ELiRF-UPV at BioNNE-R Task: Transfer Learning on Transformer-based Models for Biomedical Relation Classification

BioASQ at CLEF 2026

Pere Marco-Garcia[1,*], Vicent Ahuir[1], Marc Hurtado-Beneyto[1] and Antonio Molina-Marco[1,2]

[1]Valencian Research Institute for Artificial Intelligence - Universitat Politècnica de València (VRAIN-UPV)
[2]Valencian Graduate School and Research Network of Artificial Intelligence (ValgrAI)

Abstract

The implementation of NLP solutions in the biomedical domain is being applied to an elevated number of fields, thereby increasing the complexity of the problems posed by new, but widely applicable tasks. The BioNNE-R Shared Task is an example of it. This task involves extracting biomedical entity relations, with the added difficulty of the presence of both nested named entities and relations. The task also presents the challenge of addressing this task in English and Russian. In this paper, we present the ELiRF-UPV participation in this shared task, in the context of the BioASQ Lab at CLEF 2026. We show the different models implemented to approach this challenge as a Relation Classification task. Our systems rely on Transformer models based on BERT followed by a post-processing phase based on the dataset characteristics and the experimental results for the validation set. We also present a model trained through a Transfer Learning approach, obtaining competitive results in all tracks, achieving the best results in the English one. During the Test Phase, our best systems achieved a value of 0.5060 in F1-macro for the English Track, 0.4717 for the Russian Track, and 0.4540 for the Bilingual Track.

Keywords

BERT Models, Biomedical Domain, Relation extraction, Transfer Learning

1. Introduction

The healthcare domain can be highly encouraged by NLP research. Although machine learning models are not yet reliable enough to make diagnoses or critical decisions [1], they can be extremely useful when it comes to retrieving and presenting data extracted from scientific articles and clinical notes [2]. Nowadays, we can see a wide range of contexts where NLP techniques are used to solve healthcare domain problems, such as the detection of medical entities [3], the retrieval and presentation of relevant information from different sources [4], or the assistance with the processing of large amounts of data through the generation of informed summaries [5].

Relation extraction is of particular importance in the field of medicine due to the large number of interactions between the elements that appear in its documents. Some of these relationships may involve links between treatments and conditions, tests carried out and their results, or changes in the nomenclature of complex concepts, whether through abbreviations or specifications. Several papers in this field presented recent enhancements in the area: Zhou, S. & Yu, S., 2023[6], show how the integration of LLMs can represent an important knowledge grounding for addressing these problems, and Da Silva. D. et al., 2023[7] use a combined NER and RE approach to extract relevant information regarding patients, such as tests, dates, or intervals, from electronic health records. We also find a relevant corpus labeled for the NLP task in the biomedical domain. It's the case of NEREL-BIO [8], a

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
The code implemented for this work is not planed to be published.
pmarco@vrain.upv.es (P. Marco-Garcia); vahuir@upv.es (V. Ahuir); mhurben@upv.es (M. Hurtado-Beneyto); amolina@upv.es (A. Molina-Marco)
0009-0003-7026-3543 (P. Marco-Garcia); 0000-0001-5636-651X (V. Ahuir); 0009-0005-5651-5243 (M. Hurtado-Beneyto); 0000-0001-6537-8803 (A. Molina-Marco)


corpus in Russian and English generated from annotating Pubmed abstracts extending NEREL[9]'s previous work. From this corpuses, we can find multiple shared tasks aiming to developing models for QA in the biomedical domain [10] or entity linking from nested named entities [11].

In this work, we present our participation in the BioNNE-R[12] Shared Task at BioASQ 2026[13]: a nested relation extraction between biomedical entities in a multilingual dataset for Russian and English. In Section 2, we describe the task in detail. In Section 3, we describe the provided dataset and its characteristics. In Section 4, we present the approaches we tested to solve the task. In Section 5, we show the results obtained by the experimentation. Finally, in Section 6, we present our conclusions and the expected future work based on these results.

2. Task Description

The BioNNE-R Shared task is a competition framed in the context of the NEREL-BIO dataset[8], a project presenting an annotation on scheme and a Corpus obtained from Pubmed abstract in the Russian and English language. Several works were presented using the data provided by this project, as are the BIONNE Task[14], focused on the nested named entities recognition on biomedical context; and other work aiming for semantic indexing and QA [10].

