Multilingual and Nested Biomedical Named Entity Normalization via Candidate Retrieval and Lightweight Large Language Model Disambiguation

Notebook for the ICUE@BioNNE-L Shared Task at CLEF 2025

Antoine D. Lain[1,†], Chaeeun Lee[2,†], Simona E. Doneva[3,†], Maria Juliana Rodriguez-Cubillos[2,†], Elisa Castagnari[2,†], T. Ian Simpson[2,*] and Joram M. Posma[1,*]

[1]*Section of Bioinformatics, Division of Systems Medicine, Department of Metabolism, Digestion, and Reproduction, Faculty of Medicine, Imperial College London, London W12 0NN, United Kingdom*

[2]*School of Informatics, University of Edinburgh, 10 Crichton Street, EH8 9AB, Edinburgh, UK*

[3]*Center for Reproducible Science, University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland*

Abstract

In this work, we present our approach to the BioNNE-L 2025 task, which focuses on Named Entity Normalization (NEN) across multilingual biomedical corpora. The task involves mapping entity mentions to standardised concepts in a multilingual vocabulary, covering English, Russian, and mixed-language. Our system adopts a retrieval-based strategy, leveraging BioSyn models with tailored vocabulary subsetting to address memory constraints and enhance retrieval efficiency. For the multilingual setting, we trained a single BioSyn model and applied post-processing using a lightweight large language model (LLM) to improve top-rank accuracy by re-scoring candidates based on contextual meaning. Our approach achieved competitive results on the official leaderboard at Acc@5, with improvements in Russian monolingual performance compared to the baseline (Acc@1: 0.62 vs. 0.52 baseline) and a 2% Acc@1 gain in the multilingual task after applying LLM post-processing. These results underline the challenge of ranking in retrieval-based NEN, particularly given the considerable difference observed between the top-1 candidate and the top-5 candidates accuracy scores. Additionally, our findings demonstrate the limitations of retrieval-only systems in highly ambiguous settings, and demonstrate the value of hybrid pipelines that combine candidate retrieval with contextual disambiguation. This work focuses on low-resource multilingual biomedical NEN, especially to mitigate the risks of hallucination in resource-limited environments.

Keywords

Biomedical Natural Language Processing, Named Entity Normalization, Entity Linking, Nested Named Entity Normalization

1. Introduction

Biomedical Named Entity Normalization (BioNEN), also known as Entity Linking (EL) [1, 2] and in the biomedical domain sometimes referred to as (Bio)medical Concept Normalization (BCN/MCN) [3], is a task of Information Extraction in Biomedical Natural Language Processing (BioNLP), aiming to map entity mentions in text to standardised concepts within biomedical knowledge bases, or ontologies, such as the Unified Medical Language System (UMLS) [4]. While significant advances have been made in biomedical NEN for English texts, nested entity and multilingual concept normalization remain challenging [5].

Nested entities, in which one entity mention is embedded within another, frequently occur in biomedical literature and present difficulties for both named entity recognition and normalization systems. To illustrate this, "vertebral", "lumbar vertebral", "lumbar vertebral canal", and "lumbar vertebral

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

†These authors contributed equally.

a.lain@imperial.ac.uk (A. D. Lain); chaeeun.lee@ed.ac.uk (C. Lee); simona.doneva@uzh.ch (S. E. Doneva); juliana.rodriguez@ed.ac.uk (M. J. Rodriguez-Cubillos); e.castagnari@ed.ac.uk (E. Castagnari); ian.simpson@ed.ac.uk (T. I. Simpson); jmp111@ic.ac.uk (J. M. Posma)

canal stenosis" appear within the same longest entity, with entities hierarchically embedded inside one another. This example shows how simple mentions can be progressively contained within increasingly complex biomedical entites, complicating both detection and correct mapping to a knowledge base like UMLS.

Furthermore, with the growing volume of biomedical/clinical applications in languages other than English, there is an increasing need for robust multilingual NEN systems. Addressing these challenges in languages such as Russian, where terminology resources aligned with UMLS are limited, further complicates the task due to incomplete terminology coverage, language-specific variation, and multilingual synonymy.

The BioNNE-L Shared Task [6, 7, 8, 9], organised as part of the BioASQ challenge [10], directly addresses these challenges by providing a dataset for nested BioNEN in both English and Russian biomedical abstracts. The task involves normalising mentions of biomedical entities from the following categories: disease, chemical, and anatomy, to their corresponding UMLS concepts.

