

На правах рукописи

Алимова Ильсеяр Салимовна

**Нейросетевой механизм кросс-внимания в задачах
извлечения информации из текстов на примере
биомедицинских данных**

Специальность 05.13.11 —
«Математическое и программное обеспечение вычислительных
машин, комплексов и компьютерных сетей»

Автореферат
диссертации на соискание учёной степени
кандидата технических наук

Казань — 2021

Работа выполнена в Федеральном государственном автономном образовательном учреждении высшего образования «Казанский (Приволжский) федеральный университет «КФУ».

Научный руководитель: кандидат физико-математических наук
Тутубалина Елена Викторовна

Официальные оппоненты: **Котельников Евгений Вячеславович**,
доктор технических наук, доцент, профессор
кафедры прикладной математики и информатики,
Федеральное государственное бюджетное образовательное учреждение высшего образования «Вятский государственный университет»

Усталов Дмитрий Алексеевич,
кандидат физико-математических наук,
аналитик-разработчик программного обеспечения,
Группа анализа данных и исследований
Обособленного подразделения ООО «Яндекс.Технологии» в г. Санкт-Петербург

Ведущая организация: Автономная некоммерческая организация
высшего образования «Университет Иннополис»

Защита состоится 10 июня 2021 г. в 12 часов на заседании диссертационного совета Д 002.087.01 при Федеральном государственном бюджетном учреждении «Институт системного программирования им. В.П. Иванникова Российской академии наук» по адресу: 109004, г. Москва, ул. А. Солженицына, дом 25.

С диссертацией можно ознакомиться в библиотеке и на сайте Федерального государственного бюджетного учреждения науки «Институт системного программирования им. В.П. Иванникова Российской академии наук».

Автореферат разослан «__» _____ 2021 года.

Ученый секретарь
диссертационного совета
Д 002.087.01,
кандидат физ.-мат. наук

Зеленов С. В.

Общая характеристика работы

Актуальность темы. В настоящее время в связи с бурным развитием сети Интернет и электронных коллекций научных публикаций накоплен огромный объем неструктурированной информации, представленной текстами на естественных языках. В связи с этим становится все более востребованной разработка автоматических методов обработки текстов с целью извлечения структурированных данных, которые в дальнейшем могут быть использованы для извлечения фактов, анализа мнений пользователей, поиска информации и других задач.

Важнейшими задачами автоматического извлечения информации из текстов являются классификация и извлечение отношений. Задача классификации текста заключается в группировке текстов по определенным заранее заданным классам. Задача классификации может рассматриваться на уровне документа, параграфа, предложения и сущности. Задача извлечения отношений заключается в определении семантических связей между двумя понятиями (сущностями).

Методы классификации текста и извлечения отношений применяются для решения широкого круга прикладных задач, включающих анализ мнений пользователей о продукте, фильтрации спама, подбор контекстной рекламы, автоматическое реферирование, тегирование контента на сайте, сортировка обращений пользователей в службы поддержки, построение лингвистических баз данных и т.д. Одной из наиболее значимых областей применения методов классификации и извлечения отношений являются задачи медицинской науки, в частности, задачи фармакологии и персонализированной медицины.

В области биомедицины классификация текста применяется для поиска новой информации о побочных реакциях лекарственных веществ, не указанных в инструкции, а также для обнаружения использования лекарств с нарушением предписаний инструкции¹. Задача извлечения отношений применяется для построения биомедицинских баз данных, определения действия лекарственных веществ по отношению к системам организма, извлечения новых отношений между лекарствами и симптомами для построения гипотез. Информация, полученная в результате извлечения отношений, может также использоваться как часть входных данных для других задач, таких как извлечение событий, классификация диагнозов, принятие клинических решений и в вопросно-ответных системах.

Классификация текста и извлечение отношений между сущностями в биомедицинских текстах исследовались в трудах российских и зарубежных учёных, таких как Саркер А, Гонсалес Г., Гинн Р., Никфарджам

¹Utilizing social media data for pharmacovigilance: a review / A. Sarker [и др.] // Journal of biomedical informatics. 2015. т. 54. с. 202–212.

А., Карими С., Патки А., Кириченко С., Лиу Х., Шелманов А., Браславский П. и других авторов. Перечисленными авторами разработаны основные теоретические аспекты анализа биомедицинских текстов на естественном языке. В работах² приведен развернутый обзор существующих автоматических методов классификации текста и поиска отношений между сущностями в биомедицинских текстах. Согласно рассмотренным исследованиям, подходы машинного обучения обладают большим потенциалом для исследуемых задач, однако для реальных биомедицинских приложений результаты необходимо улучшать³. Существующие методы разработаны для текстов определенного домена: социальные медиа, электронные медицинские карты (ЭМК) или научные тексты статей и тестируются в рамках одного текстового корпуса, что является существенным недостатком в вопросе применимости моделей для реальных практических задач⁴. Кроме того, исследования в основном ведутся для английского языка и почти отсутствуют работы для русского языка. Таким образом, подтверждена актуальность разработки методов классификации текста и извлечения отношений.

Объектом исследования являются неструктурированные тексты на естественном языке, включающие: отзывы пользователей о лекарственных препаратах, тексты твитов, аннотации научных статей, электронные медицинские карты (ЭМК). **Предметом** исследования выступают задачи классификации текста и извлечения отношений между сущностями.

Целью диссертационной работы является разработка методов и программных средств для классификации сущностей и извлечения отношений между сущностями из текстов. Разрабатываемые методы и программные средства должны удовлетворять следующим требованиям:

²Detecting adverse events for patient safety research: a review of current methodologies / H. J. Murff [и др.] // Journal of biomedical informatics. 2003. т. 36, № 1/2. с. 131–143; Adverse drug reaction identification and extraction in social media: a scoping review / J. Lardon [и др.] // Journal of medical Internet research. 2015. т. 17, № 7; Text mining for adverse drug events: the promise, challenges, and state of the art / R. Harpaz [и др.] // Drug safety. 2014. т. 37, № 10. с. 777–790; *ONYE, S. C.* Review of Biomedical Relation Extraction / S. C. ONYE, A. AKKELES, N. DIMILILER // European International Journal of Science and Technology. 2017. т. 6, № 1; *Bui, Q.* Relation extraction methods for biomedical literature. т. 27 / Q. Bui. 2011; *Alshuwaier, F.* A comparative study of the current technologies and approaches of relation extraction in biomedical literature using text mining / F. Alshuwaier, A. Areshey, J. Poon // Engineering Technologies and Applied Sciences (ICETAS), 2017 4th IEEE International Conference on. IEEE. 2017. с. 1–13.

