DOI: 10.15514/ISPRAS-2026-38(2)-11

Экспериментальное исследование моделей на основе инструкций для извлечения предметно-ориентированных сущностей из студенческих отчетов

[1] *А.В. Мельникова, ORCID:0009-0006-1011-5225 <a.v.melnikova@utmn.ru>*
[1] *М.С. Воробьева, ORCID: 0000-0002-1508-4089 <m.s.vorobeva@utmn.ru>*
[1,3] *А.В. Глазкова, 0000-0001-8409-6457 <a.v.glazkova@utmn.ru>*
[2,3] *Д.А. Морозов, ORCID: 0000-0003-4464-1355 <morozowdm@gmail.com>*

[1] *Тюменский государственный университет,*
Россия, 625003, г. Тюмень, ул. А. Володарского, д. 6.

[2] *Новосибирский государственный университет,*
Россия, 630090, г. Новосибирск, ул. Пирогова, д. 1.

[3] *Национальный корпус русского языка,*
Россия, 119019, г. Москва, Гоголевский бульвар, д. 2.

Аннотация. В работе исследовалась задача извлечения из студенческих отчетов ИТ-направлений предметно-ориентированных сущностей (ПОС), являющихся ключевыми терминами, навыками, именованными сущностями, отражающими тематическую специфику текста. В качестве решений рассматривались инструмент извлечения ключевых слов rutermextract, дообученная языковая модель mBART, большие языковые модели, управляемые инструкциями (YandexGPT, Saiga, Tlite). Дообучение mBART эффективно при достаточном объеме данных. Модели на инструкциях превзошли rutermextract, перспективны при малых объемах данных, особенно Saiga, выявляющая ядро сущностей. Выявлено, что стратегия выделения ПОС в тексте точнее, чем извлечение в виде списка. Однако задача требует дополнительных исследований: ошибочное извлечение ПОС (67-89%), проявляющееся в отсутствии пересечений с эталонными ПОС, указывает на трудности моделей в отделении ядра сущности от контекста. Основные ограничения – малый корпус (2933 текста) и простые инструкции. Перспективы исследования: детализированные инструкции, оценка подходов в других областях и типах текстов.

Ключевые слова: предметно-ориентированные сущности; извлечение сущностей; обработка естественного языка; предварительно обученные языковые модели; модели на основе инструкций; генеративные языковые модели; анализ отчетных документов; обучение на основе инструкций.

Для цитирования: Мельникова А.В., Воробьева М.С., Глазкова А.В., Морозов Д.А. Экспериментальное исследование моделей на основе инструкций для извлечения предметно-ориентированных сущностей из студенческих отчетов. Труды ИСП РАН, том 38, вып. 2, 2026 г., стр. 165–182. DOI: 10.15514/ISPRAS–2026–38(2)–11.

Благодарности: Исследование выполнено при поддержке Министерства науки и высшего образования Российской Федерации в рамках государственного задания (FEWZ-2024-0052).

Experimental Study of Instruction-Based Models for Extracting Domain-Specific Entities from Student Reports

[1] *A.V. Melnikova, ORCID:0009-0006-1011-5225 <a.v.melnikova@utmn.ru>*
[1] *M.S. Vorobeva, ORCID: 0000-0002-1508-4089 <m.s.vorobeva@utmn.ru>*
[1,3] *A.V. Glazkova, 0000-0001-8409-6457 <a.v.glazkova@utmn.ru>*
[2,3] *D.A. Morozov, ORCID: 0000-0000-0000-0000 <morozowdm@gmail.com>*

[1] *University of Tyumen,*
6, Volodarskogo st., Tyumen, 625003, Russia.

[2] *Novosibirsk State University,*
1, Pirogova str., Novosibirsk, 630090, Russia

[3] *Russian National Corpus,*
2, Gogolevsky Boulevard, Moscow, 119019, Russia

Abstract. This work investigated the task of extracting domain-specific entities from student reports in the field of information technology. Domain-specific entities (DSE) represent key terms, skills, and named entities that reflect the thematic specifics of the text. The solutions evaluated included the keyword extraction tool rutermextract, a fine-tuned mBART language model, and instruction-tuned large language models (YandexGPT, Saiga, Tlite). The study found that fine-tuning mBART is effective given a sufficient volume of data. Instruction-based models outperformed rutermextract and show promise for low-data scenarios, with the Saiga model being particularly effective at identifying the core set of entities. The strategy of highlighting domain-specific entities within the text was found to be more accurate than extracting them as a simple list. However, the task requires further research: the high rate of erroneous extraction of domain-specific entities (67-89%), manifested as a complete lack of overlap with the gold-standard entities, indicates the models' difficulty in separating the core entity from its context. The main limitations of the study are the small corpus size (2,933 texts) and the use of simple instructions. Promising research directions include developing more detailed instructions and evaluating the approaches in other domains and text types.

Keywords: domain-specific entities; entity extraction; natural language processing; pre-trained language models; instruction-based models; generative language models; report document analysis; instruction tuning.

For citation: Melnikova A.V., Vorobeva M.S., Glazkova A.V., Morozov D.A. Experimental Study of Instruction-Based Models for Extracting Domain-Specific Entities from Student Reports. Trudy ISP RAN/Proc. ISP RAS, vol. 38, issue 2, 2026, pp. 165-182 (in Russian). DOI: 10.15514/ISPRAS-2026-38(2)-11.

Acknowledgements. The study was supported by the Ministry of Science and Higher Education of the Russian Federation within the framework of the state assignment (FEWZ-2024-0052).

1. Введение

Задача извлечения определенной информации из неструктурированных текстовых данных имеет фундаментальное значение для проведения комплексного анализа в широком спектре предметных областей, к ним относятся медицина (идентификация симптомов, диагнозов), юриспруденция (выявление правовых норм), сфера образования (извлечение элементов, предметной терминологии, учебных целей), а также финансовый сектор и кибербезопасность (мониторинг рыночных тенденций и детектирование угроз) [1, 2]. Извлечение предметно-ориентированных сущностей (ПОС) из текстов автоматизирует обработку информации, повышает точность поиска, улучшает понимание текста, персонализирует контент и анализирует тренды, что необходимо для эффективного принятия решений и поддержки экспертных систем. Повсеместное распространение ПОС обусловлено их способностью служить строительными блоками для построения структурированных баз знаний, что позволяет системам не только «знать», но и «рассуждать», устанавливая связи между сущностями.

В области образовательных технологий все большее значение приобретает задача извлечения предметно-ориентированных сущностей из студенческих отчетов, что позволяет проводить автоматизированный анализ цифрового следа обучающихся [3]. Под ПОС подразумеваются ключевые термины, навыки, именованные сущности и другие элементы, отражающие тематическую специфику текста. Например, в тексте *«Необходимо предусмотреть возможность будущего быстрого масштабирования с ростом нагрузки проекта, для решения этих проблем выбран Docker Swarm»* технология Docker Swarm является ПОС как выбранное решение для конкретной задачи (масштабирование) в контексте реализации проекта. Та же технология может упоминаться в виде общей справочной информации, в таком случае она не является ПОС (*«Docker Swarm – это встроенный инструмент оркестрации контейнеров, разработанный компанией Docker»*). Автоматизация выявления ПОС открывает возможности для решения ряда практических задач: от персонализации образовательных траекторий до мониторинга формирования профессиональных компетенций.

