KAZAN FEDERAL UNIVERSITY

as a manuscript

Tutubalina Elena

MODELS AND METHODS FOR AUTOMATIC PROCESSING OF UNSTRUCTURED DATA IN BIOMEDICAL DOMAIN

Dissertation Summary
for the purpose of obtaining academic degree
Doctor of Science in Computer Science

Kazan — 2023

The dissertation was prepared at Kazan Federal University.

1 Introduction

Topic of the thesis

This dissertation presents a comprehensive study that aims to improve the efficiency and effectiveness of processing unstructured data in the biomedical domain. The study introduces novel approaches for classification and information extraction, including the recognition of various entity mentions such as drugs, diseases, genes, and adverse drug reactions, as well as entity linking (also known as medical concept normalization) and relation extraction.

Natural language processing (NLP) in the biomedical domain poses several challenges due to the complexity of biomedical language and the vast amount of data generated in the field. Biomedical data comes from various sources, such as electronic health records, scientific publications, and clinical trial data, social media, which can have different formats, structures, and levels of quality. Some of the major challenges are as follows. Firstly, biomedical language is often ambiguous, with many terms having multiple meanings. E.g., "adenoid hypertrophy" ("гипертрофия аденоидов") may be linked to "nasopharyngeal tonsil hypertrophy (adenoids)" ("гипертрофия глоточных миндалин (аденоиды)") or "hypertrophy of adenoids exclusively" ("гипертрофия исключительно аденоидов"), two different concept unique identifiers (CUIs) in the Unified Medical Language System (UMLS) [1]. Some concepts have different CUIs while they are synonymous in their meaning; for example, "acholic stool" ("ахоличный стул") has a code C2675627 and "pale stool" ("светлый стул") has a code C0232720. Secondly, the medical language includes domain-specific terms and expressions that are not commonly used in everyday language. In the context of this problem, an NLP model has to be capable of translating layperson language into formal medical language. For example, the phrase "I can't fall asleep all night" ("всю ночь не могу уснуть") should be translated to "insomnia" ("бессонница"), and "head spinning a little" ("немного кружится голова") should be translated to "dizziness" ("головокружение"). This requires more than the simple matching of natural language expressions and vocabulary elements, as string matching approaches may not be effective in linking social media language to medical concepts when the words do not overlap at all. Thirdly, medical terminology is vast and continuously evolving, with different ontologies used across countries and even within different medical specialties. Medical concepts may have different types (e.g., drugs, diseases, or genes/proteins) and may be

retrieved from different single-typed ontologies. The holy grail of modern medical NLP is to effectively identify and map the same concepts across different ontologies without re-training models. Fourthly, annotated data for biomedical NLP is often limited, making it difficult to train and evaluate models effectively. Despite having a large number of resources in the general domain, many languages have not made significant progress in the biomedical field. Russian is one such example; it is one of the top 10 languages in the world and has many NLP datasets and resources, but the biomedical part of Russian NLP is underdeveloped. The Russian UMLS includes translations of Medical Dictionary for Regulatory Activities (MedDRA) [2], Logical Observation Identifiers Names and Codes (LOINC) [3], and Medical Subject Headings (MeSH) [4]. However, it only amounts to 1.8% of the English UMLS in vocabulary and 1.36% in source counts [5]. Addressing these challenges requires the development of new annotated corpora, advanced techniques, and models that can handle the complexity and variation of medical language, as well as the availability of high-quality annotated data for evaluation. The dissertation addresses these challenges by introducing new annotated corpora of texts from various sources, proposing novel evaluation strategies and advanced techniques such as Transformers [6], Bidirectional Encoder Representations from Transformers (BERT) [7], and metric learning [8–10] for optimizing the models. It demonstrates the effectiveness and robustness of the proposed approaches through extensive experiments and evaluations.

Objectives and goals of the dissertation The dissertation has three main objectives:

1. Development of NLP methods and models in the specialized domain based on deep neural networks, pre-trained models, and metric learning approaches.
2. Analysis of limitations and development of novel strategies for evaluating trained models in information retrieval and information extraction tasks.
3. Creation of new annotated corpora of texts from various sources, such as scientific abstracts, drug reviews, electronic health records, and clinical trials, in both English and Russian.

The ultimate objective is to improve the efficiency and effectiveness of biomedical search engines, pharmacovigilance systems, and medical records management and analysis in the healthcare field.

Main results

The following are the main results of this dissertation:

– New models and methods for classification and information extraction were developed:

1. Multilingual BERT-based models were analyzed for cross-domain drug and disease named entity recognition in two languages. The investigation of transfer learning strategies between four corpora demonstrated the effectiveness of pretraining on data with one or both types of transfer [11].
2. Classification-based methods were proposed with (i) a set of informative features at an entity level and a context level for relation extraction [12], and (ii) vectors of semantic similarity for entity linking [13;14]. The effectiveness of these approaches was demonstrated in multiple shared tasks, ranking first in SMM4H 2019 Task 3, SMM4H 2020 Task 3, and SMM4H 2021 Task 1c [13; 14]. The semantic similarity vectors also proved effective with a proposed encoder-decoder architecture that ranked first in CLEF eHealth 2017 Task 1 [15].
3. DILBERT (Drug and disease Interpretation Learning with Biomedical Entity Representation Transformer) was introduced. The model optimizes the relative similarity of mentions and concept names from a terminology via metric learning. It was shown that the model is robust to vocabulary switches and can recognize concepts that were not present in the training set [16;17].
4. A multimodal model combining BERT-based models for language understanding and molecular property prediction was proposed to improve the classification of tweets as potential sources of adverse drug events or drug reactions. The model achieved first and second place rankings on SMM4H 2021 Task 2 and Task 1a, respectively [18].
5. Two neural pipelines were developed: (i) a pipeline consisting of two models as a biomedical search engine that showed superior performance over a traditional search model on a manually annotated dataset of abstracts for disease and gene queries [19], and (ii) a pipeline for the classification, extraction and normalization of adverse drug events on

realistic, imbalanced data. The identification of optimal training ratios and undersampling methods was also explored [20].

– New annotated corpora were developed for information extraction. The following are some of the new corpora developed:

6. The Russian Drug Reaction Corpus (RuDReC), a partially annotated corpus of consumer reviews in Russian about pharmaceutical products, and RuDR-BERT models for named entity recognition and sentence classification tasks were introduced [21].

7. Two annotated datasets were developed for clinical concept normalization: a dataset of clinical trials in English for drug and disease normalization [16; 17], and a RuCCoN corpus, a new dataset of electronic health records in Russian, with entities linked to the UMLS [22].

8. NEREL-BIO, an annotation scheme and corpus of PubMed abstracts in Russian and English with general-domain and biomedical entity types, was introduced. The corpus includes the provision of an annotation for nested named entities [23].

– New evaluation strategies were proposed, as follows:

9. The limitations of existing benchmarks for biomedical entity linking were analyzed, and several novel evaluation strategies were proposed: (i) a novel *stratified* sampling split [13], (ii) *in-terminology* and *cross-terminology* evaluation [24]. Additionally, benchmarks were established for the cross-lingual task using clinical reports, clinical guidelines, and medical research papers. A test set *filtering* procedure was designed to analyze the "hard cases" of entity linking approaching zero-shot cross-lingual transfer learning [25].

10. The limitations of existing benchmarks of scientific abstracts and electronic health records for relation extraction were analyzed. To address performance differences in *in-domain* and *out-of-domain* setup, a cross-attention neural model was proposed that exhibits better cross-domain performance [26].

Author's contribution includes the problem formulations, the development of aforementioned methods and models for processing unstructured data, the design of annotation schemes for the aforementioned corpora and evaluation

strategies, analysis of results; the first versions of programs implementing the proposed methods and models for classification and named entity recognition and their evaluation were personally developed by the author of the dissertation; the current versions of software modules implementing the methods proposed in the dissertation within various hardware and software architectures were written under the direct supervision of the author of the dissertation.

The scientific novelty of the proposed research lies in the development of new annotated corpora for various texts, the development of novel deep learning architectures and models for biomedical information extraction and classification of texts in several languages, and novel evaluation strategies. The improvement of the quality of the developed methods in comparison with existing methods has been confirmed experimentally using standard quality metrics of natural language text analysis systems. It is experimentally shown that the developed methods are applicable to texts from various sources. The first studies were conducted to solve the problem of extracting mentions of drug effects and biomedical nested named entities for the Russian language.

The scope of dissertation is covered in 42 publications [11–52].

According to regulations of the Dissertation Council in Computer Sciences of Higher School of Economics, at least ten papers are listed below. In this list, papers are specifically mentioned: [12; 13; 17; 18; 20; 21; 23; 26] in Q1-journals; [11; 16; 19; 22; 24] in proceedings of CORE A/A* conferences, [14; 15; 25] in conference proceedings indexed on Scopus. The defense is performed based on at least seven of them (namely, the first nine from the list of first-tier publications).

Publications and probation of the work

First-tier publications

1. Miftahutdinov Z., Alimova I., **Tutubalina E.** On biomedical named entity recognition: experiments in interlingual transfer for clinical and social media texts //European Conference on Information Retrieval (ECIR). – 12036 LNCS, Springer, Cham, 2020. – pages 281-288. [Scopus, WOS, CORE A conf.]

   Contribution of the dissertation's author: main co-author; the author of this thesis formulated the scientific problem, developed neural models for named entity recognition, the first versions of programs that implement the proposed models, and performed an experimental evaluation.

2. **Tutubalina E.**, Kadurin A., Miftahutdinov Z. Fair Evaluation in Concept Normalization: a Large-scale Comparative Analysis for BERT-based Models //Proceedings of the 28th International Conference on Computational Linguistics (COLING). – 2020. – pages 6710-6716. [Scopus, CORE A conf.]
   Contribution of the dissertation's author: main co-author; the author of this thesis formulated the scientific problem, proposed novel evaluation strategies, the first versions of programs that implement the proposed evaluation, and performed an experimental evaluation of models.

3. Sakhovskiy A., **Tutubalina E.** Multimodal model with text and drug embeddings for adverse drug reaction classification //Journal of Biomedical Informatics. – 2022. – Vol. 135. – pages 104182. (Q1, Impact Factor 8.0) [Scopus, WOS]
   Contribution of the dissertation's author: main co-author; the author of this thesis formulated the scientific problem, proposed a multimodal model in collaboration with S.A., and guided the research.

4. **Tutubalina E.**, Miftahutdinov, Z., Muravlev, V., Shneyderman, A. A Comprehensive Evaluation of Biomedical Entity-centric Search //Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track. – 2022. – pages 596-605. [Scopus, CORE A conf.]
   Contribution of the dissertation's author: main co-author; the author of this thesis formulated the scientific problem, guided the annotation process, developed an information retrieval system, the first versions of programs that implement the proposed system, and conducted experiments.

5. Miftahutdinov Z., Kadurin A., Kudrin R., **Tutubalina E.** Drug and Disease Interpretation Learning with Biomedical Entity Representation Transformer //European Conference on Information Retrieval (ECIR). – 12656 LNCS, Springer, Cham, 2021. [Scopus, Core A conf.].
   Contribution of the dissertation's author: main co-author; the author of this thesis formulated the scientific problem, proposed evaluation methodology, proposed a DILBERT model in collaboration with M.Z., designed experiments, and guided the research.

6. Miftahutdinov Z., Kadurin A., Kudrin R., **Tutubalina E.** Medical concept normalization in clinical trials with drug and disease representation learning //Bioinformatics. – 2021. – V. 37. – №. 21. – pages 3856-3864 (Q1, Impact Factor 6.931) [Scopus, WOS]

Contribution of the dissertation's author: main co-author; same as [16] (this is the journal paper based on the conference version [16]).

7. **Tutubalina E.**, Alimova I., Miftahutdinov Z., Sakhovskiy A., Malykh V., and Nikolenko S. The Russian Drug Reaction Corpus and Neural Models for Drug Reactions and Effectiveness Detection in User Reviews //Bioinformatics. — 2020. — 07. DOI: 10.1093/bioinformatics/btaa675 (Q1, Impact Factor 6.931) [Scopus, WOS]
   Contribution of the dissertation's author: main co-author; the author of this thesis formulated the scientific problem, wrote a program for data collection, developed neural models for classification and named entity recognition, the first versions of programs that implement the proposed models, proposed an annotation scheme, guided the annotation process, and partially conducted experiments.

8. Nesterov, A., Zubkova G., Miftahutdinov Z., Kokh, V., **Tutubalina E.**, Shelmanov A., Alekseev A., Avetisian M., Chertok A., and Nikolenko S. RuCCoN: Clinical Concept Normalization in Russian //Proceedings of the Annual Meeting of the Association for Computational Linguistics. – 2022. – pages 239-245. [Scopus, Core A* conf.]
   Contribution of the dissertation's author: the author of this thesis formulated the scientific problem, wrote a program for additional training data collection, proposed several types of test sets for various settings, and partially conducted experiments.

9. Loukachevitch N., Manandhar S., Elina Baral E., Rozhkov, I., Braslavski P., Ivanov V., Batura T., and **Tutubalina E.** NEREL-BIO: A Dataset of Biomedical Abstracts Annotated with Nested Named Entities// Bioinformatics. — 2023. — 04. — btad161. (Q1, Impact Factor 6.931) [Scopus, WOS]
   Contribution of the dissertation's author: main co-author; the author of this thesis designed an annotation scheme in collaboration with L.N., wrote a program for data collection, set up annotation tools, developed models for initial data annotation as well as the program code that implement these models.

10. Alimova I., **Tutubalina E.** Multiple features for clinical relation extraction: a machine learning approach //Journal of biomedical informatics. – 2020. – T. 103. – pages 103382 (Q1, Impact Factor 8.0) [Scopus, WOS]

Contribution of the dissertation's author: main co-author; the author of this thesis formulated the scientific problem, proposed a feature-based model in collaboration with A.I., and guided the research.

