An assessment of partial annotation learning for biomedical entity recognition*

Liangping Ding[1], Giovanni Colavizza[2], Zhixiong Zhang[3]

[1]*l.ding@uva.nl*
National Science Library, Chinese Academy of Sciences
Department of Library Information and Archives Management, University of Chinese Academy of Science
Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, (The Netherlands)

[2]*g.colavizza@uva.nl*
Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, (The Netherlands)

[3]*zhangzhx@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences
Department of Library Information and Archives Management, University of Chinese Academy of Science, Beijing, (China)

Abstract
Partial annotation learning is useful for entity recognition when there are missing entity annotations. In our work, we systematically study partial annotation learning methods for biomedical entity recognition over different simulated scenarios for missing entity annotations. We harmonize 15 biomedical NER corpora encompassing five entity types to serve as golden standard. To explore the effectiveness of partial annotation learning methods, we compare two commonly used partial annotation learning models with the state-of-the-art biomedical entity recognition model PubMedBERT tagger. Our experiments show that partial annotation learning methods can effectively learn from biomedical corpora with even significant fractions of entity annotations missing, suggesting further work in this direction would be promising.

Introduction
Entity Recognition is a fundamental task in information extraction. For fully annotated NER datasets, this problem has been basically solved by fine-tuning pretrained language models (Devlin et al., 2019). Nevertheless, a lack of fully annotated datasets is a common issue facing scientific entity recognition that relies on experts with domain expertise. In practice, high-quality annotations are expensive and laborious to obtain at scale, setting obstacles for the application of NER models in low-resource scenarios.

To reduce the dependency on expert annotations, distant supervision (Liang et al., 2020) and exploratory expert (Effland & Collins, 2021) approaches have been proposed. However, these all result in partially annotated datasets with high precision but low recall for entity spans. More generally, datasets often suffer from unlabeled entity problems (Li et al., 2021), which means that large amounts of entity annotations are missing, as exemplified by the entity "SARS-CoV-2" in Figure 1.In such scenarios, assuming that unlabeled tags are non-entities (the O tag) likely degrades the performance of NER models.

*We acknowledge the support of Tian-Yuan Huang for drawing the figures. This work is supported by the China Scholarship Council (CSC).

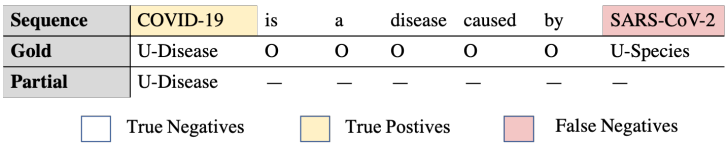

Figure 1. Example illustrating the unlabeled entity problem in NER. The Sequence row shows the token sequence of the input text, and the Gold row reflects the golden truth label path under the BILOU encoding scheme, and the Partial row shows the partially annotated label path, in which the symbol '—' represents the unknown label

Partially Annotation Learning (PAL) methods have been shown to alleviate this problem effectively in previous studies. The basic idea is to model the missing labels as latent variables, deduce all possible label paths and calculate the marginal tag likelihood (Effland & Collins, 2021; Jie et al., 2019; Mayhew et al., 2019; Tsuboi et al., 2008). Most studies focused on evaluating the effectiveness of partial annotation learning on traditional NER benchmark datasets such as CoNLL2003, dealing with the most common named entity types like person, location and organization. The effectiveness of partial annotation learning methods on the more challenging scientific NER benchmark datasets has not been assessed. Furthermore, no single study exists which comprehensively evaluates the validity of partial annotation learning and conducts an in-depth assessment on the impact of missing entity ratio and simulated annotation scenario to the model performance. Although Effland & Collins (2021) compared traditional NER model to partial annotation learning model under different missing entity ratios, they set the interval of missing entity ratios arbitrarily and the number of entity annotations $M \in \{100 (0.4\%), 500 (2.1\%), 1K (4.3\%), 5K (21.3\%), 10K (42.6\%)\}$, making it harder to appreciate the performance degradation systematically.

