Федеральное государственное автономное
образовательное учреждение высшего образования
«Казанский (Приволжский) федеральный университет»

На правах рукописи

Тутубалина Елена Викторовна

МОДЕЛИ И МЕТОДЫ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ НЕСТРУКТУРИРОВАННЫХ ДАННЫХ В БИОМЕДИЦИНСКОЙ ОБЛАСТИ

РЕЗЮМЕ ДИССЕРТАЦИИ
на соискание учёной степени
доктора компьютерных наук

Казань — 2023

Диссертационная работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Казанский (Приволжский) федеральный университет».

1 Введение

Тема диссертации

Актуальность темы. Данная диссертация представляет собой комплексное исследование, направленное на улучшение эффективности моделей и методов обработки неструктурированных данных в биомедицинской области. В рамках исследования представлены новые подходы к классификации и извлечению информации, которые включают в себя распознавание различных упоминаний сущностей, таких как лекарства, заболевания, гены, нежелательные реакции на лекарства, а также связывание сущностей (также известное как нормализация медицинских концептов) и извлечение отношений.

Методы и модели обработки естественного языка (Natural Language Processing, NLP) в биомедицинской области сталкиваются со множеством проблем из-за сложности биомедицинского языка и огромного объема генерируемых данных. Биомедицинские данные поступают из различных источников, таких как электронные медицинские записи, научные публикации, данные клинических испытаний, сообщения о лечении в сети Интернет, созданные пользователями. Тексты могут иметь разные форматы, структуры и уровни качества. Можно выделить следующие проблемы специализированной автоматической обработки текстов и биомедицинского поиска. Во-первых, биомедицинский язык часто является многозначным, и многие термины имеют несколько значений. Например, термин “гипертрофия аденоидов” может быть связан с “гипертрофия глоточных миндалин (аденоиды)” или “гипертрофия исключительно аденоидов”, которые представляют два уникальных идентификатора концентов (CUI) в базе медицинских знаний Unified Medical Language System (UMLS) [1]. Некоторые понятия имеют различные идентификаторы в UMLS, несмотря на то что они являются синонимами по своему значению. Например, термин “ахоличный стул” имеет код C2675627, а “светлый стул” - код C0232720. Во-вторых, медицинский язык включает специфические термины и выражения, которые часто не используются в повседневной речи. В контексте данной проблемы NLP модель должна быть способна переводить язык обычных людей в формальный медицинский язык. Например, фраза “всю ночь не могу уснуть” должна быть переведена как “бессонница”, а “немного кружится голова” - как “головокружение”. Это требует более сложных методов, чем простое сопоставление естественных языковых выражений и элементов

словаря, так как методы поиска по словарю могут быть неэффективны при связывании языка из сообщений пользователей о лечении с медицинскими концептами. В-третьих, медицинская терминология обширна и постоянно развивается, с использованием различных онтологий в разных странах и даже в различных медицинских специальностях. Медицинские концепты могут иметь различные типы (например, лекарства, заболевания или гены/белки) и могут быть извлечены из различных онтологий, ограниченных одним типом. Главная цель современного биомедицинского NLP - эффективно идентифицировать и сопоставлять концепты в разных онтологиях без повторного обучения моделей. В-четвертых, аннотированные наборы данных для биомедицинского NLP часто отсутствуют или ограничены, что затрудняет эффективное обучение и оценку моделей. Несмотря на то, что в общей области имеется большое количество ресурсов, многие языки не достигли значительного прогресса в биомедицинской области. Русский язык - один из примеров такого языка; он входит в десятку самых распространенных языков в мире и имеет множество наборов данных и лингвистических ресурсов общей тематики, но биомедицинские ресурсы на русском языке недостаточны развиты. Русская версия UMLS включает переводы MedDRA [2], LOINC [3] и MeSH [4]. Однако русская версия составляет всего 1,8% от английской версии UMLS по словарному запасу и 1,36% по количеству источников [5]. Решение этих проблем требует разработки новых аннотированных корпусов, NLP методов и моделей, способных учитывать сложность и вариативность медицинского языка, а также создания методологии для оценки качества. В диссертации эти проблемы решаются путем создания новых аннотированных корпусов текстов из различных источников, создания новых стратегий оценки и разработки новых моделей на основе архитектуры Трансформер [6], BERT [7] и подходов на основе метрического обучения [8–10]. Для оценки качества предложенных методов и моделей проводились обширные эксперименты на биомедицинских корпусах текстов из различных источников. В результате экспериментов было показано, что предложенные методы и модели классификации и извлечения информации позволяют значительно улучшить качество извлечения, а также качество информационного поиска заболеваний и генов.

Целями и задачами исследования, проведенного в диссертации, являются:

1. Разработка методов и моделей решения различных задач автоматической обработки текстов в специализированной области на основе глубоких нейронных сетей, предобученных моделей и подходов на основе метрического обучения;
2. Анализ ограничений и разработка новых стратегий оценки обученных моделей в задачах информационного поиска и извлечения информации;
3. Создание новых аннотированных корпусов текстов из различных источников, таких как научные абстракты, отзывы о препаратах, электронные медицинские записи и клинические исследования, как на английском, так и на русском языках.

Конечной целью исследований является повышение эффективности информационного поиска на основе разработанных методов и моделей в области биомедицины, систем фармакологического надзора, управления и анализа медицинских записей в сфере здравоохранения.

Основные результаты

Основные положения, выносимые на защиту:

– Разработаны новые модели и методы классификации и извлечения информации:

1. Многоязычные модели архитектуры BERT были проанализированы для кросс-доменного распознавания сущностей лекарств и болезней на двух языках. Исследование стратегий переноса обучения между четырьмя корпусами показало эффективность предварительного обучения на данных с одним или обоими типами переноса [11].
2. Методы, основанные на классификационном подходе, с (i) с набором информативных признаков на уровне сущностей и контекста для извлечения отношений [12], (ii) с признаки семантической близости для связывания именованных сущностей [13; 14]. Эффективность этих подходов была продемонстрирована в рамках соревнований SMM4H 2019 Task 3, SMM4H 2020 Task 3 и SMM4H 2021 Task 1c [13; 14], показав наилучшие результаты. Признаки семантической близости также оказались эффективными в глубоких нейронных сетях архитектуры “кодировщик–декодировщик”;

предложенная модель показывала наилучшие результаты в рамках соревнования CLEF eHealth 2017 Task 1 [15].

3. Модель DILBERT (Drug and disease Interpretation Learning with Biomedical Entity Representation Transformer) для связывания именованных сущностей, оптимизирующая относительную схожесть сущностей и концептов на основе метрического обучения. Модель устойчива к изменениям в словаре и способна распознавать концепты, не присутствовавшие в обучающей выборке [16; 17].
4. Мультимодальная модель на основе представлений двух моделей архитектуры BERT для языкового моделирования и предсказания молекулярных свойств в рамках задачи классификации твитов как потенциальных источников нежелательных реакций на лекарства. Модель показывала первые и вторые результаты в рамках соревнований SMM4H 2021 Task 2 и Task 1a, соответственно [18].
5. Две многокомпонентные системы: (i) система для биомедицинского информационного поиска, состоящая из двух моделей, которая показала более высокую производительность по сравнению с традиционной моделью поиска на вручную размеченном наборе данных абстрактов для запросов о заболеваниях и генах [19], и (ii) система для классификации, извлечения и нормализации нежелательных реакций на лекарства на реалистичных, несбалансированных данных; были исследованы подбор оптимальных коэффициентов обучения и метод неполной выборки (undersampling) [20].

– Были представлены новые размеченные корпуса для извлечения информации. Среди них можно выделить следующие:

6. Был создан новый корпус реакций на лекарственные средства (RuDReC), частично аннотированный корпус отзывов потребителей о фармацевтических продуктах на русском языке, а также модели RuDR-BERT для задач распознавания именованных сущностей и классификации предложений [21].
7. Для нормализации концептов были созданы два корпуса: корпус клинических испытаний на английском языке для нормализации лекарств и заболеваний [16; 17], а также корпус RuCCoN - набор электронных медицинских записей на русском языке, в котором сущности связаны с UMLS [22].

8. Был создан NEREL-BIO, корпус научных абстрактов на русском и английском языках и схема аннотирования вложенных именованных сущностей, включающие общие и биомедицинские типы сущностей [23].

– Предложены новые стратегии оценки эффективности моделей.

9. Проведен анализ ограничений существующих датасетов для связывания биомедицинских сущностей, и были предложены новые стратегии оценки моделей: (i) метод *стратифицированной* выборки [13], (ii) стратегии оценки *внутри терминологии* и *между терминологиями* [24]. Кроме того, были проведены эксперименты в рамках межъязыковой задачи, используя клинические тексты и исследовательские статьи. Была разработана процедура *фильтрации* тестового набора для анализа "сложных случаев" связывания сущностей в условиях zero-shot постановки межъязыкового переноса обучения [25].
10. Проведен анализ ограничений существующих датасетов для извлечения отношений в научных статьях и электронных медицинских записях. Для устранения различий в эффективности моделей *внутри домена* и *вне домена*, был предложена нейронная сеть с кросс-вниманием, демонстрирующая наилучшие результаты в кросс-доменных экспериментах [26].

Личный вклад автора заключается в постановке задач, разработке перечисленных выше методов и моделей обработки неструктурированных данных, разработке схем аннотирования перечисленных выше корпусов и стратегий оценки эффективности, анализе и обобщении результатов; первые версии программ, реализующих предложенные методы и модели классификации и извлечения биомедицинских сущностей и оценку эффективности, написаны автором диссертации лично; текущие версии программных модулей, реализующие предложенные в диссертации методы в рамках различных программно-аппаратных архитектур, написаны под непосредственным контролем автора диссертации.

Научная новизна работы. Научная новизна предложенного исследования заключается в разработке новых аннотированных корпусов для различных текстов, создании новых глубоких архитектур и моделей машинного обучения для извлечения и классификации биомедицинской информации

на английском и русских языках и в использовании новых стратегий оценки эффективности моделей. Улучшение показателей качества разработанных методов по сравнению с существующими подтверждено экспериментально с использованием стандартных метрик качества систем анализа текста на естественном языке. Экспериментально показано, что разработанные методы применимы к текстам из различных источников. Первые исследования были проведены для решения задачи извлечения упоминаний о лекарственных эффектах и биомедицинских вложенных именованных сущностях для русского языка.

По теме диссертации опубликовано 42 публикации [11–52].

В соответствии с требованиями Диссертационного совета по компьютерным наукам Высшей школы экономики, не менее десять статей приводятся ниже. В этом списке стоит отдельно выделить публикации: в журналах первого квартиля Q1: [12; 13; 17; 18; 20; 21; 23; 26]; в материалах конференций CORE A/A* [11; 16; 19; 22; 24]; в трудах конференций, индексированных международной базой Scopus [14; 15; 25]. Защита производится на основе не менее семи из них (в частности, первые девять из списка публикаций повышенного уровня).

Публикации по теме диссертации

Публикации первого уровня

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.]

   Вклад автора данной диссертации заключается в том, что она является главным соавтором исследования, сформулировала научную проблему, разработала нейронные модели для распознавания именованных сущностей и первые версии программ, реализующих предложенные модели, провела экспериментальную оценку.

