BIBERT-Pipe on Biomedical Nested Named Entity Linking at BioASQ 2025*

Notebook for the MSM Lab at CLEF 2025

Chunyu Li[1,*,†], Xindi Zheng[1,†] and Siqi Liu[1,†]

[1]*Individual Researcher*

Abstract

Entity linking (EL) for biomedical text is typically benchmarked on *English-only corpora with flat mentions*, leaving the more realistic scenario of *nested* and *multilingual* mentions largely unexplored. We present our system for the **BioNNE 2025 Multilingual Biomedical Nested Named Entity Linking** shared task (English & Russian), closing this gap with a lightweight pipeline that keeps the original EL model intact and modifies only three task-aligned components: **Two-stage retrieval-ranking.** We leverage the same base encoder model in both stages: the retrieval stage uses the original pre-trained model, while the ranking stage applies domain-specific fine-tuning. **Boundary cues.** In the ranking stage, we wrap each mention with learnable `[Ms]` / `[Me]` tags, providing the encoder with an explicit, language-agnostic span before robustness to overlap and nesting. **Dataset augmentation.** We also automatically expand the ranking training corpus with three complementary data sources, enhancing coverage without extra manual annotation. On the BioNNE 2025 leaderboard, our two stage system, bilingual bert (**BIBERT-Pipe**), ranks **third** in the multilingual track, demonstrating the effectiveness and competitiveness of these minimal yet principled modifications. Code are publicly available at https://github.com/Kaggle-Competitions-Code/BioNNE-L.

Keywords

Biomedical entity linking, Bilingual

1. Introduction

Biomedical entity linking (BEL) – also known as named entity normalization or grounding – is the task of mapping entity mentions in the text to entries in a reference knowledge base. In the biomedical domain, EL plays a vital role in text mining by standardizing mentions of diseases, genes, drugs, and other entities to canonical identifiers [1]. This normalization resolves synonymy and ambiguity: for example, the abbreviation "WSS" could refer to Wrinkly Skin Syndrome or Weaver-Smith Syndrome, and linking it to the correct concept ID disambiguates the intended meaning [2]. By grounding mentions to KB concepts (e.g., UMLS or Wikidata entries), EL enables effective information integration, improves literature search (e.g., concept-based PubMed indexing), and facilitates downstream tasks such as relation extraction and question answering.

While early BEL research has made significant progress in English-only settings with flat (non-overlapping) mentions, real-world biomedical documents often exhibit *nested entities* and appear in *multiple languages*—posing persistent challenges that remain under-addressed.

Nested mentions—where one entity is embedded within or overlaps another—are prevalent in biomedical literature. For example, in *"EGFR exon 19 deletion mutation"*, the terms *"EGFR"* and *"exon 19 deletion"* refer to distinct concepts, both requiring normalization. Ignoring nested structures can lead to incomplete or incorrect linking. Meanwhile, the increasing volume of biomedical text in non-English

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

*You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

*Corresponding author.

†These authors contributed equally.

li.chunyu0412@gmail.com (C. Li); xindizhe@gmail.com (X. Zheng); liusiqisq0412@163.com (S. Liu)

https://eeyorelee.github.io/ (C. Li); https://github.com/momoxia (X. Zheng); https://github.com/cmwls (S. Liu)

0009-0000-3036-0275 (C. Li); 0009-0007-0974-788X (X. Zheng); 0009-0001-0372-7155 (S. Liu)

languages highlights the importance of multilingual BEL. Studies have shown that models trained in English exhibit significant performance drops when applied to languages like Spanish or Russian [3, 4].

Several technical barriers exacerbate these challenges: (i) the lack of annotated multilingual data, especially in low-resource biomedical languages; (ii) inconsistencies in concept coverage across languages in knowledge bases; and (iii) the inherent ambiguity and granularity of biomedical terminology. Existing EL pipelines are typically not equipped to handle these complexities simultaneously.