In this work, extensive experiments have been conducted to verify the effectiveness of partial annotation learning methods for biomedical NER. We evaluated the state-of-the-art Biomedical NER tagger PubMedBERT (Gu et al., 2022) and two partial annotation learning methods BiLSTM-Partial-CRF (Jie et al., 2019) and EER-PubMedBERT (Effland & Collins, 2021) for five biomedical entity types. Further, we proposed a more realistic yet challenging entity removal scheme to simulate the unlabeled entity problem and investigated the impact of missing entity ratio and entity removal scheme to the model performance.

Materials and methods

Corpora compilation and pre-processing

In order to obtain solid evaluation results, we took consideration of the influence of corpora size and other implicit factors such as the distribution of entity mentions, sentence length, and performed our evaluations on five entity types: genes/proteins, chemicals, diseases, cell lines and species. Weber et al. (2020, 2021) successively proposed HUNER and HunFlair, stand-alone biomedical entity recognition taggers covering the above-mentioned entity types. Following their work, we relied on a total of 15 gold standard corpora to construct the fully annotated datasets for our experiments, excluding eight corpora[1] that we don't have access to, and the BioSemantics corpus which contains a large number of very long sentences.

[1]Arizona Disease, BioInfer, CLL, GELLUS, IEPA, LINNEAUS, Osiris v1.2, Variome.

Table 1. Statistics of the corpora

Entity Type	*Splits*	*Sentences*	*Annotations*	*Surface Forms*
Genes/Proteins	train	74905	82803	28109
	dev	12620	14442	6531
	test	36662	42386	16319
Chemicals	train	99037	114575	29908
	dev	16201	18392	6863
	test	49766	56596	16834
Diseases	train	16895	14216	4238
	dev	2754	2337	1009
	test	8738	7494	2564
Cell Lines	train	12592	2500	1419
	dev	2001	450	290
	test	6146	1248	797
Species	train	15195	5290	1567
	dev	2555	811	286
	test	7431	2891	917

Analogously to the data preprocessing pipeline of HUNER for each entity type, we aggregated the corresponding corpora which contain annotations for the respective entity type to learn a type-specific model and converted them into the standard CoNLL2003 format. In addition, we re-used the train/dev/test splits introduced by HUNER to split each resulting corpora for each entity type with a ratio of 60:10:30 among the splits. Subsequently, we converted the BIO encoding scheme in the standard CoNLL2003 format into BILOU (beginning, inside, last, outside, unit) encoding scheme, which was observed to outperform the widely adopted BIO encoding scheme for NER. Table 1 highlights important statistics of the corpora for five entity types. Figure 2 shows the distribution of the number of tokens in sentences among the training corpora for five entity types and the figure has used a base-10 logarithmic transformation on *x* axis.

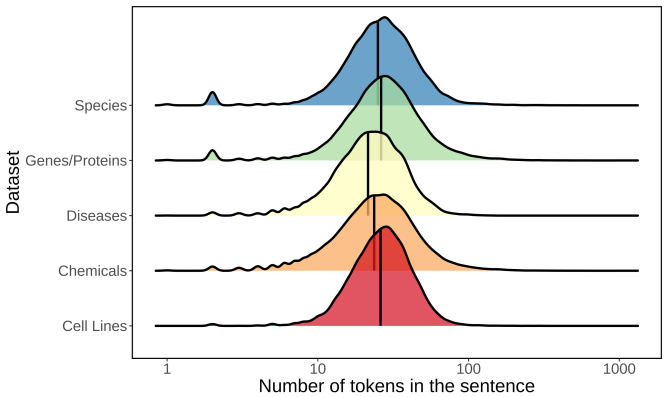

Figure 2. Distribution of number of tokens in sentences among the training corpora

Annotation scenarios simulation

In this research, our goal is to explore the capability of partial annotation learning models to effectively mitigate the unlabeled entity problem (Li et al., 2021). We assumed that partial annotations for NER task can be obtained by removing entity annotations from the fully annotated datasets and considered two entity removal schemes to simulate unlabeled entity problems in real-world situations.