В настоящем исследовании рассматривается альтернативный подход – использование моделей, основанных на инструкциях (instruction-based models), которые не требуют полного дообучения и способны решать задачу извлечения ПОС, генерируя ответ на соответствующую инструкцию (промпт). В частности, сравниваются две стратегии формирования инструкций: извлечение ПОС, при котором сущности извлекаются в виде списка, и выделение ПОС, где исходный текст переписывается с явным выделением сущностей специальными маркерами по аналогии с задачей поиска именованных сущностей. Эти подходы позволяют оценить, насколько качество извлечения зависит от формата инструкции, и выявить наиболее эффективный метод для работы с образовательными документами.

Актуальность работы обусловлена дефицитом открытых размеченных корпусов текстов для извлечения ПОС и необходимостью разработки решений, способных демонстрировать устойчивую эффективность при минимальном объеме обучающих данных.

В работе проводится сравнение эффективности моделей, основанных на инструкциях, с дообученной моделью mBART, показавшей лучший результат в предыдущем исследовании по извлечению ПОС [4] и потому включенной в сравнение в качестве эталонного решения. Кроме того, для сравнения используется инструмент rutermextract [5] для подбора ключевых слов, демонстрирующий возможности простых статистических подходов, не требующих обучения на целевом наборе данных. Модели, основанные на инструкциях, представляют между этими двумя подходами компромисс, адаптированный под конкретную задачу, но не требующий больших объемов размеченных данных для дообучения.

В рамках данного экспериментального исследования представлен расширенный набор данных с размеченными ПОС, а также проведен сравнительный анализ моделей, основанных на инструкциях, с дообученными трансформерами и традиционными методами. Результаты показывали, что модели, основанные на инструкциях, демонстрируют существенное преимущество перед методами извлечения ключевых слов, хотя и уступают тонко настроенным трансформерным моделям по рассмотренным метрикам. Кроме того, в исследовании показано, какие подходы к формированию инструкций для моделей являются более эффективными для извлечения ПОС. Полученные результаты позволяют сформулировать практические рекомендации по выбору моделей в зависимости от специфики задачи.

2. Современное состояние проблемы

Задача извлечения ПОС связана с задачами извлечения терминов, навыков и ключевых слов из текстов на естественном языке. Для решения данных задач в анализе русскоязычных

текстов применяются как традиционные лингвистические методы, так и современные подходы с использованием машинного обучения.

Традиционные методы извлечения терминов включают подходы, основанные на лексико-грамматических шаблонах [6-7] и статистической оценке [8]. В более поздних работах применяются методы глубокого обучения. Так, в работе [9] предлагается нейросетевая архитектура на основании представлений текстов из модели BERT [10]. Авторами предлагается подход к слабо контролируемому обучению на основе сбора словаря терминов, разметки текстов этими терминами и обучению модели на размеченных таким образом текстах. В статье [11] представлено сравнение дообученных моделей трансформерной архитектуры на материале текстов научных статей. Лучший результат для извлечения терминов показали модели, дополненные множеством эвристик. Авторы работы [12] используют BERT в качестве бинарного классификатора терминов-кандидатов, предварительно извлеченных из текста. В рамках соревнования по извлечению терминов из русскоязычных текстов RuTermEval [13] наиболее высокое качество показала архитектура, изначально предложенная для поиска именованных сущностей (NER) и использующая векторные представления текстов из предварительно обученной языковой модели и принцип контрастного обучения (F-мера 79,4%) [14].

В извлечении навыков из текстов вакансий доминируют предварительно обученные языковые модели (BERT, SkillNER), однако их качество снижается при работе с терминами, отсутствующими в обучающих данных [3, 15]. Для русскоязычных текстов применяют машинный перевод [15], синтаксический анализ [16], а сравнение архитектур, представленное в работе [17], выявило преимущество моделей, обученных по принципу NER (F-мера 81%).

Несмотря на активное использование традиционных подходов к извлечению ключевых слов [18-19], в настоящее время методы подбора ключевых слов развиваются в сторону их генерации с помощью нейронных сетей [20]. В работе [21] показано, что мультиязычная модель mT5 превосходит классические подходы по метрикам BERTScore (76,89%) и F-мера (11,24%). В статье [22] ChatGPT и аналоги сравниваются с классическими статистическими методами и методами машинного обучения, демонстрируя потенциал генеративных моделей.

Тренд использования моделей глубокого обучения, в том числе моделей, основанных на инструкциях, является общим для российских и зарубежных исследований. В работе [23] проводится сравнение моделей, основанных на инструкциях, с дообученными моделями и подходами, основанными на правилах, для задачи извлечения терминов из юридических документов. Показано, что, хотя модели, основанные на инструкциях, уступают дообученным моделям, их качество значительно повышается при использовании примеров в инструкции. В исследовании [24] сравнивается два подхода к формированию инструкции к модели извлечения навыков из англоязычных текстов. Показано, что выделение навыков с помощью тегов в исходном тексте работает эффективнее, чем извлечение навыков в виде списка.

Таким образом, в последние несколько лет доминирующим трендом стал переход от классических подходов к моделям глубокого обучения. Перспективным направлением в рассмотренных задачах являются модели, основанные на инструкциях, которые показывают значительный потенциал при работе со специализированными текстами, несмотря на некоторое отставание от дообученных аналогов в отдельных задачах.

3. Постановка задачи

Дано:

- неструктурированный текст D, представляющий последовательность слов: $D = (w_1, w_2, \dots, w_n)$, где n - количество слов в тексте;

- предметная область O, определяющая множество предметно-ориентированных сущностей: $O = \{o_1, o_2, \dots, o_m\}$, где m - количество предметно-ориентированных сущностей.

Найти:

Множество предметно-ориентированных сущностей $E = \{e_1, e_2, \dots, e_k\}$, где k - количество предметно-ориентированных сущностей из неструктурированного текста D, которые соответствуют ПОС предметной области O.

Каждая сущность e_i – это слово или последовательность слов, встречающееся в тексте D, то есть для каждого $e_i \in E$ существуют два индекса - начальный ($start$) и конечный (end), такие что:

- индексы $start$ и end находятся в диапазоне от 1 до n: $start, end \in [1, n]$;
- для индексов $start$ и end допустимо соотношение: $start \leq end$;
- сущность e_i является последовательностью слов из текста D, начиная со слова под номером $start$ и заканчивая словом под номером end:

$$e_i = concat\,(w_{start}, w_{start+1}, \dots, w_{end}).$$

Каждая сущность e_i релевантна предметной области O, то есть e_i напрямую совпадает (или является допустимым вариантом написания/формой) с хотя бы одним элементом o_j из предметной области O, что проверяется соответствием двух строковых значений, которое может включать точное совпадение, нормализованное совпадение (регистр символов алфавита, стемминг, лемматизация), совпадение по синонимам или аббревиатурам.

