Генерация врачебных заключений и классификация по Bethesda с использованием глубокого обучения

Е.В. Боброва, А.Ж. Маканов, С.С. Основин, Е.В. Дюльдин, Б.М. Шифман, К.С. Зайцев

Аннотация. **Целью настоящей статьи является исследование подходов к интеллектуальной обработке русскоязычной текстовой медицинской информации (NLP) цитологического описания заболеваний для решения задач классификации, генерации текста медицинского заключения и аугментации описаний при их острой нехватке. За последнее десятилетие область биомедицины в нашей стране не претерпела значительных изменений. Подходы для анализа проблем пациентов в большинстве своём основаны на ручной обработке и экспертных знаниях врачей. В работе рассмотрено создание конвейера машинного обучения, содержащего полный цикл предобработки данных и обучения моделей в области выявления заболеваний щитовидной железы по методике классификации Bethesda. Для проектирования архитектуры моделей глубокого обучения были использованы последовательные и трансформенные нейронные сети. Предложены подходы по очистке и предобработке «сырых» врачебных описаний к требуемому виду. Полученные результаты показали, что последовательные нейронные сети имеют большую точность на малых наборах данных, а трансформенные архитектуры превосходят другие при генерации врачебных заключений на больших объемах данных. Полученное в исследовании решение может быть использовано, как дополнительный справочный инструмент при работе врача-цитолога щитовидной железы.**

Ключевые слова **– биомедицина, глубокое обучение, трансформер, эмбеддинг, щитовидная железа, цитология, пункционная биопсия, классификация Bethesda, цитопатология щитовидной железы, опухоли щитовидной железы, узловой зоб.**

I. ВВЕДЕНИЕ

Попытки интегрировать методы обработки естественного языка (NLP) в области, где долгое время существовали только классические подходы, становятся с каждым годом всё более явными. Использование новых подходов обосновано тем, что классические методики часто не дают желаемых результатов, и требуют постоянных доработок. Поэтому все чаще и настойчивее внедряются современные технологии обработки текстов естественного языка в ранее не охваченные ими области.

Одной из таких областей является медицина, где для фиксации отдельных результатов используется естественный язык. Имея описанные врачом на естественном языке фиксируемые признаки заболевания требуется решить задачи классификации заболевания по принятой шкале и автоматической генерации текста врачебного заключения с указанием всех параметров, используемых для последующего лечения.

Рассмотрим постановку и решение задач работы с медицинскими текстами естественного языка на примере классификации и генерации врачебного заключения по описанию в области цитологических исследований щитовидной железы.

Исходными данными являются:

- международная система классификации Bethesda (The Bethesda System for Reporting Thyroid Cytopathology, сокр. TBSRTC) [1-2], согласно которой результат исследования материала, полученного в ходе тонкоигольной аспирационной биопсии образования щитовидной железы описывается в виде одной из 6 диагностических категорий. Каждой категории соответствует определенный риск злокачественности (от 4 до 97%), определяющий тактику дальнейшего ведения пациента.

- корпус пар реальных текстов (описание - заключение) сделанных врачами-цитологами. Правда в этом корпусе во многих парах отсутствует первый элемент – описание, в связи с его размещением в другом поле вместе с заключением.

Необходимо построить последовательность математических моделей (конвейер) для формирования врачебного заключения по описанию с указанием категории Bethesda. Результаты решения этой задачи могут быть использованы в различных приложениях: для быстрой консультации клиницистам в разных уголках страны; как обучающая программа студентам-медикам в виртуальных тренажерах.

Задача в своей общей постановке в разных областях деятельности решалась несколькими исследовательскими группами. Поэтому резонно рассмотреть наиболее близкие подходы к классификации и генерации токенов текста, проблемам временной сложности, пространственного хранения и обработки сырых массивов текста.

В настоящее время существует несколько обзорных работ, которые предлагают алгоритмы классификации текстов или выдвигают интересные идеи по оценке корпусов медицинских текстов. Так, в работе [3] обсуждается использование сложных

подходов на основе автоэнкодеров и применения сверточных сетей для классификации медицинских текстов. Интересной является идея смешивания полученных и сгенерированных данных при обучении, как метода повышения качества итоговой модели. Для нас это важно, так как для увеличения корпуса сырых данных отсутствующих медицинских описаний приходится решать задачу их аугментации по имеющимся медицинским заключениям.

В работе [4], посвященной обучению с нуля алгоритмов классификации текстов и получению наилучшего представления только на схожих данных, описаны подходы получения результатов обучения на сырых данных в противовес широко распространённым подходам дообучения архитектур, наподобие BERT (Bidirectional Encoder Representations from Transformers).

Авторы статей [5-7] привлекают к решению аналитических медицинских задач архитектуры, основанные на последовательных и трансформенных сетях, которые можно применить и к нашей задаче, для передачи знаний в кросс-доменной области.

Одной из задач настоящей работы является классификация по тексту описания заболевания - формирование категории Bethesda [8] с использованием глубокого обучения. Используя новые подходы по предобработке текста и обучения моделей, как канонических трансформеров, так и их модификаций, мы будем пытаться увеличивать точность предсказаний относительно заданного целевого порога. Решение этой задачи лежит на пути создания конвейера по предобработке данных и выделению ключевой информации из сырых корпусов медицинских текстов. Модель конвейера является и классификатором целевой метки Bethesda, и генератором врачебного заключения по врачебному описанию проблем пациента.

Структура настоящего исследования представлена в виде последовательности секций. В первой секции описаны подходы в предобработке исходных данных. Во второй - создается базовая статистическая модель, по которой будут оцениваться более продвинутые архитектуры. Следующие секции описывают последовательные подходы, направленные на выдачу коротких заключений по описанию. Далее рассмотрены трансформенные технологии и варианты их обучения. В Заключении подводятся итоги, и делается валидация результирующих моделей.