11. **Tutubalina E.**, Miftahutdinov Z., Nikolenko S., & Malykh V. Medical concept normalization in social media posts with recurrent neural networks //Journal of biomedical informatics. – 2018. – Vol. 84. – pages 93-102. (Q1, Impact Factor 8.0) [Scopus, WOS]
Contribution of the dissertation's author: main co-author; the author of this thesis formulated the scientific problem, proposed a classification model in collaboration with M.Z., designed evaluation, and guided the research.

12. Alimova I., **Tutubalina E.**, Nikolenko S. I. Cross-Domain Limitations of Neural Models on Biomedical Relation Classification //IEEE Access. – 2021. – Vol. 10. – pages 1432-1439. (Q1, Impact Factor 3.476) [Scopus, WOS]
Contribution of the dissertation's author: the author of this thesis formulated the scientific problem, proposed a neural model in collaboration with A.I., designed evaluation, and guided the research.

13. Magge A., **Tutubalina E.**, Miftahutdinov Z., Alimova I., Dirkson A., Verberne S., Weissenbacher D., Graciela Gonzalez-Hernandez G.. DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter //Journal of the American Medical Informatics Association. – 2021. – Vol. 28. – №. 10. – pages 2184-2192. (Q1, Impact Factor 7.942) [Scopus, WOS]
Contribution of the dissertation's author: the author of this thesis proposed and developed two information extraction models, and the first versions of programs that implement the proposed models.

Second-tier publications:

14. Miftahutdinov Z., **Tutubalina E.** Deep learning for ICD coding: Looking for medical concepts in clinical documents in English and in French //International Conference of the Cross-Language Evaluation Forum for European Languages. – Springer, Cham, 2018. – pages 203-215. [Scopus, WOS]
Contribution of the dissertation's author: main co-author; the author of this thesis formulated the scientific problem, proposed an encoder-decoder architecture with semantic similarity features in collaboration with M.Z., designed experiments, and guided the research.

15. Alekseev A., Miftahutdinov Z.,**Tutubalina E.**, Shelmanov A., Ivanov V., Kokh V., Nesterov A., Avetisian M., Chertok A., Nikolenko S. Medical Crossing: a Cross-lingual Evaluation of Clinical Entity Linking // 2022 Language Resources and Evaluation Conference, LREC 2022. — 2022. — pages 4212–4220. [Scopus, Core C conf.].
Contribution of the dissertation's author: the author of this thesis formulated the scientific problem, proposed novel evaluation strategies, the first versions of programs that implement the proposed evaluation, and guided the research.

16. Mftahutdinov Z., **Tutubalina E.** Deep neural models for medical concept normalization in user-generated texts // ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop. — 2019. — pages 393–399. [Scopus, WOS]
Contribution of the dissertation's author: main co-author; same as [13] (in this work, the experimental part from [13] is expanded).

Invited talks at conferences and seminars:

1. 27th Annual Conference on Intelligent Systems for Molecular Biology & 18th European Conference on Computational Biology ISMB/ECCB 2019, Basel, Switzerland, 21.07-25.07.2019, "Towards the Semantic Interpretation of User-Generated Texts about Drug Therapy";
2. Lecture from the cycle "On the edge of science", Moscow, Russia, 23.11.2021, "How to train artificial intelligence to identify adverse drug effects from social media posts";
3. International Scientific Conference "Machine Learning and Artificial Intelligence Technologies" (MLW 2021), Sochi, Russia, 25.11.2021, "Drug and Disease Interpretation Learning";
4. Open conference on artificial intelligence Opentalk.AI 2020, Moscow, Russia, 19.02-21.02.2020, "Processing messages from social media about side effects of drugs";
5. Educational Intensive "Archipelago 20.35", Innopolis, Russia, 11.11.2020, "Processing messages from social networks about side effects of drugs";
6. 4th Social Media Mining for Health Applications Workshop & Shared Task (SMM4H 2019), Florence, Italy, 28.07-03.08.2019, "KFU NLP Team

at SMM4H 2019 Tasks: Want to Extract Adverse Drugs Reactions from Tweets? BERT to The Rescue";

7. Data Fest 2018, 28.04.2018, "What's hurting you? Application of NLP methods in drug discovery";
8. 3rd Kazan Summer School on Chemoinformatics, Kazan, Russia, 5.07-7.07.2017, "Text Mining in Biomedical Research".

Contributed reports at conferences and seminars:

9. 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022), Abu Dhabi, OAE, 7.12-11.12.2022, "A Comprehensive Evaluation of Biomedical Entity-centric Search";
10. 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022), Ireland, Dublin, 22.05-27.05.2022, "RuCCoN: Clinical Concept Normalization in Russian";
11. 13th Language Resources and Evaluation Conference (LREC 2022), Marseill, France, 21.06-23.06.2022, "Medical Crossing: a Cross-lingual Evaluation of Clinical Entity Linking";
12. 7th Social Media Mining for Health Applications Workshop & Shared Task (SMM4H 2022), online, 17.10.2022, "SMM4H 2022 Task 2: Dataset for stance and premise detection in tweets about health mandates related to COVID-19";
13. 43rd European Conference on Information retrieval (ECIR 2021), online, 28.03-1.04.2021, "Drug and Disease Interpretation Learning with Biomedical Entity Representation Transformer";
14. 6th Social Media Mining for Health Applications Workshop & Shared Task (SMM4H 2021), online, 10.06.2021, "KFU NLP Team at SMM4H 2021 Tasks: Cross-lingual and Cross-modal BERT-based Models for Adverse Drug Effects";
15. Widening Natural Language Processing Workshop (WiNLP 2021), online, 21.11.2021, "Adverse Drug Reaction Classification of Tweets with Fusion of Text and Drug Representations";
16. Ivannikov ISP RAS Open Conference 2021, 02.12-03.12.2021, Moscow, Russia, "Cross-Lingual Transfer in Drug-Related Information Extraction from User-Generated Texts";

17. 28th International Conference on Computational Linguistics (COLING 2020), online, 8.12-12.12.2020, “Fair Evaluation in Concept Normalization: a Large-scale Comparative Analysis for BERT-based Models”;
18. 5th Social Media Mining for Health Applications Workshop & Shared Task (SMM4H 2020), online, 12.12.2020, “KFU NLP Team at SMM4H 2020 Tasks: Cross-lingual Transfer Learning with Pretrained Language Models for Drug Reactions”;
19. 42nd European Conference on Information retrieval (ECIR 2020), online, 14.04-17.04.2020, “On Biomedical Named Entity Recognition: Experiments in Interlingual Transfer for Clinical and Social Media Texts”;
20. 8th International Conference on Analysis of Images, Social networks and Texts (AIST 2019), Kazan, Russia, 17.07-19.07.2019, “Biomedical Entities Impact on Rating Prediction for Psychiatric Drugs”;
21. Google NLP Summit 2019, 24.06-26.06.2019, “Towards the Semantic Interpretation of User-Generated Texts about Drug Therapy”;
22. 57th Conference of the Association for Computational Linguistics (ACL 2019), Florence, Italy, 28.07-03.08.2019, “Deep Neural Models for Medical Concept Normalization in User-Generated Texts”;
23. 57th Conference of the Association for Computational Linguistics (ACL 2019), Florence, Italy, 28.07-03.08.2019, “Detecting Adverse Drug Reactions from Biomedical Texts With Neural Networks”;
24. 21th International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID 2019), Kazan, Russia, 15.10-18.10.2019, “A comparative study on feature selection in relation extraction from electronic health records”;
25. VI International Conference “Information technologies, telecommunications and control systems” (ITTCS 2019), Innopolis, Russia, 6.12.2019, “Comparative Analysis of Context Representation Models in the Relation Extraction Task from Biomedical Texts”;
26. Ivannikov ISP RAS Open Conference 2018, 21.11-22.11.2018, Moscow, Russia, “Comparative analysis of neural networks in the problem of classification of side effects at the level of entities in English texts”;

27. 9th International Conference and Labs of the Evaluation Forum (CLEF 2018), Avignon, France, 10.09-14.09.2018, “Deep Learning for ICD Coding: Looking for Medical Concepts in Clinical Documents in English and in French”;

28. Machine Learning for Health Workshop (ML4H 2018), Montreal, Canada, 2.12-08.12.2018, “Sequence Learning with RNNs for Medical Concept Normalization in User-Generated Texts”;

29. Artificial Intelligence and Natural Language Conference (AINL 2018), St. Petersburg, Russia, 17.10-19.10.2018, “Interactive Attention Network for Adverse Drug Reaction Classification”;

30. Russian Summer School in Information Retrieval (RuSSIR 2018), Kazan, Russia, 27.08-31.08.2018, “Using semantic analysis of texts for the identification of drugs with similar therapeutic effect”;

31. International Conference on Computational Linguistics and Intellectual Technologies “Dialog”, Moscow, Russia, 30.05-02.06.2018, “Leveraging Deep Neural Networks and Semantic Similarity Measures for Medical Concept Normalization in User Reviews”;

32. Ivannikov ISP RAS Open Conference 2017, 30.11-1.12.2017, Moscow, Russia, “A machine learning approach to classification of drug reviews in Russian”;

33. 8th International Conference and Labs of the Evaluation Forum (CLEF 2017), Dublin, Ireland, 11.09-14.09.2017, “KFU at clef ehealth 2017 task 1: Icd-10 coding of english death certificates with recurrent neural networks”;

34. IEEE 30th Neumann Colloquium (NC 2017), Budapest, Hungary, 24-25.11.2017, “End-to-end deep framework for disease named entity recognition using social media data”;

35. International Conference on Computational Linguistics and Intellectual Technologies “Dialog”, Moscow, Russia, 31.05-3.06.2017, “Identifying disease-related expressions in reviews using conditional random fields”.

2 New models and methods for classification and information extraction

New models and methods for classification and information extraction (IE) were proposed and developed by the author of this dissertation [11–20]. Consolidating knowledge about drugs and diseases across different sub-domains is crucial for effective biomedical applications, especially considering the vast amount of biomedical texts that require analysis. Therefore, the use of automated NLP methods is imperative for efficient information retrieval (IR) or IE. In particular, the following key scientific problems, addressed in this chapter, are discussed:

- The first problem, as highlighted in [11], is the significant human effort required to annotate sufficient training examples for each language or subdomain in modern supervised models. Furthermore, Named Entity Recognition (NER) models may exhibit exceptionally poor performance when faced with domain shift or language shift, which is another major challenge in biomedical NLP.
- Current neural network-based approaches for detecting adverse drug events (ADEs) from texts, as discussed in [18], are limited in their ability to leverage drug structure and mainly rely on capturing textual information from user posts about drugs.
- The third problem, as studied in [16; 17], is the cross-terminology mapping of entity mentions to a given lexicon without additional re-training. This is a common challenge in the biomedical domain, where different terminologies and ontologies are used to represent biomedical concepts.
- Another challenge, discussed in [19], is the effective retrieval and analysis of biomedical texts that focus on specific entities such as diseases, genes, and chemicals. With the overwhelming amount of text data produced in the biomedical field, coupled with the limitations of state-of-the-art IR approaches based on dense or sparse embeddings, there is a need for an entity-centric search engine design and evaluation.

To address the problems highlighted above, several new models and methods are developed:

- To address the first problem, multilingual transfer learning between electronic health records (EHRs) and user-generated texts (UGTs) in different

languages is explored, with the goal of investigating whether knowledge can be transferred from a high-resource language, such as English, to a low-resource language, such as Russian, to perform NER of biomedical entities [11]. This approach leverages the multilingual capabilities of pretrained models and incorporates transfer learning.

- To address the second problem, a novel method to utilize both textual and molecular information for ADE classification is proposed [18]. To fuse the drug and tweet representations, two strategies are explored, including using a co-attention mechanism to integrate features of different modalities.
- To address the third problem, a Drug and disease Interpretation Learning with Biomedical Entity Representation Transformer (DILBERT) model is proposed that uses metric learning and negative sampling to obtain entity and concept embeddings. The DILBERT model enables the creation of a shared semantic vector space for entities and concepts from the knowledge base, allowing for cross-terminology entity linking (EL) without the need for re-training [16; 17].
- To address the fourth problem, a BERT-based IE system is designed as an entity-centric search engine [19]. The system consists of two sub-modules, namely the NER sub-module and the EL sub-module, which are applied successively. The NER sub-module is responsible for identifying entities of interest, while the EL sub-module links the extracted entities with concepts from relevant knowledge bases using the DILBERT model. To evaluate the approach, a novel search collection of PubMed abstracts for disease and gene queries is developed, along with corresponding relevance judgments.

The key results and conclusions on transfer learning research [11], which explores multilingual transfer learning for NER in the biomedical domain, are as follows:

- Based on the evaluation results, it can be concluded that the multi-BERT approach exhibits the best transfer capabilities in the zero-shot setting when the training and test sets are either in the same language or in the same domain.
- Transfer learning is shown to effectively reduce the amount of labeled data required to achieve high performance. Specifically, trained models were able to achieve 98-99% of the full dataset performance on both types of entities after training on only 10-25% of sentences.

The key results and conclusions on the multimodal research [18] are as follows:

- The proposed approach is effective in utilizing both textual and molecular information for ADE classification, and achieves state-of-the-art performance on several benchmark datasets in English, French, and Russian.
- Experiments show that the molecular information obtained from neural networks is more beneficial for ADE classification than traditional molecular descriptors.

The key results and conclusions of studies on an information extraction system [16; 17; 19] are as follows:

- Experiments show that the DILBERT model significantly outperforms baseline and state-of-the-art architectures for biomedical EL. Moreover, this model is effective in knowledge transfer from the scientific literature to clinical trial data using a novel annotated dataset for drug and disease linking for evaluation.
- The neural IE architecture shows superior performance in a zero-shot setup for search with both disease and gene concept queries. Furthermore, the IE system can effectively handle out-of-domain abstracts, indicating its potential to be applied to a wide range of biomedical entities.

2.1 Cross-lingual and cross-domain NER with transfer learning

The results of this section are based on the paper [11].

Experiments are conducted using four datasets: English corpora CADEC [53] and n2c2 [54], a dataset comprising EHRs in Russian, and an author's original dataset consisting of UGTs in Russian. Each corpus is defined by two parameters: (i) language: English (EN) or Russian (RU); and (ii) domain: EHRs or UGTs.