In this case, the objective of the task involves the relation extraction of nested named entities, that is, entities that contain other entities within their boundaries, in the biomedical domain. Relations can be found at multiple levels of entities and also between overlapped entities. The data provided by the organizer can be found in English and in Russian and was extracted from the NEREL-BIO dataset. Each labeled sample includes the following fields: document-id, relation, head-text, head-span, head-type, tail-text, tail-span, tail-type. We can find 8 different types of entities in the annotation and 14 relation types. They will be better described in the following section.

The BioNNE-R challenge is divided into 3 subtasks: two monolingual tracks for English and Russian, and a Bilingual track that combines the data available for both languages. The predictions provided by a system must be for a specific task and can't be used for multiple tracks. The metric used for evaluation is F1-macro on the relations predicted averaged over 14 relation types. The relation must present an exact span match, providing the same doc id, location of the entities related, their types, and the type of the relation.

The task includes an OpenNRE-based model that uses bert-base-multilingual-cased with entity markers as baseline for the challenge. It handles negative sampling and enumerates all candidate entity pairs at the inference time. You can find more details of their implementation published on Github[1].

3. Dataset

The dataset provided consists of a series of documents labeled with entities and the relations between them. The texts provided are abstracts of scientific papers from the biomedical domain. Those documents are annotated with named entities of the biomedical domain which could be nested, multiple times in some cases, providing the document where the entity appears, the label of the entity, the text span that composes the entity, and the text span on the original text. Entities can be of one of the following types: ANATOMY, CHEM, DEVICE, DISO, FINDING, INJURY_POISONING, LABPROC, and PHYS.

For the training and validation set, we also have available the relations present in each document, describing the relationship between two existing entities and assigning them a head and tail role. The relations are distributed in 14 classes: ABBREVIATION, AFFECTS, ALTERNATIVE_NAME, APPLIED_TO, ASSOCIATED_WITH, FINDING_OF, HAS_CAUSE, ORIGINS_FROM, PART_OF, PHYSIOLOGY_OF, SUBCLASS_OF, TO_DETECT_OR_STUDY, TREATED_USING, and USED_IN. In our approach, since we used negative sampling, which we explain further below, to train the model, we also included an additional "no_relation" label.

[1]https://github.com/nerel-ds/NEREL-BIO/tree/master/BioNNE-R/baseline

The dataset is divided by language, English and Russian. As we can see in Table 1, we have a disparity in the number of samples available for each language. In the English dataset, the number of documents is almost the same in both the training (55) and validation (50) partitions. Meanwhile, the Russian dataset has a higher number of documents and labeled data in the training partition (717).

Table 1
Distribution of documents, entities and relations in each partition.

Partition	# Docs	# Entities	# Relations
Train$_{EN}$	55	3,861	3,193
Dev$_{EN}$	50	3,478	2,967
Train$_{RU}$	717	37,879	23,773
Dev$_{RU}$	50	3,210	2,521

We also studied the distribution of samples per relation type. We show in Table 2 how, in both languages, we deal with a highly unbalanced dataset. The 5 most abundant classes represent the majority of the samples (around 70% in both languages), while the remaining 9 classes account for less than 30% of the annotated relations.

Table 2
Distribution of the number of samples for each relation class in each language.

	English		Russian	
Relation Class	# of samples	% of samples	# of samples	% of samples
APPLIED_TO	97	1.57%	329	1.25%
USED_IN	168	2.73%	463	1.76%
ABBREVIATION	171	2.78%	929	3.53%
TREATED_USING	189	3.07%	328	1.25%
ALTERNATIVE_NAME	193	3.13%	746	2.84%
ORIGINS_FROM	199	3.23%	633	2.41%
FINDING_OF	204	3.31%	1441	5.48%
TO_DETECT_OR_STUDY	280	4.55%	855	3.25%
PHYSIOLOGY_OF	328	5.32%	1258	4.78%
PART_OF	452	7.34%	1691	6.43%
ASSOCIATED_WITH	558	9.06%	3537	13.45%
AFFECTS	755	12.26%	2781	10.58%
HAS_CAUSE	1066	17.31%	3117	11.85%
SUBCLASS_OF	1500	24.35%	8186	31.13%
Total	6160	100.00%	26294	100.00%

We also obtained the distribution of relations according to the number of consecutive sentences between the entities involved. In both language partitions, we observed that the majority of relations were inside a unique sentence. In English, the 95.83% of relations followed this description, and in Russian, it was the 96.67% of the samples. We also observed the pairs of types of entities involves according to each relation label. We observed that the majority of entities involved in each relations class tended to correspond to a small set of pairs of entity types.

