https://doi.org/10.29001/2073-8552-2025-40-2-201-210
УДК 004.912:004.413:004.4'242

Разработка сервиса для автоматического извлечения именованных сущностей из неструктурированных медицинских русскоязычных текстов

Л.В. Ронжин, П.А. Астанин, С.Е. Раузина, П.А. Ядгарова, Т.В. Зарубина

Российский национальный исследовательский медицинский университет имени Н.И. Пирогова (РНИМУ им. Н.И. Пирогова), 117513, Российская Федерация, Москва, ул. Островитянова, 1

Аннотация

Введение. В настоящее время значительная часть медицинских данных формируется и хранится в неструктурированном (текстовом) виде. Одним из способов обработки неструктурированной информации является извлечение именованных сущностей (NER – Named entity recognition). В классическом представлении решение задачи NER при работе с медицинскими текстами сводится к поиску объектов или понятий, имеющих определенный контекст и связанных с упоминаемыми в тексте действиями или событиями. В качестве конечного множества терминов для решения подобной задачи может быть использована Унифицированная национальная медицинская номенклатура (УНМН), разрабатываемая с 2022 г. на основе международных и федеральных справочников, а также других источников. На момент выполнения исследования в открытой научной литературе не было найдено сведений о существовании инструмента для решения задачи NER при работе с неструктурированными медицинскими текстами на русском языке.

Цель исследования: разработка инструмента для извлечения именованных сущностей из русскоязычных медицинских текстов.

Материал и методы. В качестве терминологического свода для решения задачи распознавания именованных сущностей использовалась УНМН. В алгоритмы предобработки текста включены сегментация текста, токенизация и синтаксический разбор предложений, лемматизация и морфологический анализ слов. Тестирование инструмента проводилось на клинических рекомендациях (КР), актуальных на момент проведения исследования. Основной метрикой качества считалась доля автоматически верно распознанных терминов относительно экспертной разметки.

Результаты. В ходе исследования был разработан Аннотатор медицинских текстов – сервис, предназначенный для решения задачи NER с последующими разметкой и категоризацией извлекаемых терминов УНМН. Данный сервис основан на использовании больших языковых моделей и собственных лингвистических правил. Аннотатор медицинских текстов может применяться для анализа текстов на русском языке с использованием любой терминологической системы. Аннотатор медицинских текстов является гибридным инструментом, обеспечивающим автоматическое извлечение до 93% терминов из свободного текста актуальных КР. Качество работы данного сервиса сопоставимо с зарубежными инструментами для решения задачи NER при работе с текстами на английском языке: cTAKES с точностью в 91% и MetaMap – с F1-score в 88% соответственно.

Заключение. В статье представлен гибридный сервис для распознавания именованных объектов в неструктурированных медицинских текстах. Сервис был апробирован путем извлечения терминов УНМН из актуальных клинических рекомендаций с последующей проверкой медицинскими экспертами. Полученные результаты демонстрируют потенциал как этого инструмента, так и Унифицированной национальной медицинской номенклатуры.

Ключевые слова: NLP; обработка естественного языка; NER; извлечение именованных сущностей; УНМН; концепт; база знаний; онтология.

Финансирование: настоящее исследование финансируется из средств Программы стратегического академического лидерства «Приоритет-2030». Номер субсидии: 075-15-2021-1325 (от 30 сентября 2021 г.).

Для цитирования: Ронжин Л.В., Астанин П.А., Раузина С.Е., Ядгарова П.А., Зарубина Т.В. Разработка сервиса для автоматического извлечения именованных сущностей из неструктурированных медицинских русскоязычных текстов. *Сибирский журнал клинической и экспериментальной медицины*. 2025;40(2):201–210. https://doi.org/10.29001/2073-8552-2025-40-2-201-210

Астанин Павел Андреевич, e-mail: med_cyber@mail.ru.

Development of a service for automatically extraction of medical concepts from Russian unstructured texts

Lev V. Ronzhin, Pavel A. Astanin, Svetlana E. Rauzina, Polina A. Yadgarova, Tatyana V. Zarubina

Pirogov Russian National Research Medical University, 1 Ostrovityanova str., Moscow, 117513, Russian Federation

Abstract

Introduction. A significant part of medical data is currently generated and stored in an unstructured (textual) form. One way to process unstructured information is named entity recognition (NER). In the classical view, solving the NER problem within medical texts involves identifying objects or concepts that have a specific context related to the actions or events mentioned in the text. The National Unified Terminological System (NUTS) has been developed since 2022 based on international and federal medical thesauri and other sources. It can be used as the term set for solving problems of this type. At the time of the study, there was no available information in the scientific literature about tools solving NER problem in unstructured Russian-language medical texts.

Aim: To develop a tool for extracting named entities from Russian-language medical texts.

Material and Methods. Named entity recognition is performed using the NUTS as the terminological framework. The preprocessing pipeline includes full text segmentation, sentences tokenization and dependency parsing, words lemmatization and morphological analysis. The Annotation tool has been evaluated on clinical guidelines. The primary evaluation metric is the ratio of correctly identified terms to the total number of experts' extracted terms.

Results. As part of this study, the Annotation tool for medical texts has been developed. It is an automatized tool for extraction and categorization NUTS terms. This service is based on combined use large language models and rules. The Annotation tool can analyze texts in any language of the Indo-European group using any terminological system.

The Annotation tool is hybrid and extracts automatically up to 93% of terms from the actual unstructured guidelines texts. The quality of this service is comparable to international NER tools for English-language texts: cTAKES with 91% accuracy and MetaMap with an F1-score of 88%.

Conclusion. The article presents the Annotation tool - a hybrid service for named entity recognition within unstructured medical texts. The service was validated by extraction of NUTS terms in current clinical guidelines, with subsequent verification by medical experts. The obtained results demonstrate the promising potential of both this tool and the National Unified terminology system (NUTS).

Keywords: NLP; natural language processing; NER; named entity recognition; NUTS; concept; knowledge base; ontology.

Funding: this study was supported by the federal academic leadership program "Priority 2030" (Grant No. 075-15-2021-1333, dated September 30, 2021).

For citation: Ronzhin L.V., Astanin P.A., Rauzina S.E., Yadgarova P.A., Zarubina T.V. Development of a service for automatically extraction of medical concepts from Russian unstructured texts. *Siberian Journal of Clinical and Experimental Medicine*. 2025;40(2):201–210. https://doi.org/10.29001/2073-8552-2025-40-2-201-210

Введение

В связи с широким внедрением информационных систем в сферу здравоохранения в последние годы наблюдается ускоряющийся рост объема накапливаемых медицинских данных, значительная доля которых продолжает формироваться в неструктурированном виде [1, 2]. Одним из путей решения этой проблемы является разработка исходно структурированных электронных медицинских документов. Это длительный и трудоемкий процесс, связанный с изучением нормативно-правовой базы и медицинской литературы, а также привлечением высококвалифицированных экспертов для детальной проработки предметной области. Альтернативным, но не менее перспективным направлением является обработка естественного языка (NLP – Natural Language Processing), основанная на использовании лингвистических правил и алгоритмов машинного обучения. Развитие инструментов NLP требует привлечения значительных технических ресурсов и больших данных.