NER model is based on BERT_{base} [7] with a softmax layer and the Adam optimizer with polynomial decay to update the learning rate on each epoch, with warm-up steps at the beginning. Word labels were encoded using the BIO tag scheme, and the model was trained on a sentence level. Specifically, multilingual Cased (Multi-BERT)is used, which is pretrained on 104 languages. LSTM-CRF with word embeddigs from the Saber library [55] and BioBERT [56] are used as baselines.

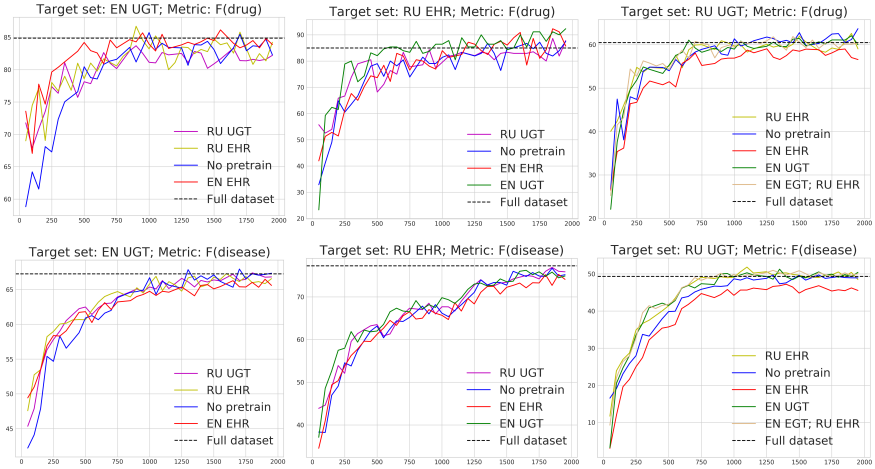

Figure 1 — Performance of Multi-BERT models with pre-training on the source dataset (a corpus's name in a legend) or without pre-training ("No pretrain" line) for the EN UGT, RU UGT, RU EHR datasets. Y-axis: F1-scores for detection of Drug or Disease mentions, X-axis: the number of sentences used for training.

Each of the datasets was randomly split into 70% training set and 30% test set. 720 models are trained on one machine with 8 NVIDIA P40 GPUs. The in-corpus (IC) and out-of-corpus (OOC) performances of all models were compared on the test sets using the CoNLL script for evaluating precision (P), recall (R), and F1-score (F) with exactly matching criteria.

Across all datasets, BERT-based models outperformed LSTM-CRF in terms of the IC performance. Furthermore, the difference in performance between BioBERT and Multi-BERT was not statistically significant (two-tailed t-test with $p \leq 0.05$). All models achieved significantly higher performance in detecting drugs than in detecting diseases. This may be attributed to boundary issues in multi--word expressions, as indicated by the average length of entities.

Zero-Shot Transfer To assess the efficacy of the BERT-based NER model trained on one corpus for detecting drugs and diseases in another language or domain in the zero-shot setting, Multi-BERT was trained on one corpus and then applied it to another language/domain without further training.

For drug recognition, the best generalizability was achieved by training on EHRs and evaluating on UGTs in English. When tested on the EN UGT corpus, the model achieved OOC F1 of 77.08% and 36.31% when trained on the EN EHR and RU UGT corpora, respectively, while the IC performance was an F1 of 84.88%. It is worth noting that the number of sentences in the EN EHR corpus is nine times greater than in the EN UGT corpus, and 78% of Drug tokens in the EN UGT corpus are found in the EN EHR set. When evaluated on the RU UGT corpus, the model obtained F1-scores of 26.31% and 34.78% when trained on the EN UGT and EN EHR corpora, respectively, while the IC performance was an F1-score of 60.45%.

For disease recognition, Multi-BERT exhibited much poorer generalization to corpora other than the one it was trained on. When evaluated on the RU UGT corpus, the model obtained OOC F1-scores of 24.12% and 30.86% when trained on the EN UGT and RU EHR corpora, respectively, while the IC performance was an F1-score of 49.35%. When tested on the EN UGT corpus, the model achieved F1-scores of 37.94% and 4.32% when trained on the RU UGT and EN EHR corpora, respectively, while the IC performance was an F1-score of 67.25%. This may be attributed to well-known differences between layperson language and professional medical terminology.

Few-Shot Transfer Next, it was investigated whether the NER model performs as well as a model trained on much larger datasets given a small number of training examples. The investigation began by randomly sampling 50 sentences from a “target” training set, training the pretrained model on this subsampled dataset, and testing it on the "target" test set. The sample size was then increased by 50 sentences of the “target” training set, and the process was repeated up to 2000 sentences of the training set. In each round, training was started from scratch to avoid overfitting, as suggested in [57].

For each pretraining setup, the size of the subset was recorded when the model achieved at least 99% of the F1-measure achieved on the full dataset. Results for the RU UGT, RU EHR, and EN UGT datasets are presented in Fig. 1. Multi-BERT pretrained on the EN UGT set and trained on 2000 sentences from the EN EHR corpus (2.81% of the full corpus) obtained 92% F1 and 76% F1 of the full dataset performance on drugs and diseases, respectively. As shown in Fig. 1, models with transfer knowledge outperformed models without the pretraining phase, even in cases where both domain and language shifted between “source” and “target” sets. Using the transfer learning strategy required up to 550 fewer

sentences than training from scratch. Notably, models required only 10% and 23% of the EN UGT and RU URT corpora, respectively, to achieve results as good as full dataset performances. Note that this observation is particularly significant for low-resource languages and new domains (e.g., social media, clinical trials). Additionally, it was observed that the performance of models with pretraining setup trained on different numbers of sentences became more stable in terms of deviations between F1-scores.

In summary, the following question was investigated: *can additional training on an existing dataset be helpful for the biomedical NER performance of a multilingual BERT-based NER model on a new dataset with a small number of labeled examples if the domain, the language, or both shift between these datasets?* As expected, models with pretraining on data in the same language or domain achieved better results in zero-shot or few-shot settings. The model with the best pretraining achieved 99% of the full dataset performance using only 23.56% of the training data on the RU URT corpus, while the model with pretraining on data with two shifts (the EN EHR set) used 26.1% of the training data. Hence, pretraining on data with two shifts can be effective.

2.2 Multimodal model with text and drug embeddings

The results of this section are based on the paper [18].

The popularity of social media as a source of health-related information has increased significantly over the past decade. One well-studied research area is pharmacovigilance from social media data, which focuses on discovering ADEs from user-generated texts. Note that the terms ADEs and adverse drug reactions (ADRs) are often used interchangeably.

A novel method to utilize both textual and molecular information for ADE classification is proposed, taking inspiration from multimodal studies. The impact of using different molecular representation approaches, including traditional molecular descriptors and neural encoders, is investigated.

Let us consider a text T that can be represented as a pair of textual modality t and a drug modality that is a set of k_T drug mentions: $T = (t, D_T)$, where $D_T = (d_1, d_2, ..., d_{k_T})$. To obtain two unimodal representations of T, a random drug d_i is sampled from D_T and encode the t and d_i using two encoders: (i) a

textual encoder M_{text} and (ii) a drug encoder M_{drug}:

$$u_{text}^T = M_{text}(t) \quad u_{drug}^T = M_{drug}(d_i)$$

where $u_{text}^T \in \mathbb{R}^{d_t}$ is the embedding of textual modality t and $u_{drug}^T \in \mathbb{R}^{d_d}$ is the embedding of drug d_i; d_t and d_d are the dimensionalities of the obtained uni-modal vector representations. The multimodal binary text classification task can be formulated as:

$$f_{cl}(f_{mod}(u_{text}^T, u_{drug}^T))$$

where $f_{mod} : \mathbb{R}^{d_t+d_d} \rightarrow \mathbb{R}^{d_{bi}}$ is a modality combination function that provides a bi-modal representation of T, $f_{cl} : \mathbb{R}^{d_{bi}} \rightarrow \mathbb{R}$ is a fully connected classification network with sigmoid output activation. d_{bi} is the dimensionality of the bi-modal vector representation.

The final embedding of the classification token (CLS token) is used as the textual representation of an input text. In the uni-modal approach, the textual representation u_{text}^T is usually directly fed to a classifier, whereas for bi-modal models, the textual modality with the drug modality are combined. Two ways are evaluated to combine text and drug modalities: (i) concatenation of text and drug embedding; (ii) the scaled dot-product attention used in Transformer-based models. In a first way, textual embedding and drug embedding are concatenated:

$$f_{mod}^{concat}(u_{text}^T, u_{drug}^T) = [u_{text}^T \oplus u_{drug}^T]$$

The attention mechanism allows to learn a multimodal embedding of sample T as a linear combination of two modalities u_{text}^T and u_{drug}^T. For text classification:

$$f_{mod}^{att}(u_{text}^T, u_{drug}^T) = \alpha \cdot u_{text}^T + (1 - \alpha) \cdot u_{drug}^T$$

where α is the weight of textual modality obtained using the attention mechanism.

To obtain drug information, drug names are mapped to their identifiers in DrugBank [58]. The following methods are employed as drug encoders: 1) ATC classification, which is the most widely recognized classification system for drugs. A drug is encoded as a sparse 14-dimensional vector with zero values for ATC codes to which a drug does not belong. 2) Molecular descriptors, which are a set of numeric values that characterize the physicochemical and structural properties of a drug molecule. Molecular-descriptor calculation library Mordred [59] is used to calculate 2 thousand descriptors for each drug from Drugbank. 3) MolBERT [60],

Table 1 — Evaluation results for SMM4H 2021 Task 2 on classification of Russian tweets. EnRuDR-BERT is trained, a bilingual model pretrained on a RuDReC corpus.

Model	**P±std**	**R±std**	**F_1±std**
Text-only models			
TF-IDF+SVM	0.158	0.281	0.202
Fasttext+CNN	0.356	0.465	0.404
BERT	0.548±0.063	0.516±0.072	0.524±0.020
Models with concatenated modalities			
BERT+ATC categories	0.529±0.054	0.521±0.058	0.519±0.090
BERT+descriptors	0.543±0.052	0.514 ±0.059	0.523±0.020
BERT+ChemBERTa	0.552±0.061	0.513±0.076	0.524±0.023
BERT+MolBERT	0.538±0.040	0.531±0.040	**0.532±0.014**
Models with cross-attention			
BERT+ATC categories	0.509±0.063	0.542±0.059	0.519 ±0.011
BERT+descriptors	0.527±0.076	0.518±0.073	0.514±0.022
BERT+ChemBERTa	0.515 ±0.049	0.537 ±0.063	0.521 ±0.020
BERT+MolBERT	0.514±0.051	0.553±0.055	0.529±0.011

a BERT-based encoder specifically designed for encoding molecular data. 4) ChemBERTa [61], another BERT-based encoder that is pretrained on chemical data.

To classify data, a fully-connected classification network is used with one hidden layer that utilizes GeLU [62] activations. The output layer uses sigmoid activation and binary cross-entropy is used as the loss function.

Three datasets are used for experiments: the French dataset from the Social Media Mining for Health (#SMM4H) 2020 task 1b and the English and Russian datasets from the SMM4H 2021 tasks. For each dataset, 10 BERT-based models are trained with different initializations of classifier weights. The mean quality and standard deviation of F_1-scores (F_1), precision (P), and recall (R) for the *ADE* class are computed.

Table 1 shows the performance of uni-modal and bi-modal models on the Russian tweet corpus of SMM4H 2021. First, the two best-performing models in terms of F_1-score are bi-modal models that utilize MolBERT drug embeddings. The concatenation-based model slightly outperforms the attention-based model (by

Table 2 — Official SMM4H 2021 results on the official test set.

Model	P	R	F_1
Task 2, Russian			
SMM4H 2021 Task 2 submission	0.58	0.57	**0.57**
Second place	0.54	0.57	0.52
Task 1a, English			
SMM4H 2021 Task 1a submission	0.552	0.681	**0.61**
First/second place, Team 4	0.515	0.752	**0.61**

0.3%) and unimodal BERT (by 0.8%). Second, bi-modal models with concatenated modalities outperform those with cross-attention. Finally, the final SMM4H 2021 submission combined the results of ten multimodal models with the same settings using a simple voting scheme, with the aim of enhancing the robustness of the final system. As illustrated in Table 2, this model achieved state-of-the-art results, exhibiting a 4% improvement in terms of F_1-measure over a single model.

After analyzing the results on the French set for SMM4H 2020 Task 1b, several conclusions can be drawn. Firstly, the highest F_1-scores were achieved by bi-modal models with concatenated modalities. Specifically, the bi-modal model with MolBERT drug embedding surpassed the best official results of the SMM4H 2020 competition and the text-only model, achieving a new state-of-the-art with an 8.5% improvement in terms of F_1. Secondly, the performance of models with cross-attention is significantly lower than those with modalities concatenation. However, cross-attention models with ChemBERTa and MolBERT embeddings slightly outperformed the text-only model by 1.8% and 0.9%, respectively. Finally, uni-modal and bi-modal BERT-based models significantly outperform unimodal SVM and CNN baselines, thus demonstrating the superiority of complex neural networks over simple neural models and traditional machine learning algorithms in the task of ADE text classification.

The top-performing bi-modal models in terms of F_1-score for both the Russian and English corpus of SMM4H 2021 are selected to determine how incorporating an additional modality affects classification performance on tweets that mention drugs from different therapeutic groups. Experiments showed that the performance of the bi-modal classification approach is not dependent on only ATC groups' distribution of input data.

In conclusion, the task of detecting ADEs in user-generated tweets about drugs has been investigated. The proposed unified approach combines several state-of-the-art models, including BERT for text representation and MolBERT for drug representation. The models demonstrated significant improvements in ADE classification of French texts of SMM4H 2020 Task 1b (achieving an 8% improvement) and achieved state-of-the-art results on recent SMM4H 2021 Task 1a and Task 2 for English and Russian texts, respectively. It is worth noting that these impressive results were obtained using only well-known BERT-based pre-trained models for individual components. The source code for the models are freely available at `https://github.com/Andoree/smm4h_2021_classification`.