In this paper, we present a system developed for the BioNNE-L Shared Task that uses BioSyn [11], a biomedical entity normalization model based on biomedical entity representations with synonym marginalisation. As part of our approach, we applied a pre-processing pipeline incorporating FastText embeddings [12] combined with cosine similarity and morphological distance measures to enhance candidate concept retrieval from the vocabulary provided. Following initial candidate generation and ranking by BioSyn, we implemented a post-processing module using a lightweight LLM called DeepSeek R1 Distill Llama 8B [13] with reasoning ability to re-arrange the top 5 candidate predictions based on context and coherence.

2. Related Work

2.1. Available Biomedical Named Entity Normalization Corpora

The development and evaluation of BioNEN systems depend on annotated corpora linking entities mentioned in text to standardised concepts in biomedical knowledge bases such as UMLS. Some datasets have been made available for this purpose in English. The NCBI Disease corpus [14] provides disease mentions in PubMed abstracts, each mapped to either MeSH [1] or OMIM [15] identifiers. Similarly, BC5CDR [16] covers both chemical and disease entities within biomedical abstracts, offering annotations for both Named Entity Recognition (NER) and entity normalization. For gene normalization, BC2GN [17] links gene mentions to Entrez Gene identifiers [18], and TAC 2017 ADR [19] offers annotations for adverse drug reactions with normalization to MedDRA concepts. These corpora have enabled the training and evaluation of NEN systems. However, multilingual biomedical NEN corpora remain limited.

2.2. Biomedical Named Entity Normalization Methods

BioNEN has progressed from early rule-based and dictionary lookup approaches to embedding-based models that leverage dense vector representations of both entity mentions and candidate concepts. Traditional systems often relied on exact string matching, heuristic post-processing, and handcrafted rules. Modern methods use deep learning to compute contextual and semantic similarity. Among these, BioSyn represents an embedding-based approach. It uses an encoder setup with a language model, for example BioBERT [20], to encode mentions and candidate names into a shared vector space, computing similarity via dot-product. BioSyn further integrates a synonym marginalisation mechanism to better capture variant expressions of the same biomedical concept. Other NEN methods include SapBERT [21], which applies self-alignment pre-training for concept-level embedding alignment, and NoteContrast [22], a contrastive pre-trained model optimised for clinical coding. While most recent developments have focused on English biomedical text, research into multilingual NEN, particularly for nested entity structures, remains limited. The application of Large Language Models (LLMs) to

[1] https://www.nlm.nih.gov/mesh/meshhome.html

BioNEN has been used in a previous challenge [23, 24, 25]. The objective of the BioCreative VIII Track 3 challenge was to extract discontinuous phenotypic key medical findings embedded within EHR texts and subsequently normalize these findings to their Human Phenotype Ontology (HPO) terms. Recent LLM-based approaches typically frame NEN as a retrieval or ranking problem, where mention strings are mapped to a candidate set of concept identifiers from a knowledge base. Studies [24, 25, 5] have explored the use of in-context learning and prompt-based strategies to adapt LLMs like GPT-3.5, GPT-4 and lightweight LLMs for biomedical concept normalization tasks. However, lightweight LLMs remain limited when in a few-shot scenario, when the vocabulary is extremely large, or when the vocabulary contains extremely similar terms [5].

2.3. BERT-based Models in English, Russian, and Multilingual Contexts

Transformer-based models [26] pre-trained on biomedical corpora have substantially improved both NER and NEN tasks. For English biomedical applications, models such as BioBERT, SciBERT [27], and SapBERT are well known, offering domain-specific embeddings fine-tuned on PubMed abstracts and PMC full-text articles.

In the Russian biomedical domain, resources have been comparatively limited. However, recent efforts have produced models such as Gherman/bert-base-NER-Russian, fine-tuned for Russian NER, and nesemenpolkov/msu-wiki-ner [28], a multilingual BERT model fine-tuned on Russian entity recognition datasets. These models have been evaluated in general biomedical NER contexts and hold potential for adaptation to NEN. There is also RuDR-BERT[2] [29], which is pre-trained on 1.4 million health-related user-generated texts collected from various Internet sources, including social media.

