BiodivBERT: a Pre-Trained Language Model for the Biodiversity Domain

Nora Abdelmageed[1,2,3,*], Felicitas Löffler[1,3] and Birgitta König-Ries[1,2,3]

[1]Heinz Nixdorf Chair for Distributed Information Systems
[2]Michael Stifel Center Jena, Germany
[3]Friedrich Schiller University Jena, Jena, Germany

Abstract

Information Extraction in the Life Sciences is getting increasing attention due to the constantly growing amount of data and text. The advancements of deep learning models further accelerate this development. However, applying these models to domain-specific data is crucial as applied domains often require different entity type extractions than general ones. This paper introduces BiodivBERT, the first pre-trained language model for the biodiversity domain. We constructed two pre-training corpora (abstracts and abstracts + full text) based on a keyword search strategy from two leading publishers in the Life Sciences. In addition, we fine-tuned BiodivBERT on two downstream tasks, i.e., Named Entity Recognition (NER) and Relation Extraction (RE), using various state-of-the-art benchmarks. The results show that BiodivBERT outperforms the state-of-the-art approaches. Moreover, we discuss a potential application of BiodivBERT for ontology auto-population. We publicly release data and code for both pre-training and fine-tuning.

Keywords

Biodiversity, Language Model, BERT, Pre-training, Fine-tuning

1. Introduction

Motivated by the predicted impending loss of biodiversity and the consequences of this loss for humanity [1], research in the biodiversity domain has recently witnessed accelerated growth. For instance, the Biodiversity Heritage Library (BHL)[1] currently holds over 55 million digitized pages of legacy biology text from the $15^{th} - 21^{st}$ centuries, representing a massive amount of textual content [2]. Moreover, Google Scholar returns more than 85,000 hits for a search using the term "biodiversity" from 2021 till the time of writing, November 2022. Thus, text mining tools are an open demand in the field to leverage this untapped wealth. Recent progress in the development of deep learning models for Natural Language Processing (NLP) promises to answer this demand. However, directly applying such NLP techniques to biodiversity texts is not promising. Modern word representation models such as Word2Vec [3], GloVe [4], ELMo [5],

SWAT4HCLS 2023: The 14th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences

*Corresponding author.
nora.abdelmageed@uni-jena.de (N. Abdelmageed); felicitas.loeffler@uni-jena.de (F. Löffler); birgitta.koenig-ries@uni-jena.de (B. König-Ries)
0000-0002-1405-6860 (N. Abdelmageed); 0000-0001-6423-7427 (F. Löffler); 0000-0002-2382-9722 (B. König-Ries)

[1]https://www.biodiversitylibrary.org/

and BERT[6] are trained and tested on general domain texts (e.g., Wikipedia). However, domain-specific, texts contain many instances of domain-specific entity types. E.g., Helianthus (genus of sunflower species), calcareous grassland, or growth rate. Thus, it is difficult to estimate the performance of general-purpose models on domain-specific datasets. Approaches to improve the performance of cutting-edge approaches like BERT on domain-specific benchmarks have first been developed for the (bio-)medical domain with BioBERT [7] and clinicalBERT [8]. BioBERT is initialized with BERT weights and pre-trained on biomedical corpora that are based on PubMed[2] and PubMed Central (PMC)[3]. It showed a significant improvement on three downstream tasks, namely Named Entity Recognition (NER), Relation Extraction (RE), and Question Answering (QA). To the best of our knowledge, there is no language model for the biodiversity domain that supports the extraction of named entities and relations from textual data.

The contributions of our paper are as follows: i) BiodivBERT is the first biodiversity-specific BERT-based model pre-trained on biodiversity corpora. ii) We show that pre-training BERT on biodiversity corpora improves its performance on two downstream tasks, NER and RE. iii) We discuss potential applications for BiodivBERT, e.g., ontology population. v) We make our pre-training corpora [9], pre-processed datasets [10], the pre-trained weights [11] of BiodivBERT, and the source code for pre-training and fine-tuning publicly available[4].