³Overview of the First Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE 1.0) / A. Jagannatha [и др.] // Drug safety. 2018. с. 1–13.

⁴*Ramponi, A.* Cross-Domain Evaluation of Edge Detection for Biomedical Event Extraction / A. Ramponi, B. Plank, R. Lombardo // Proceedings of The 12th Language Resources and Evaluation Conference. 2020. с. 1982–1989.

- более высокие оценки качества классификации сущностей и извлечения отношений предложенных методов по сравнению с существующими моделями;
- переносимость методов на тексты различных языков, в данной диссертационной работе рассматриваются тексты на русском и английском языках;
- переносимость методов на тексты различных доменов с различной языковой моделью, в данной работе рассматриваются тексты отзывов пользователей о лекарствах, твиты о здоровье, аннотации научных статей по теме биомедицины и ЭМК.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. Разработать методы на основе нейронной сети с кросс-вниманием для задачи классификации сущностей;
2. Провести исследования для извлечения наиболее информативных признаков для задачи классификации сущностей;
3. Интегрировать наиболее информативные признаки в разработанную модель на основе нейронных сетей для задачи классификации сущностей;
4. Разработать методы на основе нейронной сети с кросс-вниманием для задачи извлечения отношений между сущностями;
5. Получить размеченные текстовые коллекции на русском и английском языках, состоящие из отзывов пользователей о лекарственных препаратах, твитов о здоровье, научных статей по теме биомедицины и ЭМК;
6. Реализовать предложенные методы в виде программного средства и провести экспериментальные исследования с целью определения качества работы методов и моделей с использованием коллекций текстовых документов.

Научная новизна данной диссертационной работы заключается в следующем:

1. Предложена модель классификации сущностей на основе нейронной сети с механизмом кросс-внимания, отличающаяся от существующих моделей набором дополнительных информативных признаков.
2. Предложена модель классификации отношений сущностей на основе нейронной сети с механизмом кросс-внимания, отличающаяся от существующих моделей разделением контекста и сущностей на отдельные подсети формирования контекстных векторных представлений.

Практическая значимость. В диссертации разработана программная система классификации сущностей и извлечения отношений, основанная на предложенных нейросетевых моделях с кросс-вниманием, которая

предназначена для использования в качестве инструмента автоматического анализа текстовых корпусов. Разработанная система может быть использована как для автоматического извлечения информации из биомедицинских текстов, так и из текстов другой тематики.

Методология и методы исследования. В данной диссертационной работе применялись методы обработки естественного языка, машинного обучения, теории вероятностей и оптимизации.

Основные положения, выносимые на защиту:

1. Предложен и реализован новый метод классификации сущностей, основанный на нейронной сети с механизмом кросс-внимания и набором информативных признаков.
2. Предложен и реализован новый метод извлечения отношений между сущностями, основанный на нейронной сети с механизмом кросс-внимания и с разделением контекста и сущностей на отдельные подсети формирования контекстных векторных представлений.
3. Разработано программное обеспечение SAFEC для задачи классификации сущностей и проведено экспериментальное исследование, обосновывающее улучшение качества предложенных методов по сравнению с существующими алгоритмами в рамках корпусов из одного домена и корпусов из разных доменов.
4. Разработан программный комплекс CARE для задачи извлечения отношений между сущностями и проведено экспериментальное исследование, обосновывающее улучшение качества предложенных методов по сравнению с существующими алгоритмами в рамках текстовых корпусов из разных доменов.

Достоверность подтверждается корректным применением выбранного математического аппарата, экспериментами, проведенными в соответствии с общепринятыми стандартами, взаимосвязью данных экспериментов и научных выводов, сделанных в работе, результатами апробации алгоритмов и разработанной программной системы.

Апробация работы. Основные результаты работы докладывались на следующих конференциях:

1. Летней школе по информационному поиску RuSSIR (г. Екатеринбург, Россия, 21–25 августа 2017 г.);
2. 6-й международной конференции по анализу изображений, сетей и текстов АИСТ (г. Москва, Россия, 27–29 июля 2017 г.);
3. 8-й открытой конференции ИСП РАН имени В.П. Иванникова (г. Москва, Россия, 30 ноября - 1 декабря, 2017 г.);
4. 52-й ежегодной конференции “ESCI Annual Scientific Meeting of the European Society for Clinical Investigation” (г. Барселона, Испания, 30 мая - 1 июня, 2018 г.);

5. Международной научной конференции “Artificial Intelligence and Natural Language Conference” (г. Санкт-Петербург, Россия, 17-19 октября, 2018 г.);
6. 9-й открытой конференции ИСП РАН имени В.П. Иванникова (г. Москва, Россия, 22 - 23 ноября, 2018 г.);
7. Восточноевропейской летней школе по машинному обучению EEML (г. Бухарест, Румыния, 1-6 июля, 2019 г.)
8. 8-й международной конференции по анализу изображений, сетей и текстов АИСТ (г. Казань, Россия, 17–19 июля 2019 г.);
9. 57-й конференции “Association for Computational Linguistics” (г. Флоренция, Италия, 28 июля - 3 августа 2019 г.);
10. 1-м международном саммите “EurNLP” (г. Лондон, Великобритания, 11 октября 2019 г.);
11. 22-й международной научной конференции “Data Analytics and Management in Data Intensive Domains” (г. Казань, Россия, 17-18 октября 2019 г.).

Личный вклад. Все представленные в диссертации результаты получены лично автором.

Публикации. Основные результаты по теме диссертации изложены в 14 печатных изданиях, 2 из которых изданы в журналах, рекомендованных ВАК, 10 — в периодических научных журналах, индексируемых Web of Science и Scopus, 2 — в тезисах докладов.

В работах [1–8] автором проведено исследование предметной области, выполнен основной объем теоретических и экспериментальных исследований, изложенных в публикациях, Тутубалиной Е.В. и Соловьеву В.Д. принадлежит постановка задачи и практические рекомендации для выполнения работы. В работе [9] автором диссертации был проведен анализ полученных результатов, осуществлен поиск соответствующих примеров и структурирование исследования. В работе [10] автором были проведены эксперименты по классификации твитов на наличие побочных эффектов. В работе [11] автор принимал участие в аннотировании и подготовке финальной версии корпуса к публикации. В работах [12; 13] автором были проведены все экспериментальные исследования.