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.]

Вклад автора данной диссертации заключается в том, что она является главным соавтором исследования, сформулировала научную проблему, предложила новые стратегии оценки, разработала первые версии программ, реализующих предложенную оценку, и провела экспериментальную оценку моделей.

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]
   Вклад автора данной диссертации заключается в том, что она является главным соавтором исследования, сформулировала научную проблему, предложила мультимодальную модель совместно с С.А. и руководила исследованием.

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.]
   Вклад автора данной диссертации заключается в том, что она является главным соавтором исследования, сформулировала научную проблему, руководила процессом аннотирования, разработала систему информационного поиска и первую версию программы, реализующую предложенную систему, и провела эксперименты.

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.].
   Вклад автора данной диссертации заключается в том, что она является главным соавтором исследования, сформулировала научную проблему, предложила методику оценки, разработала модель DILBERT совместно с М.З., спроектировала эксперименты и руководила исследованием.

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]

Вклад автора данной диссертации заключается в том, что она является главным соавтором исследования; эта статья является расширенным вариантом конференционной версии [16], и вклад в работу совпадает с тем, что указано для [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. (Q1, Impact Factor 6.931) [Scopus, WOS]
   Вклад автора данной диссертации заключается в том, что она является главным соавтором исследования, сформулировала научную проблему, написала программный код для сбора данных, разработала нейронные модели для классификации и распознавания именованных сущностей и первые версии программ, реализующих предложенные модели, предложила схему аннотирования, руководила процессом аннотирования и частично проводила эксперименты.

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.]
   Вклад автора данной диссертации заключается в том, что она сформулировала научную проблему, написала программный код для сбора дополнительных обучающих данных, предложила несколько типов тестовых наборов и частично провела эксперименты.

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. [Scopus, WOS]
   Вклад автора данной диссертации заключается в том, что она является главным соавтором исследования, разработала схему аннотирования в сотрудничестве с Л.Н., написала программный код для сбора данных, настроила инструменты разметки, разработала модели для первоначальной аннотации данных, написала программный код, реализующих предложенные модели.

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]
Вклад автора данной диссертации заключается в том, что она является главным соавтором исследования, сформулировала научную проблему, предложила модель на основе характеристик совместно с А.И. и руководила исследованием.

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]
Вклад автора данной диссертации заключается в том, что она является главным соавтором исследования, сформулировала научную проблему, предложила модель классификации в сотрудничестве с М.З., разработала оценку и руководил исследованием.

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]
Вклад автора данной диссертации заключается в том, что она сформулировала научную проблему, предложила нейронную модель в сотрудничестве с А.И., разработала оценку и руководила исследованием.

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]
Вклад автора данной диссертации заключается в том, что она предложила и разработала две модели извлечения информации и первые версии программ, реализующих модели.

Публикации второго уровня:

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]

Вклад автора данной диссертации заключается в том, что она является главным соавтором исследования, сформулировала научную проблему, предложила модель архитектуры "кодировщик–декодировщик" с признаками семантической близости в сотрудничестве с М.З., спроектировала эксперименты и руководила исследованием.

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.].
    Вклад автора данной диссертации заключается в том, что она сформулировал научную проблему, предложила новые стратегии оценки, первые версии программ, реализующих предложенную оценку, и руководила исследованием.

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]
    Вклад автора данной диссертации заключается в том, что она является главным соавтором исследования; то же, что и [13] (в этой работе расширена экспериментальная часть из [13]).

Приглашенные доклады на конференциях и семинарах:

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, from 5.07-7.07.2017, “Text Mining in Biomedical Research”.

Доклады на конференциях и семинарах:

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 Новые модели и методы классификации и извлечения информации

Новые модели и методы классификации и извлечения информации были предложены и разработаны автором диссертации [11–20]. Систематизация знаний о лекарствах и заболеваниях в различных подобластях является важным условием для эффективных биомедицинских приложений, особенно учитывая огромное количество биомедицинских текстов, требующих анализа. В связи с этим использование автоматизированных методов обработки естественного языка (NLP) является необходимым для эффективного поиска информации (information retrieval, IR) или извлечения информации (information extraction, IE). В частности, в этой главе обсуждаются следующие ключевые научные проблемы:

– Первая проблема, описанная в [11], связана с необходимостью существенных усилий разметчиков для разметки достаточного количества обучающих примеров для каждого языка или подобласти (поддомена) для современных моделей с учителем. Кроме того, современные модели распознавания именованных сущностей (named entity recognition, NER) могут демонстрировать крайне низкое качество при изменении подобласти или языка, что является еще одной важной проблемой для обработки биомедицинских текстов.

– Нейросетевые подходы для обнаружения нежелательных побочных эффектов (ADEs) из текста ограничены в своей возможности использования структуры лекарств и в основном полагаются на анализ текстовой информации из сообщений пользователей [18].

– Третья проблема, изучаемая в [16; 17], связана с кросс-терминологическим связыванием сущностей с концептами заданной терминологии без дополнительного переобучения. Это общая проблема в биомедицинской области, где используются различные терминологии и онтологии для представления биомедицинских концептов.

– Еще одной проблемой, обсуждаемой в [19], является эффективный поиск и анализ биомедицинских текстов, сфокусированных на сущностях, таких как заболевания, гены и химические соединения. С огромным количеством текстовых биомедицинских данных и ограничениями современных подходов к информационному поиску на основе плотных или

разреженных векторных представлений, необходимы разработка и оценка новой поисковой системы, ориентированной на сущности.

Для решения обозначенных выше проблем предложен ряд новых моделей и методов:

- Для решения первой проблемы в рамках задачи NER исследуется многоязычный перенос обучения между электронными медицинскими картами (EHR) и текстами, созданными пользователями (UGT) на разных языках, с целью выяснить, можно ли осуществлять перенос знаний с высокоресурсного языка (английский) на тексты низкоресурсного языка (русский) [11]. Подход использует многоязычные возможности предварительно обученных моделей и техники переноса обучения (transfer learning).
- Для решения второй проблем предложен новый метод использования как текстовой, так и молекулярной информации для классификации текстов о наличии ADE [18]. Для объединения представлений лекарств и твитов исследуются две стратегии, включая использование механизма со-внимания для интеграции признаков разных модальностей.
- Для решения третьей проблемы предложена модель Drug and disease Interpretation Learning with Biomedical Entity Representation Transformer (DILBERT), которая использует метрическое обучение (metric learning) и негативное сэмплирование для получения представлений сущностей концептов. Модель DILBERT позволяет создать общее семантическое пространство векторов для сущностей и концептов из базы знаний для кросс-терминологического связывания (entity linking, EL) без необходимости повторного обучения [16; 17].
- Для решения четвертой проблемы предлагается основанную на BERT систему извлечения информации (IE), спроектированную как поисковую систему [19]. Система состоит из двух подмодулей, а именно подмодуля NER и подмодуля EL, которые применяются последовательно. Подмодуль NER отвечает за идентификацию интересующих сущностей, в то время как подмодуль EL связывает извлеченные сущности с концептами из соответствующих баз знаний с использованием модели DILBERT. Для оценки качества поисковой системы создана новая размеченная коллекция абстрактов PubMed с запросами о заболеваниях и генах, размеченная оценками релевантности.

Основные результаты и выводы исследования подхода многоязычного обучения для NER в биомедицинской области [11] следующие:

– Исходя из результатов оценки качества моделей, можно заключить, что многоязычная модель multi-BERT обладает наилучшими возможностями для передачи знаний в условиях zero-shot постановки, когда обучающие и тестовые наборы либо на одном языке, либо в одной поддомене.

– Техника переноса обучения эффективно позволяет снизить необходимое количество размеченных данных для достижения высоких показателей качества. В частности, обученные модели смогли достигнуть от 98 до 99% показателей качества на обоих типах сущностей после обучения всего на 10-25% предложений.

Основные результаты и выводы по предложенной мультимодальной модели из [18] следующие:

– Предложенная мультимодальная модель эффективна при использовании как текстовой, так и молекулярной информации для классификации ADE и обеспечивает наилучшие показатели качества на нескольких размеченных наборах данных на английском, французском и русском языках.

– Эксперименты показывают, что молекулярная информация, полученная из нейронных сетей, более полезна для классификации ADE, чем традиционные молекулярные дескрипторы (molecular descriptors).

Основные результаты и выводы по предложенным моделям извлечения информации, описанные в [16; 17; 19], следующие:

– Эксперименты демонстрируют, что модель DILBERT значительно превосходит базовые и современные архитектуры для биомедицинского EL. Более того, в работах показано, что данная модель эффективна при переносе знаний из научной литературы на клинические испытания с использованием для оценки нового аннотированного набора данных для связывания лекарств и заболеваний.

– Нейронная архитектура IE демонстрирует высокое качество поиска по запросам заболеваний и генов в условиях zero-shot постановки. Кроме того, разработанная система IE может эффективно обрабатывать

абстракты статей вне поддомена, что указывает на ее потенциал для применения к широкому кругу биомедицинских сущностей.

2.1 Межъязыковой и междоменный NER с переносом обучения

Результаты этого раздела основаны на статье [11].

Были проведены эксперименты с использованием четырех наборов данных: англоязычные корпусы CADEC [53] и n2c2 [54], набор данных, содержащий медкарты на русском языке, а также набор данных, состоящий из пользовательских текстов на русском языке. Каждый корпус определяется двумя параметрами: (i) язык: английский (EN) или русский (RU); и (ii) область: медкарты (EHR) или пользовательские тексты (UGT).

Для NER используется языковая модель BERT [7] со слоем многопеременной логистической функции (softmax). Классы слов были закодированы с использованием схемы тегов BIO, а модель была обучена на уровне предложений. В частности, используется BERT_{base}, Multilingual Cased (Multi-BERT), предварительно обученный для 104 языков, вместе с Adam оптимизатором с полиномиальным затуханием, чтобы обновлять скорость обучения в каждую эпоху с шагами прогрева в начале. В качестве базовых моделей был использован LSTM-CRF из библиотеки Saber [55] и BioBERT [56].

Каждый набор данных был случайным образом разделён на 70% обучающей выборки и 30% тестовой выборки. Было обучено 720 моделей на одной сервере с 8 графическими картами NVIDIA P40. Внутрикорпусные (IC) и внекорпусные (OOC) характеристики всех моделей сравнивались на тестовых наборах с использованием скрипта CoNLL для оценки точности (P), полноты (R) и F1 (F).

Во всех наборах данных модели на основе BERT превзошли LSTM-CRF с точки зрения метрик IC. Кроме того, разница между BioBERT и Multi-BERT не была статистически значимой (two-tailed t-test $p \leq 0{,}05$). Все модели достигли значительно более высокого качества при обнаружении лекарств, чем при обнаружении болезней. Это может быть связано с проблемами границ в многословных выражениях, на что указывает средняя длина сущностей.