The rest of this paper is organized as follows: We give an overview of our approach in Section 2. We explain the pre-training and fine-tuning tasks in Section 3 and Section 4 respectively. We show our results in Section 5. We demonstrate a potential application of the current work in Section 6. We conclude in Section 7.

2. Approach

In this paper, we introduce BiodivBERT, a pre-trained language representation model for the biodiversity domain. The overall process of pre-training and fine-tuning BiodivBERT is shown in Figure 1. First, we initialize BiodivBERT with weights from BERT [6], which was pre-trained on general domain corpora (English Wikipedia and BooksCorpus). Then, BiodivBERT is pre-trained on our collected corpora from the biodiversity domain. The first corpus is based on abstracts (+Abs), while the other contains both abstracts and full text (+Abs+Full). To demonstrate the effectiveness of BiodivBERT in biodiversity text mining, we have fine-tuned and evaluated it on two downstream tasks, NER and RE, using various task-specific datasets.

3. Pre-training BiodivBERT

In this section, we explain our pre-training data sources, selection strategy, and data statistics. In addition, we discuss the pre-training task for BiodivBERT.

3.1. Pre-training Data

In this section, we discuss the construction of our two pre-training corpora that are based on Abstracts (+Abs), and Full text (+Abs+Full). Thus, we explain our used keywords search strategy, workflow, and the resultant corpora statistics.

[2]https://pubmed.ncbi.nlm.nih.gov/
[3]https://www.ncbi.nlm.nih.gov/pmc/
[4]https://github.com/fusion-jena/BiodivBERT

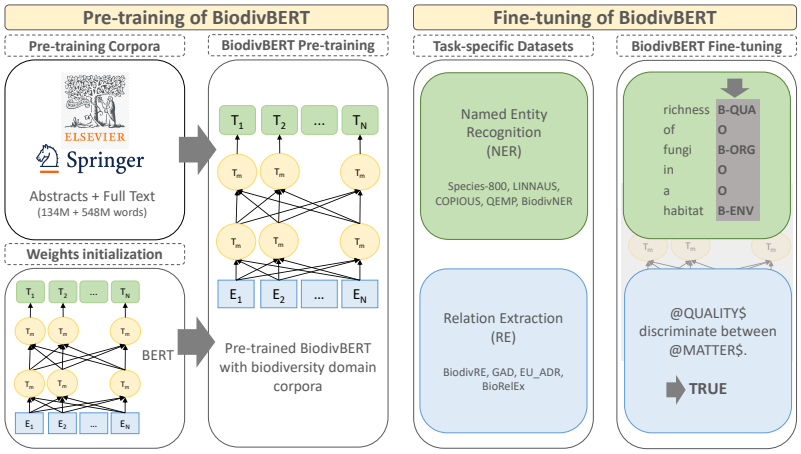

Figure 1: BiodivBERT pre-training and fine-tuning Overview.

Keywords Search: We discussed various options to crawl data for pre-training with three biodiversity experts. Our experts recommended to focus on sources that reflect recent research directions. Therefore, they suggested querying Elsevier[5] and Springer[6] rather than the BHL which contains more legacy data. In addition, these companies publish a diverse set of biodiversity-related journals. Moreover, they provide official APIs to crawl data. To crawl these massive reservoirs, our experts suggested 10 keywords that covered the domain, e.g., "biodiversity", "genetic diversity", and "taxonomic"[7] and recommended crawling data from the last three decades [1990-2020[8]].

Corpora Construction Pipeline: To show the effect of the pre-training data on the model performance through the downstream tasks, we created two pre-training corpora. One is based on abstracts (+Abs) only, while the other contains abstracts and full text of publications (+Abs+Full). Under the access rights provided by the selected publishers, we used abstracts and full texts of open-access papers and abstracts only for other publications. To construct the +Abs corpus, we used the pre-selected keywords and year range as input for both Elsevier and Springer's provided full-text search APIs. For each of them, we retrieved the corresponding DOIs for each keyword in the given year's range. We then applied a deduplication method to the result set. We applied the same procedure for the second corpus, which is based on the full texts (+Abs+Full). Elsevier provided a straightforward API to obtain the parsed full text for a given article. However, for Springer's full text, we downloaded the corresponding PDF file for each DOI, converted it to an XML format, and then extracted the text. We converted the downloaded PDFs to XML files using the GROBID service, and client [12]. We cleaned and merged the final text from both data sources. We applied both shallow and deep cleaning steps for the collected data. For instance, we filtered the sentences to include unique ones. In addition, we used regular expressions to remove URLs and DOIs.

