Overview of the BioASQ BioNNE-R Task on Nested Biomedical Relation Extraction at CLEF 2026

Igor Rozhkov[1,*], Natalia Loukachevitch[1] and Elena Tutubalina[2,3]

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

Abstract

Biomedical relation extraction is typically studied under flat named entities, keeping only relations that involve nested named entities, where one entity is contained within another, unexplored. This paper presents an overview of the BioNNE-R shared task on biomedical nested relation extraction, organized within the BioASQ 2026 workshop at CLEF. Building on the NEREL-BIO corpus and the earlier BioNNE and BioNNE-L shared tasks, BioNNE-R requires systems to extract semantic relations between nested biomedical entities in English and Russian PubMed abstracts. The task covers 8 entity types and 14 relation types and is organized into three subtasks under two evaluation tracks: a monolingual track, with separate English and Russian subtasks, and a bilingual track combining both languages. We received 20 Codabench registrations, and 7 teams submitted predictions during the evaluation phase. Most participating systems framed the task as pair classification with fine-tuned multilingual and biomedical transformer encoders; the best systems reached macro-F1 scores of approximately 0.50-0.53, substantially above the baseline given, with a document-level graph-based approach achieving the top results on the Russian and bilingual subtasks. These results indicate that relation extraction over nested biomedical entities remains a challenging problem with considerable room for improvement.

Keywords

BioNLP, Biomedical NLP, Relation Extraction, Nested Entities, Multilingual

1. Introduction

Biomedical information extraction aims to transform unstructured scientific texts into structured representations that can support biomedical search, clinical decision making, and knowledge base construction [1, 2, 3, 4, 5, 6]. Among the core tasks in biomedical natural language processing, relation extraction plays a crucial role by identifying semantic relations between biomedical entities such as disorders, chemicals, anatomical structures, and diagnostic procedures [7].

Traditional biomedical relation extraction systems usually operate on flat named entities, where entity mentions do not overlap and are treated independently [8]. However, biomedical texts frequently contain nested named entities, in which one entity is embedded within another [9, 10]. For example, a disorder mention may contain an anatomical structure or a chemical mention as its subpart. Such nested structures introduce substantial complexity for relation extraction systems because relations may involve both outer and inner entity mentions.

Recent advances in biomedical natural language processing have therefore increasingly focused on nested information extraction, including nested named entity recognition and nested entity linking [10, 9, 11, 12, 13, 14]. These studies demonstrated that modeling nested structures improves extraction quality and enables more accurate representation of biomedical concepts and their interactions.

To support research in multilingual nested biomedical information extraction, the NEREL-BIO corpus was introduced as a dataset of English and Russian PubMed abstracts annotated with nested biomedical entities [15]. The corpus extends the general-domain NEREL framework [16] to the biomedical domain and preserves annotations of nested and overlapping entities. NEREL-BIO includes annotations for

CLEF 2026 Working Notes, September 21 – 24 2026, Jena, Germany
*Corresponding author.
fulstocky@gmail.com (I. Rozhkov); louk_nat@mail.ru (N. Loukachevitch); tutubalinaev@gmail.com (E. Tutubalina)
0000-0002-9018-3757 (I. Rozhkov); 0000-0002-1883-4121 (N. Loukachevitch); 0000-0001-7936-0284 (E. Tutubalina)