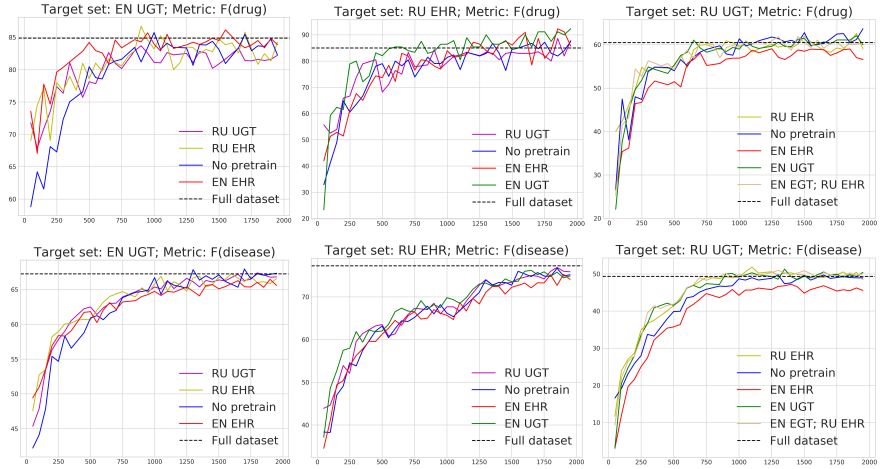

Рис. 1 — Результаты моделей Multi-BERT с обучением на исходном наборе данных (название корпуса в легенде) или без предварительного обучения (“No pretrain”) для наборов EN UGT, RU UGT, RU EHR. Ось Y: F1 для распознавания сущностей лекарств или заболеваний, ось X: количество предложений, использованных для обучения.

Zero-Shot Для оценки эффективности модели NER на основе BERT, обученной на одном корпусе для обнаружения лекарств и заболеваний на другом языке или в другом поддомене в zero-shot постановке, обучена Multi-BERT на одном корпусе и затем оценена по отношению к другому языку/поддомену без дополнительного обучения.

Для распознавания лекарств лучшую обобщаемость удалось достичь при обучении на медкартах и оценке на тестовом наборе отзывов на английском языке. При оценивании на данном наборе EN UGT модель достигла OOC F1-меры со значением 77.08% и 36.31%, когда была обучена на корпусах EN EHR и RU UGT соответственно, в то время как IC F1 84.88%. Следует отметить, что количество предложений в английском корпусе EHR в девять раз больше, чем в наборе UGT, и 78% токенов лекарств в английском наборе UGT встречаются в английском корпусе EHR. При оценке на RU UGT модель показала F1 со значениями 26.31% и 34.78% при обучении на EN UGT и EN EHR, соответственно, в то время как IC F1-мера 60.45%.

Для распознавания заболеваний Multi-BERT показал более плохую обобщаемость на корпусах, отличных от тех, на которых модель была обучена. При оценке на корпусе RU UGT модель показала OOC F1 24.12% и 30.86%, когда она была обучена на корпусах EN UGT и RU EHR, соответственно, в то время как внутрикорпусная F1 49.35%. При тестировании на корпусе EN UGT модель достигла F1 37.94% и 4.32%, когда она была обучена на корпусах RU UGT и EN EHR, соответственно, в то время как IC F1 67.25%. Это может быть связано с различиями между языком пользователей и профессиональной медицинской терминологией.

Затем было исследовано, насколько хорошо модель NER работает по сравнению с моделью, обученной на гораздо больших наборах данных, используя небольшое количество обучающих примеров. Начиная со случайной выборки из 50 предложений из "целевого" набора обучающих данных, дообучается модель на этом подвыборке данных и оценивается на "целевом" тестовом наборе. Затем увеличивается размер на 50 предложений из набора обучающих "целевых" данных, повторяется процесс до 2,000 предложений из набора обучающих данных. В каждом раунде модель обучается заново, чтобы избежать переобучения, как предложено в [57].

Был записан размер подмножества, при котором модель достигала по крайней мере 99% F1, полученной на полном наборе данных. Результаты для наборов данных RU UGT, RU EHR и EN UGT представлены на рис. 1. Multi-BERT, предварительно обученный на наборе EN UGT и обученный на 2,000 предложениях из корпуса EN EHR (2.81% от полного корпуса), достиг 92% F1 и 76% F1 на лекарствах и болезнях, соответственно. Как показано на рис. 1, модели с переносом знаний показывали лучшие результаты, чем модели без предварительного этапа обучения, даже в случаях, когда как поддомен, так и язык изменялись между "исходными" и "целевыми" наборами. Для использования стратегии обучения переноса знаний требовалось до 550 меньше предложений, чем для обучения с нуля. Заметим, что моделям потребовалось только 10% и 23% от наборов данных EN UGT и RU URT, соответственно, чтобы достичь результатов, хотя бы таких же хороших, как на полном наборе данных. Стоит отметить, что это наблюдение особенно важно для низкоресурсных языков и новых областей (например, отзывов пользователей или клинических исследований). Кроме того, качество моделей с предварительным этапом обучения, обученных на разном количестве предложений, становилась более стабильным в терминах отклонений между F1.

В заключении данной работы делается следующий вывод. Был исследован следующий вопрос: *может ли дополнительное обучение на существующем наборе данных помочь в работе модели NER на основе BERT на новом наборе данных с небольшим количеством размеченных примеров, если поддомен, язык или оба изменяются между этими наборами данных?* Модели с предобучением на данных на том же языке или поддомене показали лучшие результаты в zero-shot режиме или на небольшом количестве примеров (few-shot). Модель с лучшим предварительном обучением достигла 99% качества распознавания сущностей на полном наборе данных, используя только 23.56% обучающих данных корпуса RU URT, в то время как модель с предварительным обучением на данных с двумя сдвигами (набор данных EN EHR) использовала 26.1% обучающих данных. Таким образом, даже обучение на данных с двумя сдвигами может быть эффективным.

2.2 Мультимодальная модель с текстовыми и молекулярными представлениями

Результаты этого раздела основаны на статье [18].

Популярность социальных медиа в качестве источника информации о здоровье значительно возросла за последнее десятилетие. Одним из хорошо изученных направлений исследований является фармаконадзор на основе данных социальных медиа, которая направлена на обнаружение нежелательных лекарственных явлений (adverse drug events, ADEs) по пользовательским текстам. Стоит отметить, что термин ADE и нежелательные лекарственные реакции (adverse drug reactions, ADRs) часто используются взаимозаменяемо.

Предлагается новый метод использования как текстовой, так и молекулярной информации для классификации ADE, вдохновленный мультимодальными исследованиями. Исследуется влияние использования различных подходов к молекулярному представлению, включая традиционные молекулярные дескрипторы и нейронные кодировщики.

Рассмотрим текст T, который можно представить в виде пары текстовой модальности t и модальности лекарства, представляющей собой набор k_T упоминаний лекарств: $T = (t, D_T)$, где $D_T = (d_1, d_2, ..., d_{k_T})$. Чтобы получить два унимодальных представления T, выбирается случайный d_i из D_T и кодируется t и d_i с помощью двух кодировщиков: (i) текстового кодировщика

M_{text} и (ii) кодировщика лекарств M_{drug}:

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

где $u_{text}^T \in \mathbb{R}^{d_t}$ — представление текстовой модальности t и $u_{drug}^T \in \mathbb{R}^{d_d}$ — представление лекарства d_i; d_t и d_d — размерности полученных унимодальных векторных представлений. Задачу мультимодальной бинарной классификации текста можно сформулировать так:

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

где $f_{mod} : \mathbb{R}^{d_t+d_d} \to \mathbb{R}^{d_{bi}}$ — функция комбинации модальностей, обеспечивающая бимодальное представление T, $f_{cl} : \mathbb{R}^{d_{bi}} \to \mathbb{R}$ — полносвязная классификационная сеть с сигмоидой как функцией активации. d_{bi} — размерность бимодального векторного представления.

В качестве текстового представления входного текста используется итоговое представление токена классификации (CLS токена). В унимодальном подходе текстовое представление u_{text}^T обычно подается непосредственно на вход классификатору, а для бимодальных моделей сначала соединяется текстовая модальность с молекулярной модальностью. Предложено 2 способа объединения текстовой и молекулярной модальностей: (i) конкатенация представлений текста и молекул; (ii) использование внимания с масштабированием (scaled dot-product attention) в моделях на основе трансформеров. В первом случае текстовое представление и представление молекул конкатенируются:

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

Во втором случае используем механизм внимания, чтобы обучить T в качестве линейной комбинации u_{text}^T и u_{drug}^T. Для классификации текста используется:

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

где α - вес текстовой модальности, полученный с помощью механизма внимания.

Для получения информации о лекарственных средствах названия лекарств сопоставляются с их идентификаторами в DrugBank [58]. В качестве кодировщиков лекарств используются следующие методы: 1) Классификация по АТХ, которая является наиболее широко используемой системой классификации лекарств. Лекарство кодируется как разреженный 14-мерный вектор с

нулевыми значениями для кодов ATX, к которым лекарство не принадлежит. 2) Молекулярные дескрипторы, которые представляют собой набор числовых значений, характеризующих физико-химические и структурные свойства молекулы лекарства. Используется библиотеку для расчета молекулярных дескрипторов Mordred [59] для расчета 2 тысяч дескрипторов для каждого лекарства из Drugbank. 3) MolBERT [60] - кодировщик на основе BERT, специально разработанный для кодирования молекулярных данных. 4) ChemBERTa [61] - другой кодировщик на основе BERT, предварительно обученный на химических данных.

Для классификации данных используется полносвязанная сеть классификации с одним скрытым слоем с активациями GeLU [62]. На выходном слое используется сигмоидная активация, а в качестве функции потерь применяется бинарная кросс-энтропия.

Для экспериментов были использованы три корпуса: наборы данных на французском языке соревнования Social Media Mining for Health (#SMM4H) 2020 Task 1b, а также наборы данных на английском и русском языках соревнования SMM4H 2021. Для каждого набора данных обучены 10 моделей на основе BERT с различной инициализацией весов классификатора. Затем вычисляется среднее значение и стандартное отклонение F1-меры (F_1), точности (P) и полноты (R) для класса *ADE*.

Таблица 1 демонстрирует показатели классификации одно- и двухмодальных моделей на корпусе русских твитов SMM4H 2021. Во-первых, две модели, показавшие лучшие результаты по показателю F_1, являются моделями с двумя модальностями, использующими представления препаратов MolBERT. Модель на основе конкатенации немного превосходит модель с вниманием (на 0.3%) и одномодальный BERT (на 0.8%). Во-вторых, модели с конкатенированными модальностями показывают более высокие результаты, чем модели с перекрестным вниманием. Наконец, были объединены результаты десяти мультимодальных моделей с теми же настройками, используя простую схему голосования, с целью повысить устойчивость конечной системы. Как показано в таблице 2, данная модель достигла лучших результатов по сравнению с предыдущими работами, показав улучшение на 4% по показателю F_1 по сравнению с одной моделью.

После анализа результатов на французском наборе данных SMM4H 2020 Task 1b можно сделать несколько выводов. Во-первых, наивысшие F_1 были достигнуты двухмодальными моделями с конкатенированными модальностями.