Одной из классических задач NLP является извлечение именованных сущностей (NER – Named Entity Recognition) – объектов или понятий, имеющих определенный контекст и связанных с упоминаемыми в тексте действиями или событиями. NER может ограничиваться извлечением только сущностей для автоматической разметки различных источников информации (научных статей, клинических рекомендаций (КР), учебных пособий) при создании информационно-справочных систем. Однако при учете смыслового контекста и омонимии терминов в неструктурированных источниках (включая данные

реальной клинической практики), можно извлекать не только сущности, но и связи между ними. Две сущности и предикат, определяющий характер их взаимодействия, принято называть триплетами – элементарными единицами, необходимыми для построения подавляющего большинства баз знаний (БЗ). В подобной ситуации возможно расширение области применения NER вплоть до значительной автоматизации процесса разработки гибридных консультативно-справочных систем поддержки принятия решений с элементами интерпретации решений [3–5].

Максимальная практическая польза NER раскрывается при использовании заранее сформированного множества унифицированных терминов, связываемых с извлекаемыми сущностями и используемых для индексации текста. Важнейшим требованием является максимальная сопоставимость применяемых терминов с федеральными и международными справочниками.

Автоматическое решение задачи NER с использованием унифицированной терминологии сопряжено со следующими проблемами: дублирование одних и тех же формулировок в разных справочниках, широкая синонимия и многозначность понятий, а также существование их различных производных форм. Попытки решения данных проблем предпринимались в разных странах мира на протяжении нескольких десятилетий. Наиболее успешным путем стало объединение синонимичных формулировок в концепты – целостные языковые единицы (сущности), выделенные с учетом предметной специфики и обладающие определенным смысловым значением. Подобный принцип используется в Унифицированной медицинской языковой системе (UMLS – Unified Medical Language System), агрегирующей справочники на 13 языках мира [6]. В настоящее время UMLS обеспечивает терминологическое покрытие подавляющего большинства медицинских направлений и решает проблему обратной совместимости: любая запись из справочника будет иметь сопоставление с одним или несколькими концептами. Создание русифицированного фрагмента UMLS позволит, используя многолетний накопленный международный опыт развития данной системы, унифицировать подходы к обработке медицинских неструктурированных текстов на национальном уровне [7].

Наибольший интерес в контексте описанной проблемы вызывают разработанные на основе UMLS MetaMap и cTAKES – инструменты для разметки текстов путем извлечения из них именованных сущностей (NER – Named Entity Recognition). MetaMap и cTAKES являются относительно сопоставимыми по качеству: F1-score для MetaMap составляет 88% против точности cTAKES, равной 91% [8]. Однако в настоящее время данные инструменты применяются для решения разных задач, поскольку MetaMap предназначен для автоматической разметки научной литературы, а cTAKES – для работы с реальными медицинскими данными [6].

Процесс построения цифрового образа текста, сводимый к решению задачи NER, разметке и категоризации, принято называть аннотированием. Разработка сервиса для автоматического аннотирования медицинских текстов позволит оптимизировать процессы, связанные с созданием, обработкой и анализом клинической информации. На момент выполнения настоящего исследования в научной литературе не найдено информации о существовании валидных инструментов, применимых для аннотирования русскоязычных медицинских текстов.

Цель исследования: разработка сервиса для автоматического аннотирования медицинских текстов на русском языке.

Материал и методы

Исследование проведено в рамках программы стратегического академического лидерства «Приоритет-2030» на базе Института цифровой трансформации медицины (ИЦТМ) ФГАОУ ВО «Российский национальный исследовательский медицинский университет имени Н.И. Пирогова» Минздрава России.

Для работы с данными применялись системы управления базами данных (СУБД) PostgreSQL 14 и Neo4j, для реализации алгоритмов – язык программирования Python 3.9.

В качестве терминологического свода для решения задачи NER использовалась Унифицированная национальная медицинская номенклатура (УНМН), разрабатываемая с 2022 г. на основе UMLS, федеральных справочников и других источников [7, 9]. Текущая версия УНМН имеет обратную совместимость с UMLS, содержит свыше 13 млн формулировок, агрегированных из 102 международных и 259 федеральных справочников [7]. Онтологическая модель УНМН во многом наследует особенности UMLS, главной из которых является объединение одинаковых или синонимичных формулировок в один концепт и его отнесение к одной из 127 тематических групп [10]. В настоящее время УНМН может считаться одним из крупнейших терминологических сводов на русском языке. Недавние исследования подтверждают наличие в УНМН более 1,5 млн клинически значимых терминов, что позволяет строить на ее основе решения системного уровня [11].

Для оптимизации вычислений все формулировки УНМН подвергались специальной предобработке, результат которой хранился в статичном виде и обновлялся при любом изменении концептов и терминов. Анализируемые тексты проходили предобработку в режиме реального времени. Алгоритм предобработки для терминов и текстов был одинаковым и представлял собой последовательность из четырех операций, схематично представленных на рисунке 1.

Согласно данным, представленным на рисунке 1, унифицированный алгоритм предобработки неструктурированного текста включает сегментацию (разделение текста на предложения), токенизацию (разделение предложения на токены – неделимые частицы, позволяющие оптимизировать операции поиска с переходом от посимвольного сравнения к более сложным методам), лемматизацию и морфологический анализ, а также синтаксический разбор. Каждая операция направлена на приведение текста к строго определенному стандарту. Выпадение операций из цикла предобработки приводит к значительному снижению потенциального качества NER.

Для последовательно выполняемых процедур сегментации и токенизации текста применялись семантические правила, составленные на языке регулярных выражений. Следует подчеркнуть, что выбор в пользу работы с подстроками был обусловлен возникновением большого количества ошибок при использовании на данном этапе нейронных сетей.