Объем и структура работы. Диссертация состоит из введения, четырех глав, заключения и одного приложения. Полный объем диссертации составляет 146 страниц текста, включая 22 рисунка и 28 таблиц. Список литературы содержит 182 наименования.

Содержание работы

Во **введении** обоснована актуальность диссертационной работы, сформулированы цель и задачи представляемой работы, сформулирована

научная новизна исследований, показана практическая значимость работы, представлены выносимые на защиту научные положения.

Первая глава посвящена описанию основных подходов, применяемых в задачах извлечения информации из текста. Особое внимание уделяется рассмотрению существующих методов для решения задач классификации сущностей и извлечения отношений для текстов биомедицинской тематики. Целью данной главы является анализ достоинств и недостатков существующих методов классификации текстов на уровне сущностей и извлечения отношений.

Разработку системы извлечения информации из текста можно разделить на следующие четыре этапа: препроцессинг текста, извлечение признаков, выбор модели и оценка. Препроцессинг текста включает в себя удаление шумовых данных, разбику текста на токены и приведение слов к нормальной форме. На шаге извлечения признаков текстовые данные векторизуются, то есть переводятся в числовые значения, которые характеризуют исходные данные. Наиболее распространенные способы векторизации: мешок слов и векторное представление слов. Один из самых важных шагов в задаче извлечения информации - это выбор лучшей модели. В ранних работах в качестве моделей применялись классические методы на основе машинного обучения, такие как логистическая регрессия, наивный байесовский метод, метод опорных векторов, деревья решений. В более современных работах применяются сверточные и рекуррентные нейронные сети. В последнее время наиболее распространенными являются языковые модели на основе архитектуры Трансформер. На последнем шаге разработки системы извлечения информации из текста проводится оценка качества разработанной системы.

В настоящий момент большинство исследований по классификации сущностей и извлечению отношений между сущностями из биомедицинских текстов сводятся к использованию методов машинного обучения, где требуется сформировать вектор признаков, или к применению нейронных сетей. Среди недостатков рассмотренных моделей можно выделить три основных. Во-первых, методы разработаны и протестированы для текстов определенного домена: социальные медиа, ЭМК или научные тексты статей. Различие лексиконов и синтаксических конструкций предложений в текстах с разных ресурсов может существенно повлиять на качество результатов модели. Во-вторых, существующие модели тестируются в рамках одного корпуса и практически отсутствуют работы в которых обучение и оценка моделей проводилась бы на разных корпусах. В-третьих, существующие методы, в основном, разработаны для английского языка, при этом для русского языка исследования в данной области практически не проводились. Таким образом, для обеих задач подтверждается актуальность разработки кросс-доменных методов и проведения исследований для русского языка.

Во второй главе описан метод классификации сущностей, основанный на нейронной сети с механизмом кросс-внимания и набором информативных признаков. Нейронная сеть состоит из двух слоев LSTM для представления контекста и сущности и слоев с кросс-вниманием, объединенные выходы которых передаются в полносвязный слой для классификации. В данной главе также приводятся результаты оценки качества предложенного метода на пяти англоязычных и одном русскоязычном корпусе, состоящих из текстов отзывов о лекарственных препаратах, ЭМК пациентов, научных статей и твитов о здоровье.

Формальная постановка задачи. Каждая текстовая коллекция состоит из набора текстовых документов: $D = \{d_1, d_2, \dots, d_n\}$, где $d_i = \{s_{i1}, s_{i2}, \dots, s_{i|d_i|}\}$, s_{ij} - предложение документа. Каждое предложение состоит из последовательностей токенов: $w_{ij}^k: s_{ij} = \{w_{ij}^1, w_{ij}^2, \dots, w_{ij}^n\}$, включающих в себя слова, числа, знаки препинания. Формально, обозначим как $e_{ij} = \{w_{ij}^u, \dots, w_{ij}^v\}$, $u, v \in [1, k]$ - сущность, состоящую из одного или более токенов в рамках одного предложения. Постановка задачи звучит следующим образом: необходимо определить класс сущности $a(e_{ij})$ для всех сущностей из документов текстовой коллекции $d_{ij} \in [1, \dots, |d_i|]$, $i \in [1, \dots, |D|]$. В данной работе рассматривается бинарная классификация, то есть $a \in [0, 1]$, где $a(e_{ij}) = 1$, если e_{ij} обозначает побочный эффект и $a(e_{ij}) = 0$ в противном случае. Например, в предложении “это лекарство очень хорошее, боли почти прошли”, сущность “боль” - показание к приему лекарства, соответственно, $a(\text{“боль”}) = 0$. При этом в предложении “Не рекомендую лекарство, если вы не готовы справиться с болью, которое оно вызывает”, сущность “болью” обозначает побочный эффект ($a(\text{“болью”}) = 0$).

Модель. Для решения поставленной задачи была разработана модель LSTM+CA+feat, основанная на нейронной сети LSTM+CA (Cross-attention LSTM), принимающей на вход векторное представление слов и набор наиболее информативных признаков.

Архитектура нейронной сети LSTM+CA состоит из слоя с короткой долгосрочной памятью (англ. long short-term memory; LSTM) и механизма кросс-внимания. Исходные тексты сущностей и контекстов кодируются в виде векторного представления слов и подаются на вход LSTM слоям. Далее вычисляется среднее значение векторов скрытых состояний сущности и контекста. На основе полученных значений вычисляется два вектора кросс-внимания: среднего значения сущности относительно выхода слоя LSTM для контекста и наоборот среднего значения контекста относительно выхода слоя LSTM для сущности. Полученные вектора внимания передаются на вход полносвязному слою с функцией активации softmax для дальнейшей классификации.

Для выявления наиболее значимых признаков для задачи классификации сущностей был реализован классификатор на основе метода SVM.

В качестве входных данных для модели были сгенерированы два вида признаков: на уровне сущностей, когда признак основывается исключительно на словах, входящих в сущность, и контекстные, когда признак базируется на заданном количестве слов слева и справа от сущности, в данном случае были взяты пять слов слева и справа сущности. Исходный код модели доступен в открытом репозитории⁵.

Признаки на уровне контекста:

- *Мешок слов* (bow) представляет из себя вектор равный по длине количеству уникальных слов в тексте, на каждой позиции которого указано количество вхождений слова в текст.
- *Вектор частей речи* (pos). Для каждого контекста было посчитано количество существительных, глаголов, наречий и прилагательных.
- *Тональные признаки* (sent). В качестве признаков были подсчитаны классические числовые характеристики, применяемые в задачах анализа тональности[10]. Тональные признаки основаны на наиболее распространенных оценочных словарях SentiWordNet, MPQA Subjectivity, Bing Liu’s для английского языка и RuSentiLex для русского языка.
- *Поточечная взаимная информация* (pmi) подсчитывает количество вхождений слова в позитивный и негативный контексты.

