УДК 004.622 Научная статья

https://doi.org/10.32603/2071-8985-2025-18-3-78-86

Поиск и извлечение именованных сущностей из корпуса пользовательских соглашений

М. Д. Кузнецов

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина), Санкт-Петербург, Россия

mkuznetsov7991@gmail.com

Аннотация. Анализ и майнинг данных используются для решения множества различных задач, однако для их эффективного использования необходимы качественные и объемные наборы данных. Открытое опубликование таких наборов не всегда возможно в соответствии с законодательством. Наличие персональных данных в наборах данных обусловливает необходимость их предварительной обработки и очистки. В частности, сформированный в 2024 г. набор текстовых данных PPInRussian для исследования аспектов обработки персональных данных не может быть опубликован, но имеет потенциал стать полезным инструментом как для исследователей в области компьютерной безопасности, так и для правоведов. В данной статье рассматриваются современные методы распознавания именованных сущностей, которые могут быть использованы для очистки текстового корпуса, проведено их тестирование и оценка применимости в рамках очистки юридических документов. Кроме того, предлагается методика очистки текстового корпуса, основанная на правилах, показывающая более точные результаты по сравнению со средствами более общего назначения. Применение этой методики позволит очистить корпус пользовательских соглашений, тем самым делая возможным его опубликование для заинтересованных исследователей.

Ключевые слова: распознавание именованных сущностей, пользовательское соглашение, политика безопасности, персональные данные

Для цитирования: Кузнецов М. Д. Поиск и извлечение именованных сущностей из корпуса пользовательских соглашений // Изв. СПбГЭТУ «ЛЭТИ». 2025. Т. 18, № 3. С. 78–86. doi: 10.32603/2071-8985-2025-18-3-78-86.

Original article

Recognition and Extraction of Named Entities from the User Agreements Corpus

M. D. Kuznetsov

Saint Petersburg Electrotechnical University, Saint Petersburg, Russia

mkuznetsov7991@gmail.com

Abstract. Data analysis and mining are used to solve a variety of different problems, but their effective use requires high-quality and large datasets. Open publication of such datasets is not always possible in accordance with the law. The presence of personal data in datasets necessitates their processing and cleaning before open publication. In particular, the PPInRussian text dataset created in 2024 for studying aspects of personal data processing cannot be published, but it has the potential to become a useful tool for both computer security researchers and legal scholars. This paper discusses modern methods of named entity recognition that can be used to clean a text corpus, tests them, and evaluates their applicability in the context of cleaning legal documents. In addition, the paper proposes a rule-based text corpus cleaning technique that shows more accurate results compared to more general-purpose tools. The application of this technique will clean the corpus of user agreements and, thus, make it possible to publish it for interested researchers.

Keywords: named entity recognition, user agreements, security policy, personal data

For citation: Kuznetsov M. D. Recognition and Extraction of Named Entities from the User Agreements Corpus // LETI Transactions on Electrical Engineering & Computer Science. 2025. Vol. 18, no. 3. P. 78–86. doi: 10.32603/2071-8985-2025-18-3-78-86.

Введение. Современные методы исследования и решения задач, связанных с текстовыми данными, часто полагаются на наборы данных, причем такие наборы обладают большой ценностью, так как их формирование трудоемко и включает этапы сбора, очистки и предобработки. Так, например, в 2024 г. был сформирован набор данных PPInRussian [1], содержащий 7510 пользовательских соглашений на обработку персональных данных пользователей веб-сервисов и сайтов – единственный текстовый корпус на русском языке, используемый для исследования аспектов обработки персональных данных, однако его открытое опубликование невозможно из-за юридических ограничений. В соответствии с редакцией ФЗ № 152 от 1 марта 2021 г. «О персональных данных» наличие персональных данных (именованных сущностей – ИС) служит препятствием для открытого опубликования корпуса пользовательских соглашений, что не позволяет воспользоваться им другим исследователям при необходимости. В то же время, научное сообщество проявляет интерес к пользовательским соглашениям и аспектам обработки персональных данных, причем исследования проводятся учеными из разных стран и отраслей науки [2]–[4]. Такая проблема определяет потребность в средствах очистки текстового корпуса пользовательских соглашений от именованных сущностей. Именованные сущности в данном случае – это топонимы, имена, фамилии и т. д. Распознавание таких сущностей используется для решения самых разных задач, в частности для семантического анализа, формирования баз знаний или графовых представлений каких-либо предметных областей, однако оно позволяет выявлять и удалять из текстовых данных фрагменты, которые не должны оказаться в открытом доступе.

Существующие решения, позволяющие извлекать именованные сущности из русскоязычных текстов, необходимо оценить применительно к корпусу пользовательских соглашений. В данной статье рассматриваются такие программные средства, исследуется их производительность для упомянутого набора данных, предлагается метод поиска и извлечения именованных сущностей, а также проводится оценка результатов и их сравнительный анализ.

Постановка задачи. Задача, которую необходимо решить в рамках данной статьи, состоит в поиске в текстовом корпусе пользовательских соглашений именованных сущностей с наибольшей достижимой точностью. В соответствии с результатами испытаний существующих решений необходимо разработать методику извлечения именованных сущностей, обладающую большей эффективностью по сравнению со средствами более общего назначения.