 https://doi.org/10.29001/2073-8552-2025-40-2-201-210

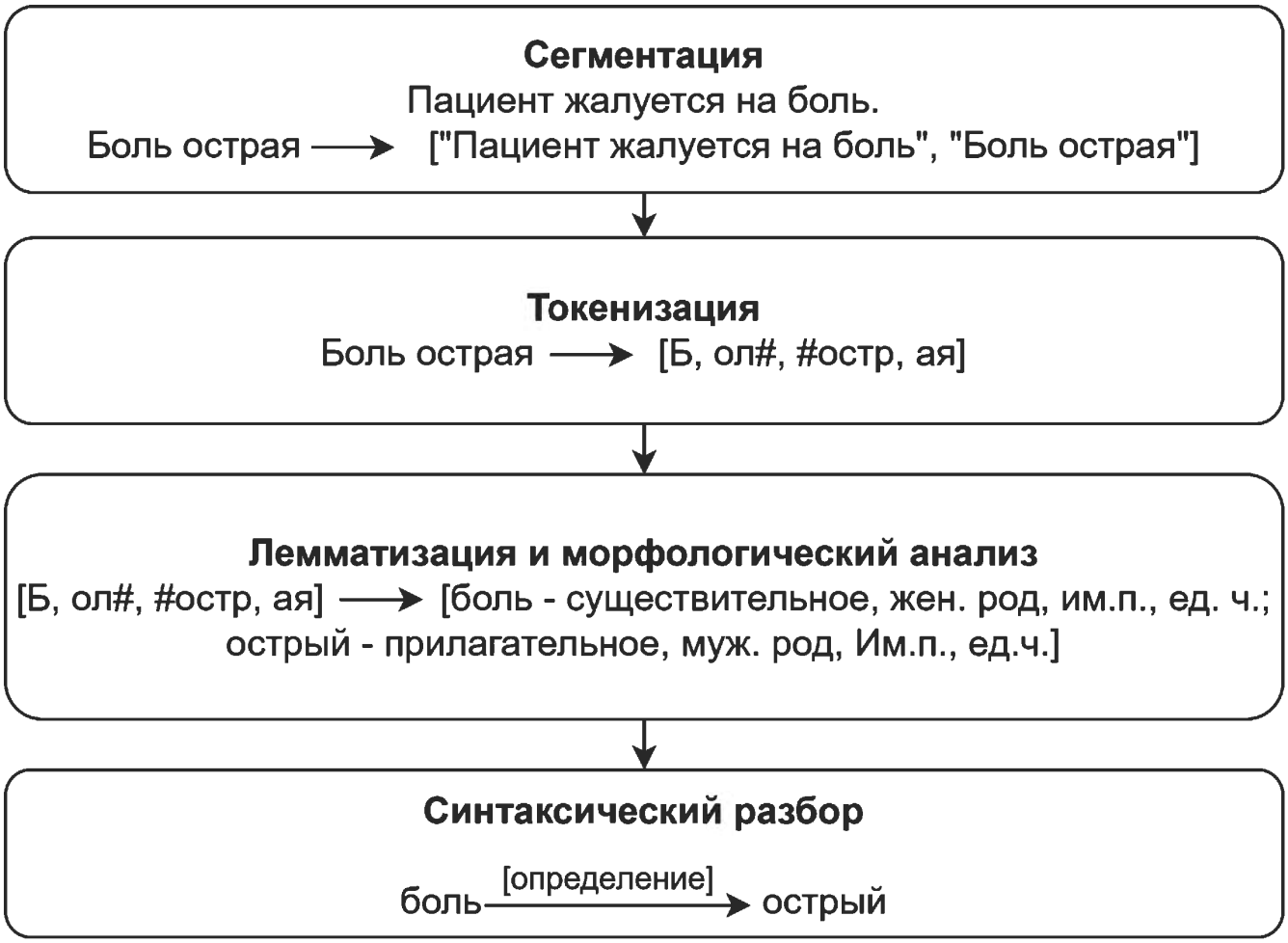

Рисунок 1. Унифицированный алгоритм предобработки неструктурированного текста
Figure 1. Unified algorithm for preprocessing unstructured text

Для устранения влияния морфологических форм на результат сравнения фрагментов текста применялась лемматизация, сохраняющая часть языковых свойств слова и направленная на приведение слова к нормальной форме – единственному числу в именительном падеже в мужском роде. Морфологический анализ и лемматизация осуществлялись гибридным способом с применением анализатора Pymorphy2, основанного на правилах и статистической частоте встречаемости словоформ, и больших языковых моделей (LLM – Large language model). На данном этапе работы с текстом в качестве LLM использовалась открытая нейронная сеть Ru-bert-case с архитектурой Transformer [12]. Итоговое решение формировалось при пересечении множеств морфологических свойств леммы, полученных в результате независимого применения двух указанных инструментов. Выбор гибридного подхода обеспечил не только учет контекста (решение задачи лингвистической неоднозначности слов и фраз), но и коррекцию неадекватных результатов работы LLM при анализе специфичных формулировок и при иных ситуациях, в которых обучающая выборка не всегда могла содержать достаточное количество примеров для покрытия всех закономерностей их применения в рамках соответствующей предметной области.

После лемматизации и морфологического разбора осуществлялась процедура синтаксического разбора или извлечения зависимостей (DP – Dependency parsing) – генерации ориентированного ациклического графа (дерева) предложения, вершинами в котором являются слова, а ребрами – связи между ними. DP является нелинейным преобразованием текста в многомерное представление, в котором мерой близости слов является их смысловая связь, а не количество слов между ними. При подобном отображении слова семантически связанные слова, расположенные на разных концах предложения, будут соединены напрямую. Следует отметить, что единой системы синтаксических связей в настоящее время не существует, однако наиболее употребимой является международная нотация Universal dependencies (UD), использованная в настоящем исследовании [13, 14].

На этапе DP применялась упомянутая ранее LLM Ru-bert-case. Для улучшения результатов DP также использовалась нейронная сеть Stanza, обученная на наборе данных Syntagrus – коллекции из предложений и удовлетворяющих требованиям UD синтаксических деревьев [15, 16]. Полученный первичный результат DP корректировался с использованием набора эмпирических правил, которые устанавливали соответствие синтаксических связей и частей речи слов, а также задавали разные веса атрибутам слова в зависимости от их источника (Pymorphy2 или LLM). Корректировка морфологических свойств и лемм производилась также на основе взвешенного пересечения множеств морфологических вариантов слова. При работе с длинными предложениями приоритет отдавался LLM для значений всех атрибутов (кроме леммы).

