УДК 004.8

Петров Александр Михайлович, магистрант,
АНО ВО «Российский новый университет»

ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ О ЛЕКАРСТВАХ И ЗАБОЛЕВАНИЯХ ИЗ МЕДИЦИНСКИХ СТАТЕЙ С ПРИМЕНЕНИЕМ ДООБУЧЕННОЙ МОДЕЛИ ROBERTA

Аннотация. Рассматривается задача автоматического извлечения информации о лекарственных препаратах и заболеваниях из научных медицинских статей. Предложен подход на основе трансформера RoBERTa, дообученного на датасетах BC5CDR, NCBI Disease и JNLPBA; F1-мера извлечения выросла с 0,67 до 0,85. Показано применение модели в контуре медицинской организации: время обработки документа сокращается с трёх часов до 20 минут, ручная проверка разметки – на 80 %.

Ключевые слова: Обработка естественного языка, извлечение именованных сущностей, RoBERTa, дообучение, медицинские тексты, трансформеры.

Введение

Объём научных публикаций в области медицины и фармакологии растёт ежегодно, и ручная обработка этого потока становится узким местом клинической практики. По данным обследования многопрофильной больницы, до 40 % рабочего времени клинических фармакологов уходит на ручное чтение рекомендаций и статей, а обработка одного документа – до трёх часов; база знаний «лекарства–заболевания» актуализируется вручную с задержкой в две–три недели.

Ключевым этапом автоматического извлечения информации из медицинских документов является распознавание именованных сущностей (Named Entity Recognition, NER) – поиск в тексте наименований лекарственных препаратов, заболеваний, симптомов и медицинских процедур. Целью настоящей работы является построение и исследование системы извлечения информации о лекарствах и заболеваниях из научных статей на основе модели-трансформера RoBERTa, дообученной на специализированных датасетах; подсистема NER рассматривается как основа автоматизации актуализации базы знаний в рамках проекта внедрения КИС медицинской организации.

Материалы и методы

Исторически задача NER решалась словарными методами, правилами и статистическими моделями, прежде всего условными случайными полями (Conditional Random Fields, CRF) [5], затем – комбинацией BiLSTM-CRF. Качественный сдвиг произошёл с появлением трансформеров [1]: механизм самовнимания оценивает семантические связи между всеми словами предложения одновременно. Модель BERT [1] обеспечила двунаправленный анализ контекста каждого токена, а RoBERTa [2] улучшила предварительное обучение за счёт динамического маскирования и отказа от предсказания следующего предложения. Доменные модели BioBERT [3] и ClinicalBERT [4], дополнительно обученные на текстах PubMed и клинической документации, подтвердили эффективность адаптации трансформеров к медицинской терминологии.

Альтернативой энкодерам служат большие языковые модели (LLM) на основе архитектуры декодера: они превосходят энкодеры на 2–8 % по F-мере на части биомедицинских датасетов [9], однако время их инференса выше в один–два порядка. Для эксплуатации в контуре организации предпочтительны модели-энкодеры; ансамблевые решения типа INSIGHTBUDDY-AI [8] применяются в облачных сервисах.

В качестве базовой выбрана модель RoBERTa [2], предварительно обученная на корпусах, включающих научную литературу. Выполнено дообучение (fine-tuning): выходной слой заменён классификатором токенов с разметкой по схеме BIO (Begin–Inside–Outside). Обучение проведено на датасетах BC5CDR [6] с разметкой химических соединений и заболеваний в 500 научных статьях, NCBI Disease [7], а также JNLPBA, BioRed и ChemProt; тексты разбивались на последовательности до 512 токенов с перекрытием, выборка разделена в пропорции 80/10/10. Гиперпараметры: скорость обучения 2e-5, размер батча 16, до 10 эпох с ранней остановкой. Качество оценивалось точностью (Precision), полнотой (Recall) и F1-мерой на уровне сущностей.

Результаты и их обсуждение

Базовая модель без дообучения показала умеренное качество: F1-мера на тестовой выборке составила 0,67; модель путала границы многословных терминов и ошибалась, когда название заболевания совпадало с общеупотребительным словом. После дообучения F1-мера достигла 0,85, то есть выросла на 18 процентных пунктов, при этом улучшились и точность, и полнота.

Полученный результат сопоставлен с показателями других подходов (рисунок 1). Подход CRF с ручным проектированием признаков достигает F1 = 0,72, что на 10–14 пунктов ниже моделей на основе предобученных трансформеров. Лучший из LLM общего назначения (Claude-2) извлекает биомедицинские сущности с F1 = 0,82 [9]; дообученная RoBERTa (0,85) приближается к ансамблевым методам (0,86) [8], оставаясь существенно дешевле в инференсе.