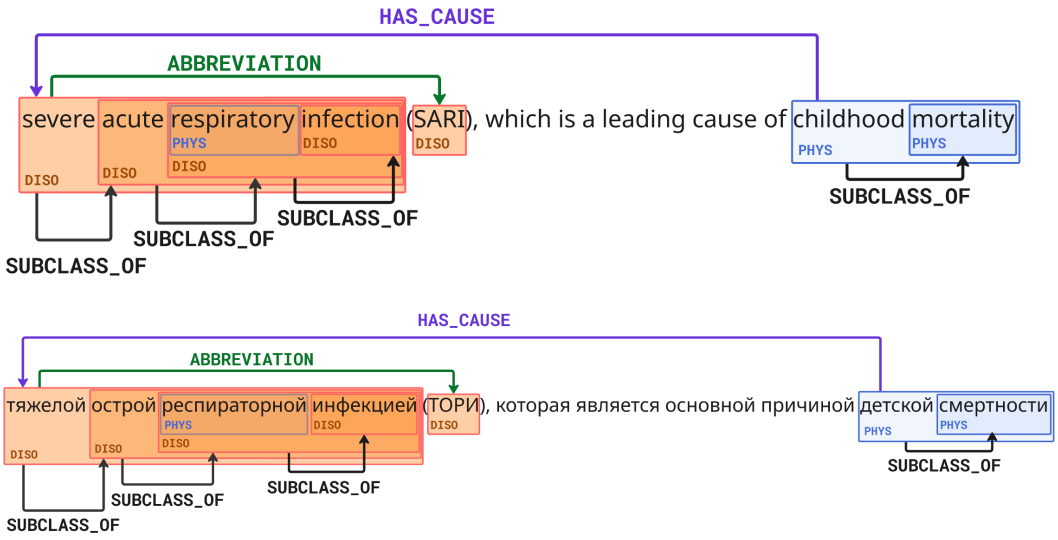

Figure 1: Example of nested named entity relations in the BioNNE-R dataset.

biomedical concepts such as disorders, chemicals, anatomical structures, physiological functions, laboratory procedures, and other medically relevant entity types.

In this paper, we present an overview of the shared task on Biomedical Nested Named Entity Relation Extraction (BioNNE-R). The BioNNE-R shared task continues the BioNNE shared task series [17, 18] organized within the BioASQ workshop at CLEF [19] and based on the NEREL-BIO corpus. The earlier editions of the series focused on multilingual nested biomedical named entity recognition (BioNNE) [17] and nested biomedical entity linking (BioNNE-L) [18], respectively. These shared tasks established benchmark datasets and evaluation settings for multilingual information extraction and highlighted the challenges associated with processing nested biomedical entities.

This year, the BioNNE-R task addresses the extraction of semantic relations between nested biomedical entities in English and Russian scientific abstracts. Participants were invited to develop systems for multilingual, English-oriented, or Russian-oriented nested relation extraction. All BioNNE-R materials can be found on the shared task's GitHub[1] and Codabench pages[2].

The paper is organized as follows. Section 2 describes the BioNNE-R task formulation and dataset construction. Section 3 presents the evaluation methodology, baselines and summarizes participating systems and experimental results. Finally, Section 4 concludes the paper and discusses future directions for multilingual nested biomedical relation extraction.

2. Dataset

Training and validation sets for the BioNNE-R competition are based on the NEREL-BIO dataset [15] and additional annotated texts for the BioNNE competition. NEREL-BIO is a corpus of PubMed abstracts written in Russian and English. It enhances the NEREL [16] dataset, which was originally developed for the general domain, with biomedical entity types defined according to the Unified Medical Language System (UMLS) [20]. All abstracts are annotated in the BRAT standoff format [21]. An example of nested named entity relations in the BioNNE-R dataset is illustrated in Figure 1.

[1] https://github.com/nerel-ds/NEREL-BIO/tree/master/BioNNE-R
[2] https://www.codabench.org/competitions/13132/

Table 1
Number of annotated documents and entities per type in the BioNNE-R dataset, broken down by language (English, Russian) and data split (train, dev, test).

	English			Russian		
	Train	Dev	Test	Train	Dev	Test
# documents	55	50	152	716	50	154
ANATOMY	1831	942	2259	8359	869	2120
CHEM	1177	547	1326	4779	527	1148
DEVICE	43	19	171	735	33	172
DISO	2214	1043	3024	11177	914	2603
FINDING	810	330	1402	4816	350	1208
INJURY_POISONING	112	72	135	643	25	115
LABPROC	370	150	411	1619	138	356
PHYS	777	375	1588	5751	354	1617
Total	7334	3478	10316	37879	3210	9339

Table 2
Number of relations per type in the BioNNE-R dataset, broken down by language (English, Russian) and data split (train, dev, test).