При наличии слова в Едином справочнике медицин-

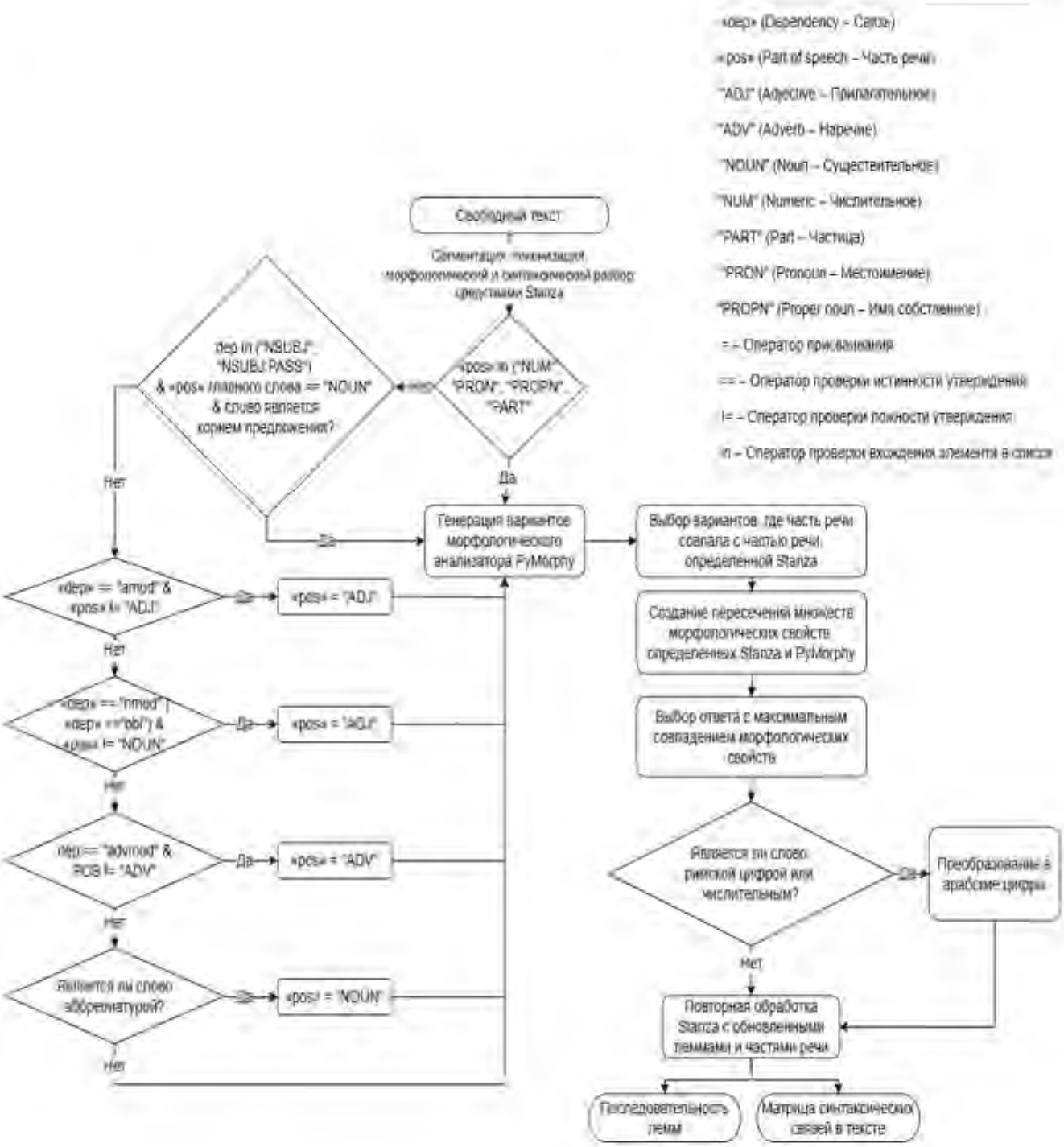

Рисунок 2. Гибридный алгоритм лемматизации и морфологического разбора с коррекцией результатов правилами
Figure 2. Hybrid algorithm for lemmatization and morphological analysis with result correction using rules

ских аббревиатур, разработанном ранее на основе анализа аннотаций к научным статьям на русском языке, ему принудительно присваивалась часть речи «PROPN» (существительное, имя собственное) [9]. Все леммы приводились к нижнему регистру.

Особое внимание уделялось нормализации числительных, для которых может существовать большое количество способов обозначения в медицинской литературе. Для стандартизации способа записи терминов УНМН и вводимого текста все порядковые, количественные числительные и римские цифры конвертировались в арабские цифры.

Также производилось удаление сочинительных союзов, некоторых частиц, знаков препинания и «стоп-слов» – служебных слов, наличие или отсутствие которых не приводит к искажению смысла текста [17]. Важно отметить, что использование фиксированного списка последних будет приводить к искажению смысла текста, поскольку в определенном контексте «стоп-слово» может стать ключевым. По этой причине удаление «стоп-слов» осуществлялось с предварительным анализом связей, входящих в узел данного слова.

Гибридный алгоритм лемматизации и морфологического разбора с коррекцией результатов правилами представлен на рисунке 2.

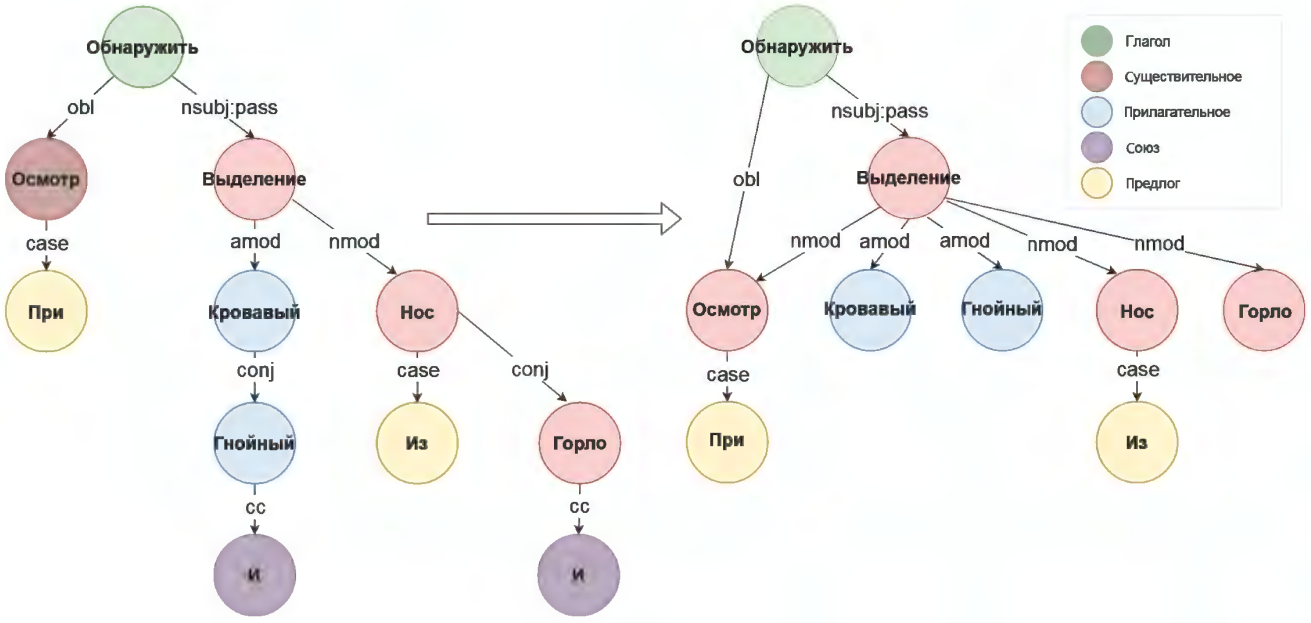

Рисунок 3. Пример синтаксического графа предложения до и после коррекции связей
Figure 3. Example of a syntactic graph of a sentence before and after the correction of relationships

После выполнения действий, представленных на рисунке 2, скорректированные леммы вновь подавались на вход нейросети, после чего выполнялся повторный синтаксический разбор. Таким образом, процедуры сегментации и токенизации текста выполнялись однократно, а лемматизации, морфологического и синтаксического разборов – двукратно.

Следует уточнить, что синтаксический граф также подвергался коррекции для обеспечения возможности последующего поиска изоморфизмов (подграфов). В первую очередь изменениям подвергалось отображение союзных («conj» – conjugated) частей предложения, поскольку в нотации UD принято конструировать ветку с союзными частями путем соединения их друг за другом в порядке их следования в предложении. Наличие подобных соединений приводило бы к значительному снижению специфичности поиска изоморфизмов терминов в таких конструкциях. В связи с этим вся цепочка узлов со связями с типом «conj» соединялась непосредственно с корневым словом связью с тем же типом, который был у связи корневого слова с первым элементом цепочки конъюгатов.