Обзор существующих решений. На данный момент существует целый ряд инструментов, позволяющих распознавать и извлекать именованные сущности из текстовых данных, которые можно разделить на 3 основные группы:

1) базовые методы – регулярные выражения и словари, основанные на заранее определенных правилах или списках сущностей, они просты в реализации, но могут иметь ограниченную точность из-за недостатка контекстного понимания, такие методы могут быть эффективно применены в условиях узкой специализации решаемых задач;

2) методы на основе машинного обучения – включают в себя использование алгоритмов кластеризации, поддержки векторов (SVM), логистической регрессии и др.; эти модели требуют большого количества размеченных данных для обучения;

3) методы глубокого обучения – с использованием нейронных сетей, особенно рекуррентных (RNN, LSTM), градиентных бустинговых моделей (например, XGBoost), а также сверточных нейронных сетей (CNN) и трансформеров (например, BERT, RoBERTa); эти модели демонстрируют высокую точность благодаря своей способности улавливать глубокий контекст, однако для их обучения необходимы еще более объемные и качественные выборки данных, чем для методов на основе машинного обучения.

Рассматривая доступные и актуальные средства распознавания именованных сущностей, можно выделить 5 основных, поддерживающих работу с текстами на русском языке. Рассмотрим их более подробно.

Библиотека Natasha [5] – это не научный проект, она нацелена на коммерческое использование, однако используется как крупными организациями (Сбербанк, Интерфакс и др.), так и исследователями. При этом основной мотивацией разработчиков

служат высокая производительность, в том числе более сжатые объемы моделей и векторных представлений. В данный проект входит ряд узконаправленных библиотек, позволяющих решать различные задачи NLP, в том числе морфологический и синтаксический анализ текста.

Еще один проект, фокусирующийся на производительности и легковесности, – SpaCy. Методы, используемые в данной библиотеке, не опубликованы, однако указано, что они во многом схожи с методами, представленными в [6]. В данной статье авторы использовали ID-CNN (Iterated Dilated CNN) – вариант сверточной нейронной сети. Библиотека обладает широким функционалом для решения задач NLP: токенизация, лемматизация, выявление именованных сущностей, разметка текста по частям речи, а также синтаксический анализ.

Открытая библиотека для обработки естественного языка DeepPavlov построена на основе библиотек глубокого обучения TensorFlow и Keras, она предоставляет набор готовых инструментов и моделей для решения задач, связанных с пониманием и генерацией текста, включая классификацию, распознавание именованных сущностей, автоматизированные ответы на вопросы и машинный перевод. Одна из важных особенностей проекта – его модульность, он предоставляет инструменты в том числе и для комбинирования моделей с помощью конвейера. В рамках проекта был опубликован ряд научных работ, посвященных различным задачам NLP [7], [8] и NER [9].

Среди разработчиков моделей глубокого обучения популярна HuggingFace Transformers [10]. Библиотека предоставляет простой и функциональный интерфейс для создания и обучения моделей глубокого обучения. Кроме того, она дает доступ к репозиторию моделей, которые можно дообучить на небольших объемах данных без значительной потери точности. Сообщество, поддерживающее библиотеку и репозиторий, весьма активно, поэтому инструментарий постоянно развивается, а модели поддерживаются в актуальном состоянии.

Проекты Stanford CoreNLP [11] и Stanza [12] также широко используются в задачах анализа естественного языка, однако, в отличие от ранее рассмотренных проектов, принадлежат к научным разработкам. CoreNLP – более зрелая библиотека, разработанная в Стенфордском университете в 2006 г., в то время как Stanza – продолжение разработок, представленная в 2020 г. – глубокая переработка CoreNLP, включающая в себя ряд улучшений, в частности улучшенную поддержку языков (60 языков), новые модели и алгоритмы для работы с текстовыми данными, а также новую архитектуру, ориентированную на масштабируемость и обработку больших объемов данных.

Предлагаемая методика поиска именованных сущностей. Использование готовых решений для распознавания именованных сущностей может иметь некоторые нежелательные эффекты, в том числе падение точности распознавания при использовании новых данных, на которых такие подходы не тестировались. Последствия такого трансфера моделей и алгоритмов могут быть непредсказуемыми, поэтому, учитывая узкую направленность работы (очистка пользовательских соглашений), появляется возможность предложить более простые и эффективные решения, которые, вероятно, решат эту задачу лучше. При этом стоит отметить проблематичность использования моделей глубокого обучения ввиду отсутствия качественного аннотированного набора данных, поэтому предлагаемая методика должна быть основана на правилах или методах кластеризации. Обсуждая простые решения, основанные на правилах, и обращая внимание на исходные данные, можно заметить, что пользовательские соглашения имеют в большинстве своем ограниченное количество вариантов указания персональных данных индивидуальных предпринимателей и информации об организациях, в том числе и адресов их регистрации. На данном этапе стоит разделить распознавание данных о лицах и топонимах, так как синтаксические различия при их упоминании достаточно велики.