2.3 Information extraction pipeline for search

The results of this section are based on the paper [19]; a novel DILBERT model for biomedical EL is based on the papers [16; 17].

The design and evaluation of a BERT-based IE system as an entity-centric search engine for a target discovery platform called `PandaOmics`[1] is presented. The following question is aimed to be answered: *given the near-excellent performance on NER and EL [14; 56], can models identify relevant publications for disease and gene queries from diverse biomedical subdomains as real-world applications?* A novel search collection of PubMed abstracts for disease and gene queries with corresponding relevance judgments has been developed. The IE pipeline is evaluated using two trained BERT-based models for NER and EL, as well as the standard document retrieval model BM25 with off-the-shelf Elasticsearch software.

The focus of this study is an extraction of two types of entities, namely disease and gene. The IE system comprises multiple pipelines, each dedicated to a specific entity type. These pipelines consist of two sub-modules, namely the NER sub-module and the EL sub-module, which are applied successively.

Named Entity Recognition BioBERT is trained on a combination of the NCBI and BC5CDR (BioCreative V CDR) datasets for disease entities [63;64], and on the DrugProt dataset for gene entities [65]. To create the combined dataset, predefined train/test subsets are utilized and merged the datasets within these

[1] `https://pandaomics.com/`

splits. The model achieved an F-measure of 88.43% and 90.39% for disease and gene entities, respectively.

Entity Linking To link a named entity with its corresponding concept in a knowledge base, the task of named entity linking is performed. A concept in this scenario refers to an element within the knowledge base that represents a specific idea or notion related to a particular field of knowledge. The DILBERT model is utilized to link the extracted entities with corresponding concepts from relevant dictionaries. Similarly to the NER component, models are trained on BC5CDR and BioCreative II GN (BC2GN) [66]. The DILBERT model enables the creation of a shared semantic vector space for entities and concepts from the knowledge base, where texts with similar meanings are located close to each other. This characteristic allows ranking of concepts based on the distance function s.

Following the notation proposed in [67], both entities and concepts are mapped to vector representations using the following equation:

$$y_m = red(T(m)); y_c = red(T(c)), \tag{1}$$

where T is a deep neural network of the transformer architecture whose weights can be updated during fine-tuning, the function $red(\cdot)$ reduces a sequence of vectors into a single vector, where m represents an entity that needs to be linked to the corresponding concept, and c denotes the concept name. There are various implementations of the $red(\cdot)$ function, such as selecting the output corresponding to the CLS token or element-wise average pooling over all vectors to obtain a fixed-size vector. Empirical evidence has established that average pooling is the optimal option for the $red(\cdot)$ function.

The relevance score of the candidate c_i for the entity m is determined by the distance function, such as the Euclidean distance, applied to the corresponding vector representations:

$$s(m, c_i) = ||y_m - y_{c_i}||, \tag{2}$$

To train the network, a triplet objective function is employed that captures the semantic similarities and differences between concepts and entities. Given a mention of the entity m, the name of the corresponding (positive) concept c_g, and the name of the non-corresponding (negative) concept c_n, the triplet objective function adjusts the neural network so that the distance between m and c_g is less

Table 3 — Statistics of the annotated datasets of clinical trials.

Mention	**#texts**	**#texts with CUIs**	**#unique texts**	**#unique texts with CUIs**
Intervention	1075	794	838	671
Condition	819	804	638	638

than the distance between m and c_n for a given threshold. The loss function is expressed as:

$$max(s(m, c_g) - s(m, c_n) + \epsilon, 0), \tag{3}$$

where ϵ is an offset that ensures that c_g is at least ϵ closer to m than to c_n. In the experiments, ϵ is set to 1.

To generate positive examples, the dictionary is limited to concepts that correspond to the entity, while the remaining part of the dictionary is used to generate negative examples [68]. Several strategies are explored for selecting positive and negative examples: (i) random sampling, (ii) hierarchy random sampling (random sampling + n parents), (iii) resampling of negative cases, (iv) resampling hierarchy (resampling + n siblings).

The metric learning approach's key feature is its ability to detect entities that do not have a suitable concept in the vocabulary. The detection methodology follows naturally from the assumption that similar elements are situated close to each other in a latent space. Therefore, if all dictionary objects are far enough from the entity, it implies the out-of-vocabulary case. Thus, if all concepts are at a distance greater than the threshold value t, it can be concluded that the entity is not corresponded by any of them. To determine the threshold, the maximum distance of true-positive cases d_{tp} and the minimum distance of false-positive cases d_{fp} is used. The threshold value is set to the weighted sum:

$$t = a_1 * d_{tp} + a_2 * d_{fp}, \tag{4}$$

Here, a_1 represents the proportion of true positive examples among entities whose closest concept is at a distance of $s \in [d_{fp}; d_{tp}]$, and a_2 denotes the proportion of false positives in the same entity set. If the given set of entities is empty, the coefficients are set to $\frac{1}{2}$.

The proposed approach was evaluated on the manually annotated clinical trial corpus. Statistics of annotated clinical trials' texts are summarized in Table 3. The experimental results are presented in Table 4. As it can be seen from

Table 4 — The metrics of the DILBERT model on the corpora of clinical trials. The results are presented for the disease (CT Condition) and drug (CT Intervention) entity types. Quality metrics are provided for a subset consisting only of entities with a single concept (single concept) and for the entire corpus (full set).

Model	**CT Condition**		**CT Intervention**	
	single concept	full set	single concept	full set
BioBERT ranking	72.60	71.74	77.83	56.97
BioSyn [69]	86.36	-	79.58	-
DILBERT with different sampling strategies				
random sampling	85.73	84.85	**82.54**	**81.16**
random + 2 parents	86.74	86.36	81.84	79.14
random + 5 parents	**87.12**	**86.74**	81.67	79.14
resampling	85.22	84.63	81.67	80.21
resampling + 5 siblings	84.84	84.26	80.62	76.16

Table 4, the DILBERT model shows the best results on CT Condition and CT Intervention corpora. The source code of DILBERT and the CT sets are available on GitHub at `https://github.com/insilicomedicine/DILBERT`.

Retrieval

Here, the dataset for entity-centric search, which includes queries and the process used to collect relevance assessments, is described. Table 5 presents the statistics of the dataset.

Table 5 — Summary of statistics of the proposed dataset.

Subset	**#queries**	**avg. number of texts per query**		
		relevant label	**nonrelevant label**	**doubtful label**
Disease CUI	73	94.86	63.57	9.78
Gene CUI	79	109.39	21.62	5.93
Ambiguous	27	45.94	11.58	0.53
Total	152	102.41	41.76	7.78

Queries In the search scenario, a user can input a gene name or symbol, such as "PSEN1" (ENSG00000080815), and retrieve all relevant publications and associated diseases, including Alzheimer's disease (EFO:0000249). An autocomplete feature suggests search terms from disease or gene dictionaries.

Pooling The standard practice of IR collection building is adopted, using a *pooling* approach to combine retrieval results from two main sources:

1. Retrieval results are obtained from Elasticsearch, and the results are pooled from these runs up to a depth of 100.
2. Retrieval results are obtained from PubMed, and the results are pooled from these runs up to a depth of 100, excluding abstracts already retrieved by the first system.

The final assessment pool consists of 23,099 query-abstract pairs, with an average of 152 abstracts per query.

Relevance Assessment For each query-abstract pair, relevance judgments are collected from two annotators with biomedical degrees. A list of queries was created by an expert annotator with a Ph.D. in biology, using logs from the target discovery platform `PandaOmics`. The queries are disease CUIs and gene CUIs. In addition, the annotators selected a list of concepts with at least one ambiguous concept name (e.g., coad refers to chronic obstructive pulmonary disease and to colon adenocarcinoma (COAD)).

Each annotator selected a disease or gene query from the list of identified identifiers, along with an abstract containing information about the publication year and journal. The abstracts were presented in a random order. Annotators were then asked to (i) judge relevance on a 3-point scale (relevant, nonrelevant, or doubtful), and (ii) categorize the reason for relevance or nonrelevance.

Search evaluation

Precision (P), recall (R), and F-measure (F) are used to evaluate the system. Precision is calculated as the fraction of relevant documents among all retrieved documents. Similarly, recall is calculated as the fraction of relevant documents from all possibly relevant documents in the dataset. Query-document pairs are used with relevant and nonrelevant labels, excluding the doubtful category.

Table 6 — IR metrics on the full set of queries and on the subset of queries with ambiguous concepts.

Model	Full Set			Ambiguous Concepts		
	P	**R**	**F**	**P**	**R**	**F**
Queries with Disease CUIs						
BERT-based	93.97	84.41	88.93	97.72	93.81	95.73
Elasticsearch BM25	82.19	83.33	82.76	75.67	96.72	84.91
Queries with Gene CUIs						
BERT-based	92.24	85.45	88.71	93.02	93.85	93.43
Elasticsearch BM25	89.92	79.93	84.63	79.58	68.88	73.85
Both						
BERT-based	92.99	84.99	88.81	94.9	93.83	94.37
Elasticsearch BM25	86.23	81.44	83.77	77.59	80.39	78.96

Table 6 presents the performance comparison of the BERT-based pipeline and BM25 on the full set of queries and the subset of concepts with ambiguous names, respectively. The results show that the BERT-based system outperforms BM25 on both sets of the dataset for both types of entities. As expected, the performance difference between the two models is more significant on the subset with ambiguous concept names. Furthermore, for the BERT-based pipeline, precision is higher than recall.

A dataset for out-of-domain abstract detection is developed to further investigate search precision. Approximately 30,000 records are included from the PubMed journal list, which publish papers not only about biological entities, but also on cultural topics, economics and econometrics, artificial intelligence, law, linguistics and language, and so on (out-of-domain categories for us). Out-of-domain journals were manually selected by the expert annotator on which the IE system should return *zero results*. From these journals, 58,790 abstracts containing at least one gene or disease concept retrieved by Elasticsearch were randomly selected by us. It was found that in 90% of these abstracts, the BERT-based system did not identify any entities.

In conclusion, a comprehensive evaluation of a biomedical entity-centric search engine that utilizes BERT models for disease and gene extraction and linking has been conducted. This engine is part of a target discovery platform, which allows users to retrieve a list of relevant publications given a disease or gene concept query.

3 New annotated corpora for information extraction

New annotated corpora with texts from various sources were proposed and developed by the author of this dissertation [16; 17; 21–23]. Developing biomedical IE systems is challenging due to the lack of comprehensive annotated datasets. In particular, the following key scientific problems, addressed in this chapter, are discussed:

- Poorly composed contexts, the ubiquitous presence of colloquialisms, shortened forms, typing/spelling mistakes, and out-of-vocabulary words introduce challenges for the effective utilization of user-generated content in the health domain, as shown in [21]. Furthermore, general-domain language models may not perform well on biomedical texts since these texts often contain technical terms, abbreviations, and domain-specific concepts.
- The inherent complexity of the domain and its terminology. The mentions of diseases, symptoms, drugs, and other concepts are highly variable, and due to the large medical vocabulary, entity linking/concept normalization becomes a challenging yet essential problem, as shown in [22].
- The third problem, addressed in [23], is the limited ability of existing datasets and NER methods to capture the complex nested entity structures that are common in biomedical texts. Biomedical texts frequently contain mentions of entities, such as diseases containing body parts or chemicals, that are nested within each other. However, most existing datasets and NER methods are designed to capture flat mention structures and are often limited to the most common entity types like drugs/chemicals and diseases.

To address the problems highlighted above, several novel datasets are presented:

- To address the first problem, the RuDReC [21] corpus was introduced, which is a partially annotated corpus of consumer reviews in Russian about pharmaceutical products, along with the RuDR-BERT models pre-trained on 1.4 million health-related comments and fine-tuned for named entity recognition and sentence classification tasks.
- To address the second problem, RuCCoN [22] was introduced, a new manually annotated dataset for clinical concept normalization in Russian. It contains over 16K entity mentions manually linked to over 2K unique concepts from the Russian language part of the UMLS. Train/test splits were

developed for different settings (stratified, zero-shot, and CUI-less) and present strong baselines obtained with state-of-the-art models.

– To address the third problem, a substantial nested named entity dataset, NEREL-BIO [23] was created, using PubMed abstracts in Russian and English. This dataset includes manually annotated entity mentions of 37 types, including nested structures with up to six layers of depth.

The key results and conclusions of the study on a corpus of user-generated texts [21] are summarized as follows:

– The dataset of health-related consumer reviews in Russian, RuDReC, is divided into two parts: (i) 1.4 million comments that can be used to train modern language models, and (ii) 500 richly annotated reviews that can be used to train downstream task-specific models.

– An annotation scheme is developed that operates on both sentence and entity levels. The sentence-level labels indicate the presence or absence of health-related issues. Furthermore, the sentences that contain health-related issues are further annotated at the entity level to identify fine-grained subtypes such as drug classes and drug forms, drug indications, and drug reactions.

– Two domain-specific BERT-based language models were trained on the raw RuDReC part.

– The evaluation of several BERT-based models on the classification and extraction of health entities.

The key results and conclusions of the study on a corpus of electronic health records [22] can be summarized as follows:

– The dataset of electronic health records in Russian, RuCCoN, where entity mentions are linked to concepts from the UMLS.

– The provision of several types of test sets for various settings, including stratified, zero-shot, and CUI-less settings.

– The evaluation of several state-of-the-art models on RuCCoN, including various fine-tuning variations, and an investigation of the necessity of labeled data in Russian for cross-lingual concept normalization from English to Russian.

The key results and conclusions of the study on a corpus of abstracts [23], are as follows:

– The dataset of biomedical abstracts in Russian and English, NEREL-BIO, which is annotated with nested entities.

– An annotation scheme which includes 17 specialized biomedical entity types and 20 general-domain entity types.

– The evaluation of several state-of-the-art models for nested NER.

Overall, these results demonstrate significant advancements in the development of resources for NLP applications in the specialized domain.

3.1 RuDReC: drug reactions in health-related user reviews

The results of this section are based on the paper [21].