	English			Russian		
	Train	Dev	Test	Train	Dev	Test
ABBREVIATION	97	74	238	862	67	244
AFFECTS	362	393	920	2493	288	774
ALTERNATIVE_NAME	95	98	285	649	97	201
APPLIED_TO	43	54	70	284	45	41
ASSOCIATED_WITH	325	233	977	3295	242	987
FINDING_OF	121	83	529	1356	85	453
HAS_CAUSE	579	487	971	2710	407	790
ORIGINS_FROM	120	79	214	566	67	135
PART_OF	204	248	428	1483	208	502
PHYSIOLOGY_OF	150	178	425	1127	131	313
SUBCLASS_OF	743	757	2475	7517	669	2187
TO_DETECT_OR_STUDY	164	116	414	771	84	239
TREATED_USING	101	88	182	282	46	61
USED_IN	89	79	222	378	85	151
Total	3193	2967	8350	23773	2521	7078

2.1. Task Definition

The BioNNE-R task is formulated as relation classification over candidate entity pairs. Given a document with gold-annotated nested entity mentions, a model must, for each pair of entities, predict a relation type or predict no relation at all. Entity mentions, including their spans and types, are provided as gold input for all tracks at each stage, including evaluation. Systems perform only relation classification over candidate pairs and are not required to detect entity spans. So, the span-matching and type-matching conditions would refer to the gold data entities, not predicted ones. Because entities may be nested, a relation can connect an outer entity to one of its internal entities, or an internal entity to an external one. The task covers 8 entity types (ANATOMY, CHEM, DEVICE, DISO, FINDING, INJURY_POISONING, LABPROC, PHYS) and 14 relation types. It is organized into three subtasks under two evaluation tracks: a monolingual track comprising an English subtask (Subtask 1) and a Russian subtask (Subtask 2), and a bilingual track (Subtask 3) requiring a single model for the combined English and Russian data. Per-language statistics for entities and relations are reported in Tables 1 and 2.

Several relation types are restricted to same-type entity pairs in the annotation schema: SUBCLASS_OF, ABBREVIATION, and ALTERNATIVE_NAME may connect any of the 8 entity types to itself, while PART_OF is further restricted to ANATOMY→ANATOMY and DEVICE →DEVICE.

This same-type constraint holds without exception across the 20,428 gold annotations of these four relation types in both languages and all splits. Other relation types are constrained to specific entity-type combinations. For example, PHYSIOLOGY_OF connects only PHYS →ANATOMY pairs, and FINDING_OF only links FINDING entities to INJURY_POISONING or PHYS entities. In total, the 14 relation types use only 39 of the 64 possible (head_type, tail_type) entity-type pairs, providing a strong schema constraint that systems can use to prune the candidate set at inference time.

Across all splits and languages, 97.7% of relations are intra-sentential, with both entity mentions in the same sentence; the remaining 2.3% require cross-sentence reasoning.

3. Experiments

3.1. Evaluation metrics

Systems are evaluated against the gold relation annotations using precision, recall, and F1-score. A predicted relation is correct when it matches a gold relation in document identifier, head and tail entity spans, head and tail entity types, and relation type. Because the relation-type distribution is highly imbalanced, macro-averaged F1 over the 14 relation types is the official ranking metric, computed independently for each subtask. In addition, we report micro-averaged precision, recall, and F1 for the baseline, together with a per-relation F1 breakdown (Table 4). Because several relation types have fewer than 100 test instances, their per-type F1 may be sensitive to sampling noise, reporting bootstrap confidence intervals is a natural refinement.

3.2. Baseline

As an official baseline, we provide a relation classifier built on the OpenNRE framework [22]. Each candidate entity pair is encoded with `bert-base-multilingual-cased` backbone [23], with entity markers inserted around the head and tail spans, and classified into one of the relation types or a no_relation class, the minimal sentence segment surrounding the two entities serves as input. Training instances are formed by enumerating gold entity pairs and adding negative examples at a 3:1 negative-to-positive ratio, with candidate pairs restricted to entity-type combinations permitted by the annotation schema. Separate models were trained for the English, Russian, and bilingual subtasks — the latter on the concatenation of the English and Russian data — for 10 epochs with a learning rate of 2e-5 and a maximum sequence length of 256. On the test set, the baseline obtains macro-F1 of 0.2825 (English), 0.3070 (Russian), 0.3027 (bilingual), and micro-F1 of 0.3122, 0.3564, and 0.3496, respectively. The gap between micro-F1 and macro-F1, together with baseline precision of roughly 0.20–0.23 against recall of 0.71–0.77, reflects substantial over-prediction, consistent with the 26.6% type-violation rate reported below.

To better understand the baseline's failure modes, we examined how often its predictions violate the entity-type constraints implicit in the annotation schema. For each relation type, we collected the set of (head_type, tail_type) pairs observed in the Russian gold annotations across all splits and counted, on the Russian test set, how many baseline predictions used entity-type pairs not present in that set. The results are reported in Table 3.