Также в процессе коррекции синтаксического графа осуществлялась симметризация (дублирование в обратном направлении) связей с типом «parataxis», обозначающим соединение двух частей сложного предложения без использования запятых и союзов (как правило, с использованием двоеточия, тире и скобок).

Отдельно следует отметить, что в одном предложении может быть только одно сказуемое, которое, согласно нотации UD, считается главным словом предложения и имеет соответствующую корневую метку. В безглагольных фразах сказуемым чаще всего является определение. Без реверса подлежащего и сказуемого поиск фраз внутри сложных предложений был бы невозможен из-за обратного направления связи. По этой причине в ряде случаев, определенных эмпирическими правилами, производился реверс подлежащего и сказуемого путем изменения направления связи между ними на противоположное. Пример коррекции связей синтаксического графа предложения представлен на рисунке 3.

Для тестирования Аннотатора медицинских текстов (далее – АМТ; Аннотатор; Сервис) использовались актуальные на момент проведения исследования КР по ряду социально значимых общетерапевтических и профильных патологий (язвенная болезнь, острые нарушения мозгового кровообращения, желудочно-кишечные кровотечения у взрослых, рак желудка, гастрит и дуоденит, хроническая обструктивная болезнь легких, внебольничная пневмония у взрослых). Из КР по перечисленным нозологиям экспертным способом извлекались клинически значимые термины. Затем извлеченные списки терминов сравнивались со списками, полученными в ходе работы АМТ. В качестве главной метрики качества аннотирования использовалась точность – отношение числа правильно выделенных Аннотатором терминов к общему числу терминов, найденных экспертами.

Результаты

В настоящем исследовании задача аннотирования неструктурированного текста сводилась к поиску подграфов терминов УНМН (изоморфизмов) в синтаксическом графе предложения методом сплошного перебора. Основной проблемой, возникшей при решении данной задачи, стала оптимизация вычислений, поскольку задача поиска графовых изоморфизмов имеет степенную сложность и не может быть реализована в приемлемые для пользователя сроки при обработке всех формулировок УНМН.

Следует отметить, что математического алгоритма, обеспечивающего снижение сложности задачи поиска изоморфизмов, в настоящее время не существует [18]. Собственные решения, строившиеся с использованием графовой СУБД Neo4j и элементов асинхронного программирования, приводили к экспоненциальному росту

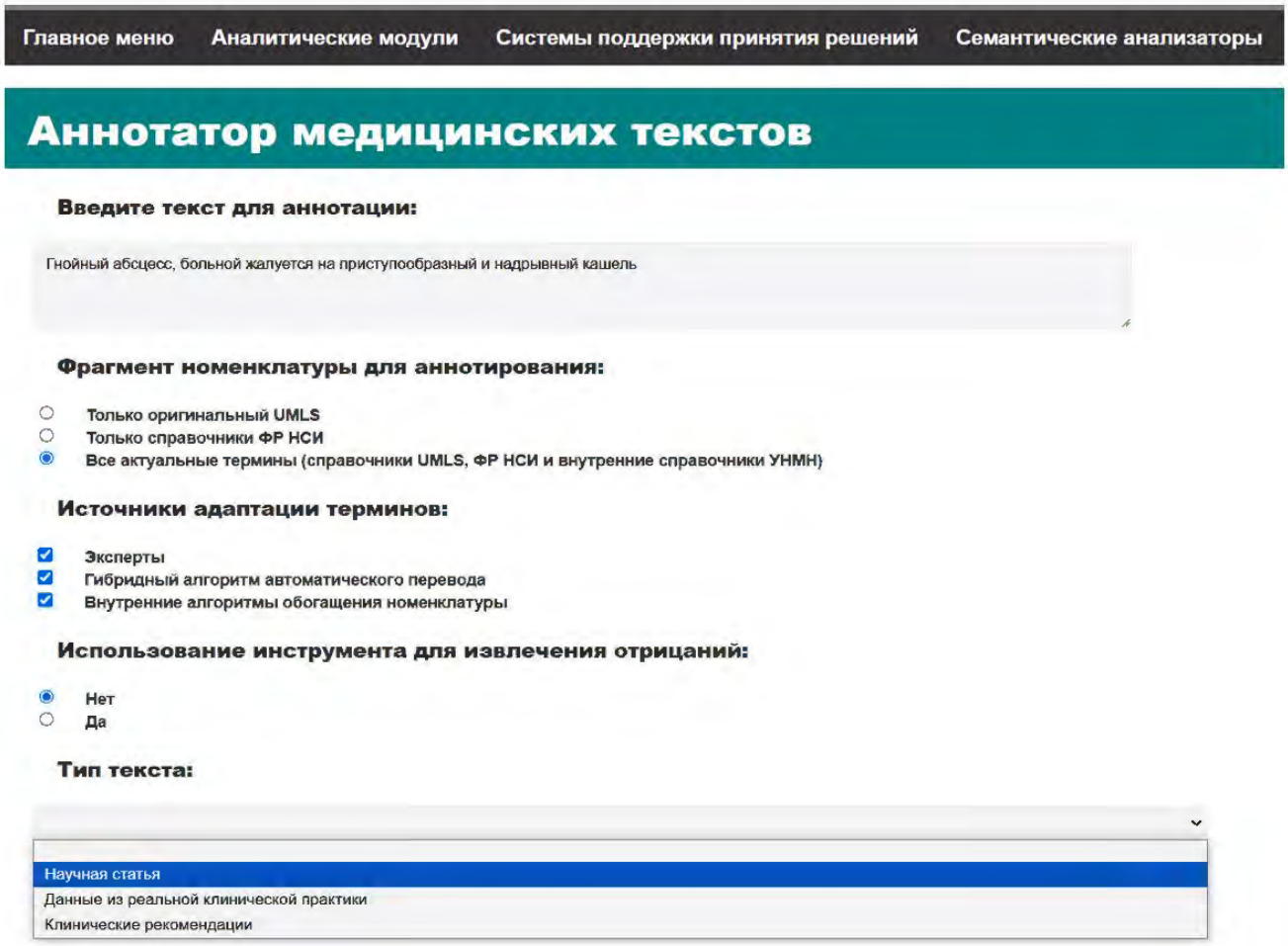

Рисунок 4. Интерфейсное решение для пользовательской работы с Аннотатором медицинских текстов
Figure 4. Interface solution for user work with Medical Text Annotator

временных затрат при увеличении размера графа предложения и числа терминов УНМН, которые необходимо найти в аннотируемом тексте.

Поскольку количество формулировок УНМН превышает 13 млн, единственным оптимизационным решением могло стать снижение числа терминов, выступающих в качестве потенциальных кандидатов для поиска. В связи с вышесказанным в список кандидатов включались только те формулировки УНМН, у которых все леммы исходно присутствовали в аннотируемом тексте. Данный способ является разновидностью полнотекстового поиска с перестановкой слов с бесконечным сдвигом и имеет максимальную чувствительность (в сравнении со всеми другими методами полнотекстового поиска).