II. ПОДГОТОВКА ДАННЫХ

При работе с медицинскими заключениями по классификации заболеваний щитовидной железы возникает задача разметки и очистки данных. Корпус реальных данных, который используется в исследовании, содержит более 27 тысяч заключений врачей, полученных в ослепленном виде (без персональных данных) за десятилетний период работы лаборатории цитологии и цитогенетики отдела патоморфологии Центра эндокринологии (2013-2023 гг). Медицинские тексты имеют слабую внутреннюю структуру, что приводит к проблемам очистки данных от лишних токенов и выделения основной информации необходимой для дальнейшей генерации признаков. Стоит заметить, что в одном врачебном описании может встретиться несколько меток Bethesda. Например, "п11-1-материал неинформативный: в мазке на фоне густого коллоида клетки периферической крови, тиреоциты не обнаружены (по Bethesda Thyroid Classification-категория I);" и далее в нем же "л11-2-пунктирован узел коллоидного в разной степени пролиферирующего зоба с кистозными изменениями (по Bethesda Thyroid Classification-категория II)."

Решая проблемы в очерёдности выделения целевой метки Bethesda и ключевых токенов описания на неразмеченном множестве данных, основной упор делается на первичное использование регулярных выражений и вероятностного поиска наиболее часто встречаемых токенов предложений. Для поиска этого паттерна применено регулярное выражения типа ‘[IV]+’, что помогает найти всевозможные комбинации метки Bethesda в рассматриваемом тексте. Предложения из исходного корпуса данных после очистки данных со сформированными метками, можно сгруппировать по классам, получив процент вхождения (таблица 1).

Таблица 1. Процентное вхождение меток в корпус

Класс (индекс)	Кол-во вхождений в класс	Процент вхождения Bethesda
1	3409	12.43
2	17799	69.42
3	1050	3.83
4	2160	7.88
5	1076	3.92
6	952	3.47

В этой таблице общее целевое поле представлено после разбиения групп сложных текстовых запросов на базовые (с одной меткой класса), что расширяет исходные группы нулевым классом, не затрагивая распределение ранее проставленных врачами меток Bethesda.

Частью процесса предобработки сырых данных для получения более объемного объектного пространства является поиск числа предложений в тексте и вероятности появления слов в каждом предложении, что является гиперпараметром при различных видах аугментации данных и валидации полученных заключений. На этапе предобработки разделены длинные заключения пациентов, которые проходят через несколько этапов решения проблем с щитовидной железой, и имеют несколько меток Bethesda. При разделении данных на логические элементы, где каждый элемент — это предложение с

меткой Bethesda, признаковое поле расширилось на 7.5 %.

Далее после разделения меток классификатора по врачебным описаниям и заключениям выделяются необходимые токены.

После очистки и создания результирующего набора данных с разделёнными метками Bethesda в промежутке от 1 до 6, получаем полное разделение данных для решения задач классификации метки Bethesda и генерации заключения по врачебному описанию проблемы щитовидной железы. Примеры заключений представлены в таблице 2.

Таблица 2. Примеры врачебных заключений

Заключение	Метка
Материал недостаточно информативный: в толстых мазках с обильной примесью эритроцитов обнаружены единичные дистрофичные тиреоциты	1
В мазках цитограмма характерная для коллоидного в разной степени пролиферирующего В-клеточного зоба с участками аденоматоза и кистозной дегенерацией	2
В мазке с гетерогенным клеточным составом среди обилия эритроцитов и коллоида обнаружены как группы полиморфных тиреоцитов с признаками зобной трансформации, так и скопления укрупненных эпителиальных клеток неправильной формы, с ядерным напластованием, разреженным хроматином, плотно расположенные в бесформенных и сосочкоподобных структурах	3
В мазке высокой клеточности с большой примесью крови - на фоне содержимого кистозно-геморрагической полости и элементов лимфоцитарного воспаления обнаружены скопления крупных эпителиальных клеток с широкой цитоплазмой, эксцентричными ядрами, выраженными дегенеративными изменениями, формирующие преимущественно смешанные структуры и расположенные разрозненно, более подозрительные в отношении фолликулярного образования щитовидной железы из В-клеток	4
в мазке на фоне кистозно-геморрагических изменений и лимфоцитарного воспаления обнаружены комплексы атипичных эпителиальных клеток с единичными внутриядерными включениями	5
в мазке обнаружены многочисленные изолированно расположенные группы клеток папиллярного рака щитовидной железы формирующего трабекулярные папиллярные структуры единичными внутирядерными псевдовключениями	6

Предварительно обработанные исходные данные (основные токены ключевых слов и очищенные сырые данные) сохраняем в формате .csv.

Этическая экспертиза. Протокол исследования рассмотрен и одобрен локальным этическим комитетом ФГ НМИЦ Эндокринологии Минздрава России (протокол № 14 от 25.07.2023)

III. СОЗДАНИЕ СТАТИСТИЧЕСКОЙ МОДЕЛИ

При построении сложной системы для классификации по Bethesda и генерации заключения на основе врачебного описания необходимо быть уверенным, что модель сможет быть результативной в плане временной и пространственной сложности.

В ранее рассмотренных работах для решения этих задач используются CNN (сверточные) и LSTM (нейронные с долгой краткосрочной памятью) сети с модификациями изначального алгоритма. Такие подходы неплохо справляются с поставленной задачей, но сложны в интерпретации, и медленны, что приводит к временным издержкам.

Для оценки временных и пространственных мощностей подходят базовые модели, на которых в дальнейшем будет сделан упор при валидации итоговой архитектуры.

Решая поэтапно задачи, появляется желание совместить подход embedding (числовой вектор, полученный из слов) и методы кластеризации для определения метки класса. Такой выбор помогает оптимизировать время выполнения запроса [9]. Мы же в настоящей работе используем подходы на основе FastText [5] и методы нормализации итоговых векторных представлений.