Across the 14 relation types, 26.6% of the baseline's predictions (6,224 of 23,392) violate the type schema. The violation rate varies sharply, from 1.5% (ALTERNATIVE_NAME, which the baseline rarely predicts) to 54.8% (ASSOCIATED_WITH, where only 5 entity-type pairs are valid but the baseline predicts 46). The four relation types restricted to same-type entity pairs, namely ABBREVIATION, ALTERNATIVE_NAME, SUBCLASS_OF, PART_OF are systematically violated by the baseline at rates of 1.5–21.1%, despite this being a hard constraint never broken in the gold data. Highly constrained relations (PHYSIOLOGY_OF, FINDING_OF, TO_DETECT_OR_STUDY) all see violation rates above 25%.

Table 3
Type-schema violations of the official multilingual baseline (mBERT) on the Russian test set, ordered by violation rate.

Relation	#Predictions	#Violations	%Violations
ASSOCIATED_WITH	4,932	2,701	54.8
USED_IN	506	225	44.5
FINDING_OF	1,088	408	37.5
APPLIED_TO	164	50	30.5
TREATED_USING	398	119	29.9
TO_DETECT_OR_STUDY	1,190	319	26.8
ORIGINS_FROM	577	148	25.6
PHYSIOLOGY_OF	976	249	25.5
ABBREVIATION	691	146	21.1
SUBCLASS_OF	5,217	1,038	19.9
AFFECTS	2,369	383	16.2
PART_OF	1,331	143	10.7
HAS_CAUSE	3,885	294	7.6
ALTERNATIVE_NAME	68	1	1.5
Total	23,392	6,224	26.6

Table 4
Per-relation F1 of the official baseline (mBERT) on the test set, by subtask.

Relation	English	Russian	Bilingual
ABBREVIATION	0.3910	0.4904	0.4369
AFFECTS	0.3633	0.3450	0.3702
ALTERNATIVE_NAME	0.0552	0.2900	0.1953
APPLIED_TO	0.1183	0.1268	0.1273
ASSOCIATED_WITH	0.1809	0.2538	0.2250
FINDING_OF	0.4567	0.4203	0.4075
HAS_CAUSE	0.1931	0.2268	0.2281
ORIGINS_FROM	0.2384	0.2472	0.2529
PART_OF	0.2527	0.3960	0.3336
PHYSIOLOGY_OF	0.3447	0.3866	0.3701
SUBCLASS_OF	0.6256	0.5534	0.6041
TO_DETECT_OR_STUDY	0.2651	0.2239	0.2516
TREATED_USING	0.2584	0.1711	0.2525
USED_IN	0.2117	0.1674	0.1833

3.3. Official Results

In total, we have received 20 Codabench registrations for the BioNNE-R task, with 7 teams submitting predictions during the evaluation phase. The official evaluation results, ordered by macro-F1, for BioNNE-R are summarized in Table 5. Below, we give an overview of these approaches.

Team **ELiRF-UPV** achieved the best results in the English-oriented track. The team adopted the OpenNRE framework, including negative sampling for the `no_relation` class. Several BERT-based models were evaluated as encoder backbones, including BERT Base [23], BioBERT [24], and BlueBERT [25]. In addition, the team experimented with sequential and bidirectional LSTM [26] layers as well as different input representations for entity pairs. The best-performing systems were based on the `bert-base-multilingual-uncased` model combined with additional sequential layers before the final softmax classifier. Due to the limited number of English training samples, the final English-oriented system employed transfer learning from a model initially trained on the Russian dataset, followed by fine-tuning on the English subset. The predictions generated for each subtask were further postprocessed by removing relations between entity type pairs not observed in the training data. The team also filtered relations between entities located in different sentences. Finally, probability thresholds were optimized

Table 5
Official BioNNE-R results: macro-F1 of each team's best submission per subtask. Best result per subtask in bold; – indicates no submission.

Team	English	Russian	Bilingual
ELiRF-UPV	**0.5060**	0.4717	0.4540
LODAC-NII	0.5016	**0.5283**	**0.5015**
aakobiakova	0.4951	0.4403	0.4716
HSE NLP	0.4733	0.5057	0.4527
rabiaozdemir	0.4620	–	–
savvafq	0.4570	0.4829	0.4849
LSI_UNED	0.4514	–	–
Baseline	0.2825	0.3070	0.3027

on the validation sets to improve the final prediction quality.