Для того чтобы очистить пользовательские соглашения от персональных данных, необходимо провести анализ данных и выявить эти сущности. Выборочный анализ некоторых пользовательских соглашений позволит определить ряд шаблонов, по которым можно производить поиск. Кроме того, необходимо провести частотный анализ лексем, который может выявить некоторые закономерности в данных, поскольку обычно персональные данные в контексте пользовательских соглашений упоминаются лишь в пределах одного документа, он также позволит сократить объемы текстовых данных для анализа.

Очистка пользовательских соглашений от топонимов – задача более сложная, решить ее с помощью правил затруднительно, поскольку формат указания адреса может сильно отличаться от до-

кумента к документу, в частности порядок указания данных, используемые разделители и т. д., поэтому допустимо использовать лексемы из открытой базы данных.

Таким образом, проведя анализ данных и определив ряд правил поиска, можно составить словарь, с помощью которого возможно прямое сопоставление токенов документов с найденными именованными сущностями. Полученные в результате анализа и сбора данных словари необходимо расширить дополнительными словоформами, полученными с помощью склонения, что позволит получить более точные результаты при распознавании. При этом для расширения полученных словарей также возможно использование публично размещенных словарей фамилий, имен, отчеств, баз данных индивидуальных предпринимателей и топонимов.

После формирования и обработки словарей именованных сущностей необходимо провести их изъятие из текстов документов. В случае с персональными данными обычно указываются токены в количестве 2–5, таким образом, если такая комбинация есть подмножество словаря именованных сущностей, она будет определена соответствующе. Полная методика поиска именованных сущностей в пользовательских соглашениях может быть представлена в виде схемы, приведенной на рис. 1.

Оценка результатов экспериментов. Очевидно, что достоверно доказать полноту найденных в документах именованных сущностей невозможно, однако имея программные средства, реализующие такие возможности, можно оценить эффективность предложенной методики. Оценка эффективности в данном случае может быть количественной, т. е. чем больше уникальных именованных сущностей было найдено, тем лучше результат, однако такой подход не будет точным, так как именованные сущности могут встречаться несколько раз. Таким образом, было бы целесообразно учесть количество уникальных лексем, составляющих именованные сущности. Имея множества распознанных лексем, составляющих именованные сущности, можно оценить их мощности и сделать выводы об эффективности различных подходов.

Программные средства и источники данных. Расширить словари для поиска, как это предлагается в методике, можно с помощью публично доступных словарей фамилий, имен, отчеств, например набор данных имен Russian (Cyrillic) full names and gender, размещенный на платформе Kaggle [13] в 2018 г., имеет 383 446 уникальных записей. Варианты имен можно получить и из баз данных индивидуальных предпринимателей [14] – база данных ЕГРЮЛ (Единый государственный реестр юридических лиц) и ЕГРИП (Единый государственный реестр инди-

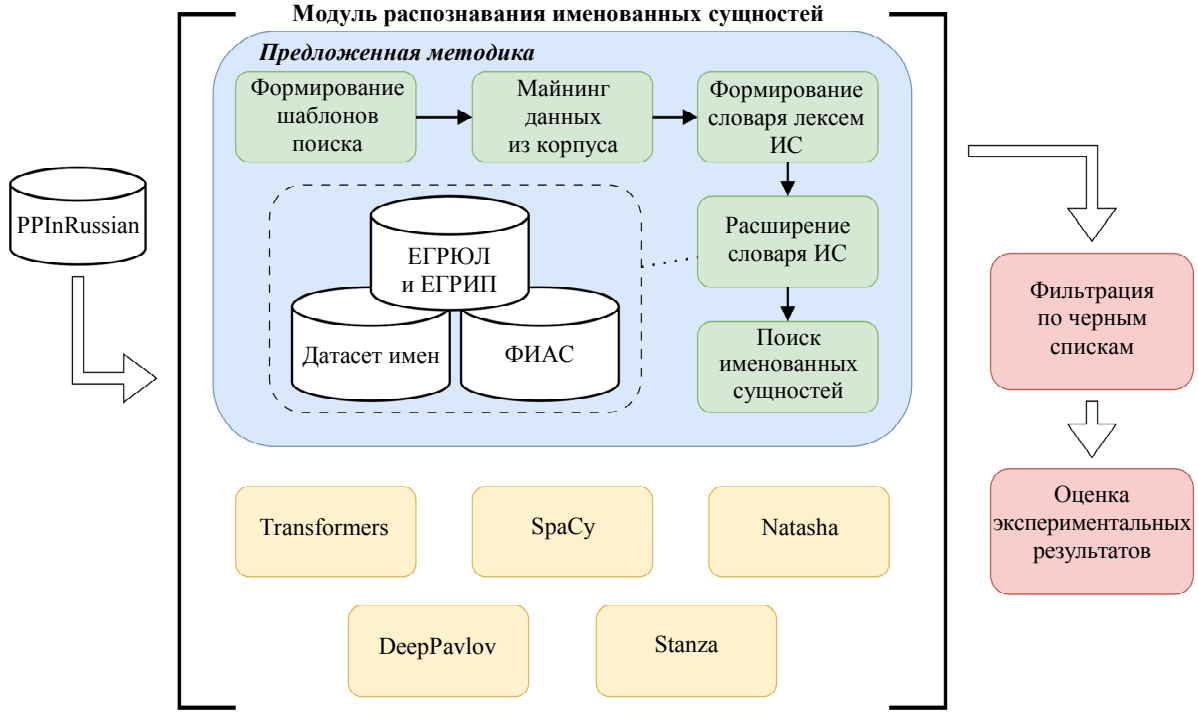

Рис. 1. Схема поиска именованных сущностей
Fig. 1. Named entity search scheme

видуальных предпринимателей). Указанный набор данных содержит записи о юридических лицах и индивидуальных предпринимателях и был сформирован в 2020 г. Всего он содержит 7040 записей об отчествах, 4874 записи об именах и 48 540 о фамилиях, полученных из исходных 22 млн 617 тыс. записей.

