Overview of BioNNE Task on Biomedical Nested Named Entity Recognition at BioASQ 2024

Vera Davydova[1], Natalia Loukachevitch[2,*] and Elena Tutubalina[3,4,*]

[1]*HSE University, Russia*
[2]*Lomonosov Moscow State University, Russia*
[3]*Kazan Federal University, Russia*
[4]*Artificial Intelligence Research Institute, Russia*

Abstract

Recognition of nested named entities, which may contain each other, can enhance the coverage of found named entities. This capability is particularly useful for tasks such as relation extraction, entity linking, and knowledge graph population. This paper presents the organizers' report on the BioNNE competition, which focused on nested named entity recognition systems in medical texts for both English and Russian. The competition includes three subtasks: Bilingual, English-oriented, and Russian-oriented. Training and validation sets were derived from a subset of the NEREL-BIO dataset, a corpus of PubMed abstracts. For the BioNNE evaluation, eight of the most common medical entity types were selected from the original dataset. Additionally, a novel test set was developed for the shared task, consisting of 154 abstracts in both English and Russian. Held within the framework of the BioASQ workshop, the competition aims to advance research in nested NER within the biomedical domain.

Keywords

BioNLP, Nested Named Entity Recognition, Biomedical Text Mining, Domain-specific Language Models

1. Introduction

Nested Named Entity Recognition extends the capabilities of standard Named Entity Recognition (NER) task by addressing the challenge of overlapping and nested entities within texts. While more complex, it provides a richer and more nuanced understanding of the entities in a document, making it invaluable for domains requiring precise and hierarchical entity extraction. One such domain is biomedical scientific texts, where nested structures are common. Identification of named entities in scientific texts is essential for extracting valuable information and progressing biomedical research. NNER involves recognizing entities that are nested within other entities. It brings additional complexity by handling the hierarchical and overlapping nature of entities within the biomedical field. Traditional NER systems often fail to adequately capture this nested structure, leading to a loss of crucial information. Recent studies [1, 2, 3] involve leveraging sequence-to-sequence models and reinforcement learning to handle these nested structures more efficiently.

While most studies are focused on flat (non-nesting) NER tasks, there are few general-domain datasets for nested entities [4, 5, 6]. The GENIA corpus [7], a popular NER dataset within the biomedical domain, includes over 100K annotations across 47 entity types, yet only 17% of the entities in the GENIA corpus are nested within another entity [8]. A recent dataset, NEREL [5], is annotated with over 56K named entities of 29 types, while it's biomedical extension, NEREL-BIO [9], is annotated with over 70k entities of 37 types.

This paper provides a comprehensive overview of the Biomedical Nested Named Entity Recognition (BioNNE) task, predominantly focusing on the challenges and advancements in nested NER across annotated PubMed abstracts. The shared task includes both English and Russian biomedical texts and was part of the BioASQ Workshop 2024 [10]. The BioASQ shared tasks aim to advance systems that exploit diverse and extensive online information to meet the information needs of biomedical scientists.

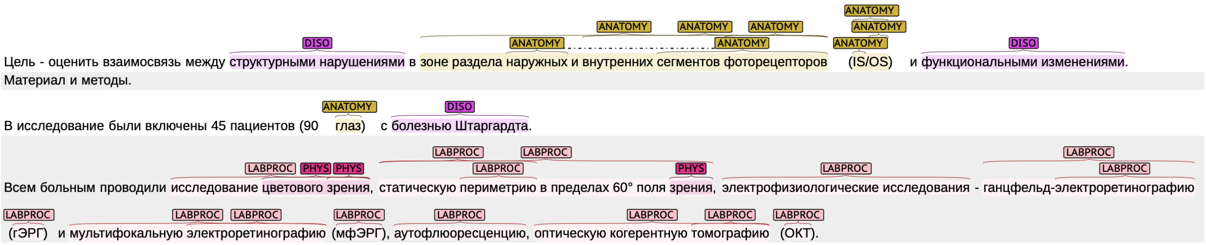
CLEF 2024: Conference and Labs of the Evaluation Forum, 9-12 September 2024, Grenoble, France