Первый метод, который мы рассмотрим, использует нормы токенов документа и кластеризацию с помощью алгоритма K-means. Этот метод позволяет достаточно быстро определить метку Bethesda. После нормировки и обучения алгоритма, делаем предсказания и сравниваем их с целевыми метками тестового набора данных.

В предложенном методе на каждом шаге выбираем наибольший кластер для максимизации метрики f-beta-score, как гармонического среднего между полнотой и точностью, что позволит получать усреднение между двумя значимыми метриками в задаче. Поскольку в алгоритме K-means нет биективного отображения кластеров на ранее полученные данные, это означает, что лучшему множеству меток целевого класса сначала присваивается метка, а затем она удаляется из пула меток. Чтобы не терять информацию, мы сначала рассматриваем наибольшие кластеры, а затем постепенно переходим к меньшим кластерам, так как при удалении кластера теряется его метка.

Следует отметить, что такому подходу присуще сильное наложение классов, поэтому при проставлении одной и той же целевой метки нескольким классам одновременно, используем жадный алгоритм отбора признаков, и получаем распределение меток по классам, как в таблице 3.

Таблица 3. Матрица кластеризации меток

	0	1	2	3	4	5	6
0	592	1131	987	1039	2085	1066	771
1	1	-	3126	-	-	1	-
2	259	2254	27	5	2	1	5
3	5	1	6041	-	-	-	-
4	93	5	2306	3	15	1	5
5	20	16	2978	3	57	8	170
6	2	2	2334	1	1	-	1

Проведя вычисление F_B (f-beta-score) по формуле вида гармонического среднего:

$$F_B = (1 + B^2) * (PR * Recall) \frac{1}{B^2 * (PR + Recall)}$$

где PR – точность, Recall – полнота, B – настраиваемый параметр смещения между точностью и полнотой, получим, что после цикла обучения и предсказания f-beta-score превосходит порог в 0.52 только в первом и четвёртом классах.

Такой подход не показывает себя хорошо на всём множестве данных, и изменение приоритетов классов ничего не меняет, так как происходит наложение значительного числа точек друг на друга.

При рассмотрении задачи кластеризация сразу отметим невозможность использования базовых реализаций DBSCAN и OPTICS [9], так как нас интересует полное поле предсказаний, а эти алгоритмы не имеют параметра настройки числа кластеров, потому что основаны на идеях точек доступности.

На этом этапе хотелось получить F_B>= 0.91. Не получив его при использовании embedding текстов, перейдём к более продвинутым методам из области NLP.

IV. РЕКУРРЕНТНЫЕ ПОДХОДЫ

Рекуррентные нейронные сети (RNN), являются очень популярными сегодня при обработке и анализе последовательностей данных, таких как тексты, временные ряды, аудио- или видеопотоки. Как известно, эти архитектуры содержат циклический слой, который позволяет передавать информацию из предыдущего по времени шага в следующий, запоминать и использовать предыдущую контекстную информацию для принятия решений в настоящем. Формально, RNN определяется следующими уравнениями:

$$h_t = \sigma h(W_x h * x_t + W_h h * h_{t-1} + b_h)$$
$$y_t = \sigma y\big(W_h y * h_t + b_y\big)$$

где x_t- входные данные на текущем шаге, h_t- "скрытое" состояние на текущем шаге, y_t- выходные данные на текущем шаге, W и b - матрицы весов и смещений, соответственно, σh и σy - функции активации для скрытого и выходного слоев, соответственно.

Одной из задач NLP, где рекуррентные нейронные сети показали себя очень эффективными, является предсказание диагноза по данным описания заболевания щитовидной железы. С помощью RNN можно анализировать результаты проведенных исследований пациента и предсказывать диагноз с высокой точностью. Исследовательская группа из университета Лейдена в Нидерландах провела исследование, в котором использовала рекуррентную нейронную сеть для определения диагноза заболеваний щитовидной железы на основе данных анализа крови пациентов [9]. Исследование показало, что RNN демонстрирует более высокую точность предсказания по сравнению с традиционными методами. В прошлом году другая группа исследователей из университета Стэнфорда и компании Google Brain разработала новую архитектуру для решения задач NLP, в том числе для предсказания диагноза по данным анализа щитовидной железы [10]. В этом исследовании рекуррентные нейронные сети использовались в сочетании с механизмом внимания, который позволяет сети сосредоточиваться на наиболее значимых элементах последовательности. Результаты исследования также показали высокую эффективность предложенного метода. Использование рекуррентных нейронных сетей в задаче предсказания диагноза по данным анализа щитовидной железы позволяет получить более точные результаты и значительно сократить время диагностики, что так же подтверждено автором статьи об использовании нейронных сетей в здравоохранении [11].

Один из самых распространенных подходов - использование рекуррентной нейронной сети (RNN) с LSTM-ячейками (Long Short-Term Memory). LSTM-ячейки позволяют сохранять информацию о предыдущих состояниях сети и хранить ее длительное время. LSTM-ячейка имеет компоненты:

Использование LSTM-ячеек позволяет рекуррентной нейронной сети сохранять и использовать информацию о предыдущих состояниях на длительное время, что делает ее особенно эффективной для обработки последовательных данных с долгосрочными зависимостями [12]. Это особенно важно, когда анализы щитовидной железы собираются несколько раз за определенный период.

Еще один подход - использование сверточных нейронных сетей (CNN) с 1D свертками. Этот подход особенно полезен при обработке данных, которые могут быть представлены в виде временных рядов, таких как несколько анализов щитовидной железы за определенный период времени, которые могут быть записаны как функции от времени:

(выход) = Conv1D(вход, фильтры) -> Активация -> Пулинг -> Распрямление -> Полносвязные слои(размер_выхода).

В этом случае 1D свертки позволяют выделять важные временные признаки, такие как пики и тренды, и использовать их для предсказания диагноза [13].