Актуальные сведения о топонимах можно получить из базы данных ФИАС [15] (Федеральной информационной адресной системы). База данных содержит актуальный список адресов по всей России, причем она также предоставляет «дельты» – изменения, произошедшие в адресном реестре ранее. База данных обновляется каждую неделю, на момент проведения исследования была использована актуальная версия, что гарантирует покрытие топонимов текстового корпуса PPInRussian.

Для реализации методики очистки текстового корпуса и проведения экспериментов применялись Python и его стандартные библиотеки, а также библиотека Pymorphy [16] и Petrovich [17] для расширения словаря различными словоформами.

В целях сравнительного анализа были использованы упомянутые ранее библиотеки распознавания именованных сущностей: DeepPavlov, Natasha, SpaCy, Stanza, Transformers.

Результаты экспериментов. Анализ текстового корпуса выявил наличие устойчивых шаблонов, с помощью которых можно выделять именованные сущности. Так, распространены следующие префиксы к ФИО: «ИП», «индивидуальный предприниматель», «администрация сайта», «руководитель» и др. Также ФИО указываются несколькими разными способами – с разным порядком и с использованием сокращений. Учитывая эти наблюдения, были составлены регулярные выражения для получения списка лексем именованных сущностей, после чего они были объединены со словарями, сформированными на основе данных из открытых источников.

Затем были проведены эксперименты по распознаванию именованных сущностей. При этом, где было возможно, был задействован графический ускоритель. Длительность обработки корпуса с помощью библиотек распознавания именованных сущностей показана в табл. 1.

Табл. 1. Длительность экспериментов по распознаванию именованных сущностей
Tab. 1. Duration statistics of experiments

Библиотека	Длительность
DeepPavlov	25 мин 30 с
Natasha	51 мин 33 с
SpaCy	9 мин 44 с
Stanza	10 ч 19 мин 50 с
Transformers	5 ч 10 мин 5 с
Предл. методика	***2 мин 13 с***

Эксперименты показали, что в результатах присутствует определенное количество ложных срабатываний, т. е. лексемы, которые не относятся к именованным сущностям, все же были отмечены как формирующие именованную сущность. Рассмотрение уникальных лексем снижает размерность – таким образом, вместо рассмотрения каждой отдельной именованной сущности достаточно проверить ~5000 уникальных лексем, входящих в состав ФИО, и ~10 000 уникальных лексем, входящих в состав топонимов. Примеры таких лексем приведены в табл. 2. Очевидно, что большинство из них не являются частями ФИО или топонимов, поэтому явно неподходящие лексемы были отсеяны с помощью черных списков.

Табл. 2. Частотный анализ лексем, попавших в черные списки
Tab. 2. Lexems included in blacklists

Лексемы в ФИО			Лексемы в топонимах		
№	Наименование	Частота	№	Наименование	Частота
1	Пользователь	68 987	1	и	7383
2	Сайт	62 803	2	в	7200
3	Администрация	14 246	3	или	6788
4	Политика	12 000	4	с	6754
5	Оператор	10 664	5	его	6222
…					
3927	Номер	1	4874	Некоммерческий	1
3928	Форкс	1	4875	Научная	1
3929	Фордевинд	1	4876	Коммунальные	1
3930	Обезличенные	1	4877	Логистических	1
3931	Некомплекта	1	4878	Поезд	1
…					

На основе списков лексем были сформированы черные списки, таким образом из выборки были удалены лексемы, извлеченные в результате ложных срабатываний. На рис. 2, *а* показано общее количество обнаруженных лексем, на рис. 2, *б* – количество верно распознанных лексем для каждого подхода, а также лексем, оказавшихся в черном списке.

Из результатов можно увидеть, что предлагаемая методика поиска именованных сущностей набрала наибольшее число уникальных лексем, причем показатель ложных срабатываний составляет примерно треть от правильных. Библиотеки Natasha и SpaCy, построенные на основе правил, показали примерно равные результаты по верным распознаваниям, однако библиотека SpaCy, набрала самое большое количество ложных срабатываний. Библиотеки, основанные на методах машинного обучения DeepPavlov и Transformers показали результаты хуже остальных библиотек, но при этом количество ложных срабатываний по отношению к верным срабатываниям оказалось самым низким. Единственной библиотекой на основе машинного обучения, которая показала результаты, сопоставимые с лучшими решениями, оказалась библиотека Stanza. В целом такие результаты библиотек на основе глубокого обучения (DeepPavlov и Transformers) можно объяснить их неприспособленностью к подобным текстам, поскольку они были обучены на текстах новостных статей. На рис. 2, *б* можно заметить, что разница в множествах лексем для библиотек DeepPavlov, Transformers и четырех лучших вантов велика, что также указывает на недостаточную эффективность этих программных средств в контексте анализа пользовательских соглашений. При этом у четырех лучших вариантов наблюдается высокая степень схожести по составу лексем и небольшая разница. Кроме того, сопоставимые по количеству и составу уникальных лексем результаты четырех лучших библиотек могут указывать на приближение к реальному количеству таковых в корпусе пользовательских соглашений. На рис. 2, *в* и *г* приведены попарные сравнения