For multilingual biomedical tasks, models like Babelscape/wikineural-multilingual-ner [28], google-bert/bert-base-multilingual-uncased [30], cambridgeltl/SapBERT-UMLS-2020AB-all-lang-from-XLMR[3] [1], GanjinZero/coder_all[4] [2], and andorei/BERGAMOT-multilingual-GAT[5] [3] provide multilingual embeddings. Although their biomedical vocabulary coverage is limited compared to domain-specific English models, these multilingual transformers offer a foundation for developing multilingual NEN systems, particularly when fine-tuned.

3. Dataset

The shared task organiser provided the participants with three datasets: two for the monolingual sub-tasks (in English and Russian), and one for the bilingual sub-task, which represents a combination of the two monolingual datasets. Each dataset was split into training, validation, and test, with a shared normalization vocabulary for candidate selection. This vocabulary consisted of concept names and their corresponding Concept Unique Identifiers (CUIs) from the UMLS. The vocabulary included a total of 4,047,990 terms, mapping to 1,510,431 unique UMLS CUIs. Of these terms, 145,803 were in Russian, with the remainder in English.

In the English monolingual dataset, the training set had 2,690 entities, the validation set 2,494 entities, and the test set 6,661 entities. The training set contained 1,119 unique CUIs, while the validation set contained 932 unique CUIs. For reference, 966 CUIs in the training set did not appear in the validation set, and 779 CUIs in the validation set were absent from the training set, resulting in a limited overlap between the training and the validation datasets. For the Russian monolingual dataset, the training set had 24,255 entities, the validation set 2,334 entities, and the test set 6,215 entities. The Russian training set contained 4,214 unique CUIs, of which 3,745 CUIs were not present in the validation set. The validation set had 875 unique CUIs, with 406 CUIs absent from the training data, indicating a close train/validation ratio overlap with the English dataset.

The bilingual dataset was constructed by combining the English and Russian monolingual datasets.

[2]https://huggingface.co/cimm-kzn/rudr-bert
[3]https://huggingface.co/cambridgeltl/SapBERT-UMLS-2020AB-all-lang-from-XLMR
[4]https://huggingface.co/GanjinZero/coder_all
[5]https://huggingface.co/andorei/BERGAMOT-multilingual-GAT

4. Methodology

Unlike nested NER, which requires identifying multiple overlapping or nested spans within a text, the BioNNE-L Shared Task dataset provides pre-annotated entity mentions, including any nested terms. This allowed us to treat each nested entity independently during normalization, without needing to infer its hierarchical relation to a larger entity mention. However, a key challenge remained as the vocabulary still contained both the nested terms and their parent entities, often with highly similar embedding values. As a result, the systems needed to learn to retrieve the most relevant concept for each mention individually, despite the presence of closely related or overlapping candidates in the vocabulary.

4.1. Monolingual English Task

For the English sub-task, the initial pre-processing step involved reducing the size of the vocabulary, which was too large to be loaded in full by BioSyn on our hardware. To achieve this, we computed both morphological and embedding similarities between each entity mention in the test set and every term in the vocabulary. We used the `FastText` English model (`cc.en.300.bin`) to generate word embeddings, and used `cosine similarity` to compute embedding-based similarity (dense similarity) alongside morphological similarity (sparse similarity). A similarity threshold of 0.7 was applied, retaining only those vocabulary entries with at least one test entity similarity above this threshold. This reduced the vocabulary size from over 4 million to 338,209 entities, potentially covering 98% of the test set entities.

Following the vocabulary reduction, we did model selection by evaluating three BioSyn variants on the validation set:

- **dmis-lab/biobert-v1.1** [6] Acc@1: 0.79
- **allenai/scibert_scivocab_uncased** [7] Acc@1: 0.78
- **cambridgeltl/SapBERT-from-PubMedBERT-fulltext** [8] Acc@1: 0.81

SapBERT had the highest *Acc@1* and *Acc@5*, and was selected as the final model for the English task. No post-processing steps were applied to the monolingual English predictions.

4.2. Monolingual Russian Task

A similar approach was adopted for the Russian sub-task. Here, we used the `FastText` Russian model (`cc.ru.300.bin`) to compute embedding and morphological similarity scores using `cosine similarity`. Applying the same threshold of 0.7 reduced the vocabulary to 294,824 entities, potentially covering 91% of the Russian test set entities.