Таблица 1 — Результаты моделей на данных SMM4H 2021 Task 2 по классификации твитов. Использована двуязычная модель EnRuDR-BERT, предварительно обученная на корпусе RuDReC.

Модель	**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
Модели с конкатенацией модальностей			
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**
Модели с механизмом внимания			
BERT+ATC	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

Таблица 2 — Официальные результаты SMM4H 2021 на официальном тестовом наборе.

Model	**P**	**R**	**F_1**
Task 2, корпус русских твитов			
Предложенная модель	0.58	0.57	**0.57**
Модель, достигнувшая второго места	0.54	0.57	0.52
Task 1a, корпус англ. твитов			
Наша модель	0.552	0.681	**0.61**
Модель, достигнувшая первого/второго места (команда #4)	0.515	0.752	**0.61**

В частности, двухмодальная модель с MolBERT превосходит лучшие официальные результаты соревнований SMM4H 2020 и одномодальные модели, достигнув наилучших результатов по показателю F_1 с улучшением на 8.5%. Во-вторых, производительность моделей с перекрестным вниманием значительно ниже, чем у моделей с конкатенированными модальностями. Однако модели с

перекрестным вниманием с ChemBERTa и MolBERT немного превзошли одномодальную модель по показателю F_1 на 1.8% и 0.9%, соответственно. Наконец, одно- и двухмодальные модели на основе BERT значительно превосходят одномодальные базовые модели SVM и CNN, демонстрируя превосходство сложных нейронных сетей над простыми моделями и традиционными алгоритмами машинного обучения в задаче классификации текстов на предмет нежелательных реакций на препараты.

Наше исследование направлено на определение того, как включение дополнительной модальности влияет на показатели классификации твитов, упоминающих препараты из разных терапевтических групп. Для этого были выбраны лучшие двухмодальные модели по показателю F_1 для английского и русского корпусов SMM4H 2021. Эксперименты показали, что показатели бимодальной классификации не зависят только от распределения входных данных по группам ATX.

В заключении делается следующий вывод. Была исследована задача обнаружения нежелательных реакций на лекарственные препараты в твитах, описанные пользователями. Предложенный подход объединяет несколько современных моделей, включая BERT для представления текста и MolBERT для представления лекарственных препаратов. Предложенные модели продемонстрировали значительное улучшение в классификации французских твитов задачи SMM4H 2020 Task 1b (достигнув улучшения на 8%) и показали лучшие результаты в рамках задач SMM4H 2021 Task 1a и Task 2 для английских и русских текстов, соответственно. Стоит отметить, что эти результаты были получены только с помощью известных предварительно обученных моделей BERT для отдельных компонентов. Исходный код моделей опубликован по адресу `https://github.com/Andoree/smm4h_2021_classification`.

2.3 Система извлечения информации для поиска

Результаты этого раздела основаны на статье [19]; новая модель DILBERT описана на статьях [16; 17].

В данном разделе представлены описание и оценка системы IE на основе BERT в качестве поисковой системы, ориентированной на сущности, для платформы `PandaOmics`[1]. В частности, исследуется следующий вопрос: *могут*

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

ли модели с высокими показателями качества в задачах NER и EL [14; 56] определять релевантные публикации на запросы о болезнях и генах в различных биомедицинских подобластях? Для оценки создана новая коллекция для поиска PubMed абстрактов по запросам о болезнях и генах с соответствующими оценками релевантности. Двухкомпонентная система IE, состоящая из обученных моделях на основе BERT для NER и EL, была оценена по сравнению со стандартной моделью информационного поиска BM25 с готовым к использованию программным обеспечением Elasticsearch.

Наше исследование сосредоточено на извлечении двух типов сущностей: болезней и генов. Наша система IE состоит из нескольких конвейерных блоков, каждый из которых посвящен определенному типу сущностей. Эти блоки состоят из двух подмодулей, которые применяются последовательно: подмодуля NER и подмодуля EL.

NER Модель для распознавания сущностей с BioBERT была обучена на комбинации наборов данных NCBI и BC5CDR (BioCreative V CDR) для сущностей, связанных с болезнями [63;64], а также на наборе данных DrugProt для генов [65]. Обученная модель достигла F-меры 88.43% и 90.39% распознавания сущностей, связанных с болезнями и генами, соответственно.

EL Чтобы связать именованную сущность с соответствующим ей концептом в базе знаний, выполняется задача связывания сущности. Концепт в этом сценарии относится к элементу в базе знаний, который представляет определенную идею или понятие, относящееся к определенной области знаний. DILBERT используется для связывания извлеченных сущностей с концептами из соответствующих словарей. Аналогично подмодулю NER, модели обучаются на BC5CDR [64] и BC2GN (BioCreative II GN) [66]. Модель DILBERT позволяет создавать общее семантическое пространство векторов для сущностей и концептов из базы знаний, где тексты с похожими смыслами находятся близко друг к другу. Эта характеристика позволяет ранжировать концепты на основе функции расстояния s.

Согласно обозначениям, предложенным в [67], сущности и концепты преобразуются в векторные представления следующим образом:

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

где T - это глубокая нейронная сеть архитектуры Трансформер, веса которой могут быть обновлены во время процесса обучения. Функция $red(\cdot)$ преобразует последовательность векторов в один вектор, где m представляет сущность, которую необходимо связать с соответствующим концептом, а c обозначает имя концепта. Существуют различные реализации функции $red(\cdot)$, такие как выбор выхода, соответствующего токену CLS, или покомпонентное усреднение всех векторов для получения вектора фиксированного размера. Эксперименты показывают, что усреднение является оптимальным вариантом для функции $red(\cdot)$.

Оценка важности кандидата c_i для сущности m определяется функцией расстояния, такой как евклидово расстояние, примененной к соответствующим векторным представлениям:

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

Для обучения нейронной сети используется триплетная (triplet) целевая функция, которая учитывает семантические сходства и различия между концептами и сущностями. Для данной целевой функции необходимо передать упоминание сущности m, имя соответствующей (позитивного) концепта c_g и имя не соответствующего (негативного) концепта c_n. Нейронная сеть обучается таким образом, чтобы расстояние между m и c_g было меньше, чем расстояние между m и c_n для заданного порога. Выражение функции потерь записывается так:

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

где ϵ — смещение, гарантирующее, что c_g как минимум на ϵ ближе к m, чем к c_n. В наших экспериментах $\epsilon = 1$.

Чтобы сгенерировать позитивные примеры, словарь ограничен концептами, соответствующим сущностям, в то время как оставшаяся часть словаря используется для генерации негативных примеров [68]. Были исследовали различные стратегии для выбора негативных примеров: (i) случайная выборка, (ii) иерархическая случайная выборка (random sampling + n parents), (iii) перевыборка негативных случаев, (iv) иерархическая перевыборка (resampling + n siblings).

Особенность метрического подхода заключается в его способности определения сущностей, которые не имеют подходящего концепта в словаре. Методология обнаружения естественным образом следует из предположения,

Таблица 3 — Статистика корпуса данных клинических испытаний.

Сущность	**#текстов**	**#текстов с CUIs**	**#уник. текстов**	**#уникальных текстов с CUIs**
Лекарство	1075	794	838	671
Болезнь	819	804	638	638

что похожие элементы расположены близко друг к другу. Если все объекты словаря находятся достаточно далеко от сущности, то это означает, что сущность не найдена в словаре. Таким образом, если все концепты находятся на расстоянии, большем, чем пороговое значение t, можно заключить, что ни одна из них не соответствует сущности. Для определения порога используются максимальное расстояние истинно положительных (true positive) примеров d_{tp} и минимальное расстояние ложно положительных (false positive) примеров d_{fp}. Пороговое значение устанавливается как взвешенная сумма:

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

где a_1 представляет долю истинно-положительных примеров среди сущностей, ближайший концепт которых находится на расстоянии $s \in [d_{fp}; d_{tp}]$, а a_2 обозначает долю ложно-положительных в том же наборе сущностей. Если множество сущностей пусто, то коэффициенты устанавливаются равными $\frac{1}{2}$.

Предложенный подход был оценен на вручную аннотированном корпусе клинических испытаний. Статистика аннотированных текстов клинических исследований представлена в таблице 3. Экспериментальные результаты представлены в таблице 4. Как видно из таблицы 4, модель DILBERT показывает наилучшие результаты на корпусах CT Condition, CT Intervention. Исходный код DILBERT и наборов CT доступен на GitHub по адресу `https://github.com/insilicomedicine/DILBERT`.

Показатели модели DILBERT на корпусе клинических исследований. Результаты представлены для типов: болезни (CT Condition) и лекарственные средства (CT Intervention). Показатели качества предоставляются для подмножества, состоящего только из сущностей с одним понятием (single concept) и для всего корпуса (full set).

Таблица 4 — Показатели модели DILBERT на корпусах клинических исследований. Результаты представлены для типов сущностей болезни (CT Condition) и лекарственного средства (CT Intervention). Показатели качества предоставляются для подмножества, состоящего только из сущностей с одним понятием (single concept) и для всего корпуса (full set).

Модель	**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 с разными стратегиями сэмплирования				
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

Поиск

Далее описывается новый набор данных для поиска по сущностям, который включает запросы и процесс, используемый для сбора оценок релевантности. Статистика набора данных представлена в таблице 5.

Таблица 5 — Статистика набора данных.

Выборка	**# запросов**	**ср. количество текстов на запрос**		
		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
Всего	152	102.41	41.76	7.78

Запросы В сценарии поиска пользователь может ввести название гена или его символ, например “PSEN1” (ENSG00000080815), и получить все соответствующие публикации и связанные заболевания, включая болезнь Альцгеймера (EFO:0000249). Функция автозаполнения предлагает варианты поисковых запросов из словарей заболеваний или генов.

Pooling Далее используется стандартная практика создания коллекций информационного поиска *pooling* (комбинация результатов) для объединения результатов поиска из двух основных источников:

1. Результаты поиска извлекаются из Elasticsearch, и результаты объединяются до глубины поиска в 100.
2. Результаты поиска извлекаются из PubMed, и результаты объединяются до глубины поиска в 100, за исключением аннотаций, уже извлеченных первой системой.

Финальный набор оценочных данных составляет 23,099 пар запрос-абстракт/аннотация, со средним количеством аннотаций на один запрос равным 152.

Релевантность оценки Для каждой пары запрос-абстракт были собраны оценки релевантности от двух аннотаторов с медицинским образованием. Эксперт-аннотатор с PhD в биологии создал список запросов, используя журналы из платформы поиска `PandaOmics`. Запросы представляют собой CUI-идентификаторы заболеваний и генов. Кроме того, аннотаторы выбрали список понятий с неоднозначным наименованием (ambiguous concepts) (например, coad относится к хронической обструктивной болезни легких и к аденокарциноме толстой кишки (COAD)).

Каждый аннотатор выбирал запрос про болезнь или ген из списка определенных идентификаторов, вместе с абстрактом, годом публикации и информацией о журнале. Аннотации были представлены в случайном порядке. Затем аннотаторы оценивали: (i) релевантность на шкале (релевантный, нерелевантный или сомнительный), и (ii) причину релевантности или нерелевантности.