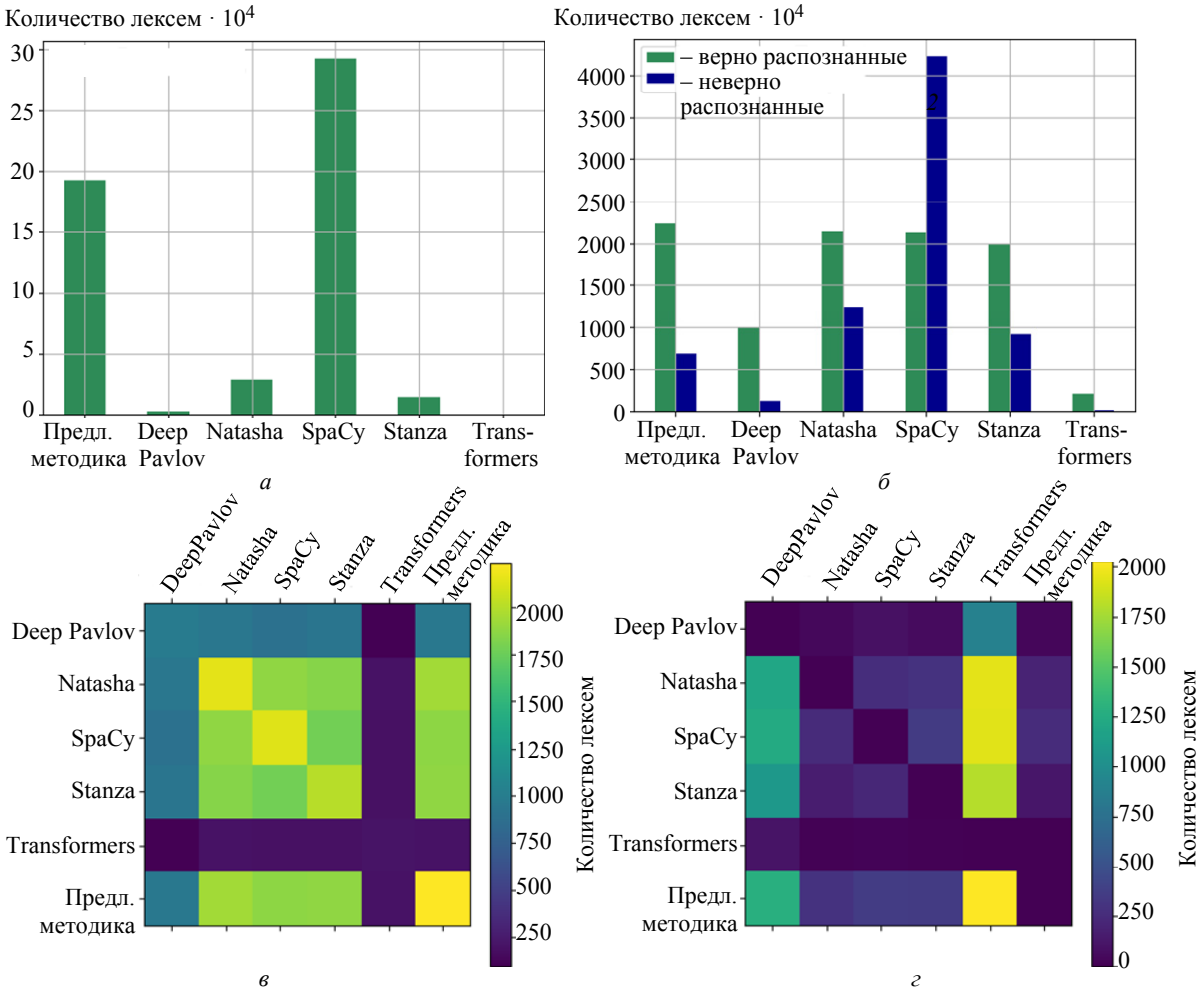

Рис. 2. Результаты экспериментов по поиску имен
Fig. 2. Results of name search experiments