Team **LODAC-NII** achieved the best results in both Russian-oriented and bilingual tracks, while obtaining top performance comparable to the **ELiRF-UPV** team in the English-oriented track. They submitted two systems, where both systems shared a candidate generation pipeline that enumerated directed entity pairs constrained by nested structure, same-sentence co-occurrence, or an 80-token context window. To address class imbalance, the team applied stratified negative sampling with 1:5 and 1:10 ratios across five hardness categories. The first system was a typed-marker BERT-based relation classifier built on PubMedBERT for the English track, RuBERT for the Russian track, and multilingual BERT for the bilingual track. The model combined representations of the `[CLS]` token, entity spans, and additional structural features within a schema-masked classification framework. The second system employed a document-level relational graph architecture in which entity mentions were represented as nodes in a heterogeneous graph connected through nesting, overlap, adjacency, and same-sentence relations. The graph representations were encoded using R-GCN and R-GAT backbones [27, 28]. The team additionally experimented with UMLS-grounded `SAME_CONCEPT` edges; however, this component was disabled in the final submission because it did not provide consistent improvements on the development set.

Team **aakobiakova** submitted systems based on the `unsloth/llama-3-8b-bnb-4bit` model [29] fine-tuned using Parameter-Efficient Fine-Tuning (PEFT) with LoRA adapters [30]. The adapters were configured with rank $r = 16$ and scaling factor `lora_alpha` $= 16$ and applied to the main attention and feed-forward projection layers. To improve relation classification performance over the base fine-tuned model's direct decoding, the team employed a log-probability-based inference strategy in which the conditional log-probability of each candidate relation label was computed. Probabilities were estimated for all possible relation types, including the `no_relation` class. Inference was further calibrated using a dedicated validation procedure consisting of two stages. First, a bias shift for the `no_relation` class was optimized to maximize macro-F1 on the calibration set. Second, the team introduced a confidence-based filtering mechanism based on the probability margin between the top-1 and top-2 predicted labels. The optimal margin thresholds were determined by grid search over values ranging from 0.0 to 1.0 with a step size of 0.02. The resulting thresholds were 0.38 for the Russian-oriented model, 0.20 for the English-oriented model, and 0.32 for the bilingual model. During inference, predictions with a confidence margin below the corresponding threshold were reassigned to the `no_relation` class, indicating insufficient confidence in the predicted relation.

Team **HSE NLP** implemented a fine-tuned transformer architecture with three principal contributions. For typed entity encoding, they replaced generic special tokens with 32 typed entity markers of the form [HEAD:TYPE] / [TAIL:TYPE], covering all 8 entity types × 2 roles × open/close, inserted at character-level span boundaries before tokenisation; a binary nesting flag was appended to the entity representation to handle approximately 15% of pairs involving nested spans. For rare-class augmentation, they applied a two-stage LLM pipeline with GPT-5-mini as generator and GPT-5.4-mini as verifier to synthesise training examples for ALTERNATIVE_NAME, APPLIED_TO, and USED_IN (917 EN and 874 RU accepted examples); augmentation improved English performance but was net negative for Russian,

where the base training set was already substantially larger. For confidence calibration, they performed a blind-dev threshold sweep to correct for approximately 30× train-to-test negative-ratio mismatch, running inference over all candidate entity pairs in the dev documents at the same enumeration density as test. For English (Subtask 1), they used `BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext` pre-trained on PubMed abstracts and full-text; for Russian (Subtask 2) and Bilingual (Subtask 3), they used `mdeberta-v3-base` [31].

Team **savvafq** adopted the OpenNRE framework with entity-marker-based sentence encoding, inserting special tokens [E1]/[/E1] and [E2]/[/E2] around head and tail entity spans to produce relation-discriminative representations. For the English track, they fine-tuned `BioLinkBERT-large` [32], a biomedical language model pre-trained on PubMed abstracts with inter-document citation structure; for the Russian and bilingual tracks, they used `XLM-RoBERTa-large` [33] to leverage its strong multilingual representations. To address the highly imbalanced relation distribution and the large number of negative candidate pairs at inference time, the team applied three post-processing techniques: entity-type masks that suppress predictions for entity-type pairs never observed in training, per-class decision threshold calibration that independently optimizes the decision boundary for each of the 14 relation types on the development set, and distance-based candidate pruning that discards entity pairs whose spans are more than 200 characters apart. For the English track, they additionally averaged the logits of two models trained with different random seeds, which provided a further improvement in macro F1.