4. System Description

Our approach is mainly based on framing the task as a classification problem. Given a pair of entities the model must decide if there exists a relation and, if it does, determine its kind. In all the systems implemented we added negative sampling to the model to represent the "no_relation" class. We also used BERT [15] models as main component of our systems. We applied a post-processing to the predictions of all the models we used. We limited the predictions to relate entities which were on the same sentence. We also discarded the predictions that related 2 types of entities by a class not seen in the dataset samples of the original training sets. Finally, we filtered the outputs of the model by the confidence

in the label prediction for all systems except the unfiltered baseline introduced as System1 below. For each model, we selected the best threshold based on the performance over the validation set.

For the English model, given that it was the smaller corpus, we modified multiple aspects of the architecture. The implemented systems are the following:

- **mBERT.** This system used a bert-base-multilingual-uncased model, used the given partitions for training and evaluations and obtained the predictions by the majority label given by the model.
- **mBERT(+thr).** Has the same implementation of mBERT, but we maximized the evaluation metric for the validation set by excluding the confidence of the prediction under a specific threshold. From now on, this implementation is used as a post-processing step for all the systems.
- **mBERT+BiLSTM.** In this system, we tested adding a BiLSTM layer to the previous model.
- **mBERT(80/20).** It starts from mBERT(+thr) implementation, but we redistributed the data provided in a training and a validation set with 80% and 20% of the samples each.
- **BioBERT(80/20).** We replicate the mBERT(80/20) implementation, but using BioBERT[16] as Bert model.
- **mBERT(RU->EN).** This system follows the same approach, but instead of starting from a generalist model, it starts it's training from a model trained with Russian samples before. With this implementation, we aim to use the task knowledge given by the Russian samples and specialize it in the English Subtask.

Other changes were tested on validation, but were discarded to be used for test predictions for their performance or other restrictions.

The best improvements in this dataset can be found with the threshold search, increasing the samples, using the BioBERT model as base model and with the transfer learning implementation. Only the threshold change could be transferred to the Russian and multilingual approaches. Due to task time constraints, no further implementations were tested for these two subtasks. mBERT(RU->EN)$_RU$ and mBERT(RU->EN)$_BI$ represent the described implementation for subtracks of Russian and bilingual respectively.

- **mBERT**$_{RU}$**.** Represents the mBERT system implementation for Russian subtrack.
- **mBERT**$_{BI}$**.** Represents the mBERT system implementation for bilingual subtrack.

Table 3
Description of the systems implemented and used to obtain the test predictions.

Name	Language	Base Model	Threshold	Partitions
mBERT	English	BERT Multilingual	No	Original
mBERT(+thr)	English	BERT Multilingual	Yes	Original
mBERT+BiLSTM	English	BERT Multilingual + BiLSTM	Yes	Original
mBERT(80/20)	English	BERT Multilingual	Yes	80 - 20
BioBERT(80/20)	English	BioBERT	Yes	80 - 20
mBERT(RU->EN)	English	BERT Multilingual + Russian training	Yes	80 - 20
mBERT$_{RU}$	Russian	BERT Multilingual	Yes	Original
mBERT$_{BI}$	Bilingual	BERT Multilingual	Yes	Original

4.1. Setup

In the experiment exploration, we also considered multiple hyperparameter settings. The systems finally presented the same hyperparameters with 10 training epochs, batch size of 32, max-length of 256 tokens, negative sampling ratio of 3, AdamW as optimizer and 2e-5 as learning rate.

The only exception to this setup was the BioBERT(80/20) model that, due to memory and time limitations, was only trained for 5 epochs and had 8 as parameter for the batch size. We also must

mention the case of the Transfer Learning implementation, mBERT(RU->EN). The Russian training ran for 10 epochs while the English fine-tuning step ran only for 3 epochs.

The threshold used to improve the validation results changed depending on the execution. However, all the values are situated between 0.95 and 0.99. All the experimentation ran accelerated by a NVIDIA GeForce RTX 4090.