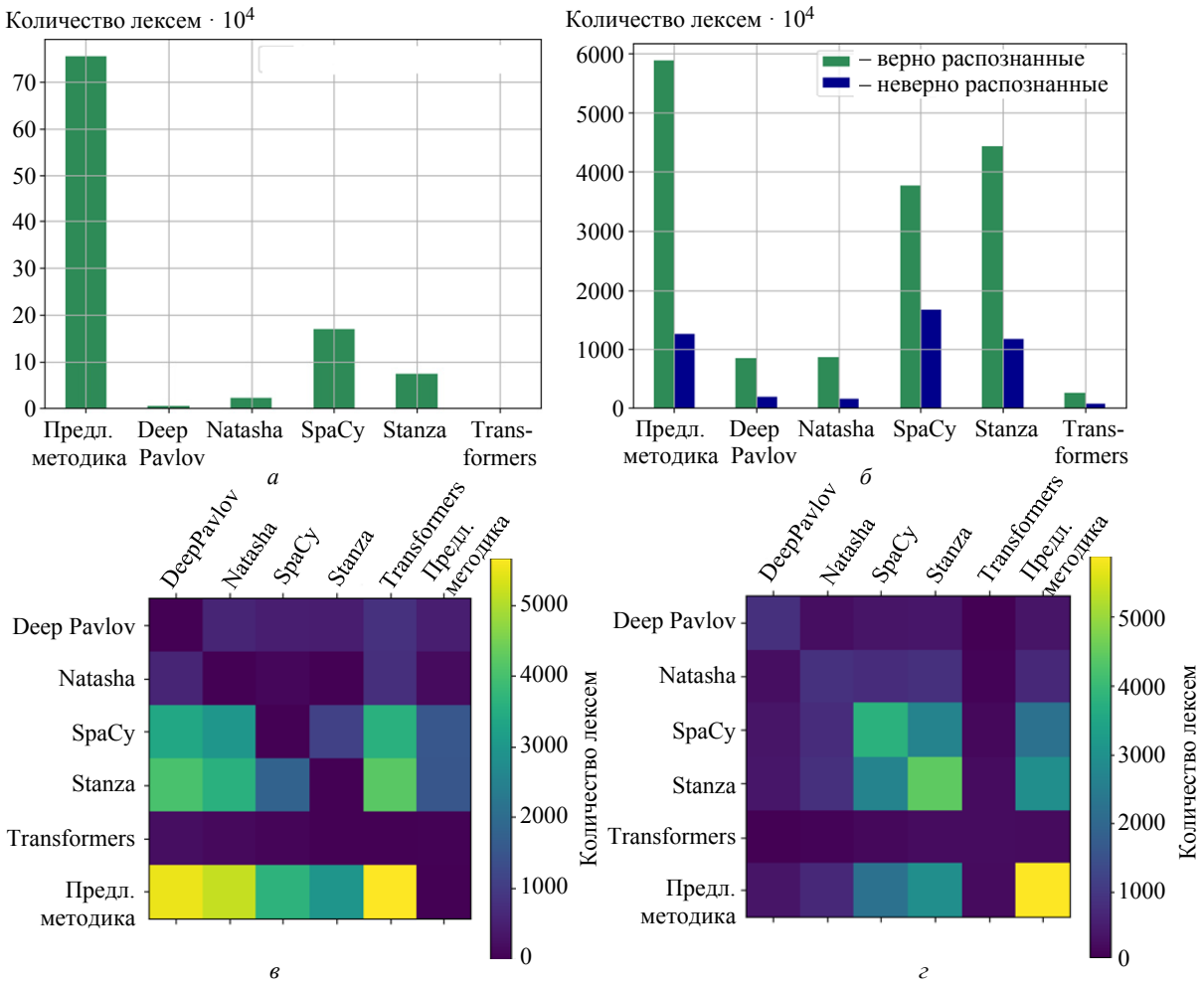

Рис. 3. Результаты экспериментов по поиску топонимов
Fig. 3. Results of toponymic search experiments

множеств уникальных лексем, формирующих именованные сущности. Матрицы были сформированы на основе запроса по лексемам full join, после чего был произведен подсчет количества кортежей без пустых значений, отражающих пересечение множеств обнаруженных лексем, а также кортежей, содержащих пустые значения, для каждой из библиотек. При этом элементы над диагональю матрицы показывают, сколько лексем было обнаружено с помощью библиотеки по оси *x*, которые не были обнаружены библиотекой по оси *y*, под диагональю – наоборот.

Распознавание топонимов было проведено по схожей методике с формированием отдельного черного списка лексем. На рис. 3, *а* также показаны общие количества обнаруженных лексем. На рис. 3, *б* результаты получились иными для топонимов – библиотека Natasha показала значительную неточность, сопоставимую с библиотекой DeepPavlov, в то же время предложенная методика, обладающая словарем, сформированным на основе базы данных ФИАС, показала ожидаемо более точные результаты. На рис. 3, *в* и *г* показаны различия и схожесть состава распознанных лексем, формирующих топонимы, как это было сделано ранее для ФИО, схожесть и разница обусловливают итоговые результаты, приведенные на рис. 3, *б*.

Выводы и заключение. Для применения и исследования аспектов сбора и обработки персональных данных необходимы соответствующие корпусы текстовых данных. На данный момент единственная опция для таких исследований – это корпус русскоязычных пользовательских соглашений PPInRussian [1]. В то же время, его открытое опубликование невозможно в связи с юридическими ограничениями, обусловленными редакцией ФЗ № 152 от 1 марта 2021 г. «О персональных данных». Единственное решение данной проблемы состоит в очистке корпуса от именованных сущностей (данных о лицах и организациях, в том числе адресов).

В статье была предложена узконаправленная методика очистки пользовательских соглашений от именованных сущностей, проведены экспери-

менты, позволяющие оценить ее эффективность по сравнению с существующими решениями. Библиотеки, основанные на моделях глубокого обучения, показали невысокие результаты в контексте анализа пользовательских соглашений, что связано с большой разницей в данных, на которых они были обучены, и пользовательскими соглашениями. В то же время библиотеки, основанные на правилах, выявили больше именованных сущностей, но показали больше ложных срабатываний.