4. Данные

Данное исследование базируется на коллекции документов из работы [4], которая была дополнена и расширена, и включает 300 студенческих отчетных документов, разделенных на 2195 фрагментов длиной в один абзац и содержащих размеченные предметно-ориентированные сущности, с общим числом 1460 уникальных ПОС.

Разметку документов выполняла команда, состоящая из экспертов-преподавателей Тюменского государственного университета и студентов ИТ-направлений. Эксперты обладают опытом преподавания на ИТ-направлениях («Математическое обеспечение и администрирование информационных систем», «Прикладная информатика», «Информационные системы и технологии», «Компьютерная безопасность», «Информационная безопасность автоматизированных систем»).

Процесс контроля качества был организован по двухуровневой схеме.

- Первичная разметка. Студенты идентифицировали и аннотировали сущности в текстовых документах.
- Верификация экспертами. Эксперты проводили выборочную проверку размеченных данных (около 30% документов). В случае обнаружения ошибок эксперт самостоятельно вносил исправления.

Разметке подлежали только те предметно-ориентированные сущности ИТ-сферы, которые были непосредственно использованы в работе, а не просто упомянуты.

Среди выделенных сущностей встречаются названия технологий/языков программирования (PHP, JavaScript, NodeJS и так далее), систем/сервисов (MySQL, PostgreSQL, Redis, Apache, Nginx и так далее), концепций/архитектур (веб-приложение, микросервис, API, серверная часть, JSON, фреймворк и так далее) и другие виды сущностей.

Для повышения репрезентативности и увеличения объема данных, а также для охвата смежных ИТ-направлений, исходный набор данных был дополнен отчетными документами

по выпускным квалификационным работам ИТ-направлений «Прикладная информатика» и «Информационные системы и технологии» за 2023/24 и 2024/25 учебные годы.

Расширение набора данных позволило увеличить количество уникальных ПОС в наборе данных (к 1460 ПОС добавилось 194 новых), для разметки текстовых отчетов использовались теги [TAG*] и [*TAG] для определения начала и конца ПОС. Итоговый размер набора данных составил 2933 текстов, описание набора данных представлено в табл. 1 и на рис. 1. Изначальный набор данных был разбит на обучающую и тестовую выборки случайным образом в соотношении 70:30, новые данные были добавлены в тестовую выборку для более точной оценки обобщающей способности моделей.

Табл. 1. Характеристики набора данных.
Table 1. Dataset characteristics.

Характеристика	Обучающая выборка	Тестовая выборка
Количество текстов	1692	1241
Средняя длина текстов (в словах)	28.04±19.54	28.00±19.40
Среднее количество предметно-ориентированных сущностей	1.51±1.07	1.84±1.48
Средняя длина предметно-ориентированных сущностей (в словах)	1.39±0.78	1.39±0.73

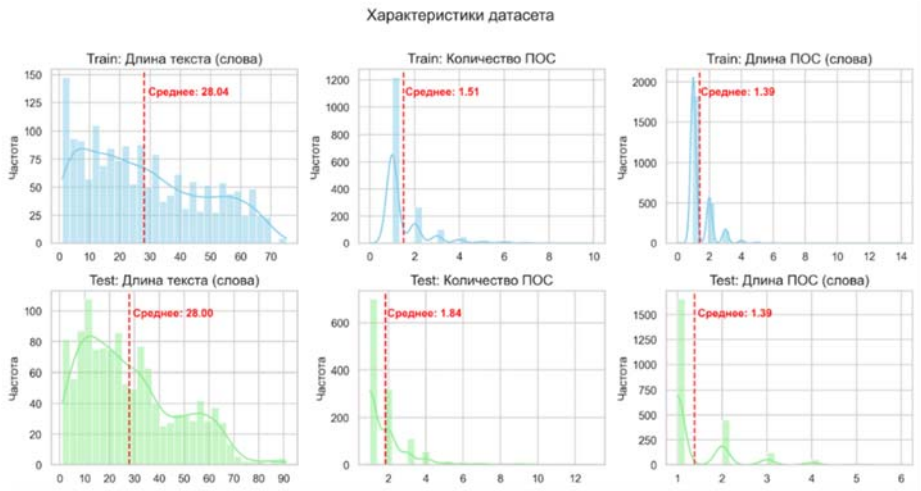

Рис. 1. Распределение значений характеристик набора данных.
Fig. 1. Distribution of dataset characteristic values.

5. Модели

В данной работе проводится сравнение трех предварительно обученных языковых моделей, обучающихся на основе инструкций (в технике prompt learning): YandexGPT, Saiga, Tlite. Модель YandexGPT [25] включает 8 млрд параметров с размером контекста 32 тыс. токенов токенизатора модели. Предварительное обучение модели проводилось в два этапа: сначала на наборе русскоязычных и англоязычных данных, в основном включающих в себя веб-документы, потом на наборе высококачественных данных из нескольких предметных

областей. Модель Saiga-Llama3 (Saiga) [26] имеет 8 млрд параметров (размер контекста – 8 тыс. токенов токенизатора модели). Модель основана на Llama-3 8B Instruct [27] и дообучена на сгенерированном русскоязычном наборе данных в формате пользовательских чатов. Tlite включает 8 млрд параметров и имеет размер контекста – 8 тыс. токенов токенизатора модели. Предварительное обучение модели было выполнено в два этапа: первый этап обучения проводился на наборе преимущественно русскоязычных данных разных предметных областей, для второго этапа использовались русскоязычные и англоязычные данные с более сбалансированным распределением доменов.

Для каждой модели были оценены два подхода к формированию инструкций для поиска предметно-ориентированных сущностей, основанные на методике, предложенной в работе [24]. Первый подход направлен на извлечение предметно-ориентированных сущностей из текста в виде списка. В рамках второго подхода от модели требовалось выделить предметно-ориентированные сущности в тексте с помощью тегов [TAG*] и [*TAG]. В каждом случае инструкция к модели сопровождалась десятью примерами. Текст инструкций и ожидаемый вывод модели показан в табл. 2.

Для запуска моделей использовался сервер, имеющий следующие характеристики: GPU – NVIDIA RTX 4090 (24 ГБ); процессор – AMD Ryzen 9 7900X (12 ядер, 24 потока, 4.7 ГГц); ОЗУ – DDR5 96 ГБ. Генерация текста выполнялась с использованием температуры модели, равной 0.1, и максимальной длины сгенерированного текста, равной 256 токенам токенизатора модели. Температура 0.1 обеспечивает почти детерминированный вывод, необходимый для корректной разметки сущностей, но одновременно вводит небольшой уровень вариативности, который помогает избежать типичных артефактов полностью детерминированного режима (в частности, генерации повторяющихся токенов). Данный выбор согласуется с подходом, использованным в предыдущих исследованиях для схожих задач структурированного извлечения сущностей [4].