[5]https://dev.elsevier.com/
[6]https://dev.springernature.com/
[7]https://github.com/fusion-jena/BiodivBERT/blob/main/keywords_search
[8]The starting year of this project.

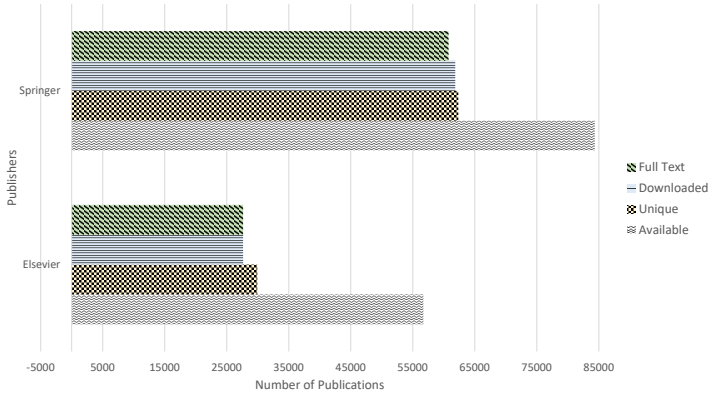

Figure 2: Data loss during full Text corpus construction.

Corpora Statistics: Table 1 gives an overview of our final corpora statistics. +Abs and +Abs+Full are around 1GB and 4GB in size, respectively. Our corpora include around 1M abstracts with 92K full publications. We faced data loss as shown in Figure 2 due to several reasons: 1) not found (404) errors for some articles because of either technical issues on the provider side or invalid DOIs. 2) Elsevier allows crawling only 6000 articles per keyword through its API. 3) GROBID failed to convert some PDF files from Springer. So, the shown numbers indicate the best we can use from both publishers under such circumstances.

Table 1
Final Pre-training Corpora Statistics.

Corpus	Data Source	#Sentences	#Words	Size
+Abs	Abstracts	5M	134M	876 MB
+Abs+Full	Full Text	25M	548M	3.81 GB

3.2. Pre-training Task

We pre-trained BiodivBERT on our domain-specific corpora (+Abs) and (+Abs+Full). We initialized BiodivBERT with the `BERT_base_cased` weights for computation efficiency and to leverage the general domain learned weights from the Wiki and books corpora by the original model. For tokenization, we used the same BERT WordPiece [13] tokenizer, which overcomes the out-of-vocab (OOV) issue. Similar to BioBERT [7], we used the cased vocabulary in our setting as it has higher performance on the downstream tasks, and we used the original vocabulary of `BERT_base_cased` for the same reason and to be compatible with both BERT and BioBERT. We compare BiodivBERT to the SOTA BERT-based models. In addition, we tested different combinations of pre-training corpora. Thus, we compare such settings in Table 2. We pre-trained BiodivBERT$_{+Abs}$, and BiodivBERT$_{+Abs+Full}$ using transformers by Hugging Face [14] library on a single V100 GPU (16 GB) for 3 and 5 days respectively. We used 512 for the maximum sequence length and 15% of the masked language model probability for hyperparameters. In addition, we used Adam's optimizer with 1e-3 learning weight and default betas. Moreover, we set the batch size to 16 and enabled the gradient accumulation with four steps for faster training.

Table 2
Pre-training Models setting corpora.

Model	Corpora
$BERT_{BASE}$[6]	Wiki + Books
$BioBERT_{v1.1}$[7]	Wiki + Books + PubMed
$BiodivBERT_{+Abs}$	Wiki + Books + Abstracts
$BiodivBERT_{+Abs+Full}$	Wiki + Books + Abstracts + FullTxt

4. Fine-tuning BiodivBERT

With minimal architectural modification, BiodivBERT can be applied to various downstream text mining tasks. In this work, we fine-tuned it on both NER and RE using P100 (16 GB) by Colab Pro[9] using various state-of-the-art datasets.