In this paper, we propose a lightweight, encoder-agnostic pipeline for multilingual, nested biomedical EL. Our method introduces three key enhancements: (i) a two-stage retrieval-ranking strategy that leverages the same base encoder model, where the retrieval stage utilizes the original pre-trained model and the ranking stage benefits from contrastive learning training; (ii) boundary cue tagging, using learnable tokens (`[Ms]` / `[Me]`) to explicitly encode span boundaries, enabling robust modeling of nested mentions; and (iii) dataset augmentation by incorporating additional complementary data sources, enriching training coverage without requiring manual annotation. Our approach maintains the original EL model architecture while significantly improving robustness across languages and nested spans. It can be seamlessly integrated with several biomedical encoders (e.g., BioLinkBERT [5], SapBERT [6]) and adapted to multilingual scenarios with minimal overhead.

Our system achieved third place in the BioNNE-L 2025 multilingual track [7], demonstrating that our proposed techniques—two-stage retrieval ranking, boundary cue tagging, and data augmentation—are not only lightweight and effective but also highly generalizable. They can be seamlessly applied to a variety of base encoders and readily integrated into multilingual biomedical EL systems. This highlights the practical value of our approach for building robust, scalable solutions to cross-lingual entity linking tasks.

2. Task Overview

To further advance research in biomedical entity linking, BioASQ 2025 [8] holds a task, BioNNE-L [7]: Nested NER in Russian and English. The BioNNE-L shared task focuses on NLP challenges in entity linking, also known as medical concept normalization (MCN), for English and Russian languages. The goal is to map biomedical entity mentions to a comprehensive set of medical concept names and their concept unique identifiers (CUIs) from the UMLS. The train, dev, and test datasets include mentions of disorders, anatomical structures, and chemicals, all mapped to concepts from the UMLS. The BioNNE-L task utilizes the MCN annotation of the NEREL-BIO dataset [9], which provides annotated mentions of disorders, anatomical structures, chemicals, diagnostic procedures, and biological functions.

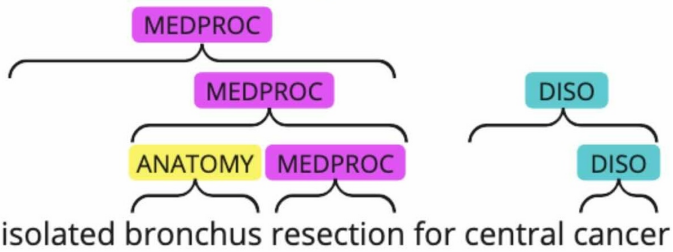

Figure 1: Example of nested named entities in NEREL-BIO. The English phrase *"isolated bronchus resection for central cancer"* is annotated with overlapping spans: the outer span (magenta) is a diagnostic procedure (MEDPROC); inside it, the token *bronchus* is an anatomical structure (yellow, ANATOMY), while *resection* is again a procedure (magenta, MEDPROC). A separate right-hand branch shows the phrase *central cancer*, where both the full span and the nested core *cancer* are labelled as a disease (cyan, DISO). This illustrates the two challenges of the task: *nesting* (entities contained within entities) and *fine-grained, type-specific* normalisation.

3. Related Work

Multilingual Biomedical Entity Linking. Multilingual BEL is an increasingly important research direction due to the global nature of biomedical literature. Traditional approaches often rely on translation to English prior to linking, but this can introduce noise and domain mismatch [4]. To overcome these limitations, recent work has focused on cross-lingual encoders and alignment techniques. SapBERT [6] uses self-alignment pretraining with UMLS synonym pairs across languages to learn language-agnostic biomedical embeddings. Guven and Lamurias [3] study bi-encoder models on English and Spanish corpora and highlight persistent performance gaps on non-English datasets.

Nested Mention Normalization. Nested named entities are a known challenge for EL systems. Standard EL models often assume flat mention boundaries and can not resolve overlapping entities. The MCN dataset [9] extends entity linking to nested mentions in both English and Russian, providing a valuable benchmark. However, few EL systems explicitly model nested mentions. Some recent work, such as Con2GEN [10] addresses multilingual biomedical entity linking using a generation-based approach with predefined prompts, effectively capturing dependencies between mentions and concepts. However, such generative methods may involve increased model complexity and computational resources compared to discriminative approaches.