For the model selection, we evaluated the following Russian-language BERT-based models:

- **Gherman/bert-base-NER-Russian** [9] Acc@1: 0.91
- **nesemenpolkov/msu-wiki-ner** [10] Acc@1: 0.90
- **KoichiYasuoka/bert-base-russian-upos** [11] Acc@1: 0.92

bert-base-russian-upos had the best validation performance in terms of *Acc@1* and *Acc@5*, and was selected for the final system. No post-processing steps were applied to the monolingual Russian predictions.

[6] https://huggingface.co/dmis-lab/biobert-v1.1
[7] https://huggingface.co/allenai/scibert_scivocab_uncased
[8] https://huggingface.co/cambridgeltl/SapBERT-from-PubMedBERT-fulltext
[9] https://huggingface.co/Gherman/bert-base-NER-Russian
[10] https://huggingface.co/nesemenpolkov/msu-wiki-ner
[11] https://huggingface.co/KoichiYasuoka/bert-base-russian-upos

4.3. Multilingual Task

For the multilingual task, the same pre-processing procedure was applied, with one modification: the similarity threshold was increased to 0.8. This adjustment was necessary as the combined English and Russian reduced vocabulary at a threshold of 0.7 remained too large to be loaded on our hardware. Applying the higher threshold reduced the vocabulary to 59,234 entities, potentially covering 84% of the multilingual test set entities.

Model selection was done using the following multilingual models:

- **Babelscape/wikineural-multilingual-ner**[12] Acc@1: 0.85
- **google-bert/bert-base-multilingual-uncased** [13] Acc@1: 0.82

wikineural-multilingual-ner was the best-performing model on the validation set for both *Acc@1* and *Acc@5*.

Unlike the monolingual tasks, a post-processing step was introduced for the multilingual task. After generating the top-5 candidate predictions with BioSyn, we used **DeepSeek-R1-Distill-Llama-8B** (via the Hugging Face Transformers library [31]) in a few-shot setting, part of the prompt can be found in Figure 1. We selected DeepSeek R1 for our LLM re-ranking step due to its reasoning capabilities, which were essential for assessing contextual differences and resolving ambiguities between morphologically similar candidates. The model was prompted to reorder the candidate list based on the surrounding textual context of the entity mention. In cases where none of the candidate CUIs appeared contextually appropriate, the model was allowed to output a special placeholder label, CUILESS, indicating that no suitable candidate was identified.

```
[
  {
    "role": "system",
    "content": "You are a world-class expert in named entity normalization (NEN) for
    clinical and biomedical text, trained to disambiguate ambiguous entities based on context.
    Your task is to select the most appropriate identifier for a target entity mention using
    the provided context, entity candidates, and their definitions."
  },
  {
    "role": "user",
    "content": "### Instructions
    - You will receive:
    - A short text snippet with the entity to normalize, marked by `<ENTITY>` `</ENTITY>` tags.
    - A list of identifier candidates in the format: `Cx: Definition` (may include synonyms,
    semantic variants, or contextual descriptions)
    - A target entity term (string) to normalize, which may be a substring of the marked text.
    - If none of the candidates fit based on the nested entity or context, return:
    ```
    <OUTPUT>CUILESS</OUTPUT>
    ```
    - Otherwise, output up to **5 candidates max**, ranked in order of best fit,
    comma-separated inside `<OUTPUT></OUTPUT>` tags.
    Example:
    ```
    <OUTPUT>C3,C1,C2</OUTPUT>
    ```
    ..."
  }
]
```

Figure 1: Example LLM Prompt for candidate re-ranking.

[12] https://huggingface.co/Babelscape/wikineural-multilingual-ner
[13] https://huggingface.co/google-bert/bert-base-multilingual-uncased

5. Experimental Setup

All experiments were done on a Linux workstation equipped with an Intel 24-core i9-13900K CPU, 192GB RAM (4×48GB), and an Nvidia GeForce RTX 4090 GPU featuring 24GB of memory.

We trained BioSyn[14] with the following training parameters: `topk` set to 20, 10 epochs, a `train_batch_size` of 16, an initial sparse weight of 0, a learning rate of 1×10^{-5}, a `max_length` of 25, and a `dense_ratio` of 0.5. A key component of the BioSyn architecture is its combination of sparse lexical retrieval (using TF-IDF scores) and dense semantic retrieval (using entity embeddings). The dense encoder component is fine-tuned using a marginal maximum likelihood (MML) objective. This joint retrieval strategy enables BioSyn to balance exact string matching with semantic similarity. We fine-tuned BioSyn separately for each setting (English, Russian, and multilingual) using domain-specific transformer encoders selected through validation set performance. Model evaluation was performed using BioSyn's *Hybrid* method, which combines the sparse and dense retrieval scores for candidate ranking.