The first scheme is Remove Annotations Randomly (RAR), previously used in Jie et al. (2019); Li et al. (2021), drops entity annotations uniformly at random. By setting the entity removal rate r, we control the number of removal entity annotations. For example, r=0.1 means that we remove 10% of all entity annotations and keep 90% of entity annotations. The drawback of this scheme is that the entity removal process is incomplete, with a diverse set of surface forms of the removal entity annotations still occurring in the dataset, which is not realistic under certain circumstances (Effland & Collins, 2021).

The second scheme is Remove All Annotations for Randomly Selected Surface Forms (RSFR), which is a more realistic yet more challenging scheme to learn from. RSFR scheme can be regarded as a simulation of distant supervised NER, wherein entity mentions not occurring in the dictionary will consistently not be annotated in the whole dataset. To simulate data for this scenario, we group annotations by their surface forms and randomly select groups of annotations to remove, as the literal meaning of this scheme. To make for a fair comparison with the RAR scheme, we downsample annotations grouped by surface forms until the number of removed entity annotations is roughly the same under both schemes at the same entity removal rate.

Overview of NER models

In this study, we evaluated both full annotation learning models and partial annotation learning models to explore the effectiveness of partial annotation learning.

For a full annotation learning model, we experimented with PubMedBERT (Abstract+Fulltext) (Gu et al., 2022), which is the state-of-the-art pretrained language model in biomedical domain. PubMedBERT (Abstract+Fulltext) was pretrained from scratch using the abstracts and full text articles from PubMed Central, which was proved to outperform BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), BioBERT (Lee et al., 2019), SciBERT (Beltagy et al., 2019) etc. on NER task based on medical language.

For partial annotation learning models, we implemented two models from prior work. One of them is BiLSTM-Partial-CRF model, proposed by Jie et al. (2019), which is a commonly used baseline model for partial annotation learning. This model is based on BiLSTM-Partial CRF followed by a self-training framework with cross-validation. Another model is EER-BERT model, proposed by Effland & Collins (2021), which is the state-of-the-art partial annotation learning model, exceeding the results from Li et al. (2021) on 7 datasets. They proposed a novel loss, the Expected Entity Ratio, based on the assumption that the number of named entity tags over the entire distribution of sentences occur at relatively stable. We adapted EER-BERT model to EER-PubMedBERT model with the modification of converting the original RoBERTa to PubMedBERT to make the comparison as fair as possible.

Experiments & results

We compared the performance of full annotation learning methods and partial annotation learning methods on synthetic datasets of 5 entity types for two entity removal schemes and several entity removal rate, detailed below.

Experimental design

For each entity type, we generated synthetic datasets from the original fully annotated dataset by randomly removing entity annotations based on the combination of entity removal rate and the entity removal scheme. The entity removal rate r is set as 0.1, 0.2, …, 0.9. For each

combination, we removed entity annotations from the original datasets with five different random seeds to account for the variance in model performance over different runs. With respect to modeling approaches, we adopted 3 models including one full annotation model PubMedBERT tagger and two partial annotation learning models BiLSTM-Partial-CRF, EER-PubMedBERT. In this way, 9 (entity removal rate) × 2 (entity removal scheme) × 5 (random seed) × 5 (entity type) × 3 (model) = 1350 experiments were conducted. Furthermore, for each entity type, we applied PubMedBERT tagger on original fully annotated dataset to provide an upper bound model performance, which we did not expect any of the other methods to outperform.