И, наконец, гибридные модели, объединяющие несколько разных архитектур нейронных сетей, которые могут быть более эффективными, чем каждая из них в отдельности [14]. Чтобы проиллюстрировать это утверждение, был проведен сравнительный анализ результатов работы рекуррентных нейронных сетей с LSTM-ячейками, сверточных нейронных сетей (CNN) с 1D свертками и гибридных моделей, объединяющих оба типа моделей. Результаты эксперимента, целью которого было выявление наиболее подходящего решения для предсказания, метки класса приведены в таблице 4.

Таблица 4. Сравнение моделей

Модель	Accuracy	Precision	Recall (%)	F1-score

	(%)	(%)		(%)
RNN + LSTM	80.7	81.5	79.9	80.7
CNN + 1D Conv	86.4	85.9	87.2	86.5
Hybrid Model	89.2	89.8	88.6	89.2

По таблице видно, что гибридная модель, которая объединяла в себе RNN и CNN, показала наилучшие результаты по всем метрикам. Однако, CNN-модель с 1D свертками довольно близко подошла по результатам, а также показала большую скорость обучения и простоту использования.

V. ПРИМЕНЕНИЕ ПРЕДОБУЧЕННЫХ МОДЕЛЕЙ

Сегодня существует большое количество моделей глубокого обучения, применяемых для обработки медицинских текстов. Наибольшее число таких моделей, работают в англоязычной области анализа текстов [15]. Но есть и модели, предобученные на русском языке. Для построения таких моделей использовались известные трансформеры BERT и RoBERTa, которые уже были предобучены на общеязыковых корпусах [16]. В настоящей работе применены две другие модели, предобученные на русскоязычной Википедии, и дополнительно на специальном корпусе данных Taiga, созданном командой SberDevice [17]. Обе эти модели были еще трансферно дообучены на открытых в общем доступе медицинских и биомедицинских текстах. Благодаря такому обучению, были получены модели RuBioBERT и RuBioRoBERTa [18-19].

Для лучшего понимания работы RuBioBERT и RuBioRoBERTa, напомним базовый принцип работы BERT. Он заключается в том, что при обучении нейронной сети маскируется слово не только в конце предложения, но и внутри него. Такая задача называется предсказанием маскированного токена (Masked Language Modeling task) [15]. Этот подход позволяет нейросети одновременно обучаться (выучивать эмбеддинги токенов) в обе стороны, чем достигается «глубокая двунаправленность» – модель учитывает контекст с двух сторон от слова. Также эта архитектура использует механизм внимания, помогающий лучше выделить контекст и взаимосвязь слов между собой.

По сути, модель RoBERTa является той же моделью BERT, но с оптимизированными гиперпараметрами и динамической маскировкой слов. Во время предобучения RoBERTa обучается только предсказанию замаскированного слова, тогда как в архитектуре BERT она также предобучалась на задаче «предсказание следующего предложения (next sentence prediction)», т.е. предсказывала, является ли в паре предложений второе предложение продолжением первого [16].

Для задачи определения связи текста заключения с классом Bethesda, выставленным врачом в заключении, использовались обе эти модели, чтобы понять какая из них работает лучше при решении задачи классификации. Для сравнения RuBioBERT и RuBioRoBERTa использовались параметры моделей (Таблица 5).

Таблица 5. Параметры RuBioBERT и RuBioRoBERTa

Параметр	Значение
Learning rate	2e-5
Weight decay	0.01
Train epochs	5
Optimization	Adam

Процесс обучения в каждой из моделей, фиксировался каждые 1000 шагов и тестировался. В результате чего были получены результаты обучения моделей (таблицы 6 и 7).

Таблица 6. Процесс обучения RuBioBERT

Step	Train Loss	Validation loss	Accuracy	F1
1000	0.32	0.24	0.946	0.945
2000	0.14	0.17	0.961	0.960
3000	0.1	0.24	0.93	0.935
4000	0.08	0.18	0.96	0.959
5000	0.06	0.22	0.95	0.951

Таблица 7. Процесс обучения RuBioRoBERTa

Step	Train Loss	Validation loss	Accuracy	F1
1000	0.26	0.20	0.953	0.952
2000	0.14	0.19	0.961	0.961
3000	0.10	0.17	0.970	0.969
4000	0.08	0.20	0.960	0.960
5000	0.06	0.19	0.953	0.961

Из таблиц видно, что RuBioRoBERTa показала более высокую точность, чем RuBioBERT.

Существенным отличием модели RuBioRoBERTa от RuBioBERT является значительное время обучения (в 3 раза выше времени оппонента) и существенно больший «вес» самой модели. В зависимости от частоты необходимого переобучения может быть выбрана и первая и вторая модели для классификации по Bethesda.

VI. АУГМЕНТАЦИЯ ТЕКСТОВЫХ ДАННЫХ

В реальном корпусе из 27000 используемых в настоящем исследовании пар данных «Описание - Заключение» реальные Описания присутствуют только в 6300 парах. Так же текст описания может быть перемещен в поле «Заключение» В остальных парах данных присутствует только Заключение. Учитывая, что при увеличении обучающей выборки растет качество модели, была дополнительно поставлена задача по аугментации недостающих описаний. Для этого, чтобы восстановить текст описания по тексту заключения, была обучена русскоязычная модель T5-base от Сбербанка.

Аугментация (на которую было затрачено гораздо больше времени, чем на fine-tuning модели T5-base) увеличила выборку пар почти в 4 раза до 24600 примеров. Анализ аугментированных данных показал, что:

- среднее количество слов в исходных (36.2) и аугментированных (51.9) текстах Описаний не сильно различаются;
- максимальное количество слов в исходных (177) и аугментированных (125) текстах.

Для анализа свойств исходных и синтетических данных было проведено тематическое моделирование LDA (Латентное размещение Дирихле) с помощью библиотеки genism, в которой начальный класс (первая категория Bethesda) помечается нулем, и далее возрастает. Количество тематик было взято равным 6, в соответствии с числом классов Bethesda

Распределение отсортировано по возрастанию вероятностей, для того чтобы можно было разглядеть общий паттерн. Распределения исходных и синтетических данных очень похожи (рисунки 2 и 3).