For the multilingual post-processing step, the same system configuration was used. The "deepseek-ai/DeepSeek-R1-Distill-Llama-8B" model was loaded via the Hugging Face Transformers library, configured with do_sample=False and dtype=torch.bfloat16 for improved inference speed and memory efficiency.

No hyperparameter tuning was performed for either the BioSyn training or DeepSeek post-processing phases.

6. Results and Discussion

6.1. System Performance

This section presents the official test set results of our system submitted to the leaderboard for the two subtasks (monolingual and multilingual), compared against the provided baselines. All results are reported as accuracy at Acc@1, meaning the correct normalization was our top candidate, and accuracy at Acc@5, meaning the correct normalization was part of our top 5 candidates.

6.1.1. Monolingual English

After applying the vocabulary subsetting strategy and training the SapBERT-based BioSyn model, our system achieved the results reported in Table 1. While the baseline performed better at Acc@1, our system barely surpassed it in Acc@5.

Table 1
Monolingual English Results

System	Acc@1	Acc@5
Baseline	0.57	0.78
Ours (BioSyn + SapBERT-from-PubMedBERT-fulltext)	0.51	0.79

6.1.2. Monolingual Russian

For the Russian subtask, applying the same vocabulary subsetting strategy with BioSyn using the `bert-base-russian-upos` model achieved the results shown in Table 2. In this case, our system outperformed the baseline on both metrics.

[14] https://github.com/dmis-lab/BioSyn

Table 2
Monolingual Russian Results

System	Acc@1	Acc@5
Baseline	0.52	0.59
Ours (BioSyn + bert-base-russian-upos)	0.62	0.72

6.1.3. Multilingual

In the multilingual setting, after vocabulary subsetting with a stricter similarity threshold, BioSyn was trained with `wikineural-multilingual-ner`. Additionally, a post-processing step was applied using "deepseek-ai/DeepSeek-R1-Distill-Llama-8B". Results before and after post-processing are presented in Table 3, together with the baseline.

Table 3
Multilingual Results

System	Acc@1	Acc@5
Baseline	0.53	0.70
Ours (BioSyn + wikineural-multilingual-ner)	0.56	0.76
Ours + LLM Post-processing	0.58	0.76

6.2. Strengths and Limitations

While BioSyn could not load the full vocabulary provided by the organisers due to hardware limitations, our Acc@5 remained competitive, despite using only about 10% of the full vocabulary size.

A limitation of BioSyn on this dataset was the considerable gap observed between Acc@1 and Acc@5 scores, averaging a 20% difference across all tasks. One likely explanation is that BioSyn does not leverage the surrounding textual context for disambiguating entities with similar morphological forms (C0003467 Anxiety, C0003469 Anxiety). As a result, candidate CUIs receive the same score regardless of context, i.e. always predicting C0003467.

One major challenge was the vocabulary size. The full English vocabulary contained 3,902,187 terms, compared to only 145,803 for Russian. The size of the English vocabulary made vocabulary reduction and candidate selection substantially more difficult for English. The high number of English terms increased the risk of overlapping or near-identical candidates with very similar embedding values, often leading to false positives among closely related concepts, as observed in the gap between Acc@1 and Acc@5. In contrast, the smaller Russian vocabulary reduced this risk and enabled more effective subsetting, which likely contributed to the higher Acc@1 observed in the Russian monolingual setting. Nevertheless, due to the limited number of biomedical Russian embedding models, the challenge of selecting the correct Russian candidate remained.

However, in the multilingual task, applying a post-processing step using "deepseek-ai/DeepSeek-R1-Distill-Llama-8B" demonstrated better performance. By providing the lightweight LLM with the top-5 candidates from BioSyn, the corresponding text, and all synonyms from the vocabulary sharing the same CUI (C0003467 Anxiety|Anxiety finding, C0003469 Anxiety|Anxiety disorder), we achieved a 2% improvement in Acc@1 on the multilingual test set.

Although lightweight LLMs are not yet competitive on NEN task [5], our pipeline helps reduce the risk of hallucination and preserves the expected output format, showing the potential for hybrid approaches combining dense retrieval-based methods with controlled generative reasoning.