Признаки на уровне сущностей:

- *Векторное представление слов* (emb). В качестве признака был использован усредненный вектор всех слов сущности, полученный из предобученной модели векторного представления слов.
- *Векторное представление на основе кластеров* (cls). Каждое слово было представлено в виде бинарного вектора длины 150, где 1 обозначает принадлежность к кластеру с соответствующим номером.
- *Семантические типы из унифицированной системы медицинского языка* (umls).

Признаки PMI, представление на основе кластеров, векторное представление слов были обучены на корпусе Health [11] для английского языка и неразмеченных отзывах о лекарственных препаратах корпуса RuDReC [6].

Эксперименты проводились на корпусах CADEC, и RuDReC. Модели оценивались на 5-ти фолдовой кросс-валидации. Результаты представлены в таблице 1. Согласно полученным результатам классификатор показал максимальный результат 80.3% F-меры на корпусе CADEC и 77.6% F-меры на корпусе RuDReC. Наиболее информативные признаки для англоязычного корпуса: мешок слов, части речи, тональные, вектора кластеризации и признаки на основе концептов из словаря UMLS. Для русскоязычного

⁵https://github.com/Ilseyar/adr_classification

Таблица 1 — Оценка информативности признаков на корпусах CADEC и RuDReC для метода на основе SVM по усредненным метрикам: P (точность), R (полнота) и F (F-мера).

Features	CADEC			RuDReC		
	P	R	F	P	R	F
bow	.827	.740	.775	.637	.631	.632
bow, pos	.824	.743	.776	.635	.642	.636
bow, pos, sent	.823	.745	.777	.633	.639	.634
bow, pos, sent, cls	.832	.776	.788	.646	.652	.648
bow, pos, sent, cls, umls	.844	.773	.803	.684	.691	.686
bow, pos, sent, cls, umls, pmi	.839	.770	.799	.705	.719	.710
bow, pos, sent, cls, umls, pmi, emb	.822	.777	.797	.772	.785	.776

корпуса также эффективными оказались признаки PMI и векторное представление слов.

Общая архитектура модели LSTM+CA+feat, представленная на рисунке 1, основана на нейронной сети LSTM+CA с набором дополнительных признаков, которые конкатенируются с финальным полносвязным слоем сети. В качестве признаков были использованы наиболее информативные признаки для задачи классификации биомедицинских сущностей: мешок слов, части речи, кластеры Брауна, основанные на тональных словарях и поточечная взаимная информация. Эксперименты показали, что векторное представление слов также дает прирост при классификации сущностей, однако, поскольку векторное представление слов используется для представления сущности и контекста, было решено не добавлять данный признак дополнительно в модель. Исходный код модели доступен в открытом репозитории⁶.

Корпуса. Оценка моделей проводилась на русскоязычном корпусе RuDReC [11] и пяти англоязычных корпусах: Twitter⁷, CADEC⁸, MADE⁹, Twimed¹⁰ и PsyTAR¹¹. Общая статистика для всех корпусов представлена

⁶https://bitbucket.org/Ilseyar/entity_classification/src/master/

⁷Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features / A. Nikfarjam [и др.] // Journal of the American Medical Informatics Association. 2015. т. 22, № 3. с. 671–681.

⁸Cadec: A corpus of adverse drug event annotations / S. Karimi [и др.] // Journal of biomedical informatics. 2015. т. 55. с. 73–81.

⁹NLP Challenges for Detecting Medication and Adverse Drug Events from Electronic Health Records (MADE1.0) / F. Liu [и др.]. 2018. URL: <https://bio-nlp.org/index.php/projects/39-nlp-challenges>.

¹⁰Alvaro, N. TwiMed: Twitter and PubMed comparable corpus of drugs, diseases, symptoms, and their relations / N. Alvaro, Y. Miyao, N. Collier // JMIR public health and surveillance. 2017. т. 3, № 2.

¹¹A systematic approach for developing a corpus of patient reported adverse drug events: A case study for SSRI and SNRI medications / M. Zolnoori [и др.] // Journal of biomedical informatics. 2019. т. 90. с. 103091.

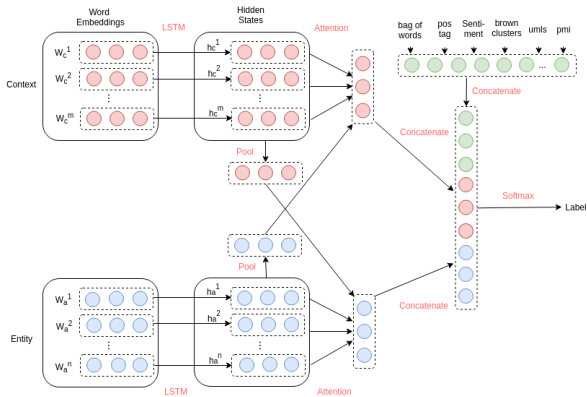


Рис. 1 — общая архитектура модели LSTM+CA+feat для классификации типов сущностей

Таблица 2 — Суммарная статистика по корпусам. ADR - количество классов с побочным эффектом, non-ADR - количество классов с отсутствием побочного эффекта

Корпус	Источник	Кол-во документов	ADR	non-ADR	Макс. длина предложения	Средняя длина предложения
Twitter	Твиттер	645	569	76	37	22
CADEC	Отзывы	1231	5770	550	236	28
MADE	ЭМК	876	1506	37077	173	21
Twimed-Pubmed	Научные статьи	1000	264	983	150	39
Twimed-Twitter	Твиттер	637	329	308	42	27
PsyTAR	Отзывы	891	4525	2987	264	32
RuDReC	Отзывы	400	1500	558	165	19

в таблице 2. Как видно из статистики, корпуса CADEC и MADE содержат большее кол-во аннотаций, чем остальные корпуса. Наибольшее количество документов в корпусе CADEC - 1231, наименьшее количество в корпусе RuDReC - 400. Максимальная средняя длина предложения в корпусе Twimed-PubMed - 39 слов, минимальная в корпусе RuDReC - 19 слов. Стоит отметить, что все корпуса, кроме Twimed-Twitter не сбалансированы. MADE - наиболее несбалансированный корпус, всего 4% примеров из класса ADR.