Contrastive and Graph-Based Learning. Contrastive learning has proven effective for biomedical EL, particularly in bi-encoder architectures. GEBERT [11] combines a Transformer with a graph neural encoder over the UMLS knowledge graph. It aligns graph node embeddings with textual descriptions through node-text contrastive learning. BERGAMOT [12] extends this with multiple contrastive losses and multilingual pretraining, improving generalization across languages and domains. SapBERT [6] trains with an InfoNCE loss to align mention and concept representations. BERGAMOT [12] extends this with multilingual graph-based contrastive learning, incorporating ontology structure. Con2GEN [10] instead adopts a controllable generation strategy to bridge mention-concept alignment using cross-lingual templates. While these methods demonstrate strong results, they often require complex training setups or extensive graph preprocessing.

Large Language Models. LLMs like ChatGPT and GPT-4 have been tested on biomedical entity link tasks [13]. While flexible, they often underperform domain-specific fine-tuned models in complex scenarios [14]. Instruction tuning and prompt engineering have been explored to close this gap [15], but performance is still limited without task-specific adaptation.

Our Contribution. In contrast to prior work, our method is efficient and explicitly designed for both multilinguality and nesting. It requires no architectural change to the encoder and is compatible with any transformer-based biomedical model. Our use of span boundary cues provides strong supervision for nested and cross-lingual linking, while dataset augmentation further improves accuracy.

4. Method

Our approach follows a two-stage paradigm: (*i*) dense retrieval to obtain a small set of plausible concepts for each mention, and (*ii*) cross–encoder ranking to pick the best concept. Although the backbone encoder may vary, the surrounding pipeline remains unchanged and is illustrated in Figure 2.

4.1. Formal Definition

Let $\mathcal{K} = \{ c_1, \ldots, c_{|\mathcal{K}|} \}$ be a biomedical knowledge base whose entries are represented by canonical names and concept unique identifiers (CUIs). Given a document D written in language $\ell \in \{\text{EN}, \text{RU}\}$

that contains a set of (possibly *nested*) entities mentioned in $\mathcal{M}(D) = \{ m_1, \ldots, m_N \}$, The goal is to find a mapping

$$\Phi : (m_i, D, \ell) \longrightarrow c^\star \in \mathcal{K}, \qquad 1 \le i \le N,$$

where $c^\star$ denotes the concept that is *semantically equivalent* to the surface form of m_i in its context.

We factor Φ into two components:

$$\textbf{Retrieval:} \quad f_{\text{ret}}(m_i, D, \ell) \longrightarrow C_i = \langle c_i^1, \ldots, c_i^k \rangle, \tag{1}$$

$$\textbf{Rank:} \quad f_{\text{rank}}(m_i, D, C_i, \ell) \longrightarrow \hat{c}_i \in C_i, \tag{2}$$

where $k \ll |\mathcal{K}|$ (we use $k = 10$). Let $c_i^\star$ be the gold concept for mention m_i. We report:

$$\text{Acc@1} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left[\hat{c}_i = c_i^\star\right], \qquad \text{Acc@}k = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left[c_i^\star \in \text{Top-}k(\hat{\pi}_i)\right], \; k \in \{5, 10\},$$

where $\hat{\pi}_i$ is the ranker–sorted list of the k retrieval candidates for the i-th mention. Thus, the rank stage simply re-orders the 10 candidates returned by the retrieval stage, and we evaluate whether the gold concept appears within the first k positions.