7. Conclusion

In this paper, we explored the performance of a retrieval-based named entity normalization system for biomedical texts, applying BioSyn models across monolingual English, monolingual Russian, and multilingual tasks. Given the constraints of loading the full candidate vocabulary, a vocabulary subsetting strategy was introduced, which allowed our system to remain competitive on accuracy at Acc@5 despite working with a reduced candidate set. The system demonstrated strong results in the Russian monolingual task compared to the baseline and showed improvements in the multilingual setting after incorporating a lightweight post-processing step using DeepSeek 8B. DeepSeek 8B was able to disambiguate among the top candidates provided by BioSyn by considering contextual meaning and synonym groups, leading to improvements in accuracy at Acc@1.

Our results show both the strengths and weaknesses of retrieval-based NEN models for multilingual biomedical text. BioSyn retrieves candidates efficiently, but without using the surrounding text, it struggles to consistently pick the correct option at Acc@1, causing a noticeable gap between Acc@1 and Acc@5. Adding a post-processing step with an LLM helped address this by re-ranking candidates based on context, improving top prediction accuracy while keeping the output format reliable. In future work, we will explore context-awareness during retrieval and test this type of hybrid approach with complex but smaller vocabulary, allowing BioSyn to load all the candidates.

Funding

J.M.P. and A.D.L. are supported by the CoDiet project. The CoDiet project is funded by the European Union under Horizon Europe grant number 101084642 and supported by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 101084642]. C.L. was supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. For the purpose of open access, the author has applied a creative commons attribution (CC BY) licence to any author accepted manuscript version arising. M.J.R.C. was supported by EASTBIO - East of Scotland Biosciences consortium, UKRI doctoral training program. E.C. was supported by the United Kingdom Research and Innovation (grant EP/Y030869/1), UKRI AI Centre for Doctoral Training in Biomedical Innovation at the University of Edinburgh. For the purpose of open access, the author has applied a creative commons attribution (CC BY) licence to any author accepted manuscript version arising.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: grammar and spelling check, paraphrase and reword, and improve writing style. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] F. Liu, I. Vulic, A. Korhonen, N. Collier, Learning domain-specialised representations for cross-lingual biomedical entity linking, ArXiv abs/2105.14398 (2021). URL: https://arxiv.org/abs/2105.14398.

[2] Z. Yuan, Z. Zhao, S. Yu, Coder: Knowledge-infused cross-lingual medical term embedding for term normalization, Journal of biomedical informatics (2020) 103983. URL: https://arxiv.org/abs/2011.02947.

[3] A. Sakhovskiy, N. Semenova, A. Kadurin, E. Tutubalina, Biomedical entity representation with graph-augmented multi-objective transformer, in: NAACL-HLT, 2024. URL: https://aclanthology.org/2024.findings-naacl.288.pdf.

[4] A. R. Aronson, Effective mapping of biomedical text to the umls metathesaurus: the metamap program, Proceedings. AMIA Symposium (2001) 17–21. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC2243666/.

[5] H. Rouhizadeh, A. Yazdani, B. Zhang, D. Teodoro, Exploring zero-shot cross-lingual biomedical concept normalization via large language models, Studies in health technology and informatics 327 (2025) 788–792. URL: https://ebooks.iospress.nl/doi/10.3233/SHTI250467.

[6] N. V. Loukachevitch, A. Sakhovskiy, E. Tutubalina, Biomedical concept normalization over nested entities with partial umls terminology in russian, in: International Conference on Language Resources and Evaluation, 2024. URL: https://aclanthology.org/2024.lrec-main.213.pdf.

[7] V. Davydova, N. V. Loukachevitch, E. Tutubalina, Overview of bionne task on biomedical nested named entity recognition at bioasq 2024, in: Conference and Labs of the Evaluation Forum, 2024. URL: https://ceur-ws.org/Vol-3740/paper-03.pdf.

[8] 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). doi:10.1093/bioinformatics/btad161, btad161.