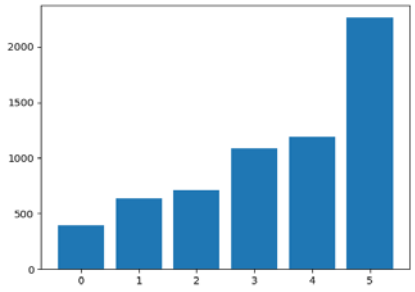


Рисунок 2. Распределение Описаний внутри тематик в исходных описаниях.

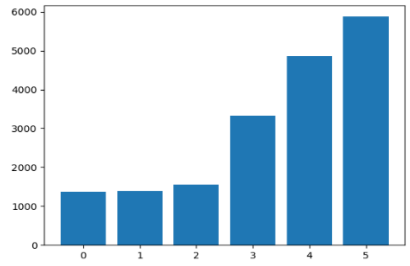


Рисунок 3 Распределение Описаний внутри тематик в аугментированных описаниях.

Помимо распределения Описания по тематикам, было построено распределение тематик по словам, чтобы выделить топ-слова, описывающие конкретную тематику (таблицы 8 и 9).

Таблица 8. Топ токены исходных описаний

Тематика	Топ-слова
0	Клетка, обнаружить, железа, щитовидный, фон
1	Зоб, коллоидный, узловой, пунктировать, элемент
2	Клетка, скопление, фон, препарат, расположить
3	Количество, небольшой, мазок, тиреоцит, фон
4	Эпителий, клетка, фолликулярный, обнаружить, фон
5	Позволять, рекомендовать, наблюдение, динамический, узловой

Таблица 9. Топ токены синтетических описаний

Тематика	Топ-слова
0	Элемент, кровь, периферический, голый, ядро
1	Эритроцит, округлый, ядро, мономорфный, относительно
2	Элемент, зоб, коллоидный, узловой, лимфоцитарный
3	Характерный, цитограмм, зоб, коллоидный, изменение, фолликулярный
4	Папиллярный, щитовидный, железа, эритроцит
5	Эритроцит, изменить, фон, группа, коллоид

Можно заметить, что лексика в тематиках почти не совпадает, однако конструктивно описания похожи

VII. ОЦЕНКА КАЧЕСТВА РЕЗУЛЬТАТОВ

Настройка (fine-tuning) русскоязычной T5-base модели на решение задачи генерации Заключения по Описанию проводилась на аугментированной (24600 примеров), и исходной (6200 примеров) выборках.

Затем, с помощью обученной модели были сгенерированы предсказанные заключения по имеющимся исходным описаниям, с помощью регулярных выражений из исходных заключений и по предсказанным заключениям были получены исходные классы Bethesda и предсказанные Bethesda. Точность (precision) классификации по описанию составила – 82.8%.

Аналогично, новая модель была использована, чтобы сгенерировать каждому описанию из аугментированного набора данных предсказанное заключение. С помощью регулярных выражений были получены метки Bethesda. Точность классификации на всем аугментированном наборе данных – 71.5%.

Точность на выборке, которая является сгенерированными заключениями на основе исходных описаний – 84.0%.

Точность классификации на выборке, состоящей только из синтетических описаний – 67.2%, что скорее всего связано с тем, что аугментация была выполнена не идеально.

Для оценки качества созданных моделей автоматической генерации рефератов врачебных заключений по тексту описания заболевания использовался набор метрик суммаризации текста ROUGE (ROUGE-1, ROUGE-2, ROUGE-L). Принято считать, что показатели ROUGE-2 и ROUGE-L хорошо подходят для задач реферирования отдельных документов, а ROUGE-1 и ROUGE-L показывают неплохие результаты на оценке коротких рефератов. Считая, что врачебное заключение может быть и коротким и средней длины, мы использовали все три показателя.

Ключевая идея метрик семейства ROUGE - оценка пересечения по n-граммам между исходным текстом

и сгенерированным (в нашем случае между текстом, написанным врачом, и сгенерированным моделью текстом). Метрика ROUGE-1 учитывает только отдельные слова (униграммы), придерживаясь философии "мешка слов", что не позволяет оценивать качество сочетаний слов. Эту проблему частично устраняют ROUGE-2 и ROUGE-L: они оценивают совпадения по биграммам и самой длинной последовательности слов соответственно. Две последние метрики более "требовательны" к генерации, чем ROUGE-1. Изменение логистической функции потерь (loss) и метрик семейства ROUGE в зависимости от шага батча – оптимизации архитектуры нейронной сети можно видеть на рисунках 4 и 5.