4.2. Framework

Retrieval stage. We experiment with five publicly available biomedical encoders, including *BioLink-BERT* [5] and *BiomedBERT* [16](`abstract/fulltext`, *bert-base-uncased*). For every mention, we wrap its span with our boundary cues `[Ms]` and `[Me]`, encode the sequence, and compute a cosine similarity to all concept representations in $\mathcal{K}$.[1]

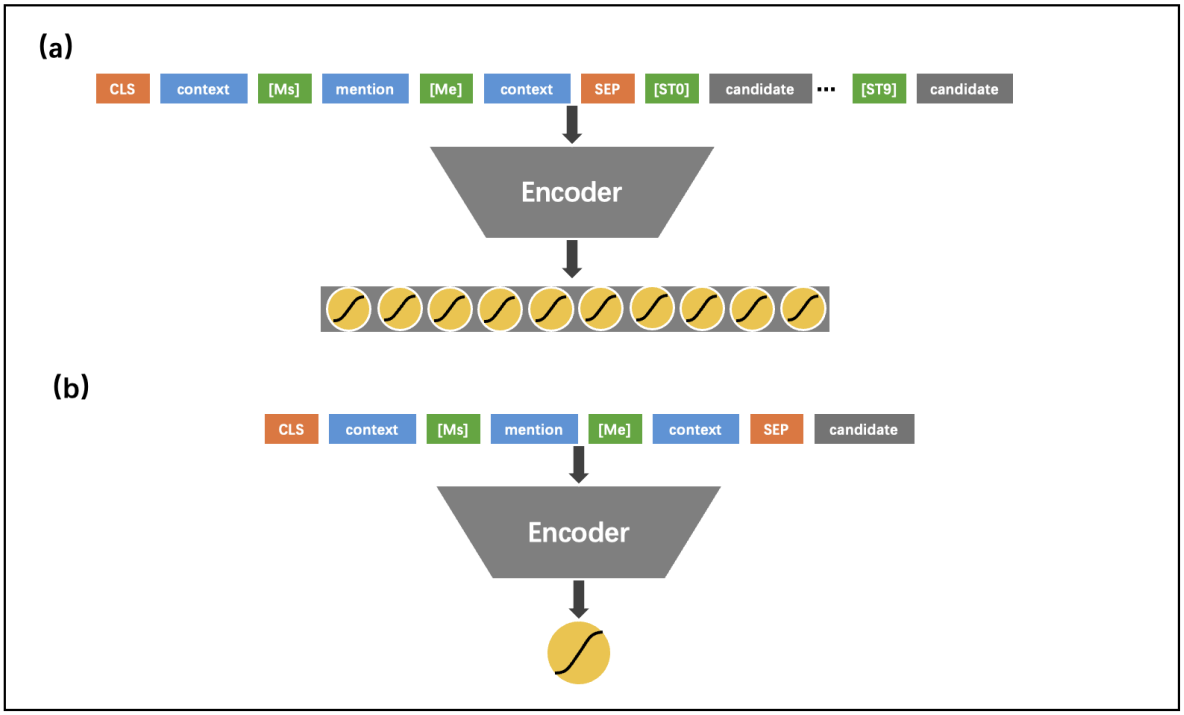

Figure 2: Pipeline of our rank stage. (a) Listwise – the k candidates returned by the retrieval stage are *concatenated* after the `[SEP]` token (`[ST0]`…`[ST`$_{k-1}$`]`), so a *single* forward pass of the encoder produces k logits, one for each candidate. **(b) Contrastive Learning** – each candidate is paired with the mention context in an independent input sequence; the encoder is applied k times and outputs a binary logit for every candidate individually.

[1] Concept vectors are pre-computed once per encoder.

Rank stage. After the retrieval stage, we need to rank the CUIs. In particular, we build our ranking model without extra model modification. It is efficient to train the rank model with the retrieval model. We use two types of architecture to feed k candidates from the retrieval stage:

1. *Listwise (LTR)* (Figure 2 a): the k candidates $\langle c_i^1 \dots c_i^k \rangle$ are concatenated after the `[SEP]` token; One forward pass yields k logits $\mathbf{z} \in \mathbb{R}^k$ trained with a listwise soft margin loss.
2. *Contrastive Learning (CL)* (Figure 2 b): Each candidate is paired with the mention context and processed independently; every pass produces a binary score $z \in \mathbb{R}$ optimized by cross-entropy.

Since the LTR scheme processes all k candidates in a single forward pass, while the CL scheme handles only one candidate per pass, the CL scheme inherently requires k times more computation during both training and inference. Despite costing k times more computation, the CL scheme eliminates cross-candidate interference. Obviously, each forward pass evaluates a single candidate against the mention context, reducing the problem to an independent binary decision that the model can learn more easily.