Results & discussions

Figure 3 shows the test performance of empirical studies, in which results are averaged across five random seeds. We can draw the following observations. Firstly, full annotation model severely suffer from the unlabeled entity problem, while partial annotation learning methods are less influenced by this issue. For all five entity types we evaluated, both partial annotation learning methods on average have better F1-scores when compared to full annotation learning model, as the removal rate gets higher. The full annotation learning model can achieve good results when there are a few missing entity annotations, while as the number of missing entity annotations increases, the advantage of partial annotation learning models begins to emerge. The RSFR entity removal scheme, which is more realistic, is also more challenging compared to the RAR scheme. For the RAR scheme, the partial annotation learning model can mitigate the unlabeled entity problem well when the entity removal rate is small, while the model performance starts to decline at the initial stage of removing entity annotations with the RSFR scheme. Thirdly, from extensive experiments on various datasets, we show that the state-of-art partial annotation learning model EER-PubMedBERT model outperforms all competitors, which performs on-par to the upper bound on the RAR scheme and can also achieve competitive results on the RSFR scheme. As we can see, even as the number of missing entity annotations increases, the performance degradation of EER-PubMedBERT model is not extreme in the RAR scheme and not tremendous in the RSFR scheme.

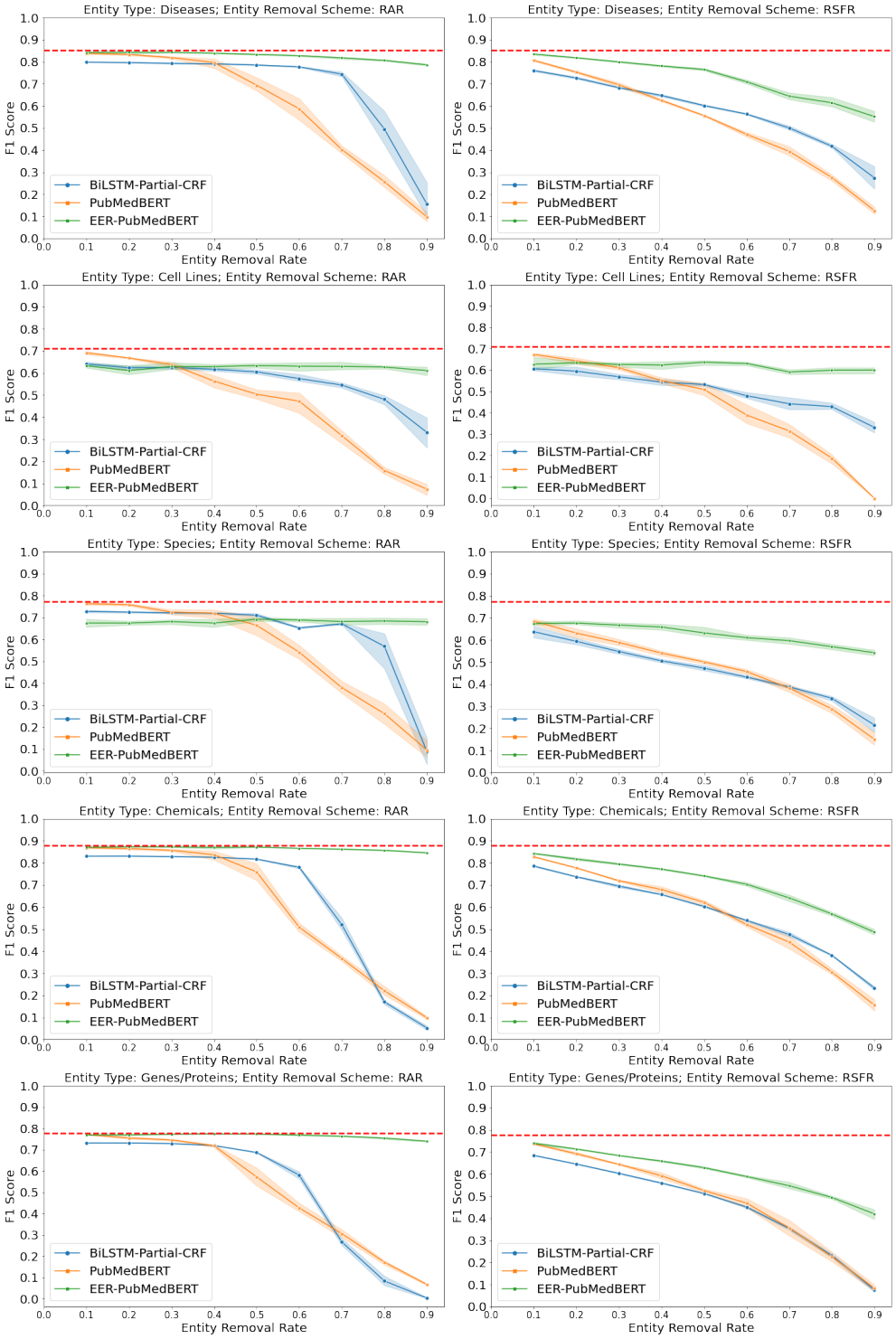

Figure 3. The test performance of empirical studies. The whole figure contains 10 (5 rows ×2 columns) subplots, where each row of subplots reflects the curves of each entity type and each column of subplots reflects the curves of each entity removal scheme. In each subplot, the horizontal axis denotes the entity removal rate, the vertical axis denotes the F1 scores, and the red dotted line defines the upper bound F1-score

Conclusion

In this work, we explore the effectiveness of partial annotation learning in Biomedical NER by comparing two partial annotation learning models, BiLSTM-Partial-CRF and EER-PubMedBERT, with the state-of-the-art PubMedBERT tagger. We investigate the effects of