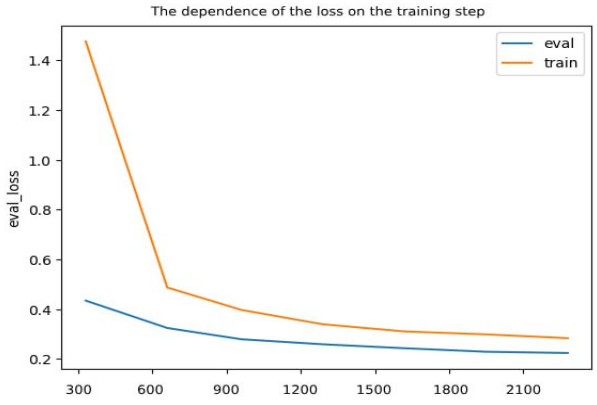
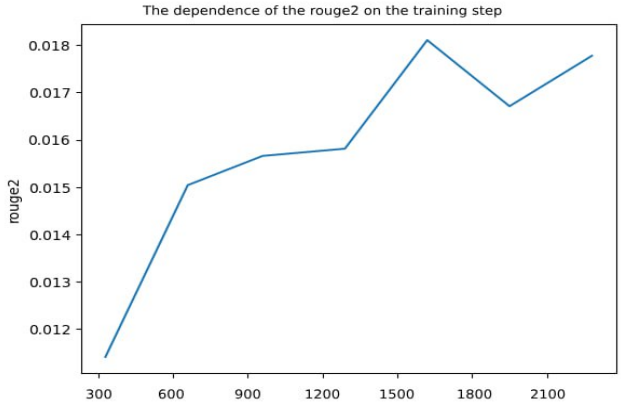
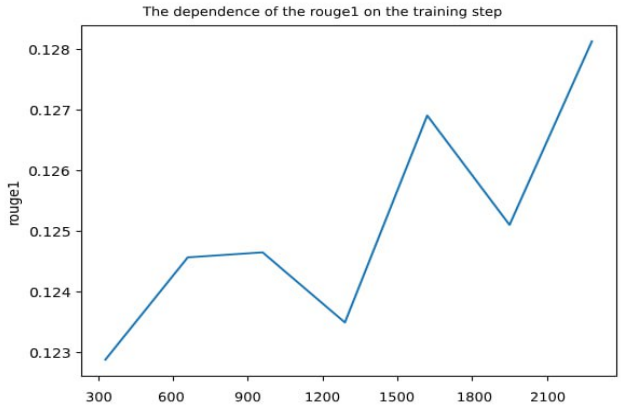


Рисунок 4. График функции потерь.





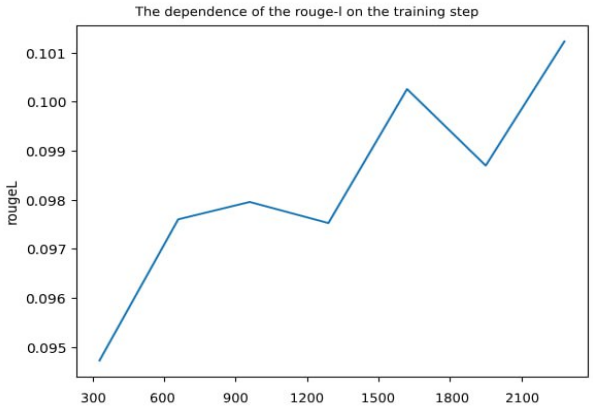


Рисунок 5. Графики метрик суммаризации ROUGE.

VIII. ОБСУЖДЕНИЕ

Методы преобразования текстов естественного языка в медицинских приложениях в настоящее время активно развиваются.

Решаемые задачи сложны в плане обработки данных и дальнейшего обучения модели, так как ошибки в доменах медицинских задач являются критичными, что заставляет более тщательно собирать данные с пониманием, каких результатов мы хотим добиться при аугментации готовыми или обученными моделями. Подходы к генерации и оценке массивов данных рассматриваются в статьях [9-10], но в нашей задаче, имея корпус готовых текстов, основное внимание было направлено на преобразование сырых данных, а не на попытки анонимизации личных данных пациентов, что было сделано не нами и заранее. Идеи предобработки легли в основу получения кластеров данных, но это в дальнейшем не показало ожидаемого ощутимого прироста целевых метрик.

Основные подходы, которые могли помочь решить поставленную задачу, опирались на последовательные генерации матричных ядер в рекуррентных сетях и архитектурах, основанных на свёртках. Эти подходы не новы, но необходимы для генерации требуемых меток исследуемого класса, частично идеи применённых методов освещены в работах [14-15].

Предложенные в статье подходы к решению задач предобработки данных, классификации и аугментации показали себя достаточно хорошо при генерации меток Bethesda, но для сложных последовательностей токенов оказались недостаточно эффективны. Это подводит к пониманию проблемы недостаточности данных в обучающей выборке. Для ее решения были предприняты попытки использовать готовые предобученные архитектуры с дообучением на целевом наборе данных, подобные методики основываются на алгоритмах, предложенных в [16].

Для получения текста полного описания, в результате генерации последовательности токенов

врачебного заключения можно использовать ранее предобученные модели Bert, но для расширения признакового пространства требуется аугментация данных и дообучение трансформенных моделей на имеющемся корпусе текстов. Именно это позволит получать более точные и полные врачебные заключения на выходе модели.

IX. ЗАКЛЮЧЕНИЕ

Результатом настоящей работы явилось создание конвейера. На первом его этапе для очистки и предобработки сырых данных использованы статистические методы, которые работают быстрее, чем нейронные сети и практически с тем же качественным результатом.

Далее были обучены модели трансформера, работающего в двух режимах, как генератор меток Bethesda и как генератор токенов заключения на основе входного описания.

Предложенный конвейер имеет интерфейс по предобработке и дообучению модели и может быть преобразован в сериализуемый .pkl файл.

Так как работа выполнялась на высокоуровневом языке python итоговые наработки могут быть масштабированы. С помощью библиотек sklearn для масштабирования интерфейса, и с помощью transformers для замены отдельных шагов преобразователя решаемых задач, что позволит оставить методы предварительной обработки данных из других медицинских областей неизменными.

В части исследовательских метрик было достигнуто улучшение на 2%, что связано с предложенными подходами обработки русскоязычных данных и совмещения методик дообучения готовых архитектур.

Настоящая работа может быть имплементирована в системы, которые опираются на автоматизацию врачебных предсказаний, и может помочь врачам делать предсказания значительно быстрее, чем при использовании ручных методов.

Рассматривая этическую составляющую данной работы, отметим, что полученный конвейер пока не является заменой врача, а позиционируется, как помощник и собеседник при формировании заключения по проблемам пациента.

БЛАГОДАРНОСТИ

Авторы выражают благодарность Высшей инжиниринговой школе НИЯУ МИФИ за помощь в возможности опубликовать результаты выполненной работы.