5. Experimental Results and Discussion

In this section, we show the performance of the different models trained for both the validation set and the test set. In Table 4, we can see the value of the metric macro_f1 obtained for each system described. On English models, we can observe a consistent increase of the performance as the changes on the implementation were being applied. The only change that didn't report an increase in the metric was the BiLSTM addition. For this reason, we did not test further implementations following this approach for the other subtracks.

With the mBERT(80/20) results, we show how the increment of training samples is critical for the training of the model, accrediting the poorer results of the English track to the little number of examples instead of the difficulty of the ability of the model to learn such task. In the transfer learning approach, mBERT(RU->EN), we confirm this hypothesis. With the prior training of the model with the Russian samples, the model learns the task before adapting it to the English language and representations, obtaining the best results of all the implementations tested. We must also mention the results of the system based on BioBERT. The performance closely mirrors the results of mBERT(RU->EN), highlighting also the relevance of domain specialized Models for specialized vocabulary management. However, on the test results, we see a bigger gap from the best result achieved.

Table 4
Performance of the systems implemented over the validation and test sets.

System	Dev Results	Test Results
mBERT	0.3478	0.3271
mBERT(+thr)	0.4004	0.3883
mBERT+BiLSTM	0.3849	0.3823
mBERT(80/20	0.5858	0.4341
BioBERT(80/20)	0.6304	0.4608
mBERT(RU -> EN)	0.6417	0.5060
mBERT$_{RU}$	0.4891	0.4717
mBERT$_{BI}$	0.4683	0.4540

In Table 5, the results shown represent the best models per track of each team based on the leaderboard of the shared task. We can see how, our English approach gets the best results among all the participants, highlighting the relevance of the transfer learning system. The other two track results remain competitive, but show some improvement margin compared to other teams' implementations.

Table 5
Best results per team on each task. Only best results included on this table were considered to elaborate the ranking.

Team	English Track	Russian Track	Bilingual Track
ELiRF-UPV	0.5060 (1)	0.4717	0.4540
lakshan-nii	0.5016 (2)	0.5283 (1)	0.5015 (1)
savvafq	0.4570	0.4829 (3)	0.4849 (2)
avaliev	0.4733	0.5057 (2)	0.4527
aakobiakova	0.4951 (3)	0.4403	0.4716 (3)
rabiaozdemir	0.4620	–	–
lsi_uned	0.4514	–	–
Baseline	0.2825	0.3070	0.3027

For knowledge purposes, and after the evaluation phase ended, we decided to use the model 6 over the Russian and bilingual dataset to know how the two phase training affected the model ability to predict the relations. The only change applied is the threshold selected in the post-processing, which was adapted to each validation set. On table 6 we observe that the performance of the Transfer Learning model overcome the basic models trained only with Russian samples. We see a relevant improvement over the development set that is not as noticeable across the test set. However, the results remain slightly superior, indicating that, after some epochs without seeing any Russian samples, the model has continued to acquire knowledge about the task whilst retaining the language component. On the bilingual approach, we see a bigger increase leading the model to produce the second-best results in the testing phase. These results suggest that training in two phases, separating the samples by language, could lead the model to extract better the finer points of the task.

Table 6
Results of the English mBERT(RU->EN) model over the validation and test sets for the Russian and bilingual tracks.

System	Dev Results	Test Results
mBERT(RU->EN)$_{RU}$	0.5149	0.4725
mBERT(RU->EN)$_{BI}$	0.5854	0.4892

6. Conclusions and Future Work

In this paper, we presented our approach to the BioNNE-R Shared Task describing the systems implemented and the differences between them based on the metric performance. We saw how the best approach in the competition for the English samples was the transfer learning system, including the post-processing steps. For the Russian and bilingual subtracks, we implemented competitive systems, but not many of the improvements spotted for the English approach were exploitable. However, the Transfer Learning approach proved to also have a good performance for the two tracks, in which it was not specialized. We also saw how the domain specific BERT model, BioBERT, obtained a strong performance, but being a monolingual model did not have a direct usage for the other two tracks.