Таким образом, с целью оптимизации работы Аннотатора для подбора кандидатов использовался высокочувствительный метод полнотекстового поиска с перестановкой слов, а для отсеивания лишних терминов и формирования итогового списка извлеченных сущностей применялся высокоточный, но затратный по времени поиск графовых изоморфизмов.

Итоговый оптимизированный алгоритм АМТ был реализован в виде микросервиса, доступного посредством программного интерфейса приложения (API – Application programming interface) и через аналитическую систему (АС) для работы с медицинскими онтологиями [11]. Сервис принимает на вход неструктурированный медицинский текст, внутренние идентификаторы источников адаптации и справочников терминов УНМН, а также расширяемый в настоящее время набор дополнительных переменных для уточнения запроса (рис. 4). Аннотатор обеспечивает категоризацию выделенных из текста формулировок путем указания их принадлежности к концептам (и соответствующим им тематическим группам) из сформированного с помощью фильтров подмножества УНМН.

В настоящее время в интерфейсе АС присутствует автоматизированное рабочее место (АРМ) аналитика. Элементы данного АРМ использовались при первичном тестировании Аннотатора, которое заключалось в сопоставлении результатов его работы с результатами решения задачи NER экспертным способом. В процессе валидации результатов тестирования принимали участие четверо врачей-экспертов с высшим образованием по специальности «Медицинская кибернетика». Двое экспертов имели ученую степень кандидата медицинских наук.

Каждый эксперт работал с разделами КР «Этиология и патогенез» и «Диагностика» (или схожими по наименованию и по смыслу разделами), произвольно выбирая по десять абзацев и выписывая из них все клинически значимые термины. Формируемые списки терминов счита-

лись эталонными и использовались при количественной оценке результатов автоматического аннотирования. Таким образом, каждый текст имел два списка выделенных из него терминов: один из них получался автоматически, а второй – экспертным способом.

Результат работы Аннотатора был тесно связан с терминологическим обогащением УНМН и всегда укладывался в один из четырех возможных сценариев (исходов, ситуаций). Первый вариант исхода предполагал наличие термина в УНМН и его успешное нахождение Сервисом в тексте. Доля данного сценария в общей структуре результатов работы АМТ определяла фактическое качество аннотирования медицинского текста с привязкой к текущей версии УНМН. Вторая ситуация предполагала наличие термина в УНМН при его отсутствии в числе найденных при аннотировании. Доля случаев, соответствующих данному исходу, позволяет дать оценку работы непосредственно алгоритма аннотирования и определить долю ошибок, связанных с неправильными сегментацией, токенизацией, лемматизацией, синтаксическим или морфологическим разборами текста. Третья ситуация объединяла случаи, при которых термин не находился Аннотатором по причине отсутствия соответствующей формулировки в УНМН. В качестве примера можно привести пару терминов «Лейциноз» и «Болезнь кленового сиропа», которые являются синонимичными и относятся к одному концепту, однако один из них мог отсутствовать в УНМН на момент тестирования Сервиса по причине отсутствия такой формулировки. Доля данного сценария позволяет дать системную количественную оценку экспертной проработки терминологии УНМН: чем меньше их доля, тем выше степень покрытия клинической области релевантными формулировками. Четвертый исход предполагал отсутствие не только необходимой формулировки, но и концепта, который присутствовал в анализируемом предложении.

Проблемы, отнесенные к третьему и четвертому сценариям, ликвидировались экспертами за счет добавления новых объектов (формулировок и концептов) с использованием редактора базы знаний – специального фрагмента автоматизированного рабочего места, представленного в аналитической системе для работы с УНМН. Важно отметить, что ошибочно извлеченные сервисом термины не всегда учитывались при подсчете долей исходов. Причины неправильного отнесения термина к сценарию были связаны преимущественно с человеческим фактором. Кроме того, качество аннотирования могло искажаться из-за омонимичных терминов (имеющих множество разных значений). Следует отметить, что текущие условия тестирования предполагали поиск составных терминов, поскольку эксперт мог выделить как цельный термин «острая боль через полчаса или час после еды», так и два отдельных термина «острая боль» и «боль после еды». По этой причине выдача Аннотатором всех трех вариантов не считалась ошибкой, если они совпадали по смыслу с содержанием текста КР.

Результаты тестирования АМТ позволили установить, что использование текущей версии УНМН обеспечивает автоматическое извлечение 63,0% (204 из 324) терминов из неструктурированного текста КР. Однако необходимо отметить, что значительная часть концептов (28,1%; n = 91) не была исходно представлена в УНМН, что привело к занижению метрик качества работы Аннотатора при рассмотрении его в качестве универсального инструмента для решения задачи NER (а не в качестве средства для разметки медицинских текстов). Кроме того, 4,0% (*n* = 13) концептов не имели в своем составе необходимой клинической формулировки. Качество решения задачи NER Сервисом (при допущении, что УНМН содержит достаточное количество клинических формулировок) составило 92,7%, поскольку 204 из 220 имевшихся формулировок были лемматизированы и извлечены из графа предложения верно.

Обсуждение

В настоящее время NER считается одним из наиболее сложных направлений NLP. Результаты решения подобного рода задач широко применимы как в научно-исследовательской деятельности, так и в реальной клинической практике.

Наиболее эффективными инструментами решения задач NLP (в частности, NER) признаются LLM, обучаемые на массивных корпусах текстов и хранящие в неинтерпретируемом виде закономерности естественного языка. Поскольку зависимость частоты употребления термина от его специфичности подчинена распределению Парето, правомерно считать, что в LLM, обучаемых на всем доступном корпусе текстов, закономерности естественного языка отражаются крайне неравномерно [19]. В связи с этим наибольшие трудности при автоматизированном анализе неструктурированной информации возникают при работе с узкоспециализированной терминологией, закономерности использования которой не всегда покрываются LLM должным образом.

В рамках настоящего исследования был создан гибридный инструмент для решения задач NER и категоризации извлекаемых терминов при работе с неструктурированными текстами. В текущий момент Аннотатор обеспечивает извлечение только терминов, однако его развитие и наполнение правилами способно преобразовать его в полноценный инструмент для извлечения связей и наполнения ими баз знаний. Сервис основан на использовании гибридного подхода (LLM и символьных правил) и является одним из первых в своем роде инструментов для решения задачи NER при работе с медицинскими текстами на русском языке. Качество работы АМТ сопоставимо с зарубежными аналогами (MetaMap, cTAKES) и при работе с текстами КР может достигать 93%. Созданный инструмент может применяться для анализа текстов на русском языке с использованием любого множества терминов.

Аннотатор позволяет находить термины, представленные в тексте разорванными фразами, что отличает его от всех существующих на сегодняшний день алгоритмов NER для русскоязычных текстов. Важно отметить, что АМТ способен использовать любую терминологическую базу (не только медицинскую), однако при использовании УНМН качество решения задачи NER будет максимально высоким ввиду масштабности данной терминологической системы. Существует возможность тонкой настройки запроса и ограничения перечня терминов для разметки на основе всего атрибутивного состава концептов и формулировок УНМН.

Использование концептов УНМН в качестве терминологического свода для аннотирования текста сохраняет обратную совместимость разметки, в связи с чем последняя не потеряет своей актуальности даже после обновления федеральных или международных справочников.