4.3. Data Augmentation.

Considering the limit of the training set, we also add additional dataset for the RU and BI tracks, including **MedMentions** [17], a manually annotated resource for the recognition of English biomedical concepts, and **MCN** [9], a novel dataset for nested entity linking in Russian. We reformat these two datasets to suit the competition and only keep the three entity types: (i) Disease (DISO), (ii) Chemical (CHEM), (iii) Anatomy (ANATOMY).

5. Experiment

Datasets. We follow the official `BioNNE 2025` split: the train set with extra dataset is for training the model, the *development set* is used for model selection; the *evaluation set* is kept blind for final ranking. We set the retrieval numbers to $k = 10$. The base dataset of this task is NEREL-BIO [18].

Metrics. For retrieval we report Acc@k; for ranking we use the cross-validated Acc@1 (CV Acc) on the development folds. Final leaderboard numbers are the organisers' Acc@1 on the hidden evaluation set.

5.1. Retrieve Stage

Table 1 compares six off-the-shelf biomedical encoders in three tracks. On the English track, **SapBERT–PubMedBERT (fulltext)** gives the strongest Acc@1 $= 0.6115$, while its *mean-token* pooling variant slightly improves the recall (Acc@10 $= 0.8184$). Russian retrieval is notably harder: even **SapBERT–XLMR–large** reaches only 0.5103 Acc@1. The bilingual runs averages those behaviours, yielding 0.5389 Acc@1 at best. These numbers indicate that the retrieval module already places the gold concept within the top-10 for more than 80%(EN) and more than 75%(BI) of mentions, leaving ample head-room for a ranker.

5.2. Rank Stage

For the ranking phase we keep the best retriever for each track: **SapBERT–PubMedBERT (mean-token)** for English, and **SapBERT–UMLS–XLMR (large)** for both the Russian and Bilingual tracks.

Table 2 analyses two cross-encoder architectures on the English dev set. The **Listwise**, although computationally cheap (one forward pass), plateaus at 0.5918 CV Acc. Switching to the **CL** design—i.e. an independent binary decision per candidate—raises accuracy to 0.6604 when additional *MedMentions* are used. We attribute the gain to two factors: (i) candidates no longer compete inside the softmax,

Table 1
Retrieval accuracy (**Acc@k**) of base encoders on the development sets for English (EN), Russian (RU), and Bilingual (BI).

Model	Acc@1	Acc@5	Acc@10
English (EN) dev set			
GEBERT	0.5898	0.7654	0.7979
BioLinkBERT-large	0.4270	0.6030	0.6210
BioLinkBERT-base	0.4720	0.6530	0.6710
SapBERT–PubMedBERT (fulltext)	**0.6115**	0.7698	0.8043
SapBERT–PubMedBERT (mean-token)	0.6038	**0.7723**	**0.8184**
Russian (RU) dev set			
SapBERT-UMLS-XLMR (base)	0.4914	0.5497	0.5686
SapBERT-UMLS-XLMR (large)	**0.5103**	**0.5613**	**0.5763**
Bilingual (BI) dev set			
SapBERT-UMLS-XLMR (base)	0.5197	0.7021	0.7331
SapBERT-UMLS-XLMR (large)	**0.5389**	**0.7171**	**0.7500**

thus reducing interference, and (ii) the binary objective is simpler, allowing the model to specialise on fine-grained lexical cues.

Bilingual results in Table 3 confirm the trend. Incorporating *MedMentions* and the MCN dataset adds a further 0.7–2.4 pp on Russian and bilingual tracks, but English still benefits the most ($+6.9$ pp).

Table 2
Rank results on the English (EN) track. The candidate set produced by the retrieval stage is identical for all systems (**Acc@1 = 0.6115**, **Acc@5 = 0.7698**, **Acc@10 = 0.8043**); we therefore report only the cross-validated accuracy (**CV Acc**) of each ranker. "Listwise" vs. "CL" denotes two different architectures. For all of the experiments in this table, the post training epoch for ranking is 5, and the learning rate is chosen from 7e-6 or 1e-5.