В результате экспериментов, проведенных в соответствии с методологией, предложенная методика показала лучшие результаты, несмотря на свою простоту, также время обработки с ее помощью было меньше, чем у существующих программных средств. Обработка текстового корпуса пользовательских соглашений по предложенной методике позволит очистить его от именованных сущностей, что в свою очередь сделает возможным его открытое опубликование.

Список литературы

1. Кузнецов М. Д., Новикова Е. С. Корпус политик конфиденциальности веб-сервисов и устройств Интернета Вещей для анализа информированности субъектов персональных данных // Информатика и автоматизация (Тр. СПИИРАН). 2025. Т. 24, № 1. С. 163–192. doi: 10.15622/ia.24.1.7.

2. Никитин А. Г. Пользовательские соглашения как правовое условие доступа к виртуальному пространству // Сб. тр. 4-й междунар. науч.-практ. конф. «Бачиловские чтения». Саратов: изд-во ООО «Амирит», 2022. С. 199–206.

3. Дубровин О. В., Ковалева И. Ю. Защита персональных данных в сети Интернет: пользовательские соглашения // Вестн. Южно-Уральского гос. ун-та. Сер. Право. 2014. Т. 14, № 2. С. 64–70.

4. Полетаева Е. Л., Самсонова Е. Д. Правовая природа пользовательского соглашения социальной сети «В Контакте» // BAIKAL RESEARCH J. 2022. Т. 13, № 4. С. 1–9. doi: 10.17150/2411-6262.2022.13(4).23.

5. Свободно распространяемая библиотека анализа текстов на естественном языке Natasha. URL: https://github.com/natasha (дата обращения: 13.09.2024).

6. Fast and accurate entity recognition with iterated dilated convolutions / E. Strubell, P. Verga, D. Belanger, A. McCallum // Conf. on Empirical Methods in Natural Language Proc. (EMNLP 2017). Copenhagen, Denmark: Association for Computational Linguistics, 2017. P. 2670–2680. doi: 10.18653/v1/D17-1283.

7. Le T. A., Arkhipov M. Y., Burtsev M. S. Application of a Hybrid Bi-LSTM-CRF Model to the task of Russian named entity recognition // 6th Conf. Artificial Intelligence and Natural Language (AINL 2017). Ser. Commun. in Comp. and Inform. Sci. (CCIS). SPb., Russia: Springer Cham, 2017. Vol. 789. P. 91–103. doi: 10.1007/978-3-319-71746-3_8.

8. Kuratov Y., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language // arXiv.org. 2019. P. 1–8. doi: 10.48550/arXiv.1905.07213 (дата обращения: 10.10.2024).

9. Anh L. T., Burtsev M. S. A deep neural network model for the task of named entity recognition // Intern. J. of Machine Learning and Comp. 2018. Vol. 9, no. 1. P. 1–6. doi: 10.18178/ijmlc.2019.9.1.758.

10. Свободно распространяемая библиотека анализа текстов на естественном языке HuggingFace Transformers. URL: https://github.com/huggingface/transformers (дата обращения: 16.09.2024).

11. The stanford CoreNLP natural language processing toolkit / Ch. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, D. McClosky // 52nd Ann. Meeting of the Association for Computational Linguistics: System Demonstrations. Baltimore, Maryland, USA: Association for Computational Linguistics, 2014. P. 55–60. doi: 10.3115/v1/P14-5010.

12. Stanza: A Python natural language processing toolkit for many human languages / P. Qi, Yu. Zhang, Yu. Zhang, Ja. Bolton, Ch. D. Manning // 58th Ann. Meeting of the Association for Computational Linguistics: System Demonstrations. Online: Association for Computational Linguistics, 2020. P. 101–108. doi: 10.18653/v1/2020.acl-demos.14.

13. Набор данных имен, фамилий и отчеств Russian (Cyrillic) full names and gender. URL: https://www.kaggle.com/datasets/rai220/russian-cyrillic-names-and-sex (дата обращения: 19.09.2024).

14. База данных наиболее частых имён, отчеств и фамилий. URL: https://www.ngodata.ru/dataset/russian names (дата обращения: 23.09.2024).

15. Федеральная информационная адресная система. URL: https://fias.nalog.ru/Frontend (дата обращения: 10.09.2024).

16. Korobov M. Morphological Analyzer and generator for Russian and Ukrainian languages // 4th Intern. Conf. Analysis of Images, Soc. Networks and Texts (AIST 2015). Ser. Commun. in Comp. and Inform. Sci. (CCIS). Yekaterinburg, Russia: Springer Cham, 2015. Vol. 542. P. 320–332.

17. Свободно распространяемый инфлектор русских антропонимов Petrovich. URL: https://github.com/petrovich (дата обращения: 24.09.2024).