Named Entity Recognition is the task of identifying the domain-specific proper nouns inside a given text. We leveraged the original BERT structure for NER such that it uses a single output layer based on the representations from its last layer to compute only token-level probabilities. We used entity-level precision, recall, and F1 score as the evaluation metrics of NER. We selected various SOTA datasets to test the performance of BiodivBERT on NER. COPIOUS [2] is based on BHL documents and has six entity types, a.k.a. tags, including, e.g., habitat and taxon names. QEMP [15] is created from datasets metadata files and contains four entity types (tags), e.g., quality and material. BiodivNER [16] (dataset [17]) is constructed from metadata files and abstracts from PubMed and has five tags, e.g., organism and phenomena. Species800 [18] and Linnaeus [19] are designed for species names that are normalised to NCBI Taxonomy database[10]. We pre-processed all of them to follow the BIO[11] format for token classification.

Table 3
Overview of the selected NER datasets.

Dataset	Tags	#Docs	#Statements	#Annotations
BiodivNER	6	150	2,398	9,982
QEMP	4	50	2,226	5,154
COPIOUS	5	668	26,277	26,007
Species800	1	800	14,756	5,330
Linnaeus	1	100	34,310	3,884

Relation Extraction is the task of classifying relations among named entities in a corpus. We utilized the sentence classifier of the original version of BERT, which uses a [CLS] token for the classification of relations. Relations in such a way use a single output layer based on a [CLS] token representation from BERT. To the best of our knowledge, BiodivRE [17] is the only available RE corpus for the biodiversity domain, so we included it in our fine-tuning setting. In addition, we included the BioRelEx [20], EU-ADR [21], and GAD [22] corpora from the biomedical domain. BiodivRE contains relations among the five entity types of BiodivNER, like occur_in and influence in a multi-class and a binary format. BioRelEx classifies the bindings

[9] https://colab.research.google.com/?utm_source=scs-index
[10] https://www.ncbi.nlm.nih.gov/taxonomy
[11] https://natural-language-understanding.fandom.com/wiki/Named_entity_recognition#BIO

between genes and diseases into three categories: exists (1), not exists (-1), and unsure (0). EU-ADR and GAD include relations between genes and diseases. In this work, we used the binary format that BiodivRE provides. For BioRelEx, we constructed a binary relation corpus by excluding the unsure relations. Moreover, similar to BioBERT, we anonymized the target named entities in a sentence using their tags, e.g., @COMPLEXPROTEIN$ and @GENE$. For EU-ADR and GAD, we used the provided pre-processed version by BioBERT's team since the original data are unavailable. Table 4 shows the selected RE datasets' statistics.

Table 4
Overview of the selected RE datasets.

Dataset	#True Statements	#False Statements	Total
BiodivRE	1,369	2,631	4,000
BioRelEx	1,379	62	1,606
GAD	25,209	22,761	53,300
EU-ADR	2,358	837	3,550

5. Results & Discussion

To gain a first impression of the performance of our approach, we ran a mask-filling task on a typical biodiversity topic on BERT, BioBERT, and BiodivBERT using the following test case: *"Diversification and [MASK] in brood pollination mutualisms."*. Table 5 shows that BiodivBERT has produced the most realistic results compared to the other two models. For example, BiodivBERT can generate both "diversity" and "evolution". Such results demonstrate the effectiveness of the pre-training data.