БИБЛИОГРАФИЯ

[1] Ali S, Cibas E. The Bethesda System for Reporting Thyroid Cytopathology. (Ali SZ, Cibas ES, eds.). Cham: Springer International Publishing; 2018. doi: https://doi.org/10.1007/978-3-319-60570-8

[2] Ali SZ, Baloch ZW, Cochand-Priollet B, Schmitt FC, Vielh P, VanderLaan PA. The 2023 Bethesda System for Reporting Thyroid Cytopathology. Thyroid®. July 2023. doi: https://doi.org/10.1089/thy.2023.0141

[3] Zixu Wang, Julia Ive, Sumithra Velupillai, Lucia Specia, Is artificial data useful for biomedical Natural Language Processing algorithms. Aug 2019.

[4] Vasilyev, Oleg & Bohannon, John. (2022). Neural Embeddings for Text. 10.48550/arXiv.2208.08386.

[5] YU GU, ROBERT TINN, HAO CHENG, MICHAEL LUCAS, NAOTO USUYAMA, XIAODONG LIU, TRISTAN NAUMANN, JIANFENG GAO, and HOIFUNG POON, Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. 16 Sep 2021.

[6] Houssein E.H., Mohamed R.E. and Ali A.A., "Machine Learning Techniques for Biomedical Natural Language Processing: A Comprehensive Review," in IEEE Access, vol. 9, pp. 140628-140653, 2021, doi: 10.1109/ACCESS.2021.3119621.

[7] Giacomo Miolo, Giulio Mantoan, Carlotta Orsenigo Department of Management, Economics and Industrial Engineering Politecnico di Milano, Italy. ELECTRAMED: A NEW PRE-TRAINED LANGUAGE REPRESENTATION MODEL FOR BIOMEDICAL NLP. Apr 2021.

[8] Naseem U, Khushi M, Reddy V, Rajendran S, Razzak I, Kim J. Bioalbert: a simple and effective pre-trained language model for biomedical named entity recognition. 2020.

[9] Dina R. Mody, Michael J. Thrall, Savitri Krishnamurthy. Diagnostic Pathology: Cytopathology (Second Edition). 2018.

[10] Lyu, C., Chen, B., Ren, Y. *et al.* Long short-term memory RNN for biomedical named entity recognition. *BMC Bioinformatics* 18, 462 (2017). https://doi.org/10.1186/s12859-017-1868-

[11] Гусев А.В. Перспективы нейронных сетей и глубокого машинного обучения в создании решений для здравоохранения // Врач и информационные технологии. 2017. №3. URL: https://cyberleninka.ru/article/n/perspektivy-neyronnyh-setey-i-glubokogo-mashinnogo-obucheniya-v-sozdanii-resheniy-dlya-zdravoohraneniya (дата обращения: 08.07.2023).

[12] Van Houdt, G., Mosquera, C. & Nápoles, G. A review on the long short-term memory model. *Artif Intell Rev* **53**, 5929–5955 (2020). https://doi.org/10.1007/s10462-020-09838-1

[13] Alzubaidi, L., Zhang, J., Humaidi, A.J. *et al.* Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* **8**, 53 (2021). https://doi.org/10.1186/s40537-021-00444-8

[14] Фартушный Э. Н., Сыч Ю. П., Фартушный И. Э., Кошечкин К. А., Лебедев Г. С. Стратификация узловых образований щитовидной железы по категориям EU-TIRADS с использованием трансферного обучения сверточных нейронных сетей // КЭТ. 2022. №2. URL: https://cyberleninka.ru/article/n/stratifikatsiya-uzlovyh-obrazovaniy-schitovidnoy-zhelezy-po-kategoriyam-eu-tirads-s-ispolzovaniem-transfernogo-obucheniya (дата обращения: 31.05.2023).

[15] Выучейская М.В., Крайнова И.Н., Грибанов А.В. Нейросетевые технологии в диагностике заболеваний (обзор) // Журнал медико-биологических исследований. 2018. №3. URL: https://cyberleninka.ru/article/n/neyrosetevye-tehnologii-v-diagnostike-zabolevaniy-obzor (дата обращения: 31.05.2023).

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, 2013b.

[17] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014.

[18] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pretraining of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010.

[21] Peng, Y., Yan, S., & Lu, Z. (2019). Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. BioNLP@ACL.
[22] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv, abs/1810.04805.
[23] Blinov, P., Reshetnikova, A., Nesterov, A., Zubkova, G., Kokh, V. (2022). RuMedBench: A Russian Medical Language Understanding Benchmark. In: Michalowski, M., Abidi, S.S.R., Abidi, S. (eds) Artificial Intelligence in Medicine. AIME 2022. Lecture Notes in Computer Science(), vol 13263. Springer, Cham. https://doi.org/10.1007/978-3-031-09342-5_38
[24] Yalunin, A., Nesterov, A., & Umerenkov, D. (2022). RuBioRoBERTa: a pre-trained biomedical language model for Russian language biomedical text mining. ArXiv, abs/2204.03951.
[25] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36, 1234 - 1240.

Статья получена 20 июля 2023.

Боброва Елизавета Витальевна, Национальный Исследовательский Ядерный Университет МИФИ, магистрант, EVBobrova@mephi.ru
Маканов Артем Жанович, Национальный Исследовательский Ядерный Университет МИФИ, бакалавр, artem.makanov@mail.ru
Основин Станислав Сергеевич, Национальный Исследовательский Ядерный Университет МИФИ, магистрант, 1300stas1300@gmail.com
Дюльдин Евгений Владимирович, Национальный Исследовательский Ядерный Университет МИФИ, аспирант, zhecos1@yandex.ru
Шифман Борис Михайлович, Национальный медицинский исследовательский центр эндокринологии, врач-эндокринолог, boris-11@mail.ru
Зайцев Константин Сергеевич, Национальный Исследовательский Ядерный Университет МИФИ, профессор, KSZajtsev@mephi.ru

Generating medical opinions and classification according to Bethesda using deep learning

E.V. Bobrova, A.Z. Makanov, S.S. Osnovin, E.V. Diuldin, B.M. Shifman, K.S. Zaytsev