Информация об авторе

Кузнецов Михаил Дмитриевич – аспирант гр. 1933, кафедра информационных систем СПбГЭТУ «ЛЭТИ».
E-mail: mkuznetsov7991@gmail.com
https://orcid.org/0000-0002-0970-8473

References

1. Kuznecov M. D., Novikova E. S. Korpus politik konfidencial'nosti veb-servisov i ustrojstv Interneta Veshhej dlja analiza informirovannosti sub#ektov personal'nyh dannyh // Informatika i avtomatizacija (Tr. SPIIRAN). 2025. T. 24, № 1. S. 163–192. doi: 10.15622/ia.24.1.7. (In Russ.).

2. Nikitin A. G. Pol'zovatel'skie soglashenija kak pravovoe uslovie dostupa k virtual'nomu prostranstvu // Sb. tr. 4-j mezhdunar. nauch.-prakt. konf. «Bachilovskie chtenija». Saratov: izd-vo OOO «Amirit», 2022. S. 199–206. (In Russ.).

3. Dubrovin O. V., Kovaleva I. Ju. Zashhita personal'nyh dannyh v seti Internet: pol'zovatel'skie soglashenija // Vestn. Juzhno-Ural'skogo gos. un-ta. Ser. Pravo. 2014. T. 14, № 2. S. 64–70. (In Russ.).

4. Poletaeva E. L., Samsonova E. D. Pravovaja priroda pol'zovatel'skogo soglashenija social'noj seti «V Kontakte» // BAIKAL RESEARCH J. 2022. T. 13, № 4. S. 1–9. doi: 10.17150/2411-6262.2022.13(4).23. (In Russ.).

5. Svobodno rasprostranjaemaja biblioteka analiza tekstov na estestvennom jazyke Natasha. URL: https://github.com/natasha (data obrashhenija: 13.09.2024). (In Russ.).

6. Fast and accurate entity recognition with iterated dilated convolutions / E. Strubell, P. Verga, D. Belanger, A. McCallum // Conf. on Empirical Methods in Natural Language Proc. (EMNLP 2017). Copenhagen, Denmark: Association for Computational Linguistics, 2017. P. 2670–2680. doi: 10.18653/v1/D17-1283.

7. Le T. A., Arkhipov M. Y., Burtsev M. S. Application of a Hybrid Bi-LSTM-CRF Model to the task of Russian named entity recognition // 6th Conf. Artificial Intelligence and Natural Language (AINL 2017). Ser. Communications in Comp. and Inform. Sci. (CCIS). SPb., Russia: Springer Cham, 2017. Vol. 789. P. 91–103. doi: 10.1007/978-3-319-71746-3_8.

8. Kuratov Y., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language // arXiv.org. 2019. P. 1–8. doi: https://doi.org/10.48550/arXiv.1905.07213. (data obrashhenija: 10.10.2024).

9. Anh L. T., Burtsev M. S. A deep neural network model for the task of named entity recognition // Intern. J. of Machine Learning and Comp. 2018. Vol. 9, no. 1. P. 1–6. doi: 10.18178/ijmlc.2019.9.1.758.

10. Svobodno rasprostranjaemaja biblioteka analiza tekstov na estestvennom jazyke HuggingFace Transformers. URL: https://github.com/huggingface/transformers (data obrashhenija: 16.09.2024). (In Russ.).

11. The stanford corenlp natural language processing toolkit / Ch. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, D. McClosky // 52nd Ann. Meeting of the Association for Computational Linguistics: System Demonstrations. Baltimore, Maryland, USA: Association for Computational Linguistics, 2014. P. 55–60. doi: 10.3115/v1/P14-5010.

12. Stanza: A Python natural language processing toolkit for many human languages / P. Qi, Yu. Zhang, Yu. Zhang, Ja. Bolton, Ch. D. Manning // 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Online: Association for Computational Linguistics, 2020. P. 101–108. doi: 10.18653/v1/2020.acl-demos.14.

13. Nabor dannyh imen, familij i otchestv Russian (Cyrillic) full names and gender. URL: https://www.kaggle.com/datasets/rai220/russian-cyrillic-names-and-sex (data obrashhenija: 19.09.2024). (In Russ.).

14. Baza dannyh naibolee chastyh imjon, otchestv i familij. URL: https://www.ngodata.ru/dataset/russian-names (data obrashhenija: 23.09.2024). (In Russ.).

15. Federal'naja informacionnaja adresnaja sistema. URL: https://fias.nalog.ru/Frontend (data obrashhenija: 10.09.2024). (In Russ.).

16. Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages // 4th Intern. Conf. Analysis of Images, Social Networks and Texts (AIST 2015). Ser. Commun. in Comp. and Inform. Sci. (CCIS). Yekaterinburg, Russia: Springer Cham, 2015. Vol. 542. P. 320–332.

17. Svobodno rasprostranjaemyj inflektor russkih antroponimov Petrovich. URL: https://github.com/petrovich (data obrashhenija: 24.09.2024). (In Russ.).

Information about the author

Mikhail D. Kuznetsov – postgraduate student gr. 1933, Department of Information systems, Saint Petersburg Electrotechnical University.
E-mail: mkuznetsov7991@gmail.com
https://orcid.org/0000-0002-0970-8473

Статья поступила в редакцию 06.11.2024; принята к публикации после рецензирования 29.01.2025; опубликована онлайн 28.03.2025.

Submitted 06.11.2024; accepted 29.01.2025; published online 28.03.2025.