This section presents the design, composition, and construction of a large dataset of user-generated texts (UGTs) about pharmaceutical products in Russian. The presented RuDReC corpus is divided into two parts: a larger raw corpus of 1.4 million health-related comments that can be used to train modern language models based on self-supervised objectives, and a smaller part containing 500 richly annotated reviews that can be used to train downstream task-specific models. The primary downstream tasks, in this case, are NER and multi-label classification. The labeling in the second part consists of two main components: sentence labels and entity labels. The review posts were split into sentences and labeled for the presence of drug indications and symptoms of a disease (DI), adverse drug reactions (ADR), drug effectiveness (DE), and drug ineffectiveness (DIE). In the entity identification phase, 6 entity types are identified and extracted: drug names, drug classes, drug forms, ADR, DI, and Findings. A total of 2,202 sentences and 4,566 entities were labeled in the corpus.

The annotation process involved two stages. In the first stage, annotators with a background in pharmaceutical sciences were asked to read 400 reviews and highlight all spans of text, including drug names and the patient's health conditions experienced before, during, or after the drug use. In the second stage, annotators were asked to screen existing annotations and annotate new texts.

The analysis of existing corpora shows two main types of entities: Drug and Disease. After several discussions, annotators defined the following Disease subtypes: (1) disease name; (2) indication (Indication); (3) positive dynamics after or during taking the drug (BNE-Pos); (4) negative dynamics after the start or some period of using the drug (ADE-Neg); (5) the drug does not work after taking the

course (NegatedADE); (6) deterioration after taking a course of the drug (Worse). As DRUG subtypes, annotators have chosen: (1) drug names, (2) drug classes, and (3) drug forms.

Metrics for computing the relaxed agreement for DISEASE and DRUG entities were used from [53]. The average agreement was approximately 70%.

After completing the first stage of the annotation process, three of the authors screened the annotations and identified several issues. There were relatively few examples of *Worse* and *ADE-Neg* types. Additionally, the *BNE-Pos* entity types contained many overly broad entities that were not related to medical concepts.

To address these issues, several changes were made to the annotation scheme. *Worse* and *ADE-Neg* with *NegatedADE* entity types were combined into a single class called *Drug Ineffectiveness* (DIE) and spanned DIE annotation on the sentence level. *BNE-Pos* entities were spanned on the sentence level and renamed to *Drug Effectiveness* (DE). Finally, *Indication* and *Disease* entity types were combined into a single *Drug Indication* (DI) type, following the CADEC corpus. At the second stage of the annotation process, two annotators continued the process according to sentence classes and entity types.

The annotated corpus contains 500 reviews, including reviews about four groups of drugs: sedatives, nootropics, immunomodulators, and antivirals. Sedatives account for 60% of the reviews. Reviews of immunomodulatory drugs have longer sentences and tokens compared to other drug groups. On average, their reviews are 30% longer, and their maximum length is twice that of other groups, although their minimum length is equivalent. The average number of sentences in Russian reviews is higher than in the English CADEC and PsyTAR corpora, with an average of 9.71 sentences per review compared to 6 sentences in the other corpora.

The total number of annotated sentences in the entire corpus is 2,202, distributed among different categories as follows: DI (949), ADR (379), FINDING (172), DE (424), and DIE (278). The total number of annotated entities in the entire corpus is 4,566, distributed among different categories as follows: DRUGNAME (1043), DRUGCLASS (330), DRUGFORM (836), DI (1401), ADR (720), FINDING (236). The analysis of part-of-speech (PoS) tags for words in entities revealed that social media users tend to use more verbs to express symptoms and ADRs compared to formal medical concepts. In the annotated portion of the

RuDReC corpus, 18.26% of the words in disease-related entities are verbs, while only 2.53% of words in the MedDRA dictionary from UMLS v. 2020AA are verbs.

User reviews were collected by web page crawling from popular medical web portals that mostly contain drug reviews about pharmaceutical products, health facilities, and pharmacies. Duplicate comments were removed, and the resulting corpus contains 1,4 million texts, 1,104,054 unique tokens, and 193,529,197 tokens in total.

The multilingual version of BERT-base (Multi-BERT) was used as initialization for training domain-specific BERT, which is referred to as **RuDR-BERT**. It was observed that 800K and 840K pretraining steps were sufficient, which roughly corresponds to a single epoch on the corpus.

Classification models were evaluated with 5-fold cross-validation in terms of the F1 score. Table 7 presents the results of RuBERT, Multi-BERT, and fine-tuned RuDR-BERT models. Based on the results, several conclusions can be drawn. Firstly, the RuDR-BERT model achieved the best results among the comparable models. Secondly, the RuBERT model outperformed the Multi-BERT model by 3.12% in terms of macro F1-score, with the highest improvement observed for DE (+4.09%) and Finding entity types (+4.19%). Thirdly, the performance of RuDR-BERT on Finding (36.24%) is significantly lower than on ADR (74.15%) and DI (85.06%). This could be due to the similarity in contexts and a much lower number of training examples.

The F1 scores computed by exact matching criteria via a CoNLL script were used to compare NER models on 5-fold cross-validation. Table 8 shows the F1-score performance of fine-tuned RuBERT, Multi-BERT, and RuDR-BERT. Based on the results, several conclusions can be drawn. Firstly, the domain-specific RuDR-BERT outperforms both RuBERT and Multi-BERT on all entity types. Secondly, RuBERT, with a vocabulary of Russian subtokens generated on Wikipedia and news, outperforms Multi-BERT. Thirdly, similar to sentence classification, the performance of RuDR-BERT on Finding is significantly lower than on ADR and DI. Finally, all models achieve higher performance for drug-related entities than for disease-related entities, which may be due to boundary problems in multi-word expressions. RuDR-BERT achieves an F1-score of 81.34% on disease-related entities and an F1-score of 94.65% on drug-related entities. The average number of tokens on drug-related entities is 1.06, while the average number of tokens on disease-related entities is 1.77.

Table 7 — Performance of fine-tuned RuDR-BERT on sentence classification with comparison to multi-BERT and RuBERT.

Model	DE	DIE	ADR	DI	Finding	Macro F1
RuBERT	67.7±2.82	62.27±3.47	66.65±2.96	81.63±2.38	28.51±4.8	61.35±3.28
Multi-BERT	63.61±4.22	60.19±3.52	63.45±2.61	79.58±4.1	24.32±2.85	58.23±3.46
RuDR-BERT	76.61±4.08	72.06±5.29	74.15±5.01	85.06±2.49	36.24±6.91	68.82±4.76

Table 8 — Performance of fine-tuned RuDR-BERT on the NER task in comparison with Multi-BERT and RuBERT

Model	ADR	DI	Finding	Drugclass	Drugform	Drugname	Macro F1
RuBERT	54.51±3.9	69.43±4.98	27.87±5.92	92.78±1.14	95.72±1.38	92.11±1.56	72.07±2.03
Multi-BERT	54.65±2.38	67.63±3.62	25.75±7.86	92.36±2.72	94.89±0.97	91.05±0.61	71.06±2.46
RuDR-BERT	60.36±2.13	72.33±2.12	33.31±7.55	94.12±2.31	95.89±1.82	93.08±1.08	74.85±2.09

In summary, this study discussed the challenges of annotating health-related Russian comments and presented several baselines for the classification and extraction of health entities. The RuDReC corpus offers opportunities for researchers to develop and evaluate text mining models for gathering meaningful information about drug effectiveness and adverse drug reactions from layperson reports. Moreover, it allows for the analysis and comparison of variations of reported patient health conditions and drug reactions of different therapeutic groups of medications. The dataset and pretrained weights of the models have been made freely available at `https://github.com/cimm-kzn/RuDReC`.

3.2 RuCCoN: clinical concepts in medical histories of patients

The results of this section are based on the paper [22].

This section describes a mapping of clinical entities from the medical histories of patients to the Russian part of UMLS. In particular, the only large-scale dataset of clinical free-text notes in Russian with NER labeling [70] was enriched by adding entity linking labeling. The dataset, created by researchers and practitioners from the Scientific Center of Children Health (SCCH), is based on medical histories of over 60 SCCH patients with allergic and pulmonary disorders and diseases. It includes discharge summaries, radiology, echocardiography, and ultrasound diagnostic reports, recommendations, and other records from various physicians. The deidentified corpus, which is freely available for research purposes, comprises 160 fully annotated texts with almost 250,000 tokens, 18,200 annotated

entities, over 7,400 attributes, and 3,500 relations with seven types of entities: "Disease", "Symptom", "Drug", "Treatment", "Body location", "Severity", and "Course".

Annotators were asked to map an entity mention to a CUI from the UMLS. The goal of entity normalization is to assign the same identifier to different synonyms of a given medical concept; e.g., "anemic heart infarction" and "myocardial infarction" refer to the same concept with CUI C0027051. Three annotators independently annotated each entity, and the Inter-Annotator Agreement (IAA) was calculated as the accuracy of the markups matched by at least two annotators over all annotated mentions. At least two annotators linked an entity to the same concept from the ontology in 13,125 cases and annotated 1,032 entities as CUI-less; IAA was 78.37%. In 3,900 cases when all annotators disagreed, the expert annotator with Ph.D. in medicine was asked to decide whether the CUI selected by one of the annotators was, in fact, correct. After this procedure, the corpus with 16,028 entities linked to 2,409 concepts and 1,293 entities linked with no concept (CUI-less) was obtained. Best represented in the annotation UMLS semantic types are *Disease or Syndrome* ($\approx 22\%$), *Body Part, Organ, or Organ Component* (17%), *Organic Chemical* (14.5%), *Finding* (7%), *Sign or Symptom* (6.5%), and *Pathologic Function* (4%). Annotation guidelines were created by an expert with Ph.D. in medicine.

Several annotation challenges are unique to low-resource languages such as Russian. These challenges include 1) the absence of Russian translations for UMLS concepts, 2) the need to combine multiple related concepts into one NER fragment, 3) redundancy in the UMLS vocabulary, and 4) complex rephrasing.

30% of the corpus was reserved for test sets using various filtering strategies. Table 9 shows the statistics for each split.

Stratified. In this case, the set was filtered such that each UMLS concept in the test set appears at least once in the training set but not the specific mention from the test set. As a result, all concepts in the test set are covered in the training set, but none of the mentions in the training set are identical to those in the test set.

Zero-shot. In this case, the set was filtered to contain only novel concepts that do not appear in the training set at all. In other words, the *stratified* test set is designed to ensure that the same concepts appear in the training, development, and test sets but with varying surface forms. The *zero-shot* test set exposes the model to unseen terms and concepts in the development and testing sets, making it more challenging than the stratified test set.

Table 9 — Dataset statistics.

Subset	**# entities**	**# unique entities**	**# concepts**
Full train	12189	5435	2031
In-KB train	11220	4934	2030
Full test	5132	2689	1232
In-KB test	4808	2464	1231
Zero-shot test	434	417	379
Stratified test	1266	1199	576
RWN med. [71]	2319	1666	635
XL-BEL [72]	681	610	510
MCN (English) [73]	13609	5979	3792

CUI-less. The purpose of this test set is to evaluate whether a linking system can avoid linking to a concept when there is no appropriate concept in the vocabulary (referred to as the "CUI-less" category in CLEF/SemEval challenges). The study refers to the subsets that include the CUI-less cases as the "full test set" and "full train set", while subsets without CUI-less mentions are known as "in-KB".

The following ranking models based on several different embeddings were used for comparison: (1) *Tf-idf*: standard sparse *tf-idf* representations constructed on character-level unigrams and bigrams; (2) *BERT*: multilingual BERT embeddings with no fine-tuning [7]; this is a cross-lingual baseline that has not been trained on biomedical texts; (3) *RuBERT*: Russian BERT embeddings [74] trained on the Russian part of Wikipedia and news data; (4) *SapBERT*: a BERT-based metric learning framework that generates hard triplets based on the UMLS for large-scale pre-training [75] and also allows for a cross-lingual variant trained on XL-BEL [72].

Additionally, several variations of fine-tuning on datasets with training sets were used via synonym marginalization as suggested by the authors of *BioSyn* [69]: (1) *SapBERT+RuCCoN*, with fine-tuning on the target train set of EHRs; (2) *SapBERT+MCN*, with tuning on the MCN set; (3) *SapBERT+WRN*, on the dataset extracted from the medical part of the RuWordNet thesaurus; (4) *SapBERT+XL-BEL*, on the the Russian part of XL-BEL; (5) *SapBERT+RuCCoN+RWMXL-BEL*, on the combination of all three sets.

As shown in Tab. 10, SapBERT outperforms other models and steadily improves results as more datasets are included for fine-tuning. SapBERT trained

Table 10 — Evaluation results with test set filtering.

Model	In-KB test		Full test		Stratified test		Zero-shot test	
	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
Tf-Idf	37.58%	46.98%	-	-	25.83%	34.20%	26.27%	41.01%
Multilingual BERT	29.01%	33.74%	29.15%	33.16%	12.32%	16.35%	15.90%	19.35%
RuBERT	25.17%	28.22%	24.05%	25.66%	11.53%	14.53%	13.82%	17.51%
SapBERT	45.84%	56.41%	37.18%	37.47%	30.02%	40.44%	29.49%	40.78%
+MCN	46.51%	56.45%	43.67%	53.23%	30.41%	40.60%	27.88%	41.47%
+WN	45.47%	55.12%	43.30%	50.19%	29.94%	39.42%	29.03%	38.48%
+XL-BEL	47.77%	58.74%	40.80%	42.30%	32.54%	42.97%	29.95%	45.16%
+RuCCoN	59.26%	68.99%	53.39%	60.02%	47.31%	61.45%	32.95%	47.47%
+RuCCoN+RWN	57.84%	68.55%	52.67%	58.79%	47.79%	63.67%	32.49%	46.31%
+RuCCoN+XL-BEL	58.78%	68.05%	53.20%	59.80%	46.52%	59.08%	33.41%	48.85%
+RuCCoN+RWN+XL-B.	58.55%	67.82%	52.65%	59.20%	50.32%	62.48%	33.41%	45.85%

on RuCCoN is notably superior to SapBERT trained on other data when tested on the full test set, but the difference diminishes on the zero-shot test, indicating that it is mostly due to specific entities labeled in the training set. This highlights the importance of labeling additional data to enhance the performance of even the most sophisticated entity linking models, which is facilitated by RuCCoN for the Russian language. Note that fine-tuning on additional medical data is generally beneficial, with SapBERT fine-tuned on English clinical notes consistently outperforming basic SapBERT across all datasets.