Team **LSI_UNED** participated in the English-oriented track and addressed the task by formulating relation extraction as a text classification problem over pairs of entities and introduced a new `NO_RELATION` label to model the absence of relations. Each candidate pair is represented using special markers that encode both entities and their corresponding types, and several BERT-based language models were fine-tuned for classification, including `bert-base-multilingual-cased`, `bert-base-uncased`, `xlm-roberta-base`, `biobert-v1.1` and `Bio_ClinicalBERT` [34]. To reduce the large number of possible entity combinations and alleviate class imbalance, a pruning strategy was applied based on valid entity-type combinations, character distance between entities, and sentence boundaries.

Team **rabiaozdemir** participated in the English-oriented track; a system description was not available at the time of writing.

In summary, as shown in Table 5, all participating systems substantially outperformed the baseline. Two teams led the competition: **ELiRF-UPV** achieved the highest macro-F1 in the English-oriented track (0.5060), while **LODAC-NII** achieved the best performance in both the Russian-oriented (0.5283) and bilingual (0.5015) tracks, with only a 0.0044 margin separating the two top English systems. Among the top-performing systems, graph-based relational modeling as employed by **LODAC-NII** and cross-lingual transfer learning as adopted by **ELiRF-UPV** proved to be effective strategies for addressing the complexity of the task. The use of large generative models, either for direct inference as in **aakobiakova**'s LLaMA-based approach or for training data augmentation as in **HSE NLP**'s GPT-based pipeline, highlights growing interest in applying generative models to biomedical relation extraction.

A more detailed comparison including training and inference cost, per-relation and per-entity-type performance, and grouping of systems by backbone and by open- versus closed-weight model use is left to future work.

4. Conclusion

This paper presented an overview of the BioNNE-R shared task on biomedical nested relation extraction at BioASQ 2026. The task evaluated extraction of relations between nested biomedical entities in English and Russian across English, Russian, and bilingual subtasks. Seven teams submitted predictions, most framed the task as pair classification with typed entity markers and fine-tuned transformer encoders, with postprocessing like schema-based type filtering, threshold calibration, and candidate pruning to handle the imbalanced relation distribution; a document-level relational graph model and LLM-based approaches were also explored. All submissions outperformed the multilingual baseline, yet the

best macro-F1 scores remained near 0.50-0.53, indicating that nested-entity relation extraction is far from solved. Several aspects of the evaluation could be strengthened in future editions. The reference baseline is deliberately minimal and omits the schema masking and per-class threshold calibration that most participants found beneficial. A stronger baseline incorporating them would better separate the contribution of the encoder from that of inference-time postprocessing. Per-relation scores, several estimated from fewer than one hundred test instances, would also benefit from confidence intervals (e.g. via bootstrap resampling), and reporting train- and development-set metrics alongside test scores would help distinguish optimization failures from generalization gaps. Covering per-relation and per-entity-type performance, training and inference cost, and the error modes of the strongest systems and above statements are within current limitations of this work. Future directions include joint modeling of nested entity structure and relations, cross-lingual transfer, handling of discontinuous entities, and LLM-based methods.

Acknowledgments

We are grateful to all participating teams for their submissions, system descriptions, and the diverse range of approaches they brought to the BioNNE-R shared task. We also thank the BioASQ organizers for their support in hosting this shared task.

Declaration on Generative AI

During the preparation of this work, author(s) used Claude (Anthropic) for: drafting content, paraphrase and reword, grammar and spelling check, and coding assistance for the evaluation and data-processing scripts. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, et al., Clinical information extraction applications: a literature review, Journal of biomedical informatics 77 (2018) 34–49.

[2] N. Perera, M. Dehmer, F. Emmert-Streib, Named entity recognition and relation detection for biomedical information extraction, Frontiers in cell and developmental biology 8 (2020) 673.

[3] U. Hahn, M. Oleynik, Medical information extraction in the age of deep learning, Yearbook of medical informatics 29 (2020) 208–220.