Table 6 and Table 7 shows the scores of fine-tuning the BiodivBERT, BERT, and BioBERT models on the two downstream tasks NER, and RE respectively. In addition, we developed a single layer of the Bidirectional Long Short Term Memory (BiLSTM) with 10% dropout as a baseline approach. We micro-averaged the results per dataset to generate the scores of all systems. We fine-tuned these models on a single P100 GPU provided by Colab Pro. First, we found that $\text{BioBERT}_{\text{v1.1}}$ obtained higher scores than $\text{BERT}_{\text{BASE}}$ on the downstream tasks, Second, $\text{BiodivBERT}_{\text{+Abs+Full}}$ and $\text{BiodivBERT}_{\text{+Abs}}$ gained the best results among all the others by achieving either first or second place on the given datasets. In detail, for NER, $\text{BiodivBERT}_{\text{+Abs+Full}}$ exceeds $\text{BioBERT}_{\text{v1.1}}$ for all datasets except QEMP ,where $\text{BiodivBERT}_{\text{+Abs}}$ exceeds $\text{BioBERT}_{\text{v1.1}}$, with 1% F1 score. In addition, we notice that the scores of the Species-related datasets (Species-800 and LINNAEUS) gained higher scores than others; we argue that is due to such datasets being easier than those with fuzzy categories to identify. E.g., QEMP and BiodivNER have a class, "QUALITY" that groups data measures that cover vast and various attributes of the biodiversity domain and would be harder to detect. For RE, we have mixed results; for example, $\text{BiodivBERT}_{\text{+Abs+Full}}$ outperforms $\text{BioBERT}_{\text{v1.1}}$ with a 2.5% F1 score for EU-ADR. However, $\text{BioBERT}_{\text{v1.1}}$ overcomes $\text{BiodivBERT}_{\text{+Abs+Full}}$ with 3% F1 score for BiodivRE. We plan to apply different fine-tuning settings on those datasets to enhance the scores.

6. Application

In this section, we discuss a possible application for BiodivBERT: A while ago, we started developing a biodiversity domain ontology, Biodivonto [23]. Such an ontology is needed for

Table 5
Fill-in mask task results by BERT-based models.

Model	Rank	Result
BERT	1	Diversification and **variation** in brood pollination mutualisms.
	2	Diversification and **variations** in brood pollination mutualisms.
	3	Diversification and **changes** in brood pollination mutualisms.
BioBERT	1	Diversification and **Image** in brood pollination mutualisms.
	2	Diversification and**im** in brood pollination mutualisms.
	3	Diversification and**vasive** in brood pollination mutualisms.
BiodivBERT	1	Diversification and **change** in brood pollination mutualisms.
	2	Diversification and **diversity** in brood pollination mutualisms.
	3	Diversification and **evolution** in brood pollination mutualisms.

Table 6
Fine-tuning scores on NER datasets. The highest score is marked in **bold** while the following score is marked underline. Evaluation Metrics (Met.) are Precision (P), Recall (R), and F1 score (F).

					BiodivBERT	
Dataset	Met.	BiLSTM	$BERT_{BASE}$	$BioBERT_{v1.1}$	+Abs	(+Abs+Full)
Spieces-800	P	0.49	0.80	**0.87**	0.81	0.79
	R	0.09	0.81	0.80	0.80	**0.84**
	F	0.16	0.80	0.80	**0.81**	**0.81**
LINNANUS	P	0.82	0.93	0.93	0.92	**0.95**
	R	0.22	0.94	0.94	0.90	**0.95**
	F	0.34	0.94	0.94	0.91	**0.95**
COPIOUS	P	0.77	0.88	0.88	0.88	**0.89**
	R	0.53	0.87	**0.89**	0.88	0.88
	F	0.63	**0.88**	**0.88**	**0.88**	**0.88**
QEMP	P	0.84	0.90	**0.91**	0.90	0.88
	R	0.53	0.73	0.76	**0.78**	0.72
	F	0.65	0.81	0.83	**0.84**	0.79
BiodivNER	P	0.66	0.85	**0.86**	**0.86**	0.85
	R	0.44	0.83	0.85	0.86	**0.88**
	F	0.53	0.84	0.86	0.86	**0.87**

many applications both by us and for other researchers in the field. BiodivOnto consists of eight core concepts including, e.g., "Organism", "Environment", "Location", and three core relations: *have*, *in*, and *influence*. So far, the population of the ontology with instances is incomplete. Adding instances manually is prohibitively expensive. Since the ontology is the basis of BiodivNERE [16] for both tasks, BiodivBERT could be used to auto-populate the ontology. For instance, *Seabass* would be classified as an "Organism". In a final step, the identified instances should be linked to existing knowledge graphs. e.g., *Seabass* would be mapped to http://www.wikidata.org/entity/Q307102 from Wikidata. This could be done, e.g. using our approach described in [24]. With this, BiodivBERT has the potential to bring us closer to our ultimate goal, the creation of a comprehensive biodiversity knowledge graph out of textual data.