Базовые модели. Для оценки эффективности разработанной модели извлечения побочных эффектов результаты сравнивались с существующими базовыми подходами классификации текста, показавшими наилучшие результаты в задаче извлечения побочных эффектов:

- SVM с наиболее информативными признаками (SVM+feat);

- сверточная нейронная сеть с двунаправленной короткой долгосрочной памятью (CNN-BiLSTM)¹²;
- модели на основе CNN¹³: сверточная нейронная сеть (CNN), рекуррентная сверточная нейронная сеть (RCNN), сверточная рекуррентная нейронная сеть (CRNN), сверточная сеть с вниманием (CNNA);
- модель на основе архитектуры Трансформер BERT¹⁴, для корпусов на английском языке применялась модель BioBERT (версия Base v1.0 (+ PubMed 200K + PMC 270K))[20], для корпуса на русском применялась модель RuBERT¹⁵.

Оценка в рамках одного корпуса. Все модели были оценены на 5-фолдовой кросс-валидации с помощью стандартных метрик оценки качества классификации: точность (P), полнота (R) и F-мера – среднее гармоническое между точностью и полнотой. Результаты экспериментов представлены в таблице 3.

Таблица 3 — Результаты метрики средней F-меры для задачи классификации в рамках одного корпуса. Жирным шрифтом выделены максимальные значения каждой метрики.

Модель	Twitter	CADEC	MADE	Twimed-Twitter	Twimed-PubMed	PsyTAR	RuDReC
SVM+feat	.749	.802	.772	.758	.834	.850	.776
CNN	.537	.722	.721	.692	.709	.857	.717
RCNN	.569	.731	.760	.722	.744	.859	.765
CRNN	.559	.739	.737	.694	.703	.837	.723
CNNA	.532	.723	.699	.669	.621	.855	.492
CNN-BiLSTM	.533	.761	.751	.736	.824	.862	.784
BERT	.589	.746	.784	.797	.872	.881	.776
LSTM+CA	.794	.815	.786	.819	.874	.812	.855
LSTM+CA+feat	.823	.821	.788	.825	.878	.888	.864

Разработанная модель LSTM+CA+feat превзошла результаты остальных моделей на всех корпусах. Среди базовых подходов на англоязычных корпусах наилучшие результаты на корпусах CADEC и Twitter

¹²Zhang, M. Adverse Drug Event Detection Using a Weakly Supervised Convolutional Neural Network and Recurrent Neural Network Model / M. Zhang, G. Geng // Information. 2019. т. 10, № 9. с. 276.

¹³Adverse Drug Reaction Classification With Deep Neural Networks / Т. Huynh [и др.] // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016. с. 877–887.

¹⁴BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin [и др.] // 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). 2019. с. 4171–4186.

¹⁵Deepavlov: Open-source library for dialogue systems / M. Burtsev [и др.] // Proceedings of ACL 2018, System Demonstrations. 2018. с. 122–127.

достигла модель SVM+feat, а на остальных корпусах - модель BioBERT. На корпусе RuDReC лучшие результаты среди базовых методов были получены моделью CNN-BiLSTM (78.4%). На корпусах CADEC и Twitter модель LSTM+CA+feat превзошла базовую модель SVM+feat на 7.4% и 1.9% F-меры соответственно. Максимальный прирост F-меры модели LSTM+CA+feat по сравнению с BioBERT достигнут на корпусе Twitter (+23.4%). Данный результат показывает, что модель LSTM+CA+feat способна лучше обучиться на несбалансированных корпусах малого размера, чем модель BioBERT. Наименьший прирост был достигнут на корпусе MADE (+0.4%). На корпусе RuDReC модель LSTM+CA+feat превзошла результаты CNN-BiLSTM на 8% F-меры. Прирост результатов на остальных корпусах по сравнению с моделью BioBERT варьируется от 0.7% до 2.8%.

Добавление признаков в модель LSTM+CA+feat позволило повысить качество классификации модели на всех корпусах. Наибольший прирост при добавлении признаков в модель был получен на корпусах PsyTAR (+7.6%) и Twitter (+2.9%). На остальных корпусах прирост по F-мере варьируется от 0.2% до 0.9% F-меры. Стоит отметить, что модель с признаками обучается быстрее и требует на 5 эпох меньше по сравнению с моделью без признаков.

Кросс-доменная оценка. В таблице 4 представлены результаты кросс-доменной оценки моделей в задаче классификации сущностей для англоязычных корпусов. В данном наборе экспериментов модели обучались на одном корпусе (source), а оценивались на целевом (target). В качестве базовой модели для сравнения была выбрана модель BioBERT, поскольку данная модель показала наилучшие результаты среди базовых моделей при оценке в рамках одного корпуса.

Модель LSTM+CA+feat превосходит модель BioBERT по среднему значению F-меры на всех корпусах, кроме MADE, где модели показывают одинаковую среднюю F-меру - 39.4%. Наибольший прирост модели LSTM+CA+feat по сравнению с BioBERT в среднем был достигнут при обучении моделей на корпусе CADEC (+4.3%), наименьший при обучении на корпусе PsyTAR (+0.2%). Стоит отметить, что при кросс-доменной оценке модель BioBERT превзошла модель LSTM+CA+feat на корпусе Twimed-PubMed, независимо от корпуса, на котором модели обучались. Это обусловлено тем, что модель BioBERT предобучалась на текстах научных статей. Максимальный прирост модели LSTM+CA+feat по сравнению с BioBERT был достигнут в эксперименте Twimed-PubMed-MADE (+19.5%).

Полученные результаты показывают, что модель LSTM+CA+feat более эффективна при оценке в рамках одного домена и в случае, когда

Таблица 4 — Результаты кросс-доменной оценки моделей для задачи классификации по метрике средней F-меры. Результаты в рамках одного домена на диагонали. Среднее - это усредненное значение F-меры для всех кросс-доменных экспериментов. Средняя потеря - это разница между результатами в рамках одного домена и среднего значения.

Жирным шрифтом выделены максимальные значения каждой метрики.