Оценка качества поиска

Для оценки используются метрики точности (P), полноты (R) и F-меры (F). Точность рассчитывается как доля релевантных документов среди всех полученных документов. Аналогично, полнота рассчитывается как доля релевантных документов из всех потенциально релевантных документов в наборе данных. Для экспериментов использованы пары запрос-абстракт с пометками релевантности и нерелевантности, исключая третью категорию.

Таблица 6 — Метрики информационного поиска на полном наборе запросов и на наборе запросов с неоднозначными понятиями.

Модель	**Полный набор**			**Ambiguous**		
	P	**R**	**F**	**P**	**R**	**F**
Запросы заболеваний						
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
Запросы генов						
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
Оба типа запросов						
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

Таблица 6 представляет сравнение показателей поиска моделей на полном наборе запросов и поднаборе запросов с неоднозначными понятиями. Наши результаты показывают, что система, основанная на BERT, превосходит BM25 на обоих наборах данных для обоих типов сущностей. Ожидаемо, разница в показателях между двумя моделями более значительна на поднаборе с неоднозначными названиями концептов. Кроме того, для системы с BERT моделями, точность выше, чем полнота.

Для дальнейшего анализа точности поиска разработан набор данных для обнаружения абстрактных запросов вне области. В этот набор было включено примерно 30 000 записей из списка журналов PubMed, которые публикуют статьи не только о биологических сущностях, но также о культурных темах, экономике и эконометрике, искусственном интеллекте, праве, лингвистике и языке и т.д. (категории вне нашего домена). Наш экспертный аннотатор вручную выбрал журналы вне области, для которых ожидается, что система IE не вернет *ни одного результата*. Из этих журналов случайным образом выбрано 58 790 абстрактов, каждый содержащий по крайней мере одно понятие гена или заболевания, извлеченное с помощью Elasticsearch. Обнаружено, что в 90% этих абстрактов система, основанная на BERT, не идентифицировала никаких сущностей.

В заключении данной работы делается следующий вывод. Была проведена всесторонняя оценка биомедицинского поискового движка, сосредоточенного на сущностях, который использует модели BERT для извлечения и связывания заболеваний и генов. Этот движок является частью платформы поиска биологических мишеней, которая позволяет пользователям получать список соответствующих публикаций при запросе концепта заболевания или гена.

3 Новые размеченные корпуса для извлечения информации

Новые размеченные корпуса текстов из различных источников были предложены и разработаны автором диссертации [16; 17; 21–23]. Разработка биомедицинских систем извлечения информации является сложной задачей из-за отсутствия аннотированных наборов данных. В частности, в этой главе обсуждаются следующие ключевые научные проблемы:

– Плохо составленные контексты, повсеместное присутствие разговорных выражений, сокращенных форм, опечаток/орфографических ошибок и слов, не входящих в словарь, создают проблемы для эффективного использования пользовательского контента в области здравоохранения, как показано в [21]. Кроме того, языковые модели общей предметной области могут плохо работать с биомедицинскими текстами, поскольку эти тексты часто содержат технические термины, аббревиатуры и концепты, специфичные для предметной области.

– Присущая предметной области сложность и ее терминология. Упоминания о болезнях, симптомах, лекарствах и других понятиях сильно варьируются, и из-за большого медицинского словаря связывание сущностей и концептов становится сложной, но существенной проблемой, как показано в [22].

– Третья проблема, рассмотренная в [23], заключается в ограниченной способности существующих наборов данных и методов NER фиксировать сложные вложенные структуры сущностей, которые часто встречаются в биомедицинских текстах. Биомедицинские тексты часто содержат упоминания о сущностях, таких как болезни, содержащие части тела или химические вещества, которые вложены друг в друга. Однако

большинство существующих наборов данных и более точных методов предназначены для захвата плоских структур сущностей по сравнению и часто ограничены наиболее распространенными типами сущностей, такими как лекарства/химические вещества и болезни.

Для решения вышеописанных проблем предложены новые корпуса:

- Для решения первой проблемы был создан корпус RuDReC [21], который представляет собой частично аннотированный корпус отзывов потребителей на русском языке о фармацевтических продуктах, наряду с моделями RuDR-BERT, предварительно обученными на 1,4 миллионах комментариев, связанных со здоровьем, и обученными для задачи распознавания именованных сущностей и классификации предложений.
- Для решения второй проблемы был создан корпус RuCCoN [22], новый вручную размеченный набор данных для нормализации клинических концептов на русском языке. Он содержит более 16 тыс. упоминаний сущностей, вручную связанных с более чем 2 тыс. уникальных концептов из русскоязычной части UMLS. Созданы обучающие/тестовые выборки для различных стратегий оценок (stratified, zero-shot, и CUI-less) и представлены результаты обученных моделей.
- Для решения третьей проблемы был создан корпус вложенных именованных сущностей NEREL-BIO [23] с использованием абстрактов PubMed на русском и английском языках. Этот набор данных включает в себя вручную размеченные сущности 37 типов, включая вложенные структуры глубиной до шести уровней. Созданы обучающие/тестовые выборки и представлены результаты обученных моделей.

Основные результаты и выводы по корпусу пользовательских текстов, описанные в [21], следующие:

- Корпус отзывов потребителей о здоровье на русском языке, RuDReC, который разделен на две части: (i) 1,4 миллиона комментариев, которые могут быть использованы для обучения современных языковых моделей, и (ii) 500 аннотированных отзывов, которые могут быть использованы для обучения двум задачам.
- Схема аннотации на уровне предложения, так и на уровне сущностей. Классы на уровне предложений указывают на наличие или отсутствие

проблем, связанных со здоровьем. Кроме того, предложения, содержащие вопросы, связанные со здоровьем, дополнительно аннотируются на уровне сущностей для определения детализированных подтипов, таких как классы лекарств и лекарственные формы, показания к применению лекарств и реакции на лекарства.

- Две языковые модели, основанные на BERT, обученные на корпусе.
- Оценка нескольких моделей, основанных на BERT, в задачах классификации и извлечения сущностей.

Основные результаты и выводы по корпусу электронных медицинских карт, описанные в [22], следующие:

- Корпус электронных медицинских карт на русском языке, RuCCoN, в котором сущности связаны с концептами из UMLS.
- Тестовые выборки для различных стратегий оценки (stratified, zero-shot, CUI-less).
- Оценка нескольких современных моделей на RuCCoN, включая различные варианты стратегий оценки, и исследование необходимости обучающих выборок для межъязыковой нормализации сущностей с английского языка на русский язык.

Основные результаты и выводы по корпусу абстрактов, описанные в [23], следующие:

- Набор данных биомедицинских абстрактов на русском и английском языках, NEREL-BIO, который содержит аннотации вложенных сущностей.
- Схема аннотации, которая включает в себя 17 специализированных типов биомедицинских сущностей и 20 типов сущностей общей предметной области.
- Оценка нескольких современных моделей для задачи распознавания вложенных сущностей.

В целом, эти результаты демонстрируют значительный прогресс в разработке ресурсов для приложений NLP в специализированной области.

3.1 RuDReC: лекарственные реакции в отзывах пользователей о здоровье

Результаты этого раздела основаны на статье [21].

В этом разделе представлены дизайн и конструирование большого корпуса пользовательских текстов (user-generated texts, UGTs) о фармацевтических продуктах на русском языке. Представленный корпус RuDReC разделен на две части: больший корпус из 1,4 миллиона комментариев, связанных со здоровьем, который может быть использован для обучения современных языковых моделей, и небольшая размеченная часть из 500 отзывов, которая может быть использована для обучения моделей NLP задачам. Основными задачами в данном случае являются классификация и NER. Разметка состоит из двух основных компонентов: разметки классов предложений и типов сущностей. Комментарии пользователей были разделены на предложения и размечены на предмет наличия показаний к применению лекарств и симптомов заболевания (drug indications and symptoms of a disease, DI), побочных реакций на лекарства (adverse drug reactions, ADR), эффективности лекарств (drug effectiveness, DE), неэффективности лекарств (drug ineffectiveness, DIE). На этапе идентификации сущностей были определены и извлечены 6 типов сущностей: названия лекарств (drug names), классы лекарств (drug classes), лекарственные формы (drug forms), ADR, DI и результаты (Findings). В общей сложности в корпусе было помечено 2,202 предложения и 4,566 сущностей.

Процесс разметки включал в себя два этапа. На первом этапе аннотаторам с фармацевтическим образованием было предложено прочитать 400 обзоров и выделить все фрагменты текста, включая названия лекарств и состояние здоровья пациента, с которыми он сталкивался до, во время или после употребления лекарств. На втором этапе аннотаторам было предложено просмотреть существующие аннотации и разметить новые тексты.

Анализ существующих корпусов выявил два основных типа сущностей: Лекарство (Drug) и Болезнь (Disease). После нескольких обсуждений аннотаторы определили следующие подтипы сущности Disease: (1) название болезни; (2) показание к применению лекарств (Indication); (3) положительная динамика после или во время приема препарата (BNE-Pos); (4) отрицательная динамика после начала или некоторого периода применения препарата (ADE-Neg); (5) препарат не действует после прохождения курса (NegatedADE); (6) ухудшение состояния после приема курса препарата (Worse). В качестве

подтипов Drug аннотаторы выбрали: (1) наименование лекарственного средства; (2) класс лекарственного средства; (3) форма лекарственного средства.

При расчете согласия аннотаторов для сущностей Disease и Drug были использованы метрики из [53]. Среднее согласие составило примерно 70%.

После завершения первого этапа процесса аннотирования трое авторов просмотрели аннотации и выявили несколько проблем. Было относительно немного примеров типов *Worse* и *ADE-Neg*. Кроме того, типы сущностей *BNE-Pos* содержали множество чрезмерно широких сущностей, которые не были связаны с медицинскими концептами.

Чтобы решить эти проблемы, в схему было внесено несколько изменений. Типы сущностей *Worse* и *ADE-Neg* с *NegatedADE* были объединены в один класс под названием *Drug Ineffectiveness* (DIE) на уровне предложения. Сущности *BNE-Pos* были объединены на уровне предложений и переименованы в *Drug Effectiveness* (DE). Наконец, типы сущностей *Indication* и *Disease* были объединены в единый тип *Drug Indication* (DI), следуя корпусу CADEC. На втором этапе процесса аннотирования два аннотатора продолжили процесс в соответствии с классами предложений и типами сущностей.

Аннотированный корпус содержит 500 отзывов, в том числе отзывы о четырех группах лекарственных средств: седативных средствах, ноотропах, иммуномодуляторах и противовирусных препаратах. 60% отзывов – о седативных препаратах. Тексты о иммуномодулирующих препаратах содержат более длинные предложения и лексемы по сравнению с другими группами препаратов. В среднем их отзывы на 30% длиннее, а их максимальная длина в два раза больше, чем у других групп, хотя их минимальная длина эквивалентна. Среднее количество предложений в русскоязычных обзорах выше, чем в английских корпусах CADEC и PsyTAR, в среднем 9.71 предложения на отзыв по сравнению с 6 предложениями в других корпусах.