Table 7
Fine-tuning scores RE datasets. The highest score is marked in **bold** while the following score is marked underline. Evaluation Metrics (Met.) are Precision (P), Recall (R), and F1 score (F).

					BiodivBERT	
Dataset	Met.	BiLSTM	$BERT_{BASE}$	$BioBERT_{v1.1}$	+Abs	(+Abs+Full)
BiodivRE	P	0.68	**0.80**	0.79	0.78	0.78
	R	0.68	**0.81**	**0.81**	0.79	0.77
	F	0.68	**0.80**	**0.80**	0.79	0.77
BioReLx	P	0.71	0.83	0.82	**0.85**	0.80
	R	0.78	**0.89**	0.70	0.75	0.74
	F	0.74	**0.86**	0.75	0.79	0.77
EU-ADR	P	0.71	0.91	0.56	**0.92**	**0.92**
	R	**0.69**	0.62	0.53	**0.69**	**0.69**
	F	0.60	0.74	0.54	**0.79**	**0.79**
GAD	P	0.66	0.77	**0.81**	0.77	0.78
	R	0.66	0.77	**0.81**	0.76	0.77
	F	0.66	0.77	**0.81**	0.77	0.78

7. Conclusions & Future Work

In this paper, we introduced BiodivBERT as a pre-trained language model on two domain-specific corpora based on modern research data from the biodiversity domain. In addition, we fine-tuned it on two downstream tasks for text mining: named entity recognition (NER) and relation extraction (RE). BiodivBERT outperforms the state-of-the-art approaches on task-specific datasets.

Future Work: We plan to pre-train and fine-tune a lightweight model, e.g., distilBERT [25]. We also plan to investigate the reasons behind the low scores on the RE task, especially with the BiodivRE and BioRelEx datasets. For example, we could try different settings for fine-tuning. In addition, we fine-tune BiodivBERT on more task-specific datasets whenever they are available. Finally, we deploy BiodivBERT for actual token and sequence prediction for the biodiversity literature to auto-populate the BiodivOnto.

Acknowledgments

The authors thank the Carl Zeiss Foundation for the financial support of the project "A Virtual Werkstatt for Digitization in the Sciences (K3, P5)" within the scope of the program line "Breakthroughs: Exploring Intelligent Systems for Digitization" - explore the basics, use applications. Our sincere thanks to Björn Barz, Luise Modersohn, Jitendra Gaikwad, Anahita Kazem, Alsayed Algergawy, Leila Feddoul, Anirudh Ashok, and Andreas Ostrowski, University Jena for their recommendations, and help.

References

[1] E. S. Brondizio, J. Settele, S. Díaz, H. T. Ngo, Global assessment report on biodiversity and ecosystem services of the intergovernmental science-policy platform on biodiversity and ecosystem services (2019). URL: https://ipbes.net/global-assessment.

[2] N. T.H. Nguyen, R. S. Gabud, S. Ananiadou, Copious: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature, Biodiversity

Data Journal 7 (2019) e29626. URL: https://doi.org/10.3897/BDJ.7.e29626. doi:10.3897/BDJ.7.e29626.

[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems 26 (2013).

[4] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[5] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2227–2237. URL: https://aclanthology.org/N18-1202. doi:10.18653/v1/N18-1202.

[6] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. doi:10.18653/v1/n19-1423.

[7] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.

[8] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, M. McDermott, Publicly available clinical BERT embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 72–78. doi:10.18653/v1/W19-1909.

[9] N. Abdelmageed, F. Löffler, B. König-Ries, Biodivbert: Pre-training corpora dois, 2022. doi:10.5281/zenodo.6555690.

