, third sentence.

15. Section 2.2: It is unclear why the same can be shared by two concepts. Providing examples will be useful here.

Response: We edited the text accordingly.. See section 2.2

16. Section 2.2 : Here the labels/synonyms is used. Is this so called names?

Response: We modified the manuscript and replaced “names” to “labels” in order not to cause any confusion.

17. Section 2.2 : Why this step is necessary? Would this introduce false negative?The explanation is needed here

Response: It was necessary for avoiding false positive matches with protein names. This is clarified in section 2.2.

18. Section 2.2: The label of DOID:138221 is “tetanic cataract” but not “tetanic cataracta”

Response: The typo is fixed

19. Section 2.4.1. : Define I-O-B (It shoul be provided here)

Response: We edited the text accordingly.. See section 2.4.1

20. Section 2.5: Figure 1 is too simple and does not reflect the whole architecture.

Response: We provided a more detailed Figure.

21. Section 3: all the results displayed in the table should be narrated in text

Response: We edited the text accordingly.. See section 3: Results.

22. Section 3: You defined the positive overlap later by “whenver the indices of the predicted annotated na dthe curated one overlaps” It is unclear to me what was exactly overlapped.

Response: indices are annotations locations in the text. We modified the text accordingly (see section 3)

23. Section 3, “Results show that the distantly supervised models (trained on abstracts and definitions plus axioms) achieved higher F1 scores ” This sentence is not clear to me.

Response: We edited the text accordingly.. See section 3.

24. Section 3, “the models incorporating context such as axioms and definitions improved the pertormance” This claim is not supported by the results

Response: We clarified our claim (the models incorporating context such as axioms and definitions improved the pertormance upon the models that lack context). See section 3.

25. Section 3, 184 FP: Where does the number of 184 come from ?

Response: 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 those 184 FPs for further manual analysis.

We modified the text accordingly, please see section 4.

26. Zero-short needs to be explain. It is a specitic term that non-ILYprofessionalwouldnt understand.

Response: We edited the text accordingly. See section 4.

27. “For example,"Angelman Syndrome"in PMID:8786067which does not correspond to any labels ynonyms in HPO and does not exist in the corpus was annotated by the definitions and axioms model” This sentence needs rewrite.

Response: We edited the text accordingly. See section 4.

28. Table 3 and 4, “The naming of a row "supervised BioBERT" will introduce a false impression that all other methods listed in the table do not use BioBERT and do not use supervise method. But you use weakly or distantly supervised method and utilize BioBERt in all methods. Suggest to change the “Supervised BioBERT” to “supervised method” and adda combined column to show other rows are weakly supervised.

Response: We edited the text accordingly. See Tables 4 and 5 (table numbering has changed since we introduced one more table)

Reviewer 3

General comment: important work for the machine learning area that reflects annotation methods using ontologies to offer a curated corpus, with quality, to build a competitive NER model to the state of the art.

I - Main issue:

- What is the research problem of the article? In the introduction, it would be good to be explicit and clear about the problem and the general objective of the research. The purpose is only mentioned in the discussion section. In the Introduction of the article, it was only noticed the explanation of the research hypothesis.

Response: We explained our research problem more explicitly in the last paragraph of "Introduction"

II- Conceptual issues:

- Elucidate about machine learning and supervised methods; bring works that relate ontologies such as semantic annotation to these practices.

Response: We modified "Introduction" accordingly

- In the last paragraph, page 1, explain [make it clear] why supervised methods require curated corpora.

Response: We modified "Introduction" accordingly

III- General issues:

- Figure 1 not available. Needs revision.

Response: We replaced this figure with a more detailed one.

Decision: the article is suitable for publication, as long as it is reviewed in the aspects mentioned above.