Base Model	**CV Acc**	**Approach**	**Training Data** / k
BioLinkBERT-base	0.5871	LTR	train, k=5
BioLinkBERT-large	0.5699	LTR	train, k=5
BERT-base-uncased	0.5683	LTR	train, k=5
BiomedBERT-abstract	0.5871	LTR	train, k=5
BiomedBERT-abstract	0.5918	CL	train, k=5
BiomedBERT-abstract	**0.6604**	CL	MedMentions + train, k=5
KRISSBERT	0.6576	CL	MedMentions + train, k=5
BiomedBERT-fulltext	0.6536	CL	MedMentions + train, k=5
BiomedBERT-fulltext	0.6532	CL	MedMentions + train, k=10

Table 3
Rank stage results on English (EN), Russian (RU), and Bilingual (BI) tracks. Details include the training dataset for the rank stage. For all of experiments of this table, the further training epoch for ranking is 5 and the learning rate is chosen from 7e-6 or 1e-5.

Base Model	**Lang**	**CV(Acc)**	**Approach**	**Details**
BiomedNLP-BiomedBERT-base-uncased-abstract	EN	0.5918	CL	train
BiomedNLP-BiomedBERT-base-uncased-abstract	EN	0.6604	CL	train+MedMentions
SapBERT-XLMR-large	RU	0.6131	CL	train
SapBERT-XLMR-large	RU	0.6204	CL	train+MedMentions+MCN,
SapBERT-XLMR-large	BI	0.6083	CL	train
SapBERT-XLMR-large	BI	0.6319	CL	train+MedMentions+MCN

5.3. Final result

Table 4 summarises leaderboard scores. Our best submissions **SapBERT-XLMR-large + MedMentions + MCN** for RU/BI, and **BiomedBERT-abstract + MedMentions** for EN achieve **0.6497**, **0.6370** and **0.6370** Acc@1 on RU, BI and EN, respectively, ranking third overall in the bilingual track.

Table 4
Final results on English (EN), Russian (RU), and Bilingual (BI) tracks on evaluation dataset. For all of experiment of this table, the further training learning rate is chosen from 7e-6 or 1e-5.

Base Model	Lang	Acc	Approach	Details
BiomedBERT-abstract	EN	0.6197	CL	MedMentions+train, epoch=2
BiomedBERT-abstract	EN	0.6273	CL	MedMentions+train+dev, epoch=2
SapBERT-XLMR-large	EN	0.6370	CL	train+MedMentions+dev, epoch=1
SapBERT-XLMR-large	RU	0.6452	CL	train, epoch=5
SapBERT-XLMR-large	RU	0.6497	CL	train+MedMentions+MCN+dev, epoch=5
SapBERT-XLMR-large	BI	0.6229	CL	train, epoch=1
SapBERT-XLMR-large	BI	0.6342	CL	train+MedMentions+MCN+dev, epoch=1

6. Ablation Study

6.1. Boundary Cues

To assess the contribution of boundary cues, we also perform an ablation study Table 5 on boundary cues, known as special tokens [Ms] and [Me], which indicate the start and end of the target entity. The improvement is most pronounced on the Russian (RU) track, where the Acc@1 increases by 6.60%. We attribute this to the richer morphology of Russian: the explicit `[Ms]` / `[Me]` markers help the model to delineate entity spans that may otherwise be obscured by inflectional endings. For the English (EN) and the Bilingual (BI) setting, the gains are more modest 1.20% and 1.24%, respectively - but still positive, confirming that boundary information remains beneficial even in languages with a relatively simpler morphology.

Table 5
Ablation study on the effect of boundary cues for the English (EN), Russian (RU), and Bilingual (BI) tracks of the evaluation set. All experiments are conduct with the same hyper-parameters on the SapBERT-XLMR-large base model; the only difference is whether the boundary-cue tokens are included. Performance is reported using Acc@1, consistent with our earlier experiments.