[10] N. Abdelmageed, F. Löffler, B. König-Ries, BiodivBERT: Pre-processed Datasets for NER and RE Downstream Tasks, 2022. doi:10.5281/zenodo.6554208.

[11] N. Abdelmageed, F. Löffler, B. König-Ries, BiodivBERT: Pre-trained weights, configuration, and training arguments, 2022. doi:10.5281/zenodo.6554141.

[12] GROBID, https://github.com/kermitt2/grobid, 2008–2021.

[13] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al., Google's neural machine translation system: Bridging the gap between human and machine translation, arXiv preprint arXiv:1609.08144 (2016).

[14] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[15] F. Löffler, N. Abdelmageed, S. Babalou, P. Kaur, B. König-Ries, Tag me if you can! se-

mantic annotation of biodiversity metadata with the QEMP corpus and the BiodivTagger, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 4557–4564. URL: https://aclanthology.org/2020.lrec-1.560.

[16] N. Abdelmageed, F. Löffler, L. Feddoul, A. Algergawy, S. Samuel, J. Gaikwad, A. Kazem, B. König-Ries, Biodivnere: Gold standard corpora for named entity recognition and relation extraction in the biodiversity domain, Biodiversity Data Journal 10 (2022) e89481. URL: https://doi.org/10.3897/BDJ.10.e89481. doi:10.3897/BDJ.10.e89481.

[17] N. Abdelmageed, F. Löffler, L. Feddoul, A. Algergawy, S. Samuel, J. Gaikwad, A. Kazem, B. König-Ries, BiodivNERE: Gold Standard Corpora for Named Entity Recognition and Relation Extraction in Biodiversity Domain, 2022. URL: https://doi.org/10.5281/zenodo.6458503. doi:10.5281/zenodo.6458503.

[18] E. Pafilis, S. P. Frankild, L. Fanini, S. Faulwetter, C. Pavloudi, A. Vasileiadou, C. Arvanitidis, L. J. Jensen, The species and organisms resources for fast and accurate identification of taxonomic names in text, PloS one 8 (2013) e65390.

[19] M. Gerner, G. Nenadic, C. M. Bergman, Linnaeus: a species name identification system for biomedical literature, BMC bioinformatics 11 (2010) 1–17.

[20] H. Khachatrian, L. Nersisyan, K. Hambardzumyan, T. Galstyan, A. Hakobyan, A. Arakelyan, A. Rzhetsky, A. Galstyan, BioRelEx 1.0: Biological relation extraction benchmark, in: Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, Florence, Italy, 2019, pp. 176–190. URL: https://aclanthology.org/W19-5019. doi:10.18653/v1/W19-5019.

[21] E. M. Van Mulligen, A. Fourrier-Reglat, D. Gurwitz, M. Molokhia, A. Nieto, G. Trifiro, J. A. Kors, L. I. Furlong, The eu-adr corpus: annotated drugs, diseases, targets, and their relationships, Journal of biomedical informatics 45 (2012) 879–884. URL: https://www.sciencedirect.com/science/article/pii/S1532046412000573.

[22] À. Bravo, J. Piñero, N. Queralt-Rosinach, M. Rautschka, L. I. Furlong, Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research, BMC bioinformatics 16 (2015) 1–17. URL: https://bmcbioinformatics.biomedcentral.com/track/pdf/10.1186/s12859-015-0472-9.pdf.

[23] N. Abdelmageed, A. Algergawy, S. Samuel, B. König-Ries, Biodivonto: Towards a core ontology for biodiversity, in: The Semantic Web: ESWC 2021 Satellite Events - Virtual Event, June 6-10, 2021, Revised Selected Papers, volume 12739, Springer, 2021, pp. 3–8. doi:10.1007/978-3-030-80418-3_1.

[24] N. Abdelmageed, S. Schindler, Jentab: Matching tabular data to knowledge graphs., in: Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab 2020) co-located with the 19th International Semantic Web Conference ISWC, 2020, pp. 40–49. URL: https://ceur-ws.org/Vol-2775/paper4.pdf.

[25] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).