[4] E. Tutubalina, Z. Miftahutdinov, V. Muravlev, A. Shneyderman, A comprehensive evaluation of biomedical entity-centric search, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track, 2022, pp. 596–605.

[5] M. Y. Landolsi, L. Hlaoua, L. B. Romdhane, Information extraction from electronic medical documents: state of the art and future research directions, Knowledge and Information Systems 65 (2022) 463.

[6] V. Arsenyan, S. Bughdaryan, F. Shaya, K. W. Small, D. Shahnazaryan, Large language models for biomedical knowledge graph construction: information extraction from emr notes, in: Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, 2024, pp. 295–317.

[7] X. Zhao, Y. Deng, M. Yang, L. Wang, R. Zhang, H. Cheng, W. Lam, Y. Shen, R. Xu, A comprehensive survey on relation extraction: Recent advances and new frontiers, ACM Computing Surveys 56 (2024) 1–39.

[8] D. Zhou, D. Zhong, Y. He, Biomedical relation extraction: from binary to complex, Computational and mathematical methods in medicine 2014 (2014) 298473.

[9] Y. Park, G. Son, M. Rho, Biomedical flat and nested named entity recognition: Methods, challenges, and advances, Applied Sciences 14 (2024) 9302.

[10] Y. Wang, H. Tong, Z. Zhu, Y. Li, Nested named entity recognition: a survey, ACM Transactions on Knowledge Discovery from Data (TKDD) 16 (2022) 1–29.

[11] T. Lv, L. Luo, J. Li, Y. Wang, Y. Pan, C. Liu, Y. Wang, Y. Jiang, H. Lv, Y. Sun, et al., A unified biomedical named entity recognition framework with large language models, in: 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2025, pp. 3913–3916.

[12] C. Li, X. Zheng, S. Liu, Bibert-pipe on biomedical nested named entity linking at bioasq 2025, arXiv preprint arXiv:2509.09725 (2025).

[13] F. Borchert, I. Llorca, M.-P. Schapranow, Improving biomedical entity linking for complex entity mentions with llm-based text simplification, Database 2024 (2024) baae067. URL: https://doi.org/10.1093/database/baae067. doi:10.1093/database/baae067. arXiv:https://academic.oup.com/database/article-pdf/doi/10.1093/database/baae067/58662801/baae067.pdf.

[14] 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), 2024, pp. 2383–2389.

[15] 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.

[16] 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.

[17] V. Davydova, N. V. Loukachevitch, E. Tutubalina, Overview of bionne task on biomedical nested named entity recognition at bioasq 2024., in: CLEF (Working Notes), 2024, pp. 28–34.

[18] A. Sakhovskiy, N. Loukachevitch, E. Tutubalina, Overview of the bioasq bionne-l task on biomedical nested entity linking in clef 2025, in: CLEF, 2025.

[19] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, I. Rozhkov, E. Tutubalina, D. Dimitriadis, V. Patsiou, 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 2026: The fourteenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, volume Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026) of *Lecture Notes in Computer Science*, Springer, 2026.

[20] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic Acids Res. 32 (2004) 267–270. URL: https://doi.org/10.1093/nar/gkh061. doi:10.1093/NAR/GKH061.

[21] 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.

[22] X. Han, T. Gao, Y. Yao, D. Ye, Z. Liu, M. Sun, Opennre: An open and extensible toolkit for neural relation extraction, in: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP): system demonstrations, 2019, pp. 169–174.

[23] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 1 (2018) 4171–4186. URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[24] 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.
[25] Y. Peng, S. Yan, Z. Lu, Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets, in: Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019), 2019, pp. 58–65.
[26] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.
[27] M. Gori, G. Monfardini, F. Scarselli, A new model for learning in graph domains, in: Proceedings. 2005 IEEE international joint conference on neural networks, 2005., volume 2, IEEE, 2005, pp. 729–734.
[28] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph attention networks, in: International Conference on Learning Representations, 2018.
[29] AI@Meta, Llama 3 model card (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
[30] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models., volume 1, 2022, p. 3.
[31] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021. arXiv:2111.09543.
[32] M. Yasunaga, J. Leskovec, P. Liang, Linkbert: Pretraining language models with document links, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 8003–8016.
[33] 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.
[34] 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. URL: https://www.aclweb.org/anthology/W19-1909. doi:10.18653/v1/W19-1909.