Следует отдельно подчеркнуть, что термины ежедневно обновляются за счет актуализации федеральных и международных справочников, входящих в ее состав, а также за счет непрерывной работы экспертов и алгоритмов автоматического обогащения УНМН.

Для использования текущей версии Аннотатора установлено два ограничения. Первое связано с максимально допустимым количеством слов в термине (не более 16), второе – с максимально допустимой длиной вводимого текста (не более 1000 слов). Правило, связанное с ограничением максимальной длины термина, формировалось эмпирически и предназначалось для удаления из свода медицинских терминов дефиниций и анкетных данных, которые также присутствуют в справочниках УНМН. В свою очередь ограничение длины вводимого текста в 1000 слов, или 2048 токенов, наследовалось от базовой модели LLM (Ru-bert-case), на основе которой строился Сервис.

Важно отметить, что текущая версия сервиса предназначена для работы с текстом, написанным строго в соответствии с правилами русского языка. Качество извлечения терминов из неструктурированного текста может заметно снижаться при наличии в нем орфографических ошибок. Менее чувствительным по отношению к пунктуационным ошибкам является Аннотатор медицинских текстов. Следует отметить, что в настоящее время устранение лексической неоднозначности реализовано на уровне контекста предложения для всех терминов УНМН, кроме аббревиатур.

К перспективам настоящего исследования относятся разработка алгоритмов коррекции ошибок в медицинских текстах и разрешения двойственности аббревиатур, совершенствование гибридного алгоритма лемматизации, формирование свода правил для распознавания полярности упоминания термина (утвердительный – affirmation или отрицательный – negation), а также обогащение УНМН новыми концептами и формулировками.

Заключение

Разработанный Аннотатор медицинских текстов позволяет автоматически извлекать до 93% терминов из неструктурированного текста КР. Качество работы данного сервиса сопоставимо с зарубежными аналогами при работе с текстом на английском языке: cTAKES с точностью в 91% и MetaMap – с F1-score в 88%.

Созданный программный продукт способен использовать любую терминологическую систему, в том числе УНМН, объединяющую более 13 млн формулировок из международных и федеральных справочников. На сегодняшний день аналогичных решений, обеспечивающих подобную автоматическую обработку медицинских текстов на русском языке с привязкой извлеченных сущностей к единой терминологической системе, в доступной литературе не обнаружено.

Литература / References

1. Гусев А.В., Зингерман Б.В., Тюфилин Д.С., Зинченко В.В. Электронные медицинские карты как источник данных реальной клинической практики. *Реальная клиническая практика: данные и доказательства*. 2022;2(2):8–20. https://doi.org/10.37489/2782-3784-myrwd-13
Gusev A.V., Zingerman B.V., Tjufilin D.S., Zinchenko V.V. Electronic medical records as a source of real-world clinical data. *Real-world data & evidence*. 2022;2(2):8–20. (In Russ.). https://doi.org/10.37489/2782-3784-myrwd-13
2. Лебедев С.В., Жукова Н.А. Слияние медицинских данных на основе онтологий. *Онтология проектирования*. 2017;7(2):145–159. https://doi.org/10.18287/2223-9537-2017-7-2-145-159
Lebedev S.V., Zhukova N.A. Ontology-driven approach to medical data fusion. *Ontology of Designing*. 2017; 7(2):145–159. (In Russ.). https://doi.org/10.18287/2223-9537-2017-7-2-145-159
3. Demner-Fushman D., Chapman W.W., McDonald C.J. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*. 2009;42(5):760–772. https://doi.org/10.1016/j.jbi.2009.08.007
4. Aronson A.R., Lang F.M. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*. 2010;17(3):229–236. https://doi.org/10.1136/jamia.2009.002733
5. Hunter L.E. Life sciences linkout. *Journal of Biomedical Informatics*. 2006;39(2):192–202. https://doi.org/10.1016/j.jbi.2005.09.006
6. Humphreys B.L, Tuttle M.S. Something new and different: The Unified Medical Language System. *Information services & use*. 2022;42(1):95–106. https://doi.org/10.3233/ISU-210138
7. Зарубина Т.В., Раузина С.Е., Астанин П.А. Создание базы медицинских знаний на основе национального метатезауруса для унификации разработки систем поддержки принятия клинических решений. *Вестник Российской академии медицинских наук*. 2024;79(2):175–192. https://doi.org/10.15690/vramn17390
Zarubina T.V., Rauzina S.E., Astanin P.A. Creation of a Medical Knowledge Base for Unify the Development of Clinical Decision Support Systems Based on the National Metathesaurus. *Annals of the Russian Academy of Medical Sciences*. 2024;79(2):175–192. (In Russ.). https://doi.org/10.15690/vramn17390
8. Reátegui R., Ratté S. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. *BMC Medical Informatics and Decision Making*. 2018;18(S3):74. https://doi.org/10.1186/s12911-018-0654-2
9. Астанин П.А., Ронжин Л.В., Федоров А.А., Раузина С.Е., Зарубина Т.В. Автоматизированная система извлечения аббревиатур терминов унифицированной национальной медицинской номенклатуры из текстов научных статей. *Врач и информационные технологии*. 2023;4:24–35. https://doi.org/10.25881/18110193_2023_4_24
Astanin P.A., Ronzhin L.V., Fedorov A.A., Rauzina S.E., Zarubina T.V. Automated abbreviations recognition system for unified national medical nomenclature filling with using Russian language unstructured text of articles. *Medical doctor and information technologies*. 2023;4:24–35. (In Russ.). https://doi.org/10.25881/18110193_2023_4_24
10. Астанин П.А., Раузина С.Е., Зарубина Т.В. Построение этиопатогенетического образа концептов метатезауруса UMLS с использованием графовых метрик. *Программные системы: теория и приложения*. 2023;14(3):59–94. https://doi.org/10.25209/2079-3316-2023-14-3-59-94
Astanin P.A., Rauzina S.E., Zarubina T.V. Computing the UMLS concepts etiopathogenetic image using graph metrics. *Program Systems: Theory and Applications*. 2023;14(3):59-94. (In Russ.). https://doi.org/10.25209/2079-3316-2023-14-3-59-94
11. Астанин П.А., Раузина С.Е., Зарубина Т.В. Автоматизированная система извлечения клинически релевантных терминов UMLS из текстов англоязычных статей на примере аксиального спондилоартрита. *Социальные аспекты здоровья населения*. 2023;69(3):14. https://doi.org/10.21045/2071-5021-2023-69-3-14
Astanin P.A., Rauzina S.E., Zarubina T.V. Automated system for recognizing clinically relevant UMLS terms in texts of the English-language articles exemplified by axial spondyloarthritis. *Social'nye aspekty zdorov'a naselenia*. 2023;69(3):14. (In Russ.). https://doi.org/10.21045/2071-5021-2023-69-3-14
12. Abdaoui A., Pradel C., Sigel G. Load what you need: Smaller versions of multilingual BERT. In: Proceedings of SustaiNLP / EMNLP; 2020. https://doi.org/10.48550/arXiv.2010.05609
13. Droganova K., Lyashevskaya O., Zeman D. Data conversion and consistency of monolingual corpora: Russian UD treebanks. Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories

(TLT 2018); December 13–14, 2018; Oslo University, Norway. Linköping Electronic Conference Proceedings 155;7:52–65.
14. Marneffe M.-C., Manning C., Nivre J., Zeman D. Universal Dependencies. *Computational Linguistics*. 2021;47(2):255–308. https://doi.org/10.1162/coli_a_00402
15. Qi P., Zhang Y., Zhang Y., Bolton J., Manning C.D. Stanza: A Python natural language processing toolkit for many human languages. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2020;101–108. https://doi.org/10.18653/v1/2020.acl-demos.14
16. Дяченко П.В., Иомдин Л.Л., Лазурский А.В., Митюшин Л.Г., Подлесская О.Ю., Сизов В.Г. и др. Современное состояние глубоко аннотированного корпуса текстов русского языка (СинТагРус). В кн.: Национальный корпус русского языка: 10 лет проекту. Труды Института русского языка им. В.В. Виноградова. М.; 2015:272–299.
Dyachenko P.V., Iomdin L.L., Lazursky A.V., Mityushin L.G., Podleskaya O.Yu., Sizov V.G., Frolova T.I., Tsinman L.L. The Current State of the Deeply Annotated Corpus of Russian Language Texts (SinTagRus). In: National Corpus of the Russian Language: 10 Years of the Project. Proceedings of the V.V. Vinogradov Institute of Russian Language. Moscow; 2015:272–299. (In Russ.).
17. Гращенко Л.А. О модельном стоп-словаре. *Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук*. 2013;1(150):40–46.
Grashhenko L.A. Application of modeling stop-list. *News of the National Academy of Sciences of Tajikistan. Department of physical, mathematical, chemical, geological and technical sciences*. 2013;1(150):40–46. (In Russ.).
18. Asiler M., Yazıcı A. BB-Graph: A subgraph isomorphism algorithm for efficiently querying big graph databases. Preprint [arXiv:1706.06654]; 2018. https://doi.org/10.1234/abcd.5678
19. Синева И.С., Головченко В.Е. Применение методов многомерного статистического анализа и NLP для классификации научных публикаций. *DSPA: Вопросы применения цифровой обработки сигналов*. 2024;14(2):44–51.
Syneva I.S., Golovchenko V.E. Application of multidimensional statistical analysis and NLP methods for classifying scientific publications. *DSPA: Digital Signal Processing*. 2024;14(2):44–51. (In Russ.).

Информация о вкладе авторов

Ронжин Л.В. – дизайн исследования, проектирование и реализация компонентов сервиса, подготовка текста рукописи.

Астанин П.А. – реализация компонентов сервиса, статистический анализ результатов, подготовка и оформление текста рукописи.

Раузина С.Е. – подготовка текста рукописи, общее руководство исследованием.

Ядгарова П.А. – подготовка текста рукописи.

Зарубина Т.В. – подготовка текста рукописи, общее руководство исследованием.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Information on author contributions

Ronzhin L.V. – study concept, design and implementation of service components, preparation of manuscript.

Astanin P.A. - implementation of service components, statistical analysis of results, preparation and formatting of manuscript.

Rauzina S.E. - preparation of the manuscript, general supervision of the research.

Yadgarova P.A. – preparation of the manuscript.

Zarubina T.V. - preparation of the manuscript, general supervision of the research.

Conflict of interest: the authors declare no conflict of interest.

Сведения об авторах

Ронжин Лев Вячеславович, аналитик лаборатории семантического анализа медицинской информации, РНИМУ им. Н.И. Пирогова, Москва, Россия, http://orcid.org/0000-0002-4653-1611.

E-mail: levronzhin@gmail.com.

Астанин Павел Андреевич, аналитик лаборатории семантического анализа медицинской информации, ассистент кафедры медицинской кибернетики и информатики им. С.А. Гаспаряна, РНИМУ им. Н.И. Пирогова, Москва, Россия, http://orcid.org/0000-0002-1854-8686.

E-mail: med_cyber@mail.ru.

Раузина Светлана Евгеньевна, канд. мед. наук, доцент, заведующий лабораторией семантического анализа медицинской информации, доцент кафедры медицинской кибернетики и информатики им. С.А. Гаспаряна, РНИМУ им. Н.И. Пирогова, Москва, Россия, http://orcid.org/0000-0002-9535-2847.

E-mail: rauzina@mail.ru.

Ядгарова Полина Алексеевна, аналитик лаборатории цифрового развития медицинского образования, РНИМУ им. Н.И. Пирогова, Москва, Россия, http://orcid.org/0000-0002-5105-3124.

E-mail: polina@yadgarova.ru.

Зарубина Татьяна Васильевна, д-р мед. наук, профессор, чл.-корр. РАН, директор Института цифровой трансформации медицины, заведующий кафедрой медицинской кибернетики и информатики им. С.А. Гаспаряна, РНИМУ им. Н.И. Пирогова, Москва, Россия, http://orcid.org/0000-0002-4403-8049.

E-mail: t_zarubina@mail.ru.

Астанин Павел Андреевич, e-mail: med_cyber@mail.ru.

Поступила 08.04.2025;
рецензия получена 30.04.2025;
принята к публикации 21.05.2025.

Information about the authors

Lev V. Ronzhin, Analyst, Laboratory of Semantic Analysis of Medical Information, Pirogov Russian National Research Medical University, Moscow, Russia, http://orcid.org/0000-0002-4653-1611.

E-mail: levronzhin@gmail.com.

Pavel A. Astanin, Analyst, Laboratory of Semantic Analysis of Medical Information; Assistant, S. A. Gasparyan Department of Medical Cybernetics and Computer Science, Pirogov Russian National Research Medical University, Moscow, Russia, http://orcid.org/0000-0002-1854-8686.

E-mail: med_cyber@mail.ru.

Svetlana E. Rauzina, PhD, Associate Professor, Head of the Laboratory of Semantic Analysis of Medical Information; Associate Professor, S. A. Gasparyan Department of Medical Cybernetics and Computer Science, Pirogov Russian National Research Medical University, Moscow, Russia, http://orcid.org/0000-0002-9535-2847.

E-mail: rauzina@mail.ru.

Polina A. Yadgarova, Analyst, Laboratory of Digital Development of Medical Education, Pirogov Russian National Research Medical University, Moscow, Russia, http://orcid.org/0000-0002-5105-3124.

E-mail: polina@yadgarova.ru.

Tatyana V. Zarubina, MD, Professor, Corresponding Member of the Russian Academy of Sciences; Director of the Institute of Digital Transformation of Medicine; Head of the S.A. Gasparyan Department of Medical Cybernetics and Computer Science, Pirogov Russian National Research Medical University, Moscow, Russia, http://orcid.org/0000-0002-4403-8049.

E-mail: t_zarubina@mail.ru.

Pavel A. Astanin, e-mail: med_cyber@mail.ru.

Received 08.04.2025;
review received 30.04.2025;
accepted for publication 21.05.2025.