*Corresponding author.

veranchos@gmail.com (V. Davydova); louk_nat@mail.ru (N. Loukachevitch); tutubalinaev@gmail.com (E. Tutubalina)
0009-0003-0845-2386 (V. Davydova); 0000-0002-1883-4121 (N. Loukachevitch); 0000-0001-7936-0284 (E. Tutubalina)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3740/paper-03.pdf



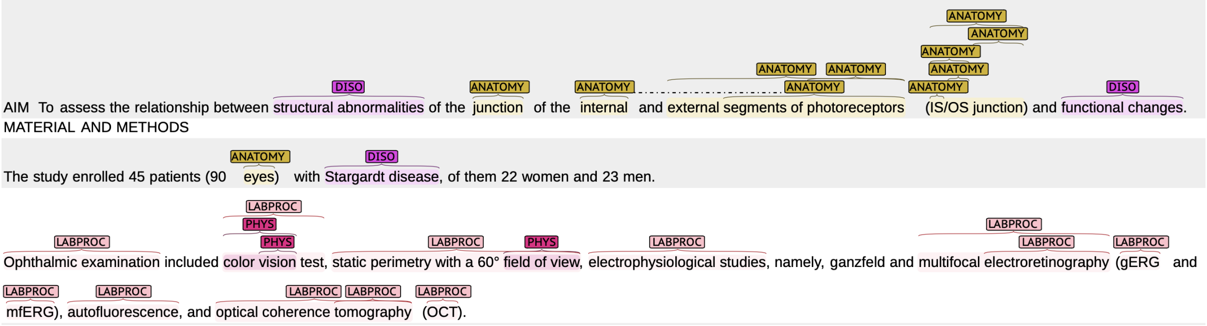
Цель - оценить взаимосвязь между структурными нарушениями в зоне раздела наружных и внутренних сегментов фоторецепторов (IS/OS) и функциональными изменениями.
Материал и методы.

В исследование были включены 45 пациентов (90 глаз) с болезнью Штаргардта.

Всем больным проводили исследование цветового зрения, статическую периметрию в пределах 60° поля зрения, электрофизиологические исследования - ганцфельд-электроретинографию (гЭРГ) и мультифокальную электроретинографию (мфЭРГ), аутофлюоресценцию, оптическую когерентную томографию (ОКТ).

ANATOMY
ANATOMY ANATOMY ANATOMY ANATOMY
DISO ANATOMY ANATOMY ANATOMY DISO
ANATOMY DISO
LABPROC LABPROC LABPROC
LABPROC PHYS PHYS LABPROC PHYS LABPROC LABPROC
LABPROC LABPROC LABPROC LABPROC LABPROC LABPROC LABPROC LABPROC

Figure 1: Sample of annotations of nested entities in the Russian abstract (PMID: 27456564).



AIM To assess the relationship between structural abnormalities of the junction of the internal and external segments of photoreceptors (IS/OS junction) and functional changes.
MATERIAL AND METHODS

The study enrolled 45 patients (90 eyes) with Stargardt disease, of them 22 women and 23 men.

Ophthalmic examination included color vision test, static perimetry with a 60° field of view, electrophysiological studies, namely, ganzfeld and multifocal electroretinography (gERG and mfERG), autofluorescence, and optical coherence tomography (OCT).

ANATOMY
ANATOMY
ANATOMY
ANATOMY ANATOMY ANATOMY
DISO ANATOMY ANATOMY ANATOMY ANATOMY DISO
ANATOMY DISO
LABPROC
PHYS
LABPROC
LABPROC PHYS LABPROC PHYS LABPROC LABPROC LABPROC
LABPROC LABPROC LABPROC LABPROC LABPROC

Figure 2: Sample of annotations of nested entities in the English abstract (PMID: 27456564).

To this end, we setup both monolingual and bilingual sub-tasks based on the NEREL-BIO dataset. This paper (i) introduces the corpus (Section 3) and (ii) reports the results of the BioNNE Shared Task 4. All the materials can be found on BioNNE GitHub page[1] and on Codalab[2] competition page.