Target→ Source↓	Модель	Twitter	CA-DEC	MA-DE	Twimed-Twitter	Twimed-PubMed	Psy-TAR	Среднее	Средняя потеря
Twitter	LSTM+CA+feat	.823	.580	.108	.638	.414	.558	.459	-.334
	BioBERT	.589	.531	.048	.626	.493	.473	.434	-.155
CADEC	LSTM+CA+feat	.719	.821	.284	.722	.670	.617	.602	-.213
	BioBERT	.592	.746	.203	.666	.781	.553	.559	-.187
MADE	LSTM+CA+feat	.242	.112	.788	.429	.696	.491	.394	-.392
	BioBERT	.269	.129	.784	.513	.743	.319	.394	-.390
Twimed-Twitter	LSTM+CA+feat	.653	.647	.247	.825	.765	.697	.602	-.217
	BioBERT	.659	.676	.119	.797	.837	.623	.583	-.214
Twimed-PubMed	LSTM+CA+feat	.518	.570	.365	.728	.878	.754	.587	-.287
	BioBERT	.491	.662	.170	.735	.872	.679	.547	-.325
PsyTAR	LSTM+CA+feat	.577	.643	.293	.602	.600	.888	.543	-.269
	BioBERT	.533	.667	.168	.624	.711	.881	.541	-.340

обучающие и тестовые данные из разных доменов, что в свою очередь доказывает большую возможность практического применения модели LSTM+CA+feat по сравнению с существующими моделями.

В третьей главе представлен метод извлечения отношений между сущностями, основанный на нейронной сети с механизмом кросс-внимания и с разделением контекста и сущностей на отдельные подсети формирования контекстных векторных представлений. Каждая подсеть состоит из слоя LSTM, формирующего семантическое представление входных текстов. К полученным выходам LSTM слоев применяется механизм кросс-внимания: вычисляется два вектора внимания относительно каждой сущности и контекста. Полученные вектора внимания конкатенируются и подаются на вход полносвязному слою с функцией активации softmax для дальнейшей классификации. В данной главе также приводятся результаты оценки качества предложенного метода на четырех англоязычных и одном русскоязычном корпусе, состоящих из текстов ЭМК пациентов и научных статей.

Формальная постановка задачи. Каждая текстовая коллекция состоит из набора документов $D = \{d_1, d_2, \dots, d_n\}$. В каждом документе выделен некоторый набор сущностей $e^1 = \{w_1^1, w_2^1, \dots, w_{|e^1|}^1\}$, $e^2 = \{w_1^2, w_2^2, \dots, w_{|e^2|}^2\}$, ... $e^n = \{w_1^n, w_2^n, \dots, w_{|e^n|}^n\}$. Сущности e_i и e_j находятся в отношении, если в тексте указано на их семантическое взаимодействие или влияние друг на друга. Постановка задачи звучит следующим образом: для каждой пары сущностей, принадлежащих одному документу: $e_i^k, e_j^k \in d_k, i, j \in [1, |e^k|], i \neq j$ необходимо определить, есть ли между ними

отношение $r(e_i^k, e_j^k) = 1$ или нет $r(e_i^k, e_j^k) = 0$. Например, в предложении “Пациент принимал 4 дня Руксиенс и в последний день Циклофосфамид” сущности “4 дня” и “Руксиенс” связаны друг с другом, в то время как сущности “4 дня” и “Циклофосфамид” не связаны. Таким образом, $r(\text{“4 дня”}, \text{“Руксиенс”}) = 1$, а $r(\text{“4 дня”}, \text{“Циклофосфамид”}) = 0$

Модель. Для поставленной задачи в модель LSTM+CA был добавлен дополнительный вход, который принимает вторую сущность. Архитектура сети LSTM+CA представлена на рисунке 2. Сеть LSTM+CA имеет три входных слоя. Два входных слоя принимают на вход закодированные векторным представлением сущности относительно которых производится классификация на наличие отношения. Третий входной слой принимает на вход контекст между сущностями, закодированный векторным представлением слов и признаком позиции. Для определения признака позиции строятся два вектора, в которых определяется относительное расстояние от каждого токена в контексте до каждой сущности. Если токен в контексте стоит в тексте после сущности, то относительная позиция является положительным числом, в противном случае позиция характеризуется отрицательным числом. На основе полученных относительных позиций строились векторные представления каждой позиции длины 5. Изначально вектора инициализировались случайным образом, затем наиболее оптимальные значения векторов подбирались в процессе обучения сети. Исходный код модели доступен в открытом репозитории¹⁶.

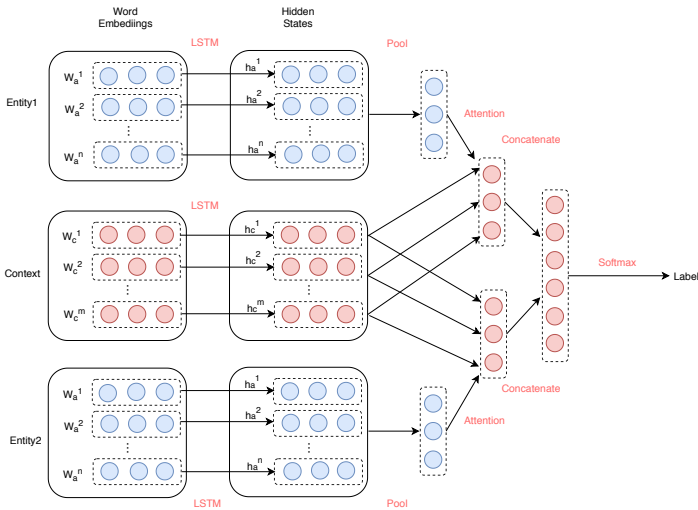


Рис. 2 — общая архитектура модели LSTM+CA для извлечения отношений

¹⁶<https://bitbucket.org/Ilseyar/relation-extraction/src/master/>

Выходы всех входных слоев проходят через LSTM слои, которые формируют семантическое представление входных текстов. Далее к полученным выходам LSTM слоев применяется механизм кросс-внимания. Для этого вычисляется среднее значение вектора скрытых слоев LSTM для каждой сущности. На основе полученных значений вычисляются вектора внимания среднего значения каждой сущности относительно выхода слоя LSTM для контекста. Полученные вектора внимания объединяются и передаются на вход полносвязному слою с функцией активации softmax для дальнейшей классификации.

Корпуса. Эксперименты по оценке эффективности методов извлечения отношений проводились на четырех англоязычных корпусах: MADE¹⁷ и i2b2¹⁸, состоящих из текстов ЭМК, на корпусах CDR¹⁹ и Phaedra²⁰, состоящих из аннотаций научных статей медицинской тематики, и на размеченном русскоязычном корпусе текстов клинических карточек RuClinical²¹. Общая статистика корпусов представлена в таблице 5. Статистика включает в себя: количество отношений, среднее и максимальное значение длины контекста между сущностями (в токенах), типы отношений.