In summary, this section introduces RuCCoN, a novel clinical concept normalization dataset in Russian, labeled by medical professionals and accompanied by several train/test splits for fair evaluation in various settings. The dataset and annotation guidelines are available at `https://github.com/AIRI-Institute/RuCCoN`.

3.3 NEREL-BIO: nested named entities in biomedical abstracts

The results of this section are based on the paper [23].

This section introduces NEREL-BIO – an annotation scheme for nested named entity recognition and a corpus of PubMed abstracts in Russian and in English. The choice of biomedical entity types for annotation in NEREL-BIO is determined by their appearance in the UMLS taxonomy and other annotated datasets in the biomedical domain. The NEREL-BIO annotation scheme incorporates 17 specialized biomedical entity types (see Tab. 12 for details) in addition to 20 entity types from the general NEREL dataset [76]. The following NEREL general entity types were incorporated: 8 basic entity types (e.g., PERSON, ORGANIZATION,

Table 11 — Statistics of NEREL-BIO.

Collection	**#Doc**	**#Entities**	**#Non-zero entity types**
Abstracts in Russian	766	66,888	37
Abstracts in English	105	10,651	32

Table 12 — Frequencies of top ten entity types with nested entities in full Russian collection and in parallel documents in English and Russian for comparison.

Entity Type	**Description**	**Full RU, in %**	**EN, in %**
FINDING	conveys the results of scientific study, experiments described	65.7	71.2
PHYS	biological function or process in organism including organism attribute (temperature) and excluding mental processes	38.3	40.7
INJURY POISONING	damage inflicted on the body as the direct or indirect result of external force including poisoning	37.7	49.0
DISO	any deviations from normal state of organism: diseases, symptoms, dysfunctions, abnormality of organ, excluding injuries or poisoning	37.3	41.2
DEVICE	manufactured objects used for medical purposes	33.9	42.5
LABPROC	testing body substances and other diagnostic procedures such as ultrasonography	30.2	34.8
MEDPROC	procedures concerned with remedial treatment of diseases, including surgical procedures	30.0	44.7
ANATOMY	organs, body part, cells and cell components, body substances	27.3	28.3
SCIPROC	scientific studies including mathematical methods or clinical studies, scales, classifiers, etc.	23.9	32.1
CHEM	chemicals including legal and illegal drugs, biological molecules	22.5	20.1

LOCATION), 7 numerical entities (e.g., DATE, AGE), and tags for characterizing persons (NATIONALITY, PROFESSION, and FAMILY), PRODUCT, and EVENT. The EVENT entity is utilized to label events like epidemics, military conflicts, and tsunamis, which are mentioned in connection with the spread of diseases or the need for additional medical care.

Note that entities annotated in NEREL-BIO can be absent in UMLS. For example, the term *left-sided congenital diaphragmatic hernia* is absent in UMLS. This phrase is annotated as follows: $[left -$

sided [*congenital* [[*diaphragmatic*]$_{ANATOMY}$ [*hernia*]$_{DISO}$]$_{DISO}$]$_{DISO}$]$_{DISO}$ Although the whole term can't be linked in UMLS, the sub-terms mapped to: Hernia (C0019270), Diaphragmatic Hernia (C0019284), Respiratory Diaphragm (C0011980), Congenital diaphragmatic hernia (C0235833).

The annotation scheme was developed through multiple rounds of preliminary annotation of parallel Russian and English abstracts. Experienced terminologists with expertise in terminological studies, including the biomedical domain, were responsible for the annotation process. Additionally, a moderator reviewed all annotated abstracts to ensure accuracy.

Table 11 summarizes the statistics of NEREL-BIO in terms of documents and entity mentions. Table 12 summarizes the frequency of nested entities in NEREL-BIO. The table includes the top ten entity types and their corresponding nestedness frequency. To calculate the nestedness frequency, the number of times an entity of a specific type appears as an outer entity (excluding multiple occurrences of the same entity) was divided by the total occurrences of the entity type in the corpus.

To conduct experiments, NEREL-BIO was divided into three subsets: train, dev, and test, with 612, 77, and 77 documents, respectively. A Machine Reading Comprehension (MRC) model [77] was fine-tuned for entity recognition experiments using the train set. As expected, depending on entity type, the performance of the MRC model varies greatly: F1 scores on ANATOMY, CHEM, DISO entities are 83.99%, 81.32%, 81.03%, respectively, while F1 scores on LABPROC, MEDPROC, DISO entities are 66.47%, 73.96%, 60.31%, respectively.

In summary, NEREL-BIO has been introduced, the first dataset of biomedical Russian abstracts annotated with nested entities. The annotation demonstrates that nested entities provide a more effective foundation for extracting relations that would otherwise be lost, as well as facilitating more complete entity linking to knowledge bases. The dataset is available at `https://github.com/nerel-ds/NEREL-BIO`.

4 New evaluation strategies

New evaluation strategies were proposed and developed by the author of this dissertation [13; 24–26]. Linking mentions of biomedical entities like chemicals, diseases, genes, and adverse drug reactions to terminologies is challenging and often requires non-syntactic interpretation. This is due to the complexity and

variability of biomedical language, which can involve a wide range of terms and abbreviations. In particular, the following key scientific problems, addressed in this chapter, are discussed:

- The first scientific problem addressed in [24; 25] is the lack of consistent and reliable evaluation strategies for entity linking/concept normalization. Methods are often evaluated on test sets of widely differing sizes and domains and a narrow subsample of concepts from specific terminologies. Additionally, reported results of neural networks vary substantially on different corpora, resulting in a range of accuracy scores.
- The second scientific problem addressed in [24] is that neural models are typically trained and evaluated on entities of the same type from a single domain. This limits the generalizability of the models and makes it difficult to reuse them for different purposes, as this requires coding to a specific terminology.

To address the problems highlighted above, several evaluation strategies are proposed:

- To tackle the first problem, one proposed strategy is to use a *stratified* sampling split to evaluate the ability of systems to recognize known concepts even with novel mentions [13]. Additionally, a test set *filtering* procedure was introduced to assess the "hard cases" of entity linking and approach zero-shot cross-lingual transfer learning [25].
- To tackle the second problem, another proposed strategy is to use both *in-terminology* and *cross-terminology* evaluations to account for the variety of biomedical entities and terminologies [24].

The key results and conclusions of this section are as follows:

- The evaluation shows a great divergence in performance between official train/test splits and with the proposed *filtered* test sets that represent refined samples of entity mentions [24; 25].
- Supervised models trained on a target domain set demonstrate significantly better performance on *stratified* test sets compared to models trained on other data [13; 22].
- Knowledge transfer can be effective between diseases, chemicals, and genes with a small average drop of accuracy in the performance on sets of scientific abstracts [24].

– The effectiveness of transfer learning varies significantly across different domains. For instance, when applied to datasets with ADRs derived from drug labels and social media, supervised models trained on other corpora exhibit a substantial decline in performance compared to models trained specifically on the target domain [24].

4.1 In-terminology and cross-terminology evaluation

The results of this section are based on the paper [24].

There are no established guidelines for evaluating models on biomedical corpora in different terminology contexts. Models are usually evaluated on narrow subsamples of concepts, and the reported results vary across corpora. Reusing trained models for different terminologies is also difficult with supervised models. To address these issues, this study compares benchmarks and neural architectures using BERT for linking entities across three domains: research abstracts, drug labels, and user-generated texts on drug therapy in English.

This study presents an extensive evaluation of five biomedical corpora manually annotated with concepts regarding diseases, chemicals, human genes, and adverse drug reactions (ADRs). Two models are utilized: (i) a baseline that ranks concepts for a given mention by comparing biomedical BERT vectors [56] with the Euclidean distance; (ii) BioSyn [69]. Models are based on BioBERT$_{base}$ v1.1 that was pre-trained on PubMed abstracts (4.5B words in total) for 1M steps.

For analysis, publicly available benchmarks with official train/dev/test splits were used: NCBI Disease corpus [63], BioCreative V CDR (BC5CDR) [64], BioCreative II GN (BC2GN) [66], TAC 2017 ADR [78], SMM4H 2017 ADR [79]. The analysis of datasets showed that approximately 80% entity mentions in the test set are textual duplicates of other entities in the test set or entities presented in train+dev sets. In order to obtain more realistic results, this study presents *refined* test sets without duplicates or exact overlaps. Note that some concepts appearing in the *refined* test set also appear in the respective training set.

BioSyn was trained on the train/dev set of each corpus with a source dictionary and evaluated on the respective test set (in-domain performance). Cross-domain evaluation includes models trained on *source* data on the test sets of all other corpora (i.e., the *target*). Both BioSyn and BioBERT ranking models retrieve the closest concept name in a target dictionary for a given mention representation during inference. Note that cross-terminology evaluation is a challenging scenario

Table 13 — Single-terminology normalization results in terms of acc@1 on the official and *refined* test sets. CDR is BC5CDR, GN is BC2GN, M4H is SMM4H

Model	**NCBI Disease**		**CDR Dis**		**CDR Chem**		**GN Gene**		**TAC ADR**		**M4H ADR**	
	test	*refined*	test	*refined*	test	*refined*	test	*refined*	test	*refined*	test	*refined*
BioSyn	90.7	72.5	93.5	74.1	96.3	83.8	90.8	85.8	95.6	83.2	83.8	60.5
BioBERT ranking	83.9	47.5	91.3	65.1	94.7	79.3	74.7	68.4	87.8	54.7	33.9	14.3
Difference	-6.8	-25.0	-1.9	-7.7	-1.6	-4.5	-16.1	-17.4	-7.8	-28.5	-49.9	-46.2

Table 14 — Comparison of BioSyn for single- and cross-terminology MCN on *refined* test sets. In-domain results are on the diagonals (with a dark gray background). Other cells contain results of a given model and differences in results between that model and the in-domain model in parentheses (by row). Light gray cells show cross-terminology experiments.

	Train set					
Test set	NCBI Dis	BC5CDR Dis	BC5CDR Chem	BC2GN Gene	TAC ADR	SMM4H ADR
NCBI Disease	72.5	67.6 (-4.9)	64.7 (-7.8)	67.2 (-5.4)	67.6 (-4.9)	48.5 (-24.0)
BC5CDR Dis	74.7 (+0.6)	74.1	73.4 (-0.8)	73.1 (-1.1)	74.9 (+0.8)	58.3 (-15.8)
BC5CDR Chem	82.4 (-1.4)	84.2 (+0.5)	83.8	82.6 (-1.2)	82.4 (-1.4)	73.9 (-9.9)
BC2GN Gene	83.1 (-2.6)	81.7 (-4.1)	83.7 (-2.1)	85.8	82.6 (-3.1)	73.2 (-12.6)
TAC ADR	74.3 (-8.9)	77.5 (-5.7)	70.1 (-13.0)	69.9 (-13.3)	83.2	51.5 (-31.7)
SMM4H ADR	27.3 (-33.2)	35.6 (-24.9)	24.8 (-35.7)	21.9 (-38.6)	30.1 (-30.4)	60.5

for developing supervised models, particularly for linking to concepts that were not encountered during training (i.e., zero-shot concepts).

The task of finding the top-k concepts for every entity mention in a text is evaluated in an IR scenario, where a dictionary of concept names and their identifiers is used. The accuracy at k is defined as 1 if the correct identifier is retrieved at rank k, otherwise 0. For composite entities, the accuracy at k is defined as 1 if every prediction for a single mention is correct.

Table 13 presents the results of the models trained and evaluated on entities of the same type from a single domain in six sets. Table 14 compares the performance of BioSyn in single- and cross-terminology normalization tasks. The models were trained on the training set from a source dataset and evaluated on the target test set with different terminology.

In order to determine whether current benchmark test sets may be leading to an overestimation of performance, the results obtained by models on both official

and *refined* test sets were compared, as shown in Table 13. The significant decrease of averaged acc@1 from 91.8% to 76.7% for BioSyn and averaged acc@1 from 77.7% to 54.9% for BioBERT ranking highlights the great need for external evaluation datasets, where the same entity mentions will not be used for both training and testing. These observations also mean that there is room for improvement in the transferability of developed methods, that is, the ability to maintain performance for entirely unseen domains or entities.

Table 13 provides insights that help answer the question of how surface characteristics of entity mentions impacting the performance of the BERT-based baseline. Based on these results, the following conclusions can be drawn. First, the simple ranking of BioBERT representations achieves strong results on CDR Disease and Chemical sets. On two refined sets with larger mentions (NCBI, TAC) and the BC2GN corpus with mentions containing numerals, the difference between BioBERT ranking and BioSyn is significant (average decrease of 23.6%). The qualitative analysis uncovered that BERT representations of mentions differing by one numeral (e.g., genes TP53 and TP63) are close in the latent space. As expected, results on SMM4H are significantly lower than on abstracts due to the gap between the language of lay public and medical professionals.

In order to determine whether a model trained on one corpus can be used for the linking of entity mentions in another type or domain in the zero-shot setting, performance differences are compared in Tables 13 and 14. The models trained on NCBI, CDR Disease, BC2GN, and TAC data perform on par with the model trained on the CDR Chemical train set (approx. 74% acc@1), while the model trained on CDR Chemical showed a 6% drop on these subsets. BioSyn trained on SMM4H achieves lower results on abstracts and drug labels than simple BioBERT ranking, while all supervised models performed better on SMM4H data than the BioBERT ranking.