2. BioNNE shared Task

The main task is to extract and classify biomedical nested named entities mentioned in the unstructured medical abstract text. The main task consists of three tracks: (i) Bilingual, (ii)English-oriented, and (iii) Russian-oriented.

- Bilingual: participants in this track were required to train a single multi-lingual NER model using training data for both Russian and English languages. The model was supposed to be used to generate prediction files for each language's dataset. Predictions from any mono-lingual model were not allowed in this track.
- English-oriented: participants in this track were required to train a nested NER model for English scientific abstracts in the biomedical domain.
- Russian-oriented: participants in this track were required to train a nested NER model for Russian scientific abstracts in the biomedical domain.

The same predictions from track (i) were not allowed in tracks (ii) and (iii). Participants were allowed to train any model architecture on any publicly available data to achieve the best performance.

3. Dataset

Training and validation sets for the BioNNE competition were based on a subset of NEREL-BIO dataset [9]. NEREL-BIO is a corpus of PubMed abstracts written in Russian and English. It enhances the NEREL [5] dataset, originally designed for the general domain, by incorporating biomedical entity

[1] https://github.com/nerel-ds/NEREL-BIO/tree/master/bio-nne/
[2] https://codalab.lisn.upsaclay.fr/competitions/16464

Table 1
The list of entity types from NEREL-BIO dataset that were used for BioNNE task.

Entity Label	Explanation	Examples
FINDING	does not have direct correspondence in UMLS, it conveys longer hospital stay, stopped the progression of the results of scientific study described in the abstract	keratoconus, stabilize the glaucoma process
PHYS	biological function or process in organism including organism attribute (temperature) and excluding mental processes	blood flow, childbirth, uterine contraction, arterial pressure, body temperature
INJURY_POISONING	damage inflicted on the body as the direct or indirect result	overdosing, burn, drowning, of external force including poisoning falling, childhood trauma
DISO	any deviations from normal state of organism: diseases, symptoms, abnormality of organ, excluding injuries or poisoning	appendicitis, haemorrhoids, magnesium deficiency dysfunctions, Diabetes Mellitus, spine pain, complication, bone cyst, acute inflammation, deep vein thrombosis
LABPROC	testing body substances and other diagnostic procedures	biochemical analysis, polymerase chain reaction test, such as ultrasonography electrocardiogram, histological
ANATOMY	comprises organs, body part, cells and cell components	eye, bone, brain, lower limb, oral cavity, blood, body substances anterior lens capsule, right ventricle, lymphocyte
CHEM	chemicals including legal and illegal drugs, biological molecules	opioid, lipoprotein, iodine, adrenalin, memantine, molecules methylprednisolone
DEVICE	manufactured objects used for medical purposes	catheter, prosthesis, tonometer, tomograph removable prosthesis, stent, metal stent

types. Biomedical entity types in NEREL-BIO are annotated according to UMLS definitions of relevant concepts. All the abstracts are annotated in the BRAT format [11].

Figures 1 and 2 present parallel examples of nested named entities in NEREL-BIO for one abstract. Table 1 provides a comprehensive list of entity types, along with their explanations and examples.

Compared to the original NEREL-BIO dataset, we fixed some annotators' errors, merged PRODUCT and DEVICE type classes into DEVICE class and selected the eight most common medical entities from the dataset: FINDING, DISO, INJURY_POISONING, PHYS, DEVICE, LABPROC, ANATOMY, CHEM. The resulting dataset comprises 662 annotated PubMed abstracts in Russian and 104 parallel abstracts in Russian and English. 104 parallel abstracts were randomly split for training and validation sets for each subtask.

A novel test set was developed for the shared task, consisting of 154 abstracts in English and Russian. To avoid manual annotation, 346 extra files were added for each language, resulting in 500 abstracts for each of the target languages. These supplementary files were excluded from the final evaluation.

Table 2 shows the number of entities represented in each part of the data set. Observations can be summarized as follows. Entities labeled as DISO and ANATOMY are the most frequent across all sets, with DISO being particularly prevalent in both training and test sets. Categories such as DEVICE and INJURY_POISONING have a much lower number of entities compared to others, highlighting potential areas where entity recognition might be more challenging due to data sparsity. The number of entities in the English test set (EN_test) and the Russian test set (RU_test) are relatively comparable, although slight variations are observed, particularly in the ANATOMY and DISO categories.