Табл. 2. Подходы к формированию инструкции для поиска ПОС.
Table 2. Approaches to the formation of instructions for searching for DSE.

Подход	Инструкция	Ожидаемый вывод
Исходный текст: «Качество работы обученных классификаторов на тестовой выборке оценено с помощью метрик Precision, Recall, F1-score (macro averaging), результаты представлены в таблице 3.»		
Извлечение ПОС	Дан фрагмент студенческого отчета. Извлеки упоминания предметно-ориентированных сущностей в виде списка, разделенного запятыми.	*Precision, F1-score, macro averaging*
Выделение ПОС	Дан фрагмент студенческого отчета. Выдели упоминания предметно-ориентированных сущностей тегами [TAG*] и [*TAG]. Примеры: Текст: … Ответ: … …	*Качество работы обученных классификаторов на тестовой выборке оценено с помощью метрик [TAG*]Precision[*TAG], [TAG*]Recall[*TAG], [TAG*]F1-score[*TAG] ([TAG*]macro averaging[*TAG]), результаты представлены в таблице 3.*

Результаты моделей, обучающихся на основе инструкций, были сравнены с результатами двух существующих решений: mBART [28] и rutermextract. mBART представляет собой

модель с архитектурой sequence-to-sequence, базирующуюся на архитектуре BART [29]. Данная модель показала лучший результат для задачи поиска предметно-ориентированных сущностей в работе [4]. Для дообучения использовались следующие параметры: количество эпох обучения – 20, максимальная длина последовательности – 256 токенов токенизатора модели, скорость обучения – 4e-5, размер батча – 8, оптимизатор – AdamW. Модель была дообучена для задачи выделения ПОС: на вход подавался исходный фрагмент без тегов [TAG*] и [*TAG], требовалось сгенерировать текст с выделенными тегами [TAG*] и [*TAG], обозначающими вхождения предметно-ориентированных сущностей. Примененный в работе инструмент rutermextract, предназначенный для извлечения ключевых слов, функционирует на основе статистического анализа текстовых данных. В ходе исследования инструмент использовался со стандартными настройками, обеспечивающими выделение всех терминов, прошедших фильтрацию на основе критерия максимальной значимости, определяемого частотой их появления в исследуемом корпусе. Термины с иерархической вложенностью в ходе обработки игнорировались, дополнительные числовые веса для терминов не присваивались.

6. Метрики

Для оценки эффективности методов были использованы мера Жаккара и метрики Strict и Relax [24].

Мера Жаккара оценивает сходство списков ПОС, фокусируясь на наличии общих элементов, а не на точном совпадении границ. Для набора текстов используется усредненное значение меры по каждому тексту.

Метрика Strict измеряет долю эталонных сущностей, которые были получены точно и полностью и показывает строгость метода извлечения или выделения. Значение метрики для каждого текста из набора рассчитывается следующим образом:

- для определения точности (Precision) вычисляется количество эталонных сущностей, полученных точно, то есть границы сущностей совпадают полностью (True Positive), и вычисляется количество полученных сущностей, которые не существуют в эталонном списке или были извлечены частично (False Positive);
- для определения полноты (Recall) вычисляется количество эталонных сущностей, которые не были извлечены или были извлечены частично (False Negative).

Например, в студенческом тексте в качестве эталонного ПОС выступает *«Docker Swarm»*. Модель может выделить только слово *«Docker»*. В таком случае полных совпадений не будет (True Positive = 0), а будет отмечено, что полученный ПОС неверный (False Positive = 1) и для эталонного ПОС не было найдено соответствующего в полученных (False Negative = 1).

Значение метрики вычисляется как гармоническое среднее между точностью и полнотой (F1 Strict) и показывает, насколько хорошо модель может определять границы извлекаемых сущностей, что особенно важно, если необходимо как можно более точно выделять ПОС и от этого может зависеть его смысл.

Метрика Relax измеряет долю эталонных сущностей, которые были извлечены хотя бы частично. Расчет значения метрики зависит от способа получения ПОС – извлечения или выделения.

Точность (Precision) и полнота (Recall) моделей извлечения ПОС рассчитываются на основе фрагментов текста, которые модель отметила специальными тегами. Местоположение каждого предсказанного фрагмента задается начальным и конечным индексом в тексте. Выделенный ПОС засчитывался как True Positive, если выделенный фрагмент частично совпадал с эталонным ПОС, как False Positive, если среди эталонных ПОС не было найдено фрагмента, пересекающегося с данным выделением. Эталонные ПОС засчитывались как

False Negative, если среди выделенных моделью ПОС не нашлось фрагмента, пересекающегося с этим эталонным ПОС.

В качестве примера рассмотрим текст с эталонными ПОС, выделенные тегами:

«Наше приложение полностью построено на [TAG]React[*TAG] + [TAG*]Redux[*TAG] – это библиотеки для [TAG*]JavaScript[*TAG].»*

Модель может выделить ПОС в тексте таким образом:

«Наше приложение полностью построено на [TAG]React + Redux[*TAG] – это библиотеки для [TAG*]JavaScript[*TAG].»*

В этом случае оба выделенных ПОС будут засчитаны как True Positive, так как для каждого будет найден эталонный ПОС (для *«React + Redux»* - пересечения с *«React»* и *«Redux»*, для *«JavaScript»* - совпадает точно).

Для оценки моделей извлечения ПОС используется подход, основанный на совпадении слов. Для каждого предсказания модели находится его максимальное пересечение по словам с любым эталонным ПОС (и наоборот). Значение метрики для каждого текста из набора рассчитывается следующим образом:

- точность (Precision) вычисляется как сумма длин всех найденных совпадений для предсказанных сущностей, деленная на общее количество слов во всех них;
- полнота (Recall) вычисляется как сумма длин всех найденных совпадений для эталонных данных, деленная на общее количество слов во всех эталонных ПОС.

Для примера возьмем текст с эталонными ПОС, в котором сумма слов всех эталонных ПОС равна 3:

«Для разработки серверной части использовался язык [TAG]PHP[*TAG], версии 7.4, [TAG*]фреймворк[*TAG] [TAG*]Laravel[*TAG] версии 8.2».*

Модель извлекла список, включающих 2 ПОС, сумма слов которых равна 5:

«PHP 7.4», «фреймворк Laravel 8.2».

Для ПОС *«PHP 7.4»* максимальным перекрытием будет *«PHP»* (1 слово), для ПОС *«фреймворк Laravel 8.2»* имеется перекрытие с сущностями *«фреймворк»* (1 слово) и *«Laravel»* (1 слово). Соответственно, для каждого эталонного ПОС есть перекрытия в извлеченных ПОС по 1 слову. Для этого текста точность - (1+1)/5 = 0.4, а полнота - (1+1+1)/3 = 1.

Значение метрики Relax, вычисляемое как гармоническое среднее между точностью и полнотой (F1 Relax), показывает, как хорошо модель улавливает ядро предметно-ориентированных сущностей, даже если не всегда точно определяет их границы.