To conclude, this study presents a comprehensive comparative evaluation of medical concept normalization datasets, including NCBI Disease, BC5CDR Disease & Chemical, BC2GN Gene, TAC 2017 ADR, and SMM4H 2017 ADR corpora. Two BERT-based models across six datasets were evaluated, with official train/test splits and *refined* test sets representing entity mentions. The evaluation revealed significant differences in performance, indicating that the state-of-the-art model BioSyn achieved up to 15% lower accuracy on the refined test set. The cross-terminology MCN task was introduced, demonstrating effective knowledge transfer between diseases, chemicals, and genes. However, models trained on four other

corpora performed poorly on TAC and SMM4H sets, with accuracy dropping by 10.2% and 33.1%, respectively. Results and source code are available on GitHub at `https://github.com/insilicomedicine/Fair-Evaluation-BERT`.

5 Conclusion

The main results of this dissertation are based on the following published papers [11–26].

The studies [21–23] focused on the development of annotation schemes for biomedical information extraction tasks, as well as the creation of annotated corpora in both English and Russian for a range of biomedical sources, including scientific abstracts, drug reviews, electronic health records, and clinical trials. Through a series of experiments, baselines for these corpora were established.

In papers [24–26], limitations of existing benchmarks in two tasks were analyzed to propose solutions for the improvement of evaluation strategies. The study [26] focused on the limitations of existing benchmarks for relation extraction in scientific abstracts and electronic health records. The work proposes a cross-attention neural model that shows better cross-domain performance. In [24], limitations of existing benchmarks for biomedical entity linking are analyzed, and novel evaluation strategies are proposed for in-terminology and cross-terminology evaluation.

The studies [11–18] proposed several neural architectures for different tasks in the biomedical domain. These include DILBERT, which optimizes the similarity of mentions and concepts via triplet loss, multilingual BERT-based models for named entity recognition, a classification-based approach to biomedical entity linking, a multimodal model for adverse drug reaction detection, a sequence-to-sequence learning framework for ICD coding, and a feature-based model for clinical relation extraction. These models and methods have demonstrated their effectiveness on several information extraction benchmarks and shared tasks.

In papers [19; 20], developed models were combined into an IE pipeline for a biomedical search over abstracts and for mining ADEs from user comments about drugs. The experiments on zero-shot retrieval described in [19] showed the neural IE architecture shows superior performance for both disease and gene concept queries. The experiments described in [20] showed that mining ADEs from Twitter posts using a pipeline architecture requires the different components to be trained based on input data imbalance to ensure optimal performance on the end-to-end resolution task.

The main results submitted for defense are as follows:

– New models and methods for classification and information extraction were developed:

1. Multilingual BERT-based models were analyzed for cross-domain drug and disease named entity recognition in two languages. The investigation of transfer learning strategies between four corpora demonstrated the effectiveness of pretraining on data with one or both types of transfer [11].
2. Classification-based methods were proposed with (i) a set of informative features at an entity level and a context level for relation extraction [12], and (ii) vectors of semantic similarity for entity linking [13;14]. The effectiveness of these approaches was demonstrated in multiple shared tasks, ranking first in SMM4H 2019 Task 3, SMM4H 2020 Task 3, and SMM4H 2021 Task 1c [13; 14]. The semantic similarity vectors also proved effective with a proposed encoder-decoder architecture that ranked first in CLEF eHealth 2017 Task 1 [15].
3. DILBERT (Drug and disease Interpretation Learning with Biomedical Entity Representation Transformer) was introduced. The model optimizes the relative similarity of mentions and concept names from a terminology via metric learning. It was shown that the model is robust to vocabulary switches and can recognize concepts that were not present in the training set [16; 17].
4. A multimodal model combining BERT-based models for language understanding and molecular property prediction was proposed to improve the classification of tweets as potential sources of adverse drug events or drug reactions. The model achieved first and second place rankings on SMM4H 2021 Task 2 and Task 1a, respectively [18].
5. Two neural pipelines were developed: (i) a pipeline consisting of two models as a biomedical search engine that showed superior performance over a traditional search model on a manually annotated dataset of abstracts for disease and gene queries [19], and (ii) a pipeline for the classification, extraction and normalization of adverse drug events on realistic, imbalanced data. The identification of optimal training ratios and undersampling methods was also explored [20].

– New annotated corpora were developed for information extraction. The following are some of the new corpora developed:

6. The Russian Drug Reaction Corpus (RuDReC), a partially annotated corpus of consumer reviews in Russian about pharmaceutical products, and RuDR-BERT models for named entity recognition and sentence classification tasks were introduced [21].
7. Two annotated datasets were developed for clinical concept normalization: a dataset of clinical trials in English for drug and disease normalization [16; 17], and a RuCCoN corpus, a new dataset of electronic health records in Russian, with entities linked to the UMLS [22].
8. NEREL-BIO, an annotation scheme and corpus of PubMed abstracts in Russian and English with general-domain and biomedical entity types, was introduced. The corpus includes the provision of an annotation for nested named entities [23].

– New evaluation strategies were proposed, as follows:

9. The limitations of existing benchmarks for biomedical entity linking were analyzed, and several novel evaluation strategies were proposed: (i) a novel *stratified* sampling split [13], (ii) *in-terminology* and *cross-terminology* evaluation [24]. Additionally, benchmarks were established for the cross-lingual task using clinical reports, clinical guidelines, and medical research papers. A test set *filtering* procedure was designed to analyze the "hard cases" of entity linking approaching zero-shot cross-lingual transfer learning [25].
10. The limitations of existing benchmarks of scientific abstracts and electronic health records for relation extraction were analyzed. To address performance differences in *in-domain* and *out-of-domain* setup, a cross-attention neural model was proposed that exhibits better cross-domain performance [26].

Acknowledgments

The author of this thesis leaded projects on biomedical NLP supported by the Russian Science Foundation grant no. 18-11-00284 (2018-2020, 2021-2022), by the Russian Foundation for Basic Research grant no. 19-07-01115 (2019-2020), by a grant from the President of the Russian Federation for young scientists-candidates of science (MK-3193.2021.1.6, 2021-2022). This research is supported by these grants, and the Russian Science Foundation grant no. 20-11-20166 (2020-2022).

Bibliography

1. *Bodenreider Olivier.* The unified medical language system (UMLS): integrating biomedical terminology // *Nucleic acids research.* — 2004. — Vol. 32, no. suppl_1. — Pp. D267–D270.

2. *Brown Elliot G, Wood Louise, Wood Sue.* The medical dictionary for regulatory activities (MedDRA) // *Drug safety.* — 1999. — Vol. 20, no. 2. — Pp. 109–117.

3. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results / Arden W Forrey, Clement J Mcdonald, Georges DeMoor et al. // *Clinical chemistry.* — 1996. — Vol. 42, no. 1. — Pp. 81–90.

4. *Coletti Margaret H, Bleich Howard L.* Medical subject headings used to search the biomedical literature // *Journal of the American Medical Informatics Association.* — 2001. — Vol. 8, no. 4. — Pp. 317–323.

5. NIH UMLS Statistics. — 2022. — URL: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html.

6. Attention is all you need / Ashish Vaswani, Noam Shazeer, Niki Parmar et al. // Advances in neural information processing systems. — 2017. — Pp. 5998–6008.

7. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, USA. — 2019. — Pp. 4171–4186.

8. Learning deep structured semantic models for web search using clickthrough data / Po-Sen Huang, Xiaodong He, Jianfeng Gao et al. // Proceedings of the 22nd ACM international conference on Information & Knowledge Management, San Francisco, USA. — 2013. — Pp. 2333–2338.

9. *Schroff Florian, Kalenichenko Dmitry, Philbin James.* Facenet: A unified embedding for face recognition and clustering // Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, USA. — 2015. — Pp. 815–823.

10. *Hoffer Elad, Ailon Nir.* Deep metric learning using triplet network // International Workshop on Similarity-Based Pattern Recognition, Copenhagen, Denmark / Springer. — 2015. — Pp. 84–92.

11. *Miftahutdinov Z., Alimova I., Tutubalina E.* On biomedical named entity recognition: Experiments in interlingual transfer for clinical and social media texts // *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* — 2020. — Vol. 12036 LNCS. — Pp. 281–288. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85084182140&doi=10.1007%2f978-3-030-45442-5_35&partnerID=40&md5=3d546446eab3f7a96da1059035620aca.

12. *Alimova I., Tutubalina E.* Multiple features for clinical relation extraction: A machine learning approach // *Journal of Biomedical Informatics.* — 2020. — Vol. 103. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85079558624&doi=10.1016%2fj.jbi.2020.103382&partnerID=40&md5=f4c92a675a4d6fa6bd4074024ea0467c.

13. Medical concept normalization in social media posts with recurrent neural networks / E. Tutubalina, Z. Miftahutdinov, S. Nikolenko, V. Malykh // *Journal of Biomedical Informatics.* — 2018. — Vol. 84. — Pp. 93–102. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85049314992&doi=10.1016%2fj.jbi.2018.06.006&partnerID=40&md5=f80bf052106ba962aedd6168d04e5b59.

14. *Miftahutdinov Z., Tutubalina E.* Deep neural models for medical concept normalization in user-generated texts // *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop.* — 2019. — Pp. 393–399. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85083989368&partnerID=40&md5=f0dc0f363ab0562cd5c078af405d5fcb.

15. *Miftahutdinov Z., Tutubalina E.* Deep learning for ICD coding: Looking for medical concepts in clinical documents in english and in French // *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* — 2018. — Vol. 11018 LNCS. — Pp. 203–215. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85052834179&doi=10.1007%2f978-3-319-98932-7_19&partnerID=40&md5=61ce8a6f8e1252c00be5ba74ef5ac436.

16. Drug and Disease Interpretation Learning with Biomedical Entity Representation Transformer / Z. Miftahutdinov, A. Kadurin, R. Kudrin, E. Tutubalina // *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. — 2021. — Vol. 12656 LNCS. — Pp. 451–466. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85107366839&doi=10.1007%2f978-3-030-72113-8_30&partnerID=40&md5=355554291932b138f5f3fcc54e773d36.

17. Medical concept normalization in clinical trials with drug and disease representation learning / Z. Miftahutdinov, A. Kadurin, R. Kudrin, E. Tutubalina // *Bioinformatics*. — 2021. — Vol. 37, no. 21. — Pp. 3856–3864. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85121339307&doi=10.1093%2fbioinformatics%2fbtab474&partnerID=40&md5=3196dcb211542f6e0ffbba60c90cbce4.

18. *Sakhovskiy A., Tutubalina E.* Multimodal model with text and drug embeddings for adverse drug reaction classification // *Journal of Biomedical Informatics*. — 2022. — Vol. 135. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85139318188&doi=10.1016%2fj.jbi.2022.104182&partnerID=40&md5=b14c26ddb7c2c02291e5da87f1b09dcf.

19. A Comprehensive Evaluation of Biomedical Entity-centric Search / Elena Tutubalina, Zulfat Miftahutdinov, Vladimir Muravlev, Anastasia Shneyderman // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP): Industry Track. — Abu Dhabi, UAE: Association for Computational Linguistics, 2022. — . — Pp. 596–605. — URL: https://aclanthology.org/2022.emnlp-industry.61.

20. DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter / A. Magge, E. Tutubalina, Z. Miftahutdinov et al. // *Journal of the American Medical Informatics Association*. — 2021. — Vol. 28, no. 10. — Pp. 2184–2192. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85116958957&doi=10.1093%2fjamia%2focab114&partnerID=40&md5=a5baad71eb80b4933dbc8141b80e5f85.

21. The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews / E. Tutubalina, I. Alimova, Z. Miftahutdinov et al. // *Bioinformatics*. — 2021. — Vol. 37, no. 2. —

Pp. 243–249. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85099813248&doi=10.1093%2fbioinformatics%2fbtaa675&partnerID=40&md5=45058e7cbc73265e95868d8384b45518.

22. RuCCoN: Clinical Concept Normalization in Russian / A. Nesterov, G. Zubkova, Z. Miftahutdinov et al. // *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. — 2022. — Pp. 239–245. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85144355127&partnerID=40&md5=54e0497a8eaafd3bcbdf53803ea3a93f.

23. NEREL-BIO: A Dataset of Biomedical Abstracts Annotated with Nested Named Entities / Natalia Loukachevitch, Suresh Manandhar, Elina Baral et al. // *Bioinformatics*. — 2023. — 04. — btad161. URL: https://doi.org/10.1093/bioinformatics/btad161.

24. *Tutubalina E., Kadurin A., Miftahutdinov Z.* Fair Evaluation in Concept Normalization: a Large-scale Comparative Analysis for BERT-based Models // *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*. — 2020. — Pp. 6710–6716. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85143411978&partnerID=40&md5=39c4d2f361c4d49247615b9e3ad7531f.

25. Medical Crossing: a Cross-lingual Evaluation of Clinical Entity Linking / A. Alekseev, Z. Miftahutdinov, E. Tutubalina et al. // *2022 Language Resources and Evaluation Conference, LREC 2022*. — 2022. — Pp. 4212–4220. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85144412484&partnerID=40&md5=62eb1911dd0d8cdd3f010887f44607a5.

26. *Alimova I., Tutubalina E., Nikolenko S.I.* Cross-Domain Limitations of Neural Models on Biomedical Relation Classification // *IEEE Access*. — 2022. — Vol. 10. — Pp. 1432–1439. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85121795127&doi=10.1109%2fACCESS.2021.3135381&partnerID=40&md5=0ad8a6a13c88b80d18b5a4dd8ca0bf1a.

27. *Tutubalina E., Nikolenko S.* Automated prediction of demographic information from medical user reviews // *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. — 2017. — Vol. 10089 LNAI. — Pp. 174–184. — URL: https://www.scopus.com/inward/record.uri?eid=

2-s2.0-85018433898&doi=10.1007%2f978-3-319-58130-9_17&partnerID=40&md5=aa049c3dd5e26a00ccae2d950d926a50.

28. *Tutubalina E., Nikolenko S.* Demographic prediction based on user reviews about medications // *Computacion y Sistemas.* — 2017. — Vol. 21, no. 2. — Pp. 227–241. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85021791555&doi=10.13053%2fCyS-21-2-2736&partnerID=40&md5=edcbaffa51097a5998b0be781c719fa5.