Table 2
Number of entities in each subset of the BioNNE 2024 dataset.

Entity Label	Refined NEREL-BIO				Novel BioNNE test set	
	RU_train	EN_train	RU_dev	EN_dev	EN_test	RU_test
number of documents	716	54	50	50	154	154
number of entities						
DISO	11169	2043	914	1012	2995	2730
ANATOMY	8346	911	869	897	2221	2157
PHYS	5742	397	354	379	1569	1570
CHEM	4741	579	527	575	1308	1203
FINDING	4808	456	350	348	1365	1388
LABPROC	1618	190	138	154	401	405
DEVICE	734	20	33	28	168	165
INJURY_POISONING	643	90	25	20	129	120
Total	37369	4686	3210	3413	10156	9738

4. Experiments

4.1. Evaluation Metric

F$_1$ was used as the main evaluation metric. It is calculated according to the following formula:

$$F_1 = \frac{1}{n} \sum_{c \in C} F_{1_{rel_c}} \tag{1}$$

where C ={“FINDING”, “DISO”, “INJURY_POISONING”, “PHYS”, “DEVICE”, “LABPROC”, “ANATOMY”, “CHEM”}, n is the size of C, $F_{1_{rel_c}}$ is macro F$_1$-score averaged over all entity classes.

4.2. Baseline Solution

We leveraged the BINDER model [12] as a baseline solution for the BioNNE task. BINDER utilizes two encoders to map text and entity types into a shared vector space. It efficiently reuses vector representations of entity types for various text spans (or vice versa), leading to accelerated training and inference speeds. Leveraging bi-encoder representations, BINDER introduces a contrastive learning framework for NER. This framework facilitates similarity between the representations of entity types and their corresponding mentions while encouraging dissimilarity with non-entity text spans. Additionally, BINDER introduces a dynamic thresholding loss in contrastive learning. During testing, it employs candidate-specific dynamic thresholds to differentiate entity spans from non-entity ones. For our backbone model, we utilized the multilingual BERGAMOT model[3] [13], which is pre-trained on the Unified Medical Language System (UMLS) (version 2020AB) using a Graph Attention Network (GAT) encoder [14]. The best-performing results were achieved with the following hyperparameters: a learning rate of 3e-5 and 5 training epochs. AdamW was used as the optimizer [15].

To address the cross-lingual transfer problem, the baseline model was trained and evaluated on various language variations: RU, EN, and RU+EN. The highest scores were achieved on the combined RU and EN subsets (see Tab. 3). Additionally, models that were trained on one language showed comparatively high results when evaluated on the other language, with training on Russian data proving to be more effective. This effectiveness can be attributed to the difference in the size of the training data. Thus, combined with the results from the participants' models, we can conclude that cross-lingual techniques can be effectively applied to the NNER task.

Table 3
Results (F_1 scores on the test sets) of bilingual and monolingual subtasks. The best result in each task is bolded.

Model	Both (Track 1)	English (Track 2)	Russian (Track 3)
fulstock	**0.7044**	**0.6181**	**0.6981**
hasin.rehana	0.5053	0.5636	0.6007
wenxinzh	-	0.3480	-
vampire	-	0.1203	-
baseline EN_train	0.4061	0.5561	0.4041
baseline RU_train	0.4866	0.4041	0.5849
baseline RU+EN_train	0.6430	0.5655	0.6732

4.3. Official BioNNE Results

We observed a strong interest in the shared task, with 26 teams registered in CodaLab. We have received 155 submissions from 5 teams. One team opted to withdraw their results from the official publication. We summarized performance for all tracks in Table 3. Below, we give an overview of these approaches.