Используемые в исследовании метрики позволяют оценить модели с точки зрения разных аспектов качества извлечения:

- Strict оценивает точность полного и идеального совпадения извлеченной сущности с эталонной;
- Relax оценивает способность модели найти сущность, даже если ее границы определены не идеально, учитывая частичное совпадение;
- мера Жаккара дает общую оценку сходства между списками на основе пересечения слов, что полезно для анализа пересечения эталонных и извлеченных сущностей без привязки к их точным границам.

7. Обсуждение результатов

Результаты сравнения моделей для извлечения и выделения ПОС представлены в табл. 3.

По представленным результатам видно, что модель mBART в большинстве случаев доминирует: показывает лучший охват эталонных сущностей, более точно определяет

границы ПОС и эффективна даже при частичной разметке сущностей в тексте, что подтверждает способность корректно идентифицировать ядро сущностей. Остальные модели часто выделяют отсутствующие в эталонной разметке сущности, но не отстают по полноте определения эталонных ПОС. Модели Tlite (list) и инструмент rutermextract характеризуются высокими значениями Recall (0.80 и 0.90 соответственно), что обусловлено их подходом к генерации различных вариантов эталонных ПОС, в результате чего достигаются совпадения для значительной части эталонных сущностей.

Табл. 3. Результаты.
Table 3. Results.

Модели	Мера Жаккара	Strict			Relax		
		Precision	Recall	F1	Precision	Recall	F1
Модели, основанные на инструкциях (извлечение ПОС)							
YandexGPT (list)	0.35	0.41	0.59	0.42	0.47	0.77	0.50
Saiga (list)	0.51	0.60	**0.66**	0.59	0.66	0.75	0.64
Tlite (list)	0.19	0.22	0.61	0.27	0.27	0.80	0.33
Модели, основанные на инструкциях (выделение ПОС)							
YandexGPT (NER)	0.40	0.45	0.50	0.44	0.52	0.60	0.52
Saiga (NER)	0.40	0.45	0.63	0.52	0.52	0.71	0.54
Tlite (NER)	0.41	0.47	0.51	0.47	0.54	0.59	0.53
Существующие решения							
Rutermextract	0.16	0.17	0.56	0.23	0.23	**0.90**	0.32
mBART	**0.61**	**0.72**	0.61	**0.63**	**0.76**	0.64	**0.66**

Среди моделей, основанных на инструкциях, для извлечения ПОС лучше всего показала Saiga, которая почти приблизилась к значениям mBART по охвату эталонных сущностей, но все так же, как и другие модели, извлекает много лишних сущностей. YandexGPT показывает стабильно средние результаты в подходах list и NER. Модель Tlite демонстрирует сильную зависимость от подхода: в конфигурации для извлечения модель показывает наихудшие показатели метрик, однако в конфигурации для выделения результаты практически не уступают другим моделям, основанным на инструкциях. Инструмент rutermextract оказался наименее эффективным, извлекая для каждого текста самое большое количество ПОС, которые не были отмечены в эталонных текстах, и ошибаясь в определении границ сущностей. Однако его максимальная полнота указывает на применимость инструмента в сценариях, где пропуск сущности недопустим, а все полученные результаты будут проверены и отфильтрованы.

Подход с извлечением ПОС может дать выигрыш в полноте, что видно из результатов, в особенности по метрике Relax, но часто проигрывает в точности и сбалансированности F1 и склонен к генерации ложных срабатываний. Подход с выделением ПОС обеспечивает более стабильные, надежные и сбалансированные результаты. Все три модели в NER-конфигурации показали близкие и предсказуемые значения F1, что делает этот подход надежным выбором.

Для анализа ошибок моделей использовались метрики True Positive, False Positive, False Negative, вычисленные для Strict. На рис. 2 визуализированы значения метрик. Анализ False Negative выявил значительные различия между моделями: в то время как mBART более точно и полно предсказывает эталонные ПОС (24%), YandexGPT (NER) пропускает значительную часть сущностей (56%). Все модели в среднем правильно предсказывают половину

эталонных ПОС (True Positive). При этом модели извлечения ПОС демонстрируют высокое значение ложных срабатываний (False Positive).

Таким образом, mBART предлагает лучший баланс, минимизируя оба типа ошибок. Модели семейства Saiga показывают высокую полноту (высокий True Positive), но ценой большего количества ложных срабатываний. Инструмент rutermextract и модель Tlite (list) требуют серьезной постобработки для фильтрации ложных результатов.

Рис. 2. Оценка извлечения ПОС: TruePositive, FalsePositive, FalseNegative по моделям.
Fig. 2. DSE extraction score: TruePositive, FalsePositive, FalseNegative by models.

По ошибочно извлеченным ПОС (False Positive) был проведен подсчет, сколько сущностей были выделены или извлечены частично (ошибка разбиения), сколько были объединены с лишними словами (ошибка слияния), сколько полностью состояли из слова, которые не присутствуют в сущности (полностью ошибочное извлечение). В табл. 4 представлены результаты подсчета.

Доминирующим типом ошибок всех представленных моделей является полное неверное извлечение предметно-ориентированных сущностей (ПОС), доля которых превышает 67% и не позволяет провести даже частичное сопоставление результатов с эталоном. Наиболее высокие показатели продемонстрировали модель Tlite (list) и инструмент rutermextract – как в абсолютной числовой величине, так и в относительном выражении.

Анализ ошибок показывает, что модели, основанные на инструкциях, систематически интегрируют ПОС с дополнительным контекстом, отсутствующим в экспертной разметке. Такие ошибки слияния больше характерны при извлечении ПОС из-за менее строгих ограничений на вывод, а модели, выделяющие ПОС, включают дополнительные слова слева или справа от эталонного ПОС. Например, добавляются уточнения (*«php» → «язык php»*) или технические версии (*«php» → «php 7.2»*). В то же время, наблюдаются ошибки разбиения – сокращение полного наименования ПОС до части исходного текста (*«node.js» → «js»*).

Полностью ошибочные извлечения включают некорректные выделения вследствие неверной интерпретации контекста упоминания сущности. Например, в тексте *«Среди бесплатных*

систем управления базами данных к рассмотрению были приняты SQLite, PostgreSQL и MySQL. В результате сравнительного анализа для реализации проекта была выбрана СУБД [TAG]PostgreSQL[*TAG]»* отмечается выделение исключительно PostgreSQL, тогда как модели зачастую ошибочно идентифицируют остальные упомянутые системы управления базами данных (SQLite, MySQL). Кроме того, распространены ситуации ложного распознавания функциональных возможностей применяемых технологий (*«messages.send», «Scipy.stats.poisson»*), наименований файлов (*«index.php», «.blade.php», «retrain.py»*) и имен собственных (*«LinkedIn», «Расмус Лердорф»*). Также встречаются извлечения смыслового окружения непосредственно рядом с ПОС (*«[TAG*]серверный язык программирования[*TAG] PHP»*). Зафиксированы случаи появления случайных извлечений сущностей, нерелевантных предметной области (*«управление сообществом», «сжатие и обрезка изображений», «уровень нагрузки»*).