Lang	w/ [Ms] and [Me]	w/o [Ms] and [Me]	Gain
EN	0.6370	0.6292	0.0078 (1.24%)
RU	0.6497	0.6095	0.0402 (6.60%)
BI	0.6342	0.6267	0.0075 (1.20%)

7. Conclusion

We present a simple yet effective two–stage pipeline for the BioNNE 2025 Bilingual Nested Entity Linking task. Keeping the *base encoder untouched*, we obtained competitive performance by addressing three task-specific bottlenecks: (i) explicit mention boundary cues (`[Ms]`/`[Me]`) indicating the position of mention, (ii) efficient rank architecture design for ranking mention and (iii) data augmentation with MedMentions / MCN boosting the final result. On the official leaderboard our system ranks **3rd** in BI track, with Acc@1 of 0.637 (BI), while training on a single Nvidia 3090.

8. Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly, to: Grammar and spelling check, paraphrase, and minor translation. After using these tools, the authors reviewed and edited the content as needed and assume full responsibility for the content of the publication.

References

[1] E. French, B. T. McInnes, An overview of biomedical entity linking throughout the years, Journal of biomedical informatics 137 (2023) 104252.

[2] S. Garda, U. Leser, Belhd: Improving biomedical entity linking with homonym disambiguation, Bioinformatics 39 (2023) btad698. doi:10.1093/bioinformatics/btad698.

[3] Z. A. Guven, A. Lamurias, Multilingual bi-encoder models for biomedical entity linking, Expert Systems 40 (2023) e13388.

[4] F. Liu, I. Vulić, A. Korhonen, N. Collier, Learning domain-specialised representations for cross-lingual biomedical entity linking, arXiv preprint arXiv:2105.14398 (2021).

[5] M. Yasunaga, J. Leskovec, P. Liang, Linkbert: Pretraining language models with document links, in: Association for Computational Linguistics (ACL), 2022.

[6] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-alignment pretraining for biomedical entity representations, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 4228–4238. URL: https://aclanthology.org/2021.naacl-main.334/.

[7] 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.

[8] 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, G. M. Di Nunzio, N. Ferro, 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: J. Carrillo-de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

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

[10] T. Zhu, Y. Qin, Q. Chen, X. Mu, C. Yu, Y. Xiang, Controllable contrastive generation for multilingual biomedical entity linking, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 5742–5753.

[11] A. Sakhovskiy, N. Semenova, A. Kadurin, E. Tutubalina, Graph-enriched biomedical entity representation transformer, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2023, pp. 109–120.

[12] 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, 2024, pp. 4626–4643.

[13] I. Jahan, M. T. R. Laskar, C. Peng, J. Huang, Evaluation of ChatGPT on biomedical tasks: A zero-shot comparison with fine-tuned generative transformers, in: D. Demner-Fushman, S. Ananiadou, K. Cohen (Eds.), Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 326–336. URL: https://aclanthology.org/2023.bionlp-1.30/. doi:10.18653/v1/2023.bionlp-1.30.

[14] Q. Chen, Y. Hu, X. Peng, Q. Xie, Q. Jin, A. Gilson, M. B. Singer, X. Ai, P.-T. Lai, Z. Wang, V. K. Keloth, K. Raja, J. Huang, H. He, F. Lin, J. Du, R. Zhang, W. J. Zheng, R. A. Adelman, Z. Lu, H. Xu, Benchmarking large language models for biomedical natural language processing applications and recommendations, Nature Communications 16 (2025) 3280. doi:10.1038/s41467-025-56989-2.

[15] Y. Ding, Q. Zeng, T. Weninger, ChatEL: Entity linking with chatbots, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 3086–3097. URL: https://aclanthology.org/2024.lrec-main.275/.

[16] 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, 2020. arXiv:arXiv:2007.15779.

[17] S. Mohan, D. Li, Medmentions: A large biomedical corpus annotated with umls concepts, 2019. URL: https://arxiv.org/abs/1902.09476. arXiv:1902.09476.

[18] N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, Biomedical concept normalization over nested entities with partial UMLS terminology in Russian, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 2383–2389. URL: https://aclanthology.org/2024.lrec-main.213/.