Team **fulstock** achieved the best results by using the BINDER model. In contrast with the baseline architecture, it has XLM-RoBERTa [16] as a backbone model. The participant experimented with different ways of entity type description (prompts) for BINDER learning. The following variants of prompts were used: keyword (name of the entity type); 2, 5 or 10 the most frequent component words for entity type in the training data, contextual prompt (an example of a sentence with the target entity), lexical prompt (an example of a sentence, in which the target entity is masked with the entity label) [17]. The model was trained during 64 epochs. Results have shown that contextual Russian named entity type description proved to be the best option for the bilingual track (achieving 0.704 in F1-score), while for the Russian-only track, one worked the best (F1-score is 0.698). In the English track, the 10 most frequent English components prompt resulted in the 0.6181 F1-score. Thus, these prompts benefited from getting first place in the BioNNE competition on all three tracks.

Team **wenxinzh** [18] used the combination of a pre-trained Mixtral model [19] and en_ner_bc5cdr_md, a spaCy NER model trained on the BC5CDR corpus [20]. They also adapted and customized rules based on semantic types of UMLS (Unified Medical Language System). First, the system uses Mixtral and en_ner_bc5cdr_md to extract potential entities for each category from the text. Then, the system finds the UMLS semantic type associated with the entities to determine their final entity types. The team applied the system to the English subtask and achieved third place in the overall results, with an F1 score of 0.348.

Team **hasin.rehana** [21] processed the BioNNE dataset by splitting each abstract into sentences and mapping the corresponding annotations to these sentences. Then, they implemented the BIO-tagging scheme, a well-known method for named entity recognition encoding. Tokens were encoded as B-TYPE for the beginning of an entity, I-TYPE for subsequent tokens of the same entity, and O for tokens that do not belong to any entity class. Overall, six levels of BIO-tagging were applied to the BioNNE dataset. The core of the model for English NNER is the pre-trained PubMedBERT, which provides contextualized word embeddings [22]. For the Russian NNER task, the team used a pre-trained SBERT-Large-NLU-RU model[4]. For Bilingual NNER, they have employed BERT-Base-Multilingual-uncased[23]. A series of six classification layers were added to the base model. Each layer was designed to output a specific level of NER tags, with each linear layer taking the hidden states from PubMedBERT and mapping them to the required number of labels for that layer. Although the original number of classes in the BioNNE dataset is eight, the total number of output classes for each classification layer is 17 to support the preprocessed BIO-tagged dataset. This includes “B-Class” and “I-Class” for each of the eight original classes, as well as “O” class for any token that does not belong to any entity class. To enrich the NER process, the team leveraged the UMLS Metathesaurus for vocabulary expansion. They utilized the MRCONSO.RRF data file within UMLS to extract relevant concepts and their child concepts based on the concept IDs

[3] https://huggingface.co/andorei/BERGAMOT-multilingual-GAT
[4] https://huggingface.co/ai-forever/sbert_large_nlu_ru

provided by the BioNNE challenge organizers, which broadened the model's ability to recognize entities by incorporating synonyms and related terms. For these experiments, the team has employed 6 NVIDIA Tesla V100 GPUs with 32GB of HBM2 VRAM each.

5. Conclusion

In this paper, we present the organizers' report on the competition for nested named entity recognition systems in the biomedical domain (BioNNE). The competition included three subtasks: Bilingual, English-oriented, and Russian-oriented. The participants were asked to extract the eight most common medical entities, both in Russian and English, which can contain each other, from PubMed abstracts. The best results in the evaluation were achieved by using the BINDER model based on bi-encoder representations and a contrastive learning framework. The winner experimented with different ways of entity type description (prompts) for BINDER learning. We hope that the outcomes of the competition will foster further research and development in nested NER for healthcare applications.

Acknowledgments

The work of E.T. has been supported by the Russian Science Foundation grant # 23-11-00358. We would like to thank all the participating teams who contributed to the success of the shared task through their interesting approaches and experiments.

References

[1] Y. Yang, X. Hu, F. Ma, S. Li, A. Liu, L. Wen, P. S. Yu, Gaussian prior reinforcement learning for nested named entity recognition, arXiv preprint arXiv:2305.12003 (2023). URL: https://arxiv.org/abs/2305.12003.