different entity removal rates and schemes on model performance. Our results on five entity types strongly confirm the usefulness of partial annotation learning, which have strong ability to supersede and still learn from partially annotated corpora.

References

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. *ArXiv:1903.10676 [Cs]*. http://arxiv.org/abs/1903.10676

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Effland, T., & Collins, M. (2021). Partially Supervised Named Entity Recognition via the Expected Entity Ratio Loss. *Transactions of the Association for Computational Linguistics*, *9*, 1320–1335. https://doi.org/10.1162/tacl_a_00429

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2022). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, *3*(1), 1–23. https://doi.org/10.1145/3458754

Jie, Z., Xie, P., Lu, W., Ding, R., & Li, L. (2019). Better Modeling of Incomplete Annotations for Named Entity Recognition. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 729–734. https://doi.org/10.18653/v1/N19-1079

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, btz682. https://doi.org/10.1093/bioinformatics/btz682

Li, Y., Liu, L., & Shi, S. (2021). EMPIRICAL ANALYSIS OF UNLABELED ENTITY PROB- LEM IN NAMED ENTITY RECOGNITION. *ICLR*, 12.

Liang, C., Yu, Y., Jiang, H., Er, S., Wang, R., Zhao, T., & Zhang, C. (2020). BOND: BERT-Assisted Open-Domain Named Entity Recognition with Distant Supervision. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1054–1064. https://doi.org/10.1145/3394486.3403149

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv:1907.11692 [Cs]*. http://arxiv.org/abs/1907.11692

Mayhew, S., Chaturvedi, S., Tsai, C.-T., & Roth, D. (2019). Named Entity Recognition with Partially Annotated Training Data. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 645–655. https://doi.org/10.18653/v1/K19-1060

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Tsuboi, Y., Kashima, H., Mori, S., Oda, H., & Matsumoto, Y. (2008). Training Conditional Random Fields Using Incomplete Annotations. *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 897–904. https://www.aclweb.org/anthology/C08-1113

Weber, L., Münchmeyer, J., Rocktäschel, T., Habibi, M., & Leser, U. (2020). HUNER: Improving biomedical NER with pretraining. *Bioinformatics*, *36*(1), 295–302. https://doi.org/10.1093/bioinformatics/btz528

Weber, L., Sänger, M., Münchmeyer, J., Habibi, M., Leser, U., & Akbik, A. (2021). HunFlair: An easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics*, *37*(17), 2792–2794. https://doi.org/10.1093/bioinformatics/btab042