[9] A. Sakhovskiy, N. Loukachevitch, E. Tutubalina, Overview of the BioASQ BioNNE-L Task on Biomedical Nested Entity Linking in CLEF 2025, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[10] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, D. Dimitriadis, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, F. N. Maria Di Nunzio, Giorgio, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering, in: L. P. A. G. S. d. H. J. M. F. P. P. R. D. S. G. F. N. F. Jorge Carrillo-de Albornoz, Julio Gonzalo (Ed.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[11] M. Sung, H. Jeon, J. Lee, J. Kang, Biomedical entity representations with synonym marginalization, in: Annual Meeting of the Association for Computational Linguistics, 2020. URL: https://aclanthology.org/2020.acl-main.335/.

[12] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2016) 135–146. URL: https://arxiv.org/abs/1607.04606.

[13] DeepSeek-AI, D. Guo, D. Yang, et al., Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, ArXiv abs/2501.12948 (2025). URL: https://arxiv.org/abs/2501.12948.

[14] R. I. Dogan, 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://www.sciencedirect.com/science/article/pii/S1532046413001974?via%3Dihub.

[15] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, V. A. McKusick, Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders, Nucleic Acids Research 33 (2004) D514 – D517. URL: https://academic.oup.com/nar/article/33/suppl_1/D514/2505259?login=false.

[16] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. 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: The Journal of Biological Databases and Curation 2016 (2016). URL: https://academic.oup.com/database/article/doi/10.1093/database/baw068/2630414.

[17] A. A. Morgan, Z. Lu, X. Wang, A. M. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, J. Hakenberg, et al., Overview of biocreative ii gene normalization, Genome biology 9 (2008) 1–19.

[18] D. R. Maglott, J. Ostell, K. D. Pruitt, T. A. Tatusova, Entrez gene: gene-centered information at ncbi, Nucleic Acids Research 35 (2006) D26 – D31. URL: https://academic.oup.com/nar/article/39/suppl_1/D52/2507756.

[19] K. Roberts, D. Demner-Fushman, J. M. Tonning, Overview of the tac 2017 adverse reaction extraction from drug labels track, Theory and Applications of Categories (2017). URL: https:

//tac.nist.gov/publications/2017/additional.papers/TAC2017.ADR_overview.proceedings.pdf.

[20] 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 (2019) 1234 – 1240. URL: https://academic.oup.com/bioinformatics/article/36/4/1234/5566506.

[21] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-alignment pretraining for biomedical entity representations, in: North American Chapter of the Association for Computational Linguistics, 2020. URL: https://arxiv.org/abs/2010.11784.

[22] P. Kailas, M. Homilius, R. C. Deo, C. A. Macrae, Notecontrast: Contrastive language-diagnostic pretraining for medical text, ArXiv abs/2412.11477 (2024). URL: https://arxiv.org/html/2412.11477v1.

[23] BioCreative VIII – Task 3: Genetic Phenotype Normalization from Dysmorphology Physical Examinations, Zenodo, 2023. URL: https://doi.org/10.5281/zenodo.10104630. doi:10.5281/zenodo.10104630.

[24] UTH-Olympia@BC8 Track 3: Adapting GPT-4 for Entity Extraction and Normalizing Responses to Detect Key Findings in Dysmorphology Physical Examination Observations, Zenodo, 2023. URL: https://doi.org/10.5281/zenodo.10104725. doi:10.5281/zenodo.10104725.

[25] H. Kim, C. Kim, J. Sohn, T. Beck, M. Rei, S. Kim, I. Simpson, J. M. Posma, A. Lain, M. Sung, J. Kang, Ku aigen icl edi@bc8 track 3: Advancing phenotype named entity recognition and normalization for dysmorphology physical examination reports, ArXiv abs/2501.09744 (2025). URL: https://arxiv.org/abs/2501.09744.

[26] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Neural Information Processing Systems, 2017. URL: https://arxiv.org/abs/1706.03762.

[27] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, in: Conference on Empirical Methods in Natural Language Processing, 2019. URL: https://aclanthology.org/D19-1371/.

[28] nesemenpolkov, Fine-tuned multilingual model for russian language ner., in: Detecting names in noisy and dirty data., Moscow, Russian Federation, 2024.

[29] E. Tutubalina, I. Alimova, Z. Miftahutdinov, A. Sakhovskiy, V. Malykh, S. Nikolenko, The Russian Drug Reaction Corpus and Neural Models for Drug Reactions and Effectiveness Detection in User Reviews, Bioinformatics (2020). URL: https://doi.org/10.1093/bioinformatics/btaa675. doi:10.1093/bioinformatics/btaa675, btaa675.

[30] 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.

[31] 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.