[2] P. Wajsburt, Y. Taillé, X. Tannier, Effect of depth order on iterative nested named entity recognition models, arXiv preprint arXiv:2104.00542 (2021). URL: https://arxiv.org/abs/2104.00542.

[3] U. Yaseen, P. Gupta, H. Schütze, Linguistically informed relation extraction and neural architectures for nested named entity recognition in bionlp-ost 2019, arXiv preprint arXiv:1910.03549 (2019). URL: https://arxiv.org/abs/1910.03549.

[4] H. Ming, J. Yang, L. Jiang, Y. Pan, N. An, Few-shot nested named entity recognition, arXiv preprint arXiv:2212.00968 (2022). URL: https://arxiv.org/abs/2212.00968.

[5] N. Loukachevitch, E. Artemova, T. Batura, P. Braslavski, V. Ivanov, S. Manandhar, A. Pugachev, I. Rozhkov, A. Shelmanov, E. Tutubalina, et al., Nerel: a russian information extraction dataset with rich annotation for nested entities, relations, and wikidata entity links, Language Resources and Evaluation (2023) 1–37.

[6] E. Artemova, M. Zmeev, N. Loukachevitch, I. Rozhkov, T. Batura, V. Ivanov, E. Tutubalina, Runne-2022 shared task: Recognizing nested named entities, Komp'juternaja Lingvistika i Intellektual'nye Tehnologii 2022 (2022) 33 – 41. doi:10.28995/2075-7182-2022-21-33-41.

[7] J.-D. Kim, T. Ohta, Y. Tateisi, J. Tsujii, Genia corpus—a semantically annotated corpus for bio-textmining, Bioinformatics 19 (2003) i180–i182.

[8] A. Katiyar, C. Cardie, Nested named entity recognition revisited, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1, 2018.

[9] N. Loukachevitch, S. Manandhar, E. Baral, I. Rozhkov, P. Braslavski, V. Ivanov, T. Batura, E. Tutubalina, NEREL-BIO: A Dataset of Biomedical Abstracts Annotated with Nested Named Entities, Bioinformatics (2023). URL: https://doi.org/10.1093/bioinformatics/btad161. doi:10.1093/bioinformatics/btad161, btad161.

[10] A. Nentidis, G. Katsimpras, A. Krithara, S. Lima-López, E. Farré-Maduell, M. Krallinger, N. Loukachevitch, V. Davydova, E. Tutubalina, G. Paliouras, Overview of BioASQ 2024: The

twelfth BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. Maria Di Nunzio, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[11] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, brat: a web-based tool for NLP-assisted text annotation, in: Proceedings of the Demonstrations Session at EACL 2012, Association for Computational Linguistics, Avignon, France, 2012.

[12] S. Zhang, H. Cheng, J. Gao, H. Poon, Optimizing bi-encoder for named entity recognition via contrastive learning, in: The Eleventh International Conference on Learning Representations, 2022.

[13] A. Sakhovskiy, N. Semenova, A. Kadurin, E. Tutubalina, Biomedical entity representation with graph-augmented multi-objective transformer, in: Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico, 2024.

[14] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph attention networks, arXiv preprint arXiv:1710.10903 (2018).

[15] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

[16] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[17] I. Rozhkov, N. Loukachevitch, Prompts in few-shot named entity recognition, Pattern Recognition and Image Analysis 33 (2023) 122–131.

[18] W. Zhou, Biomedical Nested NER with Large Language Model and UMLS Heuristics, in: CLEF Working Notes, 2024.

[19] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mixtral of experts, 2024. arXiv:2401.04088.

[20] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, Z. Lu, Biocreative V CDR task corpus: a resource for chemical disease relation extraction, Database J. Biol. Databases Curation 2016 (2016). URL: https://doi.org/10.1093/database/baw068. doi:10.1093/database/baw068.

[21] H. Rehana, B. Bansal, N. Bengisu Çam, J. Zheng, Y. He, A. Özgür, J. Hur, Nested Named Entity Recognition using Multilayer BERT-based Model, in: CLEF Working Notes, 2024.

[22] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Transactions on Computing for Healthcare (HEALTH) 3 (2021) 1–23.

[23] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.