Базовые модели. В качестве базовых моделей было рассмотрено две архитектуры нейронных сетей: сверточная сеть с механизмом внимания (Attention-based convolutional neural network; CNNA)²², BERT. Для корпусов на английском языке применялась модель BioBERT, для корпуса на русском применялась модель RuBERT.

Оценка в рамках одного корпуса. Результаты оценки моделей в рамках одного корпуса представлены в таблице 6. Наибольшие результаты по средней F-мере на всех англоязычных корпусах показала модель BioBERT. Наиболее высокие результаты были получены на корпусе MADE (95.4%). LSTM+CA превзошла результаты модели CNNA на всех корпусах. Наибольший прирост модели LSTM+CA по отношению к CNNA среди англоязычных корпусов был достигнут на корпусе Phaedra (+5%). На корпусе

¹⁷NLP Challenges for Detecting Medication and Adverse Drug Events from Electronic Health Records (MADE1.0) / F. Liu [и др.]. 2018. URL: <https://bio-nlp.org/index.php/projects/39-nlp-challenges>.

¹⁸2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text / Ö. Uzuner [и др.] // Journal of the American Medical Informatics Association. 2011. т. 18, № 5. с. 552–556.

¹⁹Overview of the BioCreative V chemical disease relation (CDR) task / C.-H. Wei [и др.] // Proceedings of the fifth BioCreative challenge evaluation workshop. т. 14. 2015.

²⁰Annotation and detection of drug effects in text for pharmacovigilance / P. Thompson [и др.] // Journal of cheminformatics. 2018. т. 10, № 1. с. 37.

²¹Shelmanov, A. Information extraction from clinical texts in Russian / A. Shelmanov, I. Smirnov, E. Vishneva // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”. т. 14. 2015. с. 537–549.

²²Shen, Y. Attention-based convolutional neural network for semantic relation extraction / Y. Shen, X.-J. Huang // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016. с. 2526–2536.

Таблица 5 — Суммарная статистика по корпусам. Максимальное и среднее расстояние посчитано в токенах между сущностями. Жирным шрифтом выделены максимальные значения каждой метрики.

Модель	Источ-ник	Кол-во отношений	Среднее расстояние	Максимальное расстояние	Типы отношений
MADE	ЭМК	27145	29.9	981	лекарство-побочный эффект /курс/доза/частота приема/ продолжительность/показания к применению; симптомы-тяжесть
CDR	Научные статьи	4513	14.8	394	заболевание-химическое вещество
PHAEDRA	Научные статьи	1136	15.0	262	пациент-заболевание; эквивалентные заболевания кореференция
i2b2	ЭМК	9413	3.7	73	медицинская проблема - лечение/ тест/медицинская проблема
RuClinical	ЭМК	1601	1.4	48	заболевание-тяжесть/стадия /часть тела; симптом-часть тела

Таблица 6 — Результаты метрики F-меры для задачи извлечения отношений в рамках одного корпуса

Корпус	CDR	Phaedra	i2b2	MADE	RuClinical
CNNA	.663	.810	.784	.906	.601
BERT	.741	.918	.829	.954	.762
LSTM+CA	.666	.860	.799	.917	.841

RuClinical модель LSTM+CA показала самые высокие результаты F-меры (84.1%), превзойдя модель RuBERT на 7.9% и модель CNNA на 24%. Таким образом, полученные результаты показывают эффективность разработанной модели для русского языка.

Результаты кросс-доменной оценки корпусов. В таблице 7 представлены результаты кросс-доменной оценки моделей в задаче извлечения отношений. В качестве базовой модели для сравнения была выбрана модель BioBERT, поскольку данная модель показала наилучшие результаты при оценке моделей в рамках одного корпуса. Разработанная модель LSTM+CA превосходит модель BioBERT по среднему значению F-меры на всех корпусах. Наибольший прирост модели LSTM+CA по сравнению с BioBERT в среднем был достигнут при обучении модели на корпусе Phaedra (+3.5%), наименьший при обучении на корпусе CDR (+0.5%). Максимальный прирост модели LSTM+CA по сравнению с BioBERT был достигнут в эксперименте CDR-MADE (+7.4%). Наименьший прирост в эксперименте CDR-Phaedra (+0.01%). Полученные результаты показывают, что модель LSTM+CA более эффективна в случае, когда обучающие и тестовые данные из разных доменов, что в свою очередь доказывает

Таблица 7 — Результаты кросс-доменной оценки моделей для задачи извлечения отношений по метрике средней F-меры. Результаты в рамках одного домена на диагонали. Среднее - это усредненное значение F-меры для всех кросс-доменных экспериментов. Средняя потеря - это разница между результатами в рамках одного домена и среднего значения.

Target→ Source↓	Модель	CDR	Phaedra	i2b2	MADE	Среднее	Средняя потеря
CDR	LSTM+CA	.666	.520	.514	.590	.541	-.125
	BioBERT	.741	.519	.574	.516	.536	-.205
Phaedra	LSTM+CA	.467	.860	.509	.471	.482	-.378
	BioBERT	.399	.918	.477	.465	.447	-.471
i2b2	LSTM+CA	.579	.628	.799	.541	.583	-.216
	BioBERT	.583	.593	.829	.523	.566	-.263
MADE	LSTM+CA	.525	.468	.606	.917	.533	-.374
	BioBERT	.579	.446	.556	.954	.527	-.427

большую возможность практического применения модели LSTM+CA по сравнению с BioBERT.

В четвертой главе приведено описание архитектуры программных комплексов SAFEC для классификации сущностей и CARE для извлечения отношений, а также временная оценка обучения и оценка затрачиваемого объема памяти созданных с применением данных пакетов моделей.

Программные комплексы предоставляют функционал для обучения и оценки моделей нейронных сетей с механизмом кросс-внимания LSTM+CA для задач классификации сущностей и извлечения отношений. Программные комплексы написаны на языке Python 3.6 и предоставлены в открытом доступе. Программные комплексы состоят из следующих общих модулей: (i) модуль слоев нейронных сетей; (ii) модуль моделей; (iii) модуль признаков; (iv) модуль загрузки данных; (v) модуль запуска обучения и тестирования.