Табл. 4. Причины ошибок False Positive при извлечении ПОС: разбиение, слияние, полностью ошибочное извлечение.
Table 4. Reasons for False Positive errors during DSE extraction: splitting, merging, completely erroneous extraction.

Модель	Ошибка разбиения, %	Ошибка слияния, %	Полностью ошибочное извлечение, %
YandexGPT (list)	1.8	12.3	85.9
Saiga (list)	3.8	11.4	84.8
Tlite (list)	2.0	11.7	86.3
YandexGPT (NER)	4.3	22.7	73.0
Saiga (NER)	5.9	8.1	86.0
Tlite (NER)	8.8	19.9	71.3
rutermextract	1.9	9.1	89.0
mBART	15.1	17.2	67.7

На рис. 3 проанализировано количество ошибок, допускаемых различными моделями, в зависимости от длины извлекаемой ПОС (от 1 до 6 слов). Распределение длин ПОС в выборке является неравномерным, что влияет на результаты анализа.

Общая тенденция показывает, что частота ошибок возрастает с увеличением длины последовательности. Наименьшее количество ошибок наблюдается при обработке коротких ПОС, состоящих из одного слова, демонстрируется уровень ошибок порядка 15–25%, что объясняется простотой и однозначностью таких сущностей.

Небольшое снижение количества ошибок для ПОС длиной в 3 слова, связано с тем, что к этой категории часто относятся четкие, стандартизированные названия, которые являются уникальными для предметной области ИТ и хорошо распознаются моделями.

Модели демонстрируют наихудшие результаты при обработке длинных ПОС (5-6 слов), что подтверждается диаграммой, указывающей на максимальное количество ошибок именно на этой длине. Это связано с тем, что такие ПОС являются редкими в проверочном наборе и часто представляют составные наименования (например, название алгоритма, содержащее имена авторов). Модели часто фиксируют лишь их фрагменты (например, только слово «алгоритм» или только фамилию), не выявляя всю сущность целиком.

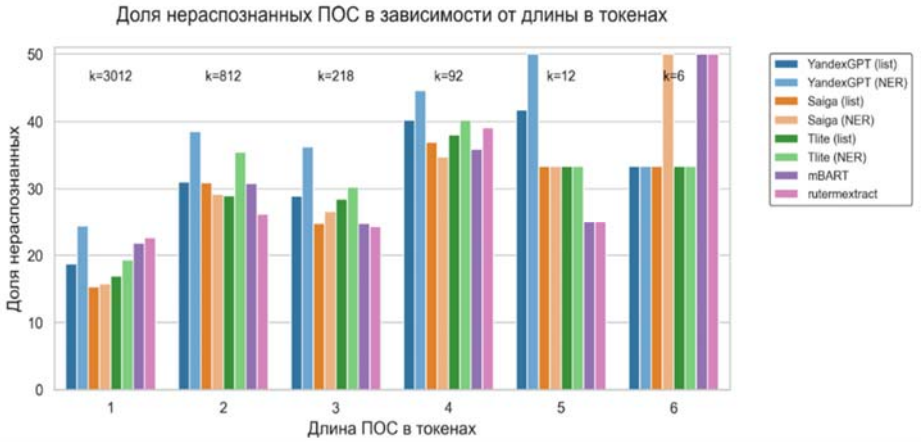

Рис. 3. Зависимость количества ошибок извлечения ПОС от длины (k - количество ПОС определенной длины).
Fig. 3. Dependence of the number of DSE extraction errors on the length (k is the number of DSE of a certain length).

8. Заключение

В настоящей работе рассмотрена задача извлечения из текста предметно-ориентированных сущностей (ПОС): ключевых терминов, навыков, именованных сущностей и других элементов, отражающих тематическую специфику текста. В качестве материала исследования были использованы отчетные документы студентов ИТ-направлений. Среди возможных подходов к решению были исследованы инструменты извлечения ключевых слов (rutermextract), различные варианты дообучения предобученных языковых моделей типа Transformer (на примере модели mBART), а также применение локально запускаемых больших языковых моделей, основанных на инструкциях (на примере моделей YandexGPT, Saiga, Tlite).

Результаты исследования позволяют сделать ряд выводов относительно эффективных подходов к решению задачи выделения ПОС. Во-первых, продемонстрирована высокая эффективность дообучения модели mBART, что подтверждает целесообразность использования тонкой настройки предобученных моделей в условиях наличия достаточного объема данных. Во-вторых, модели, основанные на инструкциях, показали значительное превосходство над инструментом rutermextract, что позволяет говорить об их потенциале в сценариях, характеризующихся нехваткой обучающих данных; в частности, модель Saiga продемонстрировала высокие значения метрики F1 Relax, свидетельствующие о способности к выделению ядра сущностей. В-третьих, стратегия выделения ПОС непосредственно в тексте оказалась более эффективной в плане точности определения границ по сравнению со стратегией извлечения в виде списка. Таким образом, модели, основанные на инструкциях, демонстрируют значительное преимущество, поскольку не требуют трудоемкого этапа обучения и привлечения больших массивов данных.

В то же время, задачу автоматизированного извлечения ПОС нельзя считать полностью решенной. Основным типом ошибок для всех моделей являлось полностью ошибочное извлечение ПОС (67-89% ложноположительных срабатываний). Модели испытывали трудности с разделением ядра сущности от контекста (например, добавление версий технологий или уточнений). Это свидетельствует о необходимости дальнейших изысканий в этой предметной области.

Ключевыми ограничениями экспериментального исследования являются сравнительно небольшой размер обучающей выборки (с учетом расширения набор данных содержит 2933 текста) и достаточно простая структура инструкций для моделей. Исследование ограничилось одним, хотя и интуитивно понятным, вариантом формулировок инструкций для каждой из стратегий (извлечения и выделения). Это может потенциально ограничивать надежность выводов об их сравнительной эффективности. Кроме того, использованная методология оценки не включала такие подходы, как скользящий контроль, что могло повлиять на стабильность и обобщающую способность полученных оценок.

Разработка более сложных, предметно-специфичных и контекстно-обогащенных инструкций, а также применение других форм валидации, представляется перспективным направлением для повышения эффективности подхода и качества его оценки. Масштабирование применения предложенных методик для извлечения ПОС в других предметных областях (например, медицине, юриспруденции, экономике, физике) и из различных классов текстовых документов (включая научные статьи, техническую документацию, новостные ленты) также представляет исследовательский интерес.

Список литературы / References

[1]. Кан А. В., Козловская Я. Д., Токолова А. А. Извлечение научно-технических фактов из отраслевых документов на основе методов семантико-синтаксического и концептуального анализа. Моделирование и анализ данных, 2024, т. 14, № 1, с. 27-40. / Kan A.V. et al. Izvlechenie nauchno-tekhnicheskikh faktov iz otraslevykh dokumentov na osnove metodov semantiko-sintaksicheskogo i kontseptual'nogo analiza [Extraction of scientific and technical facts from industry documents based on semantic-syntactic and conceptual analysis methods]. Modelirovanie i analiz dannykh [Modeling and Data Analysis], 2024, vol. 14, no. 1, pp. 27-40 (in Russian). DOI: 10.17759/mda.2024140102.