29. *Miftahutdinov Z.Sh., Tutubalina E.V., Tropsha A.E.* Identifying disease-related expressions in reviews using conditional random fields // *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii.* — 2017. — Vol. 1, no. 16. — Pp. 155–166. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85021794913&partnerID=40&md5=046ee34420ea35a0b1a64d399cfb6d9d.

30. *Tutubalina E., Nikolenko S.* Combination of Deep Recurrent Neural Networks and Conditional Random Fields for Extracting Adverse Drug Reactions from User Reviews // *Journal of Healthcare Engineering.* — 2017. — Vol. 2017. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85029668695&doi=10.1155%2f2017%2f9451342&partnerID=40&md5=756985a3cb41cf9897dac67d4454c1e0.

31. *Miftahutdinov Z., Tutubalina E.* KFU at CLEF eHealth 2017 Task 1: ICD-10 coding of English death certificates with recurrent neural networks // *CEUR Workshop Proceedings.* — 2017. — Vol. 1866. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85034788154&partnerID=40&md5=78facae035f322db6f0b19e3b5dab366.

32. Using semantic analysis of texts for the identification of drugs with similar therapeutic effects / E.V. Tutubalina, Z.S. Miftahutdinov, R.I. Nugmanov et al. // *Russian Chemical Bulletin.* — 2017. — Vol. 66, no. 11. — Pp. 2180–2189. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85043979586&doi=10.1007%2fs11172-017-2000-8&partnerID=40&md5=df4a2d92399cd902b67829cf69126371.

33. *Miftahutdinov Z., Tutubalina E.* Leveraging deep neural networks and semantic similarity measures for medical concept normalisation in user reviews // *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii.* — 2018. — no. 17. — Pp. 490–500. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85058005600&partnerID=40&md5=79de73aee59fe4364404964585277152.

34. *Miftahutdinov Z., Tutubalina E.* End-to-end deep framework for disease named entity recognition using social media data // *IEEE 30th Jubilee Neumann Colloquium, NC 2017.* — 2018. — Vol. 2018-January. — Pp. 47–52. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85047745771&doi=10.1109%2fNC.2017.8263281&partnerID=40&md5=717a2eee2dcc974aa0463cc29c554b52.

35. A machine learning approach to classification of drug reviews in Russian / I. Alimova, E. Tutubalina, J. Alferova, G. Gafiyatullina // *Proceedings - 2017 Ivannikov ISPRAS Open Conference, ISPRAS 2017.* — 2018. — Vol. 2018-January. — Pp. 64–69. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85050889586&doi=10.1109%2fISPRAS.2017.00018&partnerID=40&md5=096a8bc585b9244b3f3e41229dcecf5d.

36. *Tutubalina E., Nikolenko S.* Exploring convolutional neural networks and topic models for user profiling from drug reviews // *Multimedia Tools and Applications.* — 2018. — Vol. 77, no. 4. — Pp. 4791–4809. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85033379716&doi=10.1007%2fs11042-017-5336-z&partnerID=40&md5=b77cb2c2952844be223042bf35932f49.

37. *Alimova I., Tutubalina E.* A comparative study on feature selection in relation extraction from electronic health records // *CEUR Workshop Proceedings.* — 2019. — Vol. 2523. — Pp. 34–45. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85077496207&partnerID=40&md5=a41856a3e2a0525e35a23f96b962140b.

38. *Tutubalina E., Alimova I., Solovyev V.* Biomedical entities impact on rating prediction for psychiatric drugs // *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* — 2019. — Vol. 11832 LNCS. — Pp. 97–104. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85077506911&doi=10.1007%2f978-3-030-37334-4_9&partnerID=40&md5=c87ed7952a76208f990dbe0d21a72f6b.

39. *Alimova I., Tutubalina E.* Comparative analysis of context representation models in the relation extraction task from biomedical texts // *CEUR Workshop Proceedings.* — 2019. — Vol. 2525. —

URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85077611413&partnerID=40&md5=43f89b6daf69a3d77c7a52b7bb4953a7.

40. *Alimova I., Tutubalina E.* Detecting adverse drug reactions from biomedical texts with neural networks // *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop*. — 2019. — Pp. 415–421. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85083960585&partnerID=40&md5=ef77e7e92e54a96bdce2d72ed3dddb20.

41. *Nugmanov Ramil, Miftahutdinov Zulfat, Tutubalina Elena.* Addressing medical coding of free-text clinical records in English with deep learning // *European Journal of Clinical Investigation*. — 2019.

42. *Alimova I.S., Tutubalina E.V.* Entity-Level Classification of Adverse Drug Reaction: A Comparative Analysis of Neural Network Models // *Programming and Computer Software*. — 2019. — Vol. 45, no. 8. — Pp. 439–447. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85077861472&doi=10.1134%2fS0361768819080024&partnerID=40&md5=5e5c77ebc92c42aedd16a14214bdfcce.

43. *Alimova I., Tutubalina E.* Selection of Pseudo-Annotated Data for Adverse Drug Reaction Classification Across Drug Groups // *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. — 2022. — Vol. 13217 LNCS. — Pp. 37–44. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85142714539&doi=10.1007%2f978-3-031-16500-9_4&partnerID=40&md5=f0a2b84f3acf6ab56a0e8ad9706a5ea4.

44. A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research—An International Collaboration / Juan M. Banda, Ramya Tekumalla, Guanyu Wang et al. // *Epidemiologia*. — 2021. — Vol. 2, no. 3. — Pp. 315–324. — URL: https://www.mdpi.com/2673-3986/2/3/24.

45. *Miftahutdinov Zulfat, Alimova Ilseyar, Tutubalina Elena.* KFU NLP Team at SMM4H 2019 Tasks: Want to Extract Adverse Drugs Reactions from Tweets? BERT to The Rescue // Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task. — Florence, Italy: Association for Computational Linguistics, 2019. — . — Pp. 52–57. — URL: https://aclanthology.org/W19-3207.

46. *Miftahutdinov Zulfat, Sakhovskiy Andrey, Tutubalina Elena.* KFU NLP Team at SMM4H 2020 Tasks: Cross-lingual Transfer Learning with Pretrained Language Models for Drug Reactions // Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task. — Barcelona, Spain (Online): Association for Computational Linguistics, 2020. — . — Pp. 51–56. — URL: https://aclanthology.org/2020.smm4h-1.8.

47. *Sakhovskiy A., Miftahutdinov Z., Tutubalina E.* KFU NLP Team at SMM4H 2021 Tasks: Cross-lingual and Cross-modal BERT-based Models for Adverse Drug Effects // *Social Media Mining for Health, SMM4H 2021 - Proceedings of the 6th Workshop and Shared Tasks.* — 2021. — Pp. 39–43. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85122180382&partnerID=40&md5=5e4b7f72a179f12901b6bd7ee980c9c9.

48. Overview of the Fifth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at COLING 2020 / Ari Klein, Ilseyar Alimova, Ivan Flores et al. // Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task. — Barcelona, Spain (Online): Association for Computational Linguistics, 2020. — . — Pp. 27–36. — URL: https://aclanthology.org/2020.smm4h-1.4.

49. Overview of the Sixth Social Media Mining for Health Applications (SMM4H) Shared Tasks at NAACL 2021 / A. Magge, A.Z. Klein, A. Miranda-Escalada et al. // *Social Media Mining for Health, SMM4H 2021 - Proceedings of the 6th Workshop and Shared Tasks.* — 2021. — Pp. 21–32. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85129550090&partnerID=40&md5=3e356287e5bc8e1aa41ee8a47fb8cbaf.

50. Overview of the Seventh Social Media Mining for Health Applications (#SMM4H) Shared Tasks at COLING 2022 / Davy Weissenbacher, Juan Banda, Vera Davydova et al. // Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task. — Gyeongju, Republic of Korea: Association for Computational Linguistics, 2022. — . — Pp. 221–241. — URL: https://aclanthology.org/2022.smm4h-1.54.

51. *Davydova Vera, Tutubalina Elena.* SMM4H 2022 Task 2: Dataset for stance and premise detection in tweets about health mandates related to COVID-19 // Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task. — Gyeongju, Republic

of Korea: Association for Computational Linguistics, 2022. — . — Pp. 216–220. — URL: https://aclanthology.org/2022.smm4h-1.53.

52. *Sakhovskiy A.S. Tutubalina E.V.* Cross-lingual transfer learning in drug-related information extraction from user-generated texts // *Proceedings of the Institute for System Programming of the RAS (Proceedings of ISP RAS) (In Russ.).* — 2022. — Vol. 33, no. 6. — Pp. 217–228. — URL: https://doi.org/10.15514/ISPRAS-2021-33(6)-15.

53. Cadec: A corpus of adverse drug event annotations / Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, Chen Wang // *Journal of biomedical informatics.* — 2015. — Vol. 55. — Pp. 73–81.

54. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records / Sam Henry, Kevin Buchan, Michele Filannino et al. // *Journal of the American Medical Informatics Association.* — 2019.

55. *Giorgi JM, Bader GD.* Towards reliable named entity recognition in the biomedical domain. // *Bioinformatics (Oxford, England).* — 2019.

56. Biobert: pre-trained biomedical language representation model for biomedical text mining / Jinhyuk Lee, Wonjin Yoon, Sungdong Kim et al. // *Bioinformatics.* — 2019. — 09.

57. Active Learning with Partial Feedback / Peiyun Hu, Zachary C Lipton, Anima Anandkumar, Deva Ramanan // International Conference on Learning Representations. — 2018.

58. DrugBank: a knowledgebase for drugs, drug actions and drug targets / David S Wishart, Craig Knox, An Chi Guo et al. // *Nucleic acids research.* — 2008. — Vol. 36, no. suppl_1. — Pp. D901–D906.

59. Mordred: a molecular descriptor calculator / Hirotomo Moriwaki, Yu-Shi Tian, Norihito Kawashita, Tatsuya Takagi // *Journal of cheminformatics.* — 2018. — Vol. 10, no. 1. — Pp. 1–14.

60. Molecular representation learning with language models and domain-relevant auxiliary tasks / Benedek Fabian, Thomas Edlich, Héléna Gaspar et al. // *arXiv preprint arXiv:2011.13230.* — 2020.

61. *Chithrananda Seyone, Grand Gabe, Ramsundar Bharath.* ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction // *arXiv preprint arXiv:2010.09885.* — 2020.

62. *Hendrycks Dan, Gimpel Kevin.* Gaussian error linear units (gelus) // *arXiv preprint arXiv:1606.08415.* — 2016.

63. *Doğan Rezarta Islamaj, Leaman Robert, Lu Zhiyong.* NCBI disease corpus: a resource for disease name recognition and concept normalization // *Journal of biomedical informatics.* — 2014. — Vol. 47. — Pp. 1–10.

64. BioCreative V CDR task corpus: a resource for chemical disease relation extraction / Jiao Li, Yueping Sun, Robin J Johnson et al. // *Database.* — 2016. — Vol. 2016.

65. Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of drug-gene/protein relations / Antonio Miranda, Farrokh Mehryary, Jouni Luoma et al. // Proceedings of the seventh BioCreative challenge evaluation workshop. — 2021.

66. Overview of BioCreative II gene normalization / Alexander A Morgan, Zhiyong Lu, Xinglong Wang et al. // *Genome biology.* — 2008. — Vol. 9, no. S2. — P. S3.

67. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring / Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, Jason Weston // *CoRR abs/1905.01969. External Links: Link Cited by.* — 2019. — Vol. 2. — Pp. 2–2.

68. Distributed representations of words and phrases and their compositionality / Tomas Mikolov, Ilya Sutskever, Kai Chen et al. // Advances in neural information processing systems, Lake Tahoe, USA. — 2013. — Pp. 3111–3119.

69. Biomedical Entity Representations with Synonym Marginalization / Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, Jaewoo Kang // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Seattle, USA. — 2020. — Pp. 3641–3650.

70. *Shelmanov AO, Smirnov IV, Vishneva EA.* Information extraction from clinical texts in Russian // Computational Linguistics and Intellectual Technologies. — 2015. — Pp. 560–572.

71. Creating Russian wordnet by conversion / Natalia V Loukachevitch, German Lashevich, Anastasia A Gerasimova et al. // Computational Linguistics and Intellectual Technologies: papers from the Annual conference Dialogue. — 2016. — Pp. 405–415.

72. Learning Domain-Specialised Representations for Cross-Lingual Biomedical Entity Linking / Fangyu Liu, Ivan Vulic, Anna Korhonen, Nigel Collier // ACL/IJCNLP (2). — 2021. — Pp. 565–574. — URL: https://doi.org/10.18653/v1/2021.acl-short.72.

73. *Luo Yen-Fu, Sun Weiyi, Rumshisky Anna.* MCN: A comprehensive corpus for medical concept normalization // *Journal of biomedical informatics.* — 2019. — Vol. 92. — P. 103132.

74. *Kuratov Y, Arkhipov M.* Adaptation of deep bidirectional multilingual transformers for Russian language // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. — 2019. — Pp. 333–339.

75. Self-Alignment Pretraining for Biomedical Entity Representations / Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng et al. // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2021. — . — Pp. 4228–4238.

76. NEREL: A Russian Dataset with Nested Named Entities, Relations and Events / N. Loukachevitch, E. Artemova, T. Batura et al. // *International Conference Recent Advances in Natural Language Processing, RANLP.* — 2021. — Pp. 876–885. — URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85123612546&doi=10.26615%2f978-954-452-072-4_100&partnerID=40&md5=572fee1b89296afaa54fd184e05b617d.

77. A Unified MRC Framework for Named Entity Recognition / Xiaoya Li, Jingrong Feng, Yuxian Meng et al. // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — 2020. — Pp. 5849–5859.

78. *Roberts Kirk, Demner-Fushman Dina, Tonning Joseph M.* Overview of the TAC 2017 Adverse Reaction Extraction from Drug Labels Track. // TAC. — 2017.

79. Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task / Abeed Sarker, Maksim Belousov, Jasper Friedrichs et al. // *Journal of the American Medical Informatics Association.* — 2018. — Vol. 25, no. 10. — Pp. 1274–1283.