Общее число аннотированных предложений во всем корпусе составляет 2,202, распределенных по различным категориям следующим образом: DI (949), ADR (379), FINDING (172), DE (424) и DIE (278). Общее число аннотированных сущностей во всем корпусе составляет 4,566, распределенных по различным категориям следующим образом: DRUGNAME (1043), DRUGCLASS (330), DRUGFORM (836), DI (1401), ADR (720), FINDING (236).

Анализ частей речи (PoS) для слов в сущностях показал, что пользователи социальных сетей, как правило, используют больше глаголов для выражения симптомов и ADR по сравнению с формальными медицинскими

понятиями. В аннотированной части корпуса RuDReC 18.26% слов в сущностях, связанных с болезнями, являются глаголами в то время, как только 2.53% слов в словаре MedDRA из UMLS v. 2020AA являются глаголами.

Отзывы пользователей были собраны путем обхода веб-страниц с популярных медицинских веб-порталов, которые в основном содержат тексты пользователей о фармацевтических продуктах, активных добавках, медицинских учреждениях и аптеках. Дублирующиеся комментарии были удалены, и полученный в результате корпус содержит 1,4 миллиона текстов, 1,104,054 уникальных токена и 193,529,197 токенов.

Многоязычная версия BERT-base (MultiBERT) была использована в качестве инициализации для обучения специфичному для области BERT, который называется **RuDR-BERT**. Было замечено, что 800 тыс. и 840 тыс. шагов языкового обучения были достаточными, что примерно соответствует одной эпохе на корпусе.

Классификационные модели оценивались с помощью 5-кратной перекрестной проверки с точки зрения оценки F1. В таблице 7 представлены результаты обученных моделей RuBERT, Multi-BERT и RuDR-BERT. Основываясь на полученных результатах, можно сделать несколько выводов. Во-первых, модель RuDR-BERT достигла наилучших результатов среди сопоставимых моделей. Во-вторых, модель RuBERT превзошла модель Multi-BERT на 3.12% с точки зрения макро-оценки F1, при этом наибольшее улучшение наблюдалось для DE (+4.09%) и поиска типов сущностей (+4.19%). В-третьих, производительность RuDR-BERT при поиске (36.24%) значительно ниже, чем у ADR (74.15%) и DI (85.06%). Это может быть связано со сходством контекстов и гораздо меньшим количеством обучающих примеров.

Оценки F1, рассчитанные по критериям точного соответствия с помощью скрипта CoNLL, были использованы для сравнения NER моделей при 5-кратной перекрестной проверке. Таблица 8 показывает результаты обученных RuBERT, Multi-BERT и RuDR-BERT. Основываясь на полученных результатах, можно сделать несколько выводов. Во-первых, RuDR-BERT превосходит как RuBERT, так и Multi-BERT по всем типам сущностей. Во-вторых, RuBERT, со словарем токенов, полученных по текстам Википедии и новостях, превосходит Multi-BERT. В-третьих, аналогично классификации предложений, показатели RuDR-BERT при сущностях типа FINDING значительно ниже, чем при ADR и DI. Наконец, все модели обеспечивают более

Таблица 7 — Показатели классификации предложений обученных моделей RuDR-BERT, multi-BERT and RuBERT

Модель	DE	DIE	ADR	DI	Finding	Макро 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

Таблица 8 — Показатели распознавания сущностей обученных моделей RuDR-BERT, multi-BERT and RuBERT

Модель	ADR	DI	Finding	Drugclass	Drugform	Drugname	Макро 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

высокие показатели для сущностей, связанных с лекарствами, чем для сущностей, связанных с болезнями, что может быть связано с проблемами границ в многословных выражениях. RuDR-BERT достигает показателя F1 в 81.34% по сущностям, связанным с заболеваниями, и F1 в 94.65% по сущностям, связанным с лекарствами. Среднее число токенов для сущностей, связанных с лекарствами, составляет 1.06, в то время как среднее число токенов для сущностей, связанных с болезнями, составляет 1.77.

В заключении данной работы делается следующий вывод. В этом исследовании обсуждались проблемы аннотирования комментариев на русском языке, связанных со здоровьем, и были представлены несколько размеченных данных и моделей для классификации и выделения сущностей. Корпус RuDReC предоставляет исследователям возможности для разработки и оценки моделей интеллектуального анализа текста для сбора значимой информации об эффективности лекарств и побочных реакциях на лекарства из сообщений пользователей. Кроме того, модели позволяют анализировать и сравнивать изменения состояния здоровья пациентов, а также лекарственные реакции на различные терапевтические группы лекарств. Корпус данных и веса обученных моделей доступны по адресу `https://github.com/cimm-kzn/RuDReC`.

3.2 RuCCoN: клинические концепты в историях болезни пациентов

Результаты этого раздела основаны на статье [22].

В этом разделе описывается связывание клинических сущностей из историй болезни пациентов на русскоязычные концепты UMLS. В частности, единственный корпус клинических текстов в свободном тексте на русском языке с разметкой сущностей [70] был обогащен добавлением разметки, связывающей сущности с концептами. Корпус данных, созданный исследователями из Национального медицинского исследовательского центра здоровья детей, основан на историях болезни более 60 пациентов с аллергическими и легочными расстройствами и болезнями. Он включает в себя выписки из стационара, отчеты о рентгенологической, эхокардиографической и ультразвуковой диагностике, рекомендации и другие записи от различных врачей. Неидентифицированный корпус, который находится в свободном доступе для исследовательских целей, содержит 160 полностью аннотированных текстов с почти 250,000 лексем, 18,200 аннотированных сущностей, более 7,400 атрибутов и 3,500 связей с семью типами сущностей: 'Болезнь" ("Disease"), "Симптом" ("Symptom"), "Лекарство" ("Drug"), "Лечение" ("Treatment"), "Часть тела" ("Body location"), "Серьезность" ("Severity") и "Курс" ("Course").

Аннотаторам было предложено сопоставить упоминание сущности с идентификатором концепта (CUI) из UMLS. Целью нормализации сущностей является связывание одного и того же идентификатора с различными синонимами данного медицинского понятия; например, "анемический инфаркт сердца" и "инфаркт миокарда" относятся к одному и тому же понятию с помощью CUI C0027051.

Три аннотатора независимо разметили каждую сущность, и согласие Inter-Annotator Agreement (IAA) было рассчитано как точность разметок, сопоставленных по крайней мере двумя аннотаторами по всем аннотированным упоминаниям. По крайней мере, два аннотатора связали сущность с одним и тем же концептом из онтологии в 13,125 случаях и аннотировали 1,032 сущности как CUI-less; IAA составил 78,37%. В 3,900 случаях, когда все разметчики были не согласны, эксперт-аннотатор со степенью в области медицины решал, действительно ли CUI, выбранный одним из аннотаторов, был правильным. После этой процедуры был получен корпус с 16,028 сущностями, связанными с 2,409 концептами и 1,293 сущностями, не связанными с концептами (без CUI). Семантические типы UMLS, наиболее часто представленные в корпусе, - это *Disease or Syndrome* (22%), *Body Part, Organ, or Organ Component* (17%), *Organic Chemical* (14.5%), *Finding* (7%), *Sign or Symptom* (6,5%) и *Pathologic*

Function (4%). Руководство по разметке было создано экспертом со степенью в области медицины.

Существует несколько проблем с разметкой, которые характерны только для языков с ограниченными ресурсами, таких как русский. Эти проблемы включают в себя 1) отсутствие переводов понятий UML на русский язык, 2) необходимость объединения нескольких связанных понятий в один более точный фрагмент, 3) избыточность словаря UMLS и 4) сложное перефразирование.

30% корпуса было использовано для тестовой выборки с различными стратегиями фильтрации. Таблица 9 показывает статистику для каждого разделения.

Stratified. В этом случае выборка была отфильтрована таким образом, чтобы каждое понятие UMLS в тестовом выборке появлялось по крайней мере один раз в обучающем выборке, но не конкретное упоминание из тестовой выборки. В результате все концепты из тестовой выборки рассматриваются в обучающей выборке, но ни одно из упоминаний в обучающей выборке не идентично тем, что содержатся в тестовой выборке.

Zero-shot. В этом случае выборка была отфильтрована таким образом, чтобы содержать только новые концепты, которые вообще не отображаются в обучающем выборке. Другими словами, выборка *stratified* разработана так, чтобы одни и те же концепты появлялись для обучения, валидации и тестирования, но с различными лексическими формами. Тестовая выборка *zero-shot* предоставляет моделям доступ к новым терминам и концептам для валидации и тестирования, что делает его более сложным, чем stratified.

CUI-less. Цель состоит в том, чтобы оценить, может ли модель избежать связывания, когда в словаре нет подходящего концепта (в задачах CLEF/SemEval это категория “CUI-less”). В исследовании подмножества, включающие CUI-less случаи, называются “full test set” и “full train set”, в то время как подмножества без CUI-less случаев – “in-KB”.

Для сравнения были использованы следующие модели ранжирования: (1) *Tf-idf*: стандартные представления *tf-idf*, построенные на униграммах и биграммах на уровне символов; (2) *BERT*: многоязычные BERT представления без обучения [7]; это межъязыковая модель, которая не была обучена на биомедицинских текстах; (3) *RuBERT*: BERT [74], обученный на русскоязычной части Википедии и новостных данных; (4) *SapBERT*: модель с метрическим обучением на основе BERT, которая генерирует тройки на основе UMLS для

Таблица 9 — Статистика корпуса.

Выборка	**#сущностей**	**#уник. сущностей**	**#концептов**
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 (англ.) [73]	13609	5979	3792

Таблица 10 — Результаты оценки с фильтрацией тестовой выборки.

Модель	**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%

предварительного обучения [75], а также межъязыковой вариант, обученный на XL-BEL [72].

Кроме того, было использовано несколько вариантов обучения на обучающих выборках, как предложено авторами *BioSyn* [69]: (1) *SapBERT+RuCCoN*, с обучением на целевой выборке; (2) *SapBERT+MCN*, с обучением на MCN; (3) *SapBERT+WRN*, с обучением на наборе данных, извлеченном из медицинской части тезауруса RuWordNet; (4) *SapBERT+XL-BEL*, обучением на русскоязычной части XL-BEL; (5) *SapBERT+RuCCoN+RWMXL-BEL*, на основе комбинации всех трех наборов.

Как показано в таблице 10, SapBERT демонстрирует наилучшие показатели по сравнению с другими моделями и улучшает результаты по мере увеличения количества обучающих выборок. SapBERT, обученный на

RuCCoN, заметно превосходит SapBERT, обученный на других данных, при тестировании на полном тестовой выборке, но разница уменьшается при zero-shot тестировании, указывая на то, что это в основном связано с конкретными сущностями из обучающей выборки. Это подчеркивает важность разметки дополнительных данных для повышения показателей качества, чему способствует RuCCoN для русского языка. Обучение на дополнительных медицинских данных, как правило, полезно, поскольку SapBERT, обученный на английских клинических заметках, показывает наилучшие показатели по сравнению с базовым SapBERT.

В заключении данной работы делается следующий вывод. Был представлен RuCCoN, новый корпус для нормализации клинических концептов на русском языке, размеченный медицинскими работниками и имеющий несколько выборок для обучения/тестирования для объективной оценки в различных условиях. RuCCoN и руководство по разметке доступно по адресу `https://github.com/AIRI-Institute/RuCCoN`.