The transfer of knowledge between datasets can be crucial in tasks with small data available or in less represented languages. In our systems, we can see how the knowledge from the Russian dataset could be transferred to a task with fewer samples, as was the case of the English dataset. This area could be deeply explored in the future for extracting knowledge about transfer learning and aiming to replicate this technique in other environments. We must also insist on the importance of the domain specific models, as we have seen with the BioBERT based model. Another interesting area of research would be the implementation of multilingual domain specific models, as XLM-R fine-tuned on PubMed-style data, or the development of tools to transfer the specialized knowledge in existing models to multiple languages.

Acknowledgments

This work is supported in part by MICIU/AEI/10.13039/501100011033 and FEDER, EU, under Project PID2024-155948OB-C55. It is also partially supported by the Generalitat Valenciana under the GVA-Predoctoral Research Grant (CIACIF/2022/234).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] E. Yu, X. Chu, W. Zhang, X. Meng, Y. Yang, X. Ji, C. Wu, Large language models in medicine: Applications, challenges, and future directions, Int J Med Sci 22 (2025) 2792–2801. URL: https://www.medsci.org/v22p2792.htm. doi:10.7150/ijms.111780.

[2] V. Kocaman, F.-Y. Cheng, J. Bonis, G. Raut, P. Timsina, D. Talby, A. Kia, Exploring named entity recognition potential and the value of tailored natural language processing pipelines for radiology, pathology, and progress notes in clinical decision support: Quantitative study, JMIR AI 4 (2025). URL: https://www.sciencedirect.com/science/article/pii/S2817170525000687. doi:https://doi.org/10.2196/59251.

[3] M. S. Ullah Miah, J. Sulaiman, T. B. Sarwar, S. S. Islam, M. Rahman, M. S. Haque, Medical named entity recognition (medner): A deep learning model for recognizing medical entities (drug, disease) from scientific texts, in: IEEE EUROCON 2023 - 20th International Conference on Smart Technologies, 2023, pp. 158–162. doi:10.1109/EUROCON56442.2023.10199075.

[4] P. Richter-Pechanski, P. Wiesenbach, D. M. Schwab, C. Kiriakou, N. Geis, C. Dieterich, A. Frank, Clinical information extraction for lower-resource languages and domains with few-shot learning using pretrained language models and prompting, Natural Language Processing 31 (2025) 1210–1233. doi:10.1017/nlp.2024.52.

[5] V. Kieuvongngam, B. Tan, Y. Niu, Automatic text summarization of covid-19 medical research articles using bert and gpt-2, 2020. URL: https://arxiv.org/abs/2006.01997. arXiv:2006.01997.

[6] S. Zhou, S. Yu, High-throughput biomedical relation extraction for semi-structured web articles empowered by large language models, 2024. URL: https://arxiv.org/abs/2312.08274. arXiv:2312.08274.

[7] D. P. da Silva, W. da Rosa Fröhlich, B. H. de Mello, R. Vieira, S. J. Rigo, Exploring named entity recognition and relation extraction for ontology and medical records integration, Informatics in Medicine Unlocked 43 (2023) 101381. URL: https://www.sciencedirect.com/science/article/pii/S2352914823002277. doi:https://doi.org/10.1016/j.imu.2023.101381.

[8] N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, Biomedical concept normalization over nested entities with partial umls terminology in russian, 2024, p. 2383 – 2389.

[9] 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.

[10] A. Nentidis, A. Krithara, G. Paliouras, M. Krallinger, L. G. Sanchez, S. Lima, E. Farre, N. Loukachevitch, V. Davydova, E. Tutubalina, Bioasq at clef2024: The twelfth edition of the large-scale biomedical semantic indexing and question answering challenge, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 14612 LNCS (2024) 490 – 497. doi:10.1007/978-3-031-56069-9_67.

[11] N. Loukachevitch, P. Braslavski, V. Ivanov, T. Batura, S. Manandhar, A. Shelmanov, E. Tutubalina, Entity Linking over Nested Named Entities for Russian, in: Proceedings of LREC, 2022.

[12] I. Rozhkov, N. Loukachevitch, E. Tutubalina, Overview of the BioASQ BioNNE-R Task on Nested Biomedical Relation Extraction at CLEF 2026, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.

[13] 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.

[14] V. Davydova, N. Loukachevitch, E. Tutubalina, Overview of bionne task on biomedical nested

named entity recognition at bioasq 2024, volume 3740, 2024, p. 28 – 34.
[15] 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.
[16] 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: http://dx.doi.org/10.1093/bioinformatics/btz682. doi:10.1093/bioinformatics/btz682.