Оценки по памяти и по времени на каждом корпусе для программных комплексов SAFEC и CARE представлены в таблице 8, соответственно. Временная оценка напрямую зависит от количества обучающих примеров. Средняя скорость обучения в программном комплексе SAFEC, примерно, 500 примеров в секунду. В архитектуре нейронной сети LSTM+CA в программном комплексе CARE имеется дополнительный слой с механизмом внимания, поэтому средняя скорость обучения 400 примеров в секунду меньше, чем в SAFEC. Объем занимаемой памяти на видеокарте напрямую зависит от длины контекста. Согласно приведенным оценкам, время, затрачиваемое на обучение для обеих задач не превосходит 60 минут, что позволяет обучать сети для новых корпусов в довольно сжатые сроки и является преимуществом разработанных программных комплексов по сравнению с существующими языковыми моделями.

Таблица 8 — Оценка времени обучения модели LSTM+CA+feat в программном комплексе SAFEC и модели LSTM+CA в программном комплексе CARE. Формат времени: mm:ss.

Корпус	Кол-во обучающих примеров	Макс. длина контекста	Объем памяти на видеокарте (MiB)	Время 1 эпоха	Время всего
CAFEC					
CADEC	5056	236	763	00:12	02:00
MADE	30852	173	727	04:50	48:20
PsyTAR	5523	264	775	00:16	02:40
Twimed-PubMed	996	150	701	00:02	00:20
Twimed-Twitter	472	42	677	00:01	00:10
Twitter	512	37	679	00:01	00:10
RuDReC	1262	165	711	00:02	00:20
CARE					
CDR	8593	394	1059	00:34	05:40
PHAEDRA	8150	262	801	00:21	03:30
i2b2	83382	73	745	02:21	23:30
MADE	205604	981	3075	05:48	58:00
RuClinical	48180	48	526	00:26	04:20

В **заключении** приведены основные результаты работы, которые заключаются в следующем:

1. Предложен и реализован новый метод классификации сущностей, основанный на нейронной сети с механизмом кросс-внимания и набором информативных признаков.
2. Предложен и реализован новый метод извлечения отношений между сущностями, основанный на нейронной сети с механизмом кросс-внимания и с разделением контекста и сущностей на отдельные подсети формирования контекстных векторных представлений.
3. Разработано программное обеспечение SAFEC и проведено экспериментальное исследование, обосновывающее улучшение качества предложенных методов по сравнению с существующими алгоритмами в рамках корпусов из одного домена и корпусов из разных доменов для задачи классификации сущностей.
4. Разработано программное обеспечение CARE и проведено экспериментальное исследование, обосновывающее улучшение качества предложенных методов по сравнению с существующими алгоритмами в рамках текстовых корпусов из разных доменов для задачи извлечения отношений между сущностями.

Дальнейшие перспективы развития исследований могут быть связаны с (i) проведением мультязычных экспериментов, в которых языки

обучающего и оценочного корпуса различаются; (ii) применением подходов удаленного обучения (distant supervision); (iii) разработкой модели с применением многозадачного обучения (multi-task learning).

Публикации автора по теме диссертации

1. *Алимова, И. С.* Сравнительный анализ нейронных сетей в задаче классификации побочных эффектов на уровне сущностей в англоязычных текстах / И. С. Алимова, Е. В. Тутубалина // Труды Института системного программирования РАН. — 2018. — т. 30, № 5.
2. *Alimova, I.* Automated detection of adverse drug reactions from social media posts with machine learning / I. Alimova, E. Tutubalina // International Conference on Analysis of Images, Social Networks and Texts. — Springer. 2017. — P. 3–15.
3. A Machine Learning Approach to Classification of Drug Reviews in Russian / I. Alimova [и др.] // Ivannikov ISPRAS Open Conference (ISPRAS), 2017. — IEEE. 2017. — с. 64–69.
4. *Alimova, I.* A Comparative Study on Feature Selection in Relation Extraction from Electronic Health Records. / I. Alimova, E. Tutubalina // DAMDID/RCDL. — 2019. — с. 34–45.
5. *Alimova, I.* Detecting Adverse Drug Reactions from Biomedical Texts With Neural Networks / I. Alimova, E. Tutubalina // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. — 2019. — с. 415–421.
6. *Tutubalina, E.* Biomedical Entities Impact on Rating Prediction for Psychiatric Drugs / E. Tutubalina, I. Alimova, V. Solovyev // International Conference on Analysis of Images, Social Networks and Texts. — Springer. 2019. — с. 97–104.
7. *Alimova, I.* Multiple features for clinical relation extraction: A machine learning approach / I. Alimova, E. Tutubalina // Journal of Biomedical Informatics. — 2020. — т. 103. — с. 103382.
8. *Alimova, I.* Interactive Attention Network for Adverse Drug Reaction Classification / I. Alimova, V. Solovyev // Artificial Intelligence and Natural Language. — Springer. 2018. — с. 36–48.
9. Using semantic analysis of texts for the identification of drugs with similar therapeutic effects / E. Tutubalina [и др.] // Russian Chemical Bulletin. — 2017. — т. 66, № 11. — с. 2180–2189.

10. *Miftahutdinov, Z.* KFU NLP Team at SMM4H 2019 Tasks: Want to Extract Adverse Drugs Reactions from Tweets? BERT to The Rescue / Z. Miftahutdinov, I. Alimova, E. Tutubalina // Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task. — 2019. — с. 52–57.
11. The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews / E. Tutubalina [и др.] // Bioinformatics. — 2020. — июль. — URL: <https://doi.org/10.1093/bioinformatics/btaa675>.
12. *Nugmanov, R.* Detecting adverse drug reactions from user reviews with machine learning / R. Nugmanov, I. Alimova, E. Tutubalina // European Journal of Clinical Investigation. т. 48. — Wiley. 2018. — с. 217–218.
13. *Nugmanov, R.* Adverse drug reactions identification in social media posts and electronic health records with neural networks / R. Nugmanov, I. Alimova, E. Tutubalina // European Journal of Clinical Investigation. т. 49. — Wiley. 2019. — с. 116–117.
14. *Алимова, И.* Метод извлечения отношений из текстов биомедицинской тематики на основе контекстного представления / И. Алимова // Вестник Казанского государственного технического университета им. А.Н. Туполева. — 2019. — июль. — т. 4. — с. 84–93.

Алимова Ильсеяр Салимовна

Нейросетевой механизм кросс-внимания в задачах извлечения информации из
текстов на примере биомедицинских данных

Автореф. дис. на соискание ученой степени канд. тех. наук

Подписано в печать _____._____._____. Заказ № _____

Формат 60×90/16. Усл. печ. л. 1. Тираж 100 экз.

Типография _____