3.3 NEREL-BIO: вложенные именованные сущности в биомедицинских абстрактах

Результаты этого раздела основаны на статье [23].

В данном разделе описывается NEREL-BIO – схема разметки вложенных биомедицинских сущностей и корпус PubMed абстрактов на русском и на английском языках. Выбор типов биомедицинских сущностей для разметки в NEREL-BIO определяется их появлением в таксономии UMLS и других аннотированных корпусах данных в биомедицинской области. Схема разметки NEREL-BIO включает в себя 17 специализированных типов биомедицинских сущностей (см. табл. 12 для получения подробной информации) в дополнение к 20 типам сущностей из общего корпуса данных NEREL [76].

Были включены следующие общие типы сущностей NEREL: 8 базовых типов сущностей (например, PERSON, ORGANIZATION, LOCATION), 7 числовых сущностей (например, DATE, AGE) и теги для характеристики лиц (NATIONALITY, PROFESSION, FAMILY), PRODUCT и EVENT. Сущность EVENT используется для обозначения таких событий, как эпидемии, военные конфликты и цунами, которые упоминаются в связи с распространением болезней или необходимостью в дополнительной медицинской помощи.

Таблица 11 — Статистика корпуса NEREL-BIO.

Выборка	**#документов**	**#сущностей**	**#Ненулевое число типов**
# рус. абстрактов	766	66,888	37
# англ. абстрактов	105	10,651	32

Таблица 12 — Частоты десяти типов сущностей с вложенностью в полной коллекции на русском языке и в документах на английском.

Тип сущности	**Описание**	**тексты на русском, %**	**англ., %**
FINDING	результаты научных исследований, описанные эксперименты	65.7	71.2
PHYS	биологическая функция или процесс в организме, включая свойство организма (температуру) и исключая психические процессы	38.3	40.7
INJURY POISONING	повреждения, нанесенные организму в результате прямого или косвенного воздействия внешней силы, включая отравление	37.7	49.0
DISO	любые отклонения от нормального состояния организма: заболевания, симптомы, дисфункции, аномалии органа, исключая травмы или отравления	37.3	41.2
DEVICE	изготовленные предметы, используемые в медицинских целях	33.9	42.5
LABPROC	тестирование веществ в организме и другие диагностические процедуры, такие как ультразвуковое исследование	30.2	34.8
MEDPROC	процедуры, связанные с восстановительным лечением заболеваний, включая хирургические вмешательства	30.0	44.7
ANATOMY	органы, части тела, клетки и клеточные компоненты, вещества организма	27.3	28.3
SCIPROC	научные исследования, включая математические методы или клинические исследования, шкалы, классификаторы и т.д.	23.9	32.1
CHEM	химические вещества, включая легальные и нелегальные лекарства, биологические молекулы	22.5	20.1

Стоит отметить, что сущности, аннотированные в NEREL-BIO, могут отсутствовать в UMLS. Например, термин *левосторонняя врожденная диафрагмальная грыжа* отсутствует в UML. Данная фраза размечена следующим образом: [*левосторонняя* [*врожденная* [[*диафрагмальная*]$_{ANATOMY}$

$[грыжа]_{DISO}]_{DISO}]_{DISO}]_{DISO}$ Хотя весь термин не присутствует в UMLS, можно связать фрагменты: Грыжа (C0019270), Диафрагмальная грыжа (C0019284), Дыхательная диафрагма (C0011980), Врожденная диафрагмальная грыжа (C0235833).

Схема разметки была разработана путем многократного предварительного аннотирования параллельных текстов на русском и английском языках. За процесс аннотирования отвечали опытные терминологи, имеющие опыт в терминологических исследованиях, включая биомедицинскую область. Кроме того, модератор просмотрел все аннотированные тезисы для обеспечения точности.

Таблица 11 содержит статистику NEREL-BIO с точки зрения документов и упоминаний сущностей. Таблица 12 содержит частоту появления вложенных сущностей в NEREL-BIO. Таблица включает в себя десять основных типов сущностей и соответствующую им частоту вложенности. Чтобы рассчитать частоту вложенности, количество раз, когда сущностей определенного типа появляется как внешняя сущность (исключая многократные появления одной и той же сущности), было разделено на общее количество появлений типа сущности в корпусе.

Для проведения экспериментов корпус NEREL-BIO был разделен на три выборки: обучающая, валидационная и оценочная, с 612, 77 и 77 документами соответственно. Была обучена модель машинного понимания (MRC) [77] для NER экспериментов на обучающей выборке. Как и ожидалось, в зависимости от типа сущности показатели качества модели MRC сильно варьируется: F1 по ANATOMY, CHEM, DISO сущностям – 83.99%, 81.32%, 81.03%, соответственно, тогда как баллы F1 LABPROC, MEDPROC, DISO сущностям – 66.47%, 73.96%, 60.31%, соответственно.

В заключении данной работы делается следующий вывод. Был предложен корпус NEREL-BIO, первый корпус биомедицинских абстрактов на русском языке, размеченный вложенными сущностями. Предложенная схема разметки демонстрирует, что вложенные сущности обеспечивают более эффективную основу для извлечения связей, которые в противном случае были бы потеряны, а также облегчают более полную привязку сущностей к базам знаний. Корпус доступен по адресу `https://github.com/nerel-ds/NEREL-BIO`.

4 Новые стратегии оценки эффективности моделей

Новые стратегии оценки эффективности моделей были предложены и разработаны автором диссертации [13; 24–26]. Связывание биомедицинских сущностей, таких как химические вещества, болезни, гены и побочные реакции на лекарственные препараты, с терминологией является сложной задачей и часто требует несинтаксической интерпретации. Это связано со сложностью и вариативностью биомедицинского языка, который может включать в себя широкий спектр терминов и сокращений. В частности, в этой главе обсуждаются следующие ключевые научные проблемы:

– Первой научной проблемой, рассмотренной в [24; 25], является отсутствие последовательных и надежных стратегий оценки для связывания сущностей/нормализации концептов. Методы часто оцениваются на тестовых наборах самых разных размеров и предметных подобластей, а также на узкой подвыборке концептов из конкретных терминологий. Кроме того, результаты работы нейронных сетей существенно различаются на разных корпусах, что приводит к различным показателям качества.

– Вторая научная проблема, рассмотренная в [24], заключается в том, что нейронные модели обычно обучаются и оцениваются на сущностях одного типа из одной предметной подобласти. Это ограничивает обобщаемость моделей и затрудняет их повторное использование для различных целей, поскольку для этого требуется кодирование в соответствии с определенной терминологией.

Для решения вышеописанных проблем предложены новые стратегии:

– Для решения первой проблемы одна из предлагаемых стратегий заключается в использовании *стратифицированного* (stratified) разделения выборки для оценки способности систем распознавать известные концепты даже при новых упоминаниях сущностей [13]. Кроме того, была введена процедура *фильтрации* тестовой выборки для оценки “трудных случаев” связывания сущностей и подхода к обучению межъязыкового переноса в zero-shot постановке [25].

– Для решения второй проблемы другая предлагаемая стратегия заключается в использовании стратегии оценки внутри терминологии

(in-terminology), так и стратегии оценки между терминологиями (cross-terminology) для учета разнообразия биомедицинских сущностей и терминологий [24].

Основные результаты и выводы данного раздела следующие:

- Оценка показывает большое расхождение в показателях связывания между официальным разделением на обучающую и тестовую выборки и предложенными *отфильтрованными* наборами тестов, которые представляют уточненные выборки упоминаний сущностей [24; 25].
- Модели машинного обучения, обученные на наборе целевой предметной области, демонстрируют значительно лучшие показатели качества на stratified выборках по сравнению с моделями, обученными на других данных [13; 22].
- Перенос знаний может быть эффективным между болезнями, химическими веществами и генами с небольшим средним снижением показателей при оценке на корпусах научных абстрактов [24].
- Показатели качества моделей с техникой переноса обучения значительно различается в разных областях. Например, на корпусах инструкций лекарств и комментариев пользователей с сущностями типа ADR модели, обученные на отличных корпусах, демонстрируют существенное снижение показателей по сравнению с моделями, обученными специально для целевой области [24].

4.1 Стратегии оценки внутри терминологии и между терминологиями

Результаты этого раздела основаны на статье [24].

Не существует установленных стратегий для оценки моделей на биомедицинских корпусах в различных терминологических контекстах. Модели обычно оцениваются на основе узких подвыборок концептов, и представленные результаты различаются в разных корпусах. Повторное использование обученных моделей для различных терминологий также затруднительно при использовании моделей машинного обучения с учителем. Чтобы решить эти проблемы, в этом исследовании сравниваются корпуса и нейронные архитектуры, использующие BERT для связывания сущностей в трех областях:

абстракты исследований, инструкции лекарств и пользовательские тексты о медикаментозной терапии на английском языке.

В этом исследовании представлена обширная оценка пяти биомедицинских корпусов, вручную аннотированных концептами, касающимися болезней, химических веществ, человеческих генов и побочных реакций на лекарственные препараты (ADR). Используются две модели: (i) базовая, которая ранжирует концепты для данного упоминания сущности путем сравнения биомедицинских векторов BERT [56] с евклидовым расстоянием; (ii) BioSyn [69]. Модели основаны на BioBERT$_{base}$ v1.1, который был предварительно обучен на PubMed абстрактах (всего 4,5 миллиарда слов) за 1 миллион шагов.

Для анализа были использованы общедоступные корпуса с официальными разбиением на обучающую (train), валидационную (dev) и тестовую выборки: NCBI Disease corpus [63], BioCreative V CDR (BC5CDR) [64], BioCreative II GN (BC2GN) [66], TAC 2017 ADR [78], SMM4H 2017 ADR [79]. Анализ корпусов данных показал, что примерно 80% упоминаний сущностей в тестовой выборке являются текстовыми дубликатами других сущностей в тестовой выборке или сущностей, представленных в обучающей и валидационной выборках. Чтобы получить более реалистичные результаты, в этом исследовании представлены усовершенствованные (*refined*) выборки без дубликатов или точных совпадений. Стоит отметить, что некоторые концепты, появляющиеся в тестовом выборке *refined*, также появляются в соответствующем обучающей выборке.

BioSyn был обучен на train/dev выборках каждого корпуса с исходным словарем и протестирован на соответствующем тестовой выборке (in-domain). Междоменная оценка оценивает модели, обученные на исходных (*source*) данных, на тестовых выборках всех других целевых корпусов (т.е. *target*). Как модели ранжирования BioSyn, так и BioBERT извлекают наиболее близкое название концепта в целевом словаре для данного представления сущности. Стоит отметить, что перекрестная терминологическая оценка является сложным сценарием для разработки моделей машинного обучения с учителем, особенно для связывания с концептами, с которыми модели не встречались во время обучения (т.е. zero-shot концептами).

Задача поиска концептов топ-k для каждого упоминания сущности в тексте оценивается как сценарий информационного поиска, где используется словарь названий концептов и их идентификаторов. Точность топ-k определяется как 1, если правильный идентификатор получен в ранге k, в противном

Таблица 13 — Результаты нормализации с использованием единой терминологии в официальных и *refined* тестовых выборках. CDR - это BC5CDR, GN - это BC2GN, M4H - это SMM4H.