[2]. Chumwatana T., Hpone A. K. K. Bridging the IT skill gap with industry demands: An AI-driven text mining approach to job market trends using large language model. Journal of Theoretical and Applied Information Technology, 2025, vol. 103, no. 6, pp. 2270-2282. DOI: 10.5281/zenodo.17175903.

[3]. Мельникова А. В. и др. Разработка алгоритма формирования команд ИТ-проектов на основе данных цифрового следа студентов. Труды Института системного программирования РАН, 2024, т. 36, № 3, с. 213-224. / Melnikova A.V. et al. Razrabotka algoritma formirovaniya komand IT-proektov na osnove dannykh tsifrovogo sleda studentov [Development of an algorithm of the formation of IT project teams based on data from the digital footprint of students]. Trudy Instituta sistemnogo programmirovaniya RAN [Proceedings of the Institute for System Programming of the RAS], 2024, vol. 36, no. 3, pp. 213-224 (in Russian). DOI: 10.15514/ISPRAS-2024-36(3)-15.

[4]. Мельникова А. В., Воробьева М. С., Глазкова А. В. Сравнение предварительно обученных моделей для извлечения предметно-ориентированных сущностей из студенческих отчетных документов // Моделирование и анализ информационных систем, 2025, т. 32, № 1, с. 66-79. / Melnikova A.V., Vorobeva M.S., Glazkova A.V. Sravnenie predvaritelno obuchennykh modelei dlya izvlecheniya predmetno-orientirovannykh sushchnostei iz studencheskikh otchetnykh dokumentov [Comparison of pre-trained models for domain-specific entity extraction from student report documents]. Modelirovanie i analiz informatsionnykh sistem [Modeling and Analysis of Information Systems], 2025, vol. 32, no. 1, pp. 66-79 (in Russian). DOI: 10.18255/1818-1015-2025-1-66-79.

[5]. Rutermextract. Available at: https://github.com/igor-shevchenko/rutermextract (accessed: 29.09.2025).

[6]. Бутенко Ю. И., Николаева Н. С., Маргарян Т. Д. Структурные модели терминологических словосочетаний для разметки корпуса научно-технических текстов. Вестник Новосибирского государственного университета. Серия: Лингвистика и межкультурная коммуникация, 2021, т. 19, № 3, с. 45-56. / Butenko Yu.I., Nikolaeva N.S., Margaryan T.D. Strukturnye modeli terminologicheskikh slovosochetanii dlya razmetki korpusa nauchno-tekhnicheskikh tekstov [Structural models of terminological word combinations for marking up a corpus of scientific and technical texts]. Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Lingvistika i mezhkulturnaya kommunikatsiya [Bulletin of Novosibirsk State University. Series: Linguistics and Intercultural Communication], 2021, vol. 19, no. 3, pp. 45-56 (in Russian). DOI:10/25205/1818-7935-2021-19-3-45-56.

[7]. Бутенко Ю. И. Извлечение номенклатурных наименований из англо-и русскоязычных научно-технических текстов // Искусственный интеллект и принятие решений, 2024, № 3, с. 113-121. /

Butenko Yu.I. Izvlechenie nomenklaturnykh nazvanii iz anglo- i russkoyazychnykh nauchno-tekhnicheskikh tekstov [Nomenclature names extraction from English and Russian-language scientific and technical texts]. Iskusstvennyi intellect i prinyatie reshenii [Artificial Intelligence and Decision Making], 2024, no. 3, pp. 113-121 (in Russian). DOI: 10.14357/20718594240309.

[8]. Шелманов А. О. и др. Открытое извлечение информации из текстов. Часть II. Извлечения семантических отношений с помощью машинного обучения без учителя // Искусственный интеллект и принятие решений, 2019, № 2, с. 39-49. / Shelmanov A.O. et al. Otkrytoe izvlechenie informatsii iz tekstov. Chast II. Izvlecheniya semanticheskikh otnoshenii s pomoshchyu mashinnogo obucheniya bez uchitelya [Open information extraction from texts. Part II. Extraction of semantic relations using unsupervised machine learning] // Iskusstvennyi intellect i prinyatie reshenii [Artificial Intelligence and Decision Making], 2019, no. 2, pp. 39-49 (in Russian). DOI: 10.14357/20718594190204.

[9]. Бручес Е. П., Батура Т. В. Метод автоматического извлечения терминов из научных статей на основе слабо контролируемого обучения. Вестник Новосибирского государственного университета. Серия: Информационные технологии, 2021, т. 19, № 2, с. 5-16. / Bruches E.P., Batura T.V. Metod avtomaticheskogo izvlecheniya terminov iz nauchnykh statei na osnove slabo kontroliruemogo obucheniya [Method for automatic term extraction from scientific articles based on weak supervision] // Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Informatsionnye tekhnologii [Bulletin of Novosibirsk State University. Series: Information Technologies], 2021, vol. 19. no. 2, pp. 5-16 (in Russian). DOI: 10/25205/1818-7900-2021-19-2-5-16.

[10]. Devlin J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. 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, pp. 4171-4186. DOI: 10.18653/v1/N19-1423.

[11]. Дементьева Я. Ю., Бручес Е. П., Батура Т. В. Извлечение терминов из текстов научных статей // Программные продукты и системы, 2022, т. 35, № 4, с. 689-697. / Dementieva Ya.Yu., Bruches E.P., Batura T.V. Izvlechenie terminov iz tekstov nauchnykh statei [Terms extraction from texts of scientific papers]. Programmye produkty i sistemy [Software & Systems], 2022, vol. 35, no. 4, pp. 689-697 (in Russian). DOI: 10.15827/0236-235X.140.689-697.

[12]. Большакова Е. И., Семак В. В. Методы и средства извлечения терминов из текстов для терминологических задач. Программные продукты и системы, 2025, т. 38, № 1, с. 5-16. / Bolshakova E.I., Semak V.V. Metody i sredstva izvlecheniya terminov iz tekstov dlya terminologicheskikh zadach [Methods and means of term extraction from texts for terminological tasks]. Programmye produkty i sistemy [Software & Systems], 2025, vol. 38, no. 1, pp. 5-16 (in Russian). DOI: 10.15827/0236-235X.149.005-016.

[13]. Mamontova A., Ischenko R., Vorontsov K. RuTermEval-2024: Cross-domain Automatic Term Extraction and Classification in Russian scientific texts. Proceedings of the International Conference «Dialogue», 2025, vol. 2025, pp. 245-256. DOI: 10.28995/2075-7182-2025-23-245-256.

[14]. Rozhkov I., Loukachevitch N. Methods for Recognizing Nested Terms. Proceedings of the International Conference «Dialogue», 2025, vol. 2025, pp. 299-311. DOI: 10.48550/arXiv.2504.16007.