Abstract. The purpose of this article is to study approaches to the intelligent processing of Russian-language medical textual information (NLP) of a cytological description of situations to solve the problems of detection, generation of observation text and augmentation of descriptions in case of their acute shortage. For a decade, the field of biomedicine in our country has not changed. Approaches for analyzing patient problems are in most cases based on manual processing and expert knowledge of physicians. The paper considers the creation of a machine production pipeline, a full cycle of data and model preprocessing in the field of measuring the incidence of the thyroid gland using the Bethesda protection method. The ideas of sequential and transformable neural networks were used to design the architecture of deep learning models. Approaches to cleaning and preprocessing information about "raw" medical descriptions that require detection are also considered. The obtained results show that subsequent neural networks are of great importance on small data sets, and the transformed architectures are superior to others when generating doctor circuits on large data sets. The solution obtained in the experiment can practically be used as an additional reference tool in the work of a cytologist to determine the thyroid gland.

Keywords – biomedicine, deep learning, transformer, embedding, thyroid gland, cytology, needle biopsy, Bethesda classification, thyroid cytopathology, thyroid tumors, nodular goiter

REFERENCES

[1] Ali S, Cibas E. The Bethesda System for Reporting Thyroid Cytopathology. (Ali SZ, Cibas ES, eds.). Cham: Springer International Publishing; 2018. doi: https://doi.org/10.1007/978-3-319-60570-8

[2] Ali SZ, Baloch ZW, Cochand-Priollet B, Schmitt FC, Vielh P, VanderLaan PA. The 2023 Bethesda System for Reporting Thyroid Cytopathology. Thyroid®. July 2023. doi: https://doi.org/10.1089/thy.2023.0141

[3] Zixu Wang, Julia Ive, Sumithra Velupillai, Lucia Specia, Is artificial data useful for biomedical Natural Language Processing algorithms. Aug 2019.

[4] Vasilyev, Oleg & Bohannon, John. (2022). Neural Embeddings for Text. 10.48550/arXiv.2208.08386.

[5] YU GU, ROBERT TINN, HAO CHENG, MICHAEL LUCAS, NAOTO USUYAMA, XIAODONG LIU, TRISTAN NAUMANN, JIANFENG GAO, and HOIFUNG POON, Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. 16 Sep 2021.

[6] Houssein E. H., Mohamed R. E. and Ali A. A., "Machine Learning Techniques for Biomedical Natural Language Processing: A Comprehensive Review," in IEEE Access, vol. 9, pp. 140628-140653, 2021, doi: 10.1109/ACCESS.2021.3119621.

[7] Giacomo Miolo, Giulio Mantoan, Carlotta Orsenigo Department of Management, Economics and Industrial Engineering Politecnico di Milano, Italy. ELECTRAMED: A NEW PRE-TRAINED LANGUAGE REPRESENTATION MODEL FOR BIOMEDICAL NLP. Apr 2021.

[8] Naseem U, Khushi M, Reddy V, Rajendran S, Razzak I, Kim J. Bioalbert: a simple and effective pre-trained language model for biomedical named entity recognition. 2020.

[9] Dina R. Mody, Michael J. Thrall, Savitri Krishnamurthy. Diagnostic Pathology: Cytopathology (Second Edition). 2018.

[10] Victoria Hutterer1 , Ronny Ramlau2 and Iuliia Shatokhina3. Real-time Adaptive Optics with pyramid wavefront sensors: Accurate wavefront reconstruction using iterative methods. October 1, 2018

[11] Lyu, C., Chen, B., Ren, Y. et al. Long short-term memory RNN for biomedical named entity recognition. BMC Bioinformatics 18, 462 (2017). https://doi.org/10.1186/s12859-017-1868-

[12] Fartushny E. N., Sych Yu. P., Fartushny I. E., Koshechkin K. A., Lebedev G. S. STRATIFICATION OF THYROID NODES BY EU-TIRADS CATEGORIES USING TRANSFER LEARNING OF CONVOLUTIONAL NEURAL NETWORKS // KET. 2022. №2. URL: https://cyberleninka.ru/article/n/stratifikatsiya-uzlovyh-obrazovaniy-schitovidnoy-zhelezy-po-kategoriyam-eu-tirads-s-ispolzovaniem-transfernogo-obucheniya (date of access: 05/31/2023).

[13] Van Houdt, G., Mosquera, C. & Nápoles, G. A review on the long short-term memory model. Artif Intel Rev 53, 5929–5955 (2020). https://doi.org/10.1007/s10462-020-09838-1

[14] Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8, 53 (2021). https://doi.org/10.1186/s40537-021-00444-8

[15] Vyucheyskaya M.V., Krainova I.N., Gribanov A.V. Neural network technologies in the diagnosis of diseases (review) // Journal of Biomedical Research. 2018. №3. URL: https://cyberleninka.ru/article/n/neyrosetevye-tehnologii-v-diagnostike-zabolevaniy-obzor (date of access: 05/31/2023).

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, 2013b.

[17] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014.

[18] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pretraining of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010.

[21] Peng, Y., Yan, S., & Lu, Z. (2019). Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. BioNLP@ACL.

[22] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv, abs/1810.04805.

[23] Blinov, P., Reshetnikova, A., Nesterov, A., Zubkova, G., Kokh, V. (2022). RuMedBench: A Russian Medical Language Understanding Benchmark. In: Michalowski, M., Abidi, S.S.R., Abidi, S. (eds) Artificial Intelligence in Medicine. AIME 2022. Lecture Notes in Computer Science(), vol 13263. Springer, Cham. https://doi.org/10.1007/978-3-031-09342-5_38

[24] Yalunin, A., Nesterov, A., & Umerenkov, D. (2022). RuBioRoBERTa: a pre-trained biomedical language model for Russian language biomedical text mining. ArXiv, abs/2204.03951.
[25] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36, 1234-1240.