Модель	**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

Таблица 14 — Показатели качества BioSyn при оценке внутри терминологии и между терминологиями на *refined* тестовых выборках. Результаты внутри домена отображаются по диагоналям (с фоном темно-серый). Другие ячейки содержат результаты данной модели и различия в результатах между этой моделью и моделью в предметной области в круглых скобках (по строкам). Светло-серые ячейки показывают эксперименты с перекрестной терминологией.

	Train					
Test	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

случае 0. Для составных сущностей показатель определяется как 1, если каждое предсказание для отдельного упоминания является правильным.

Таблица 13 представляет результаты моделей, обученных и оцененных на сущностях одного типа из одного домена в шести выборках. Таблица 14 сравнивает показатели качества BioSyn в задачах нормализации внутри терминологии и между терминологиями. Модели были обучены на обучающем выборке из исходного набора данных (source) и оценены на целевом тестовой выборке (target) с использованием другой терминологии.

Чтобы определить, могут ли текущие тестовые выборки приводить к завышению показателей связывания, были сравнены результаты, полученные моделями как на официальных, так и на *refined* тестовых выборках, как показано в таблице 13. Значительное снижение усредненного показателя acc@1 с 91.8% до 76.7% для BioSyn и усредненного показателя acc@1 с 77.7% до

54.9% для BioBERT подчеркивает большую потребность во внешних оценочных выборках данных, где одни и те же упоминания сущностей не будут использоваться как для обучения, так и для тестирования. Эти наблюдения также означают, что есть возможности для улучшения переносимости разработанных методов, то есть способности поддерживать показатели качества для новых доменов или сущностей.

Таблица 13 помогает ответить на вопрос о том, как поверхностные характеристики упоминаний сущностей влияют на производительность базовой модели на основе BERT. Основываясь на этих результатах, можно сделать следующие выводы. Во-первых, простое ранжирование представлений BioBERT позволяет получить высокие результаты по заболеваниям и химических веществам. В двух refined выборках с более длинными сущностями (NCBI, TAC) и корпусе BC2GN с сущностями с цифрами разница между ранжированием BioBERT и BioSyn значительна (среднее снижение на 23.6%). Качественный анализ показал, что BERT-представления сущностей, отличающихся на одну цифру (например, гены TP53 и TP63), близки в векторном пространстве. Как и ожидалось, результаты по SMM4H значительно ниже, чем по абстрактам, из-за разницы между языком пользователей и профессионалов в области медицины.

Чтобы определить, может ли модель, обученная на одном корпусе, использоваться для связывания сущностей в другом типе или домене в zero-shot настройках, мы сравниваем различия в показателях качества в таблицах 13 и 14. Модели, обученные на данных NCBI, CDR Disease, BC2GN и TAC, работают наравне с моделью, обученной на CDR Chemical train (прибл. 74% в соответствии с acc@1), в то время как модель, обученная на CDR Chemical, показала снижение на 6% в этих подмножествах. BioSyn, обученный на SMM4H, достигает более низких результатов по абстрактам и инструкциям лекарств, чем простое ранжирование BioBERT, в то время как все модели машинного обучения с учителем показали лучшие результаты на данных SMM4H, чем базовая модель BioBERT.

В заключении данной работы делается следующий вывод. В этом исследовании представлена сравнительная оценка корпусов для нормализации медицинских концептов, включая корпуса NCBI Disease, BC5CDR Disease & Chemical, BC2GN Gene, TAC 2017 ADR, SMM4H 2017 ADR. Были оценены две модели на основе архитектуры BERT на шести корпусах, с официальными разбиениями на обучающую и тестовую выборки и новыми *refined* тестовыми

выборками. Оценка выявила значительные различия в показателях качества, указывая на то, что современная модель BioSyn достигла точности на 15% ниже на refined выборках. Была введена задача MCN между терминологиями, демонстрирующая эффективную передачу знаний между болезнями, химическими веществами и генами. Однако модели, обученные на четырех других корпусах, показали плохие результаты на корпусах TAC и SMM4H, точность снизилась на 10.2% и 33.1%, соответственно. Результаты и исходный код доступны на GitHub по адресу `https://github.com/insilicomedicine/Fair-Evaluation-BERT`.

5 Заключение

Основные положения, выносимые на защиту, основаны на опубликованных статьях [11–26].

Исследования [21–23] были сосредоточены на разработке схем аннотирования для задач извлечения биомедицинской информации, а также на создании аннотированных корпусов текстов на английском и русском языках из ряда биомедицинских источников, включая научные абстракты, пользовательские отзывы о лекарствах, электронные медицинские карты и клинические испытания. В результате серии экспериментов была произведена оценка моделей машинного обучения на основе архитектуры BERT на этих корпусах.

В статьях [24–26] были проанализированы ограничения существующих эталонных корпусов в двух задачах с целью улучшения стратегий оценки моделей в этих задачах. Исследование [26] было сосредоточено на анализе эталонных корпусов для извлечения отношений между сущностями в научных текстах и электронных медицинских картах. В работе предлагается нейронная модель с механизмом перекрестного внимания, которая показывает лучшую междоменные показатели качества. В [24] анализируются эталонные корпуса в задаче связывания биомедицинских сущностей и концептов, и предлагаются новые стратегии оценки внутри терминологии и между терминологиями.

В исследованиях [11–18] были предложены несколько нейронных архитектур для различных задач в биомедицинской области. К ним относятся модель DILBERT, которая оптимизирует сходство сущностей и концептов, многоязычные модели архитектуры BERT для распознавания именованных сущностей, основанный на классификации подход к связыванию биомедицинских сущностей, мультимодальные методы для обнаружения побочных реакций на лекарственные препараты, модель архитектуры “кодировщик–декодировщик” для МКБ (ICD) кодирования и модель машинного обучения с признаками для извлечения клинической отношений. Эти модели и методы продемонстрировали высокие показатели качества на нескольких эталонных корпусах извлечения информации и в открытых соревнованиях.

В работах [19;20] разработанные модели были объединены в систему для биомедицинского поиска по абстрактам научных статей и систему извлечения сущностей типа ADE из комментариев пользователей о лекарствах. Эксперименты по поиску в zero-shot постановке, описанные в [19], показали, что

нейронная архитектура предложенной системы демонстрирует наилучшие показатели качества как для запросов о болезнях, так и для запросов о генах. Эксперименты, описанные в [20], показали, что для извлечения сущностей типа ADE из сообщений Твиттера с использованием многокомпонентной архитектуры требуется обучение различных компонентов на основе дисбаланса входных данных, чтобы обеспечить оптимальную производительность в рамках сквозного разрешения.

Основные результаты, представленные на защиту:

– Разработаны новые модели и методы классификации и извлечения информации:

1. Многоязычные модели архитектуры BERT были проанализированы для кросс-доменного распознавания сущностей лекарств и болезней на двух языках. Исследование стратегий переноса обучения между четырьмя корпусами показало эффективность предварительного обучения на данных с одним или обоими типами переноса [11].
2. Методы, основанные на классификационном подходе, с (i) с набором информативных признаков на уровне сущностей и контекста для извлечения отношений [12], (ii) с признаки семантической близости для связывания именованных сущностей [13; 14]. Эффективность этих подходов была продемонстрирована в рамках соревнований SMM4H 2019 Task 3, SMM4H 2020 Task 3 и SMM4H 2021 Task 1c [13; 14], показав наилучшие результаты. Признаки семантической близости также оказались эффективными в глубоких нейронных сетях архитектуры “кодировщик–декодировщик”; предложенная модель показывала наилучшие результаты в рамках соревнования CLEF eHealth 2017 Task 1 [15].
3. Модель DILBERT (Drug and disease Interpretation Learning with Biomedical Entity Representation Transformer) для связывания именованных сущностей, оптимизирующая относительную схожесть сущностей и концептов на основе метрического обучения. Модель устойчива к изменениям в словаре и способна распознавать концепты, не присутствовавшие в обучающей выборке [16; 17].
4. Мультимодальная модель на основе представлений двух моделей архитектуры BERT для языкового моделирования и предсказания

молекулярных свойств в рамках задачи классификации твитов как потенциальных источников нежелательных реакций на лекарства. Модель показывала первые и вторые результаты в рамках соревнований SMM4H 2021 Task 2 и Task 1a, соответственно [18].

5. Две многокомпонентные системы: (i) система для биомедицинского информационного поиска, состоящая из двух моделей, которая показала более высокую производительность по сравнению с традиционной моделью поиска на вручную размеченном наборе данных абстрактов для запросов о заболеваниях и генах [19], и (ii) система для классификации, извлечения и нормализации нежелательных реакций на лекарства на реалистичных, несбалансированных данных; были исследованы подбор оптимальных коэффициентов обучения и метод неполной выборки (undersampling) [20].

– Были представлены новые размеченные корпуса для извлечения информации. Среди них можно выделить следующие:

6. Был создан новый корпус реакций на лекарственные средства (RuDReC), частично аннотированный корпус отзывов потребителей о фармацевтических продуктах на русском языке, а также модели RuDR-BERT для задач распознавания именованных сущностей и классификации предложений [21].

7. Для нормализации концептов были созданы два корпуса: корпус клинических испытаний на английском языке для нормализации лекарств и заболеваний [16; 17], а также корпус RuCCoN - набор электронных медицинских записей на русском языке, в котором сущности связаны с UMLS [22].

8. Был создан NEREL-BIO, корпус научных абстрактов на русском и английском языках и схема аннотирования вложенных именованных сущностей, включающие общие и биомедицинские типы сущностей [23].

– Предложены новые стратегии оценки эффективности моделей.

9. Проведен анализ ограничений существующих датасетов для связывания биомедицинских сущностей, и были предложены новые стратегии оценки моделей: (i) метод *стратифицированной* выборки [13], (ii) стратегии оценки *внутри терминологии* и *между*

терминологиями [24]. Кроме того, были проведены эксперименты в рамках межъязыковой задачи, используя клинические тексты и исследовательские статьи. Была разработана процедура *фильтрации* тестового набора для анализа "сложных случаев" связывания сущностей в условиях zero-shot постановки межъязыкового переноса обучения [25].

10. Проведен анализ ограничений существующих датасетов для извлечения отношений в научных статьях и электронных медицинских записях. Для устранения различий в эффективности моделей *внутри домена* и *вне домена*, был предложена нейронная сеть с кросс-вниманием, демонстрирующая наилучшие результаты в кросс-доменных экспериментах [26].

Благодарность

Автор данной диссертации руководила проектами по биомедицинскому NLP, поддержанными грантом Российского научного фонда #18-11-00284 (2018-2020, 2021-2022), грантом Российского фонда фундаментальных исследований #19-07-01115 (2019-2020), грантом Президента Российской Федерации для молодых ученых-кандидатов наук (МК-3193.2021.1.6, 2021-2022). Данное исследование поддержано этими грантами и грантом Российского научного фонда #20-11-20166 (2020-2022).

Литература

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.