[15]. Korytov P. V. et al. Analysis of approaches for identifying key skills in vacancies. 2024 XXVII International Conference on Soft Computing and Measurements (SCM). IEEE, 2024, pp. 242-245. DOI: 10.1109/SCM62608.2024.10554269.

[16]. Николаев И. Е. Метод извлечения знаний и навыков/компетенций из текстов требований вакансий. Онтология проектирования, 2023, т. 13, № 2 (48), с. 282-293. / Nikolaev I.E. Metod izvlecheniya znanii i navykov/kompetentsii iz tekstov trebovanii vakansii [Knowledge and skills extraction from the job requirements texts]. Ontologiya proektirovaniya [Ontology of Designing], 2023, vol. 13, No. 2 (48), pp. 282-293 (in Russian). DOI: 10.18287/2223-9537-2023-13-2-282-293.

[17]. Matkin N. et al. Comparative Analysis of Encoder-Based NER and Large Language Models for Skill Extraction from Russian Job Vacancies // Analysis of Images, Social Networks and Texts: 12th International Conference, AIST 2024, Bishkek, Kyrgyzstan, October 17–19, 2024, Revised Selected Papers. Springer Nature, 2025, vol. 2364, pp. 45-51. DOI: 10.48550/arXiv.2407.19816.

[18]. Khokhlova M., Koryshev M. Keyness Analysis and Its Representation in Russian Academic Papers on Computational Linguistics: Evaluation of Algorithms. RASLAN, 2022, pp. 25-33.

[19]. Митрофанова О. А., Гаврилик Д. А. Эксперименты по автоматическому выделению ключевых выражений в стилистически разнородных корпусах русскоязычных текстов. Terra Linguistica, 2022, т. 13, № 4, с. 22-40. / Mitrofanova O.A., Gavrilik D.A. Eksperimenty po avtomaticheskomu vydeleniyu klyuchevykh vyrazhenii v stilisticheski raznorodnykh korpusakh russkoyazychnykh tekstov [Experiments on automatic keyphrase extraction in stylistically heterogeneous corpus of Russian texts]. Terra Linguistica, 2022, vol. 13, no. 4, pp. 22-40 (in Russian). DOI: 10.18721/JHSS.13402.

[20]. Song M., Feng Y., Jing L. A survey on recent advances in keyphrase extraction from pre-trained language models. Findings of the association for computational linguistics: EACL 2023, 2023, pp. 2153-2164. DOI: 10.18653/v1/2023.findings-eacl.161.

[21]. Glazkova A. V. et al. Keyword generation for Russian-language scientific texts using the mT5 model. Automatic Control and Computer Sciences, 2024, vol. 58, no. 7, pp. 995-1002. DOI: 10.3103/S014641162470041X.

[22]. Гусева Д. Д., Митрофанова О. А. Ключевые выражения в русскоязычных научно-популярных текстах: сравнение восприятия устной и письменной речи с результатами автоматического анализа. Terra Linguistica, 2024, т. 15, № 1, с. 20-35. / Guseva D.D., Mitrofanova O.A. Klyuchevyye vyrazheniya v russkoyazychnykh nauchno-populyarnykh tekstakh: sravneniye vospriyatiya ustnoy i pis'mennoy rechi s rezultatami avtomaticheskogo analiza [Keyphrases in Russian-language popular science texts: comparison of oral and written speech perception with the results of automatic analysis]. Terra Linguistica, 2024, vol. 15, no. 1, pp. 20-35 (in Russian). DOI: 10.18721/JHSS.15102.

[23]. Breton J. et al. Leveraging LLMs for legal terms extraction with limited annotated data. Artificial Intelligence and Law. 2025, pp. 1-27. DOI: 10.1007/s10506-025-09448-8.

[24]. Nguyen K. C. et al. Rethinking Skill Extraction in the Job Market Domain using Large Language Models. 1st Workshop on Natural Language Processing for Human Resources, NLP4HR 2024. Association for Computational Linguistics, ACL Anthology, 2024, pp. 27-42. Available at: https://arxiv.org/pdf/2402.03832, accessed 29.09.2025.

[25]. Yandex YandexGPT-5-Lite-8B-pretrain. Available at: https://huggingface.co/yandex/YandexGPT-5-Lite-8B-pretrain, accessed: 29.09.2025.

[26]. saiga_llama3_8b Available at: https://huggingface.co/IlyaGusev/saiga_llama3_8b, accessed: 29.09.2025.

[27]. Grattafiori A. et al. The Llama 3 herd of models, 2024. Available at: https://arxiv.org/pdf/2407.21783, accessed 29.09.2025.

[28]. Tang Y. et al. Multilingual translation from denoising pre-training. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 3450-3466. DOI: 10.18653/v1/2021.findings-acl.304.

[29]. Lewis M. et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7871-7880. DOI: 10.18653/v1/2020.acl-main.703.

Информация об авторах / Information about authors

Антонина Владимировна МЕЛЬНИКОВА – старший преподаватель Школы компьютерных наук Тюменского государственного университета. Сфера научных интересов: обработка естественного языка, машинное обучение, анализ текстовых данных, извлечение информации.

Antonina Vladimirovna MELNIKOVA – Senior Lecturer, School of Computer Science, University of Tyumen. Research interests: natural language processing, machine learning, text data analysis, information extraction.

Марина Сергеевна ВОРОБЬЕВА – кандидат технических наук. Профессор Школы компьютерных наук Тюменского государственного университета. Сфера научных интересов: исследование методов и технологий машинного обучения для сопровождения образовательного процесса, анализ данных цифрового следа студента, извлечение образовательной информации, разработка и внедрение образовательных технологий.

Marina Sergeevna VOROBEVA – Cand. Sci. (Tech.). Professor, School of Computer Science, University of Tyumen. Research interests: research of machine learning methods and technologies for educational process support, analysis of student digital footprint data, educational information extraction, development and implementation of educational technologies.

Анна Валерьевна ГЛАЗКОВА – кандидат технических наук. Доцент Школы компьютерных наук Тюменского государственного университета. Сфера научных интересов: обработка естественного языка, машинное обучение, компьютерная лингвистика, цифровые гуманитарные науки.

Anna Valerievna GLAZKOVA – Cand. Sci. (Tech.). Associate Professor, School of Computer Science, University of Tyumen. Research interests: natural language processing, machine learning, computational linguistics, digital humanities.

Дмитрий Алексеевич МОРОЗОВ – кандидат технических наук. Младший научный сотрудник Лаборатории прикладных цифровых технологий Новосибирского государственного университета. Технический директор Национального корпуса русского языка. Сфера научных интересов: машинное обучение, обработка естественного языка, корпусная лингвистика, алгоритмы токенизации, автоматическая морфо-синтаксическая разметка текстов.

Dmitry Alekseevich MOROZOV – Cand. Sci. (Tech.). Junior Research Fellow, Laboratory of Applied Digital Technologies, Novosibirsk State University. Technical Director, Russian National Corpus. Research interests: machine learning, natural language processing, corpus linguistics, tokenization algorithms, automatic morpho-syntactic text annotation.