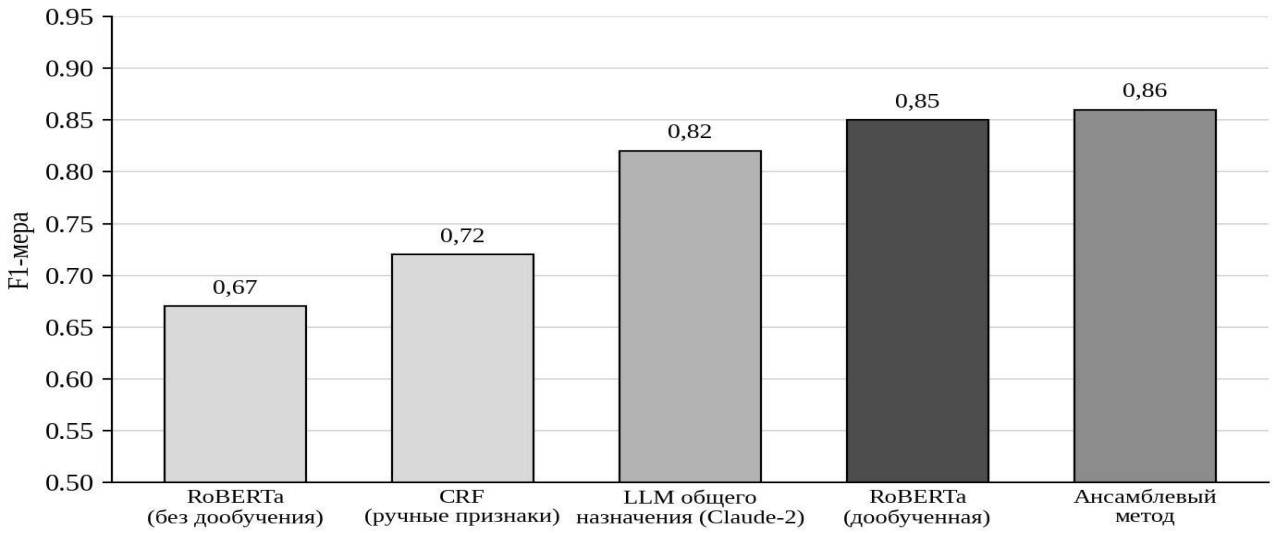

Рисунок 1. Сравнение подходов к извлечению медицинских сущностей по F1-мере

В few-shot-режиме динамический отбор примеров превзошёл статистический во всех 12 комбинациях (4 датасета × 3 режима); рост числа примеров с 5 до 50 повышает F1-меру на BioNLP13GE с 44,1 до 70,3 – это обеспечивает быструю адаптацию подсистемы к внутренним протоколам без полного переобучения. Типовые ошибки модели связаны с редко употребляемыми аббревиатурами (например, CAD – «ишемическая болезнь сердца» или «компьютерное проектирование»), границами сущностей у препаратов с разными дозировками и редкими заболеваниями, представленными в обучающей выборке менее чем десятью примерами.

Внедрение результатов

Модель использована для реинжиниринга процесса актуализации базы знаний больницы. В процессе AS IS ручное чтение публикаций занимает до трёх часов на документ,

каждая запись проходит сплошную двойную проверку, база ведётся в Excel без единой схемы аннотации, а согласование справок требует трёх подписей в четырёх подразделениях.

В процессе TO BE чтение заменено автоматической функцией NER-извлечения с порогом уверенности 0,9; эксперту подлежит только предразметка с уверенностью ниже порога, что сокращает ручные операции на 80 %. Введён один владелец процесса – врач-клинический фармаколог как единая точка контакта; справки доставляются автоматически через API медицинской информационной системы (МИС). Показатели процесса приведены в таблице 1.

Таблица 1

Показатели процесса до и после реинжиниринга

Показатель	**AS IS**	**TO BE**
Время обработки одного документа	до 3 часов	20 минут
Задержка актуализации базы знаний	2–3 недели	не более 1 рабочего дня
Количество шагов процесса	12	7
Количество передач ответственности	5	2
Количество контрольных проверок	3	1
Трудозатраты на обработку документов (120 документов в месяц)	до 360 чел.-ч	около 40 чел.-ч

Структурные показатели процесса AS IS и TO BE представлены в таблице 1 и на рисунке 2. Целевые KPI: F1-мера не менее 0,85, время обработки документа не более 20 минут, задержка актуализации базы знаний не более одного рабочего дня, доля авторазметки не менее 80 %; экономия трудозатрат – около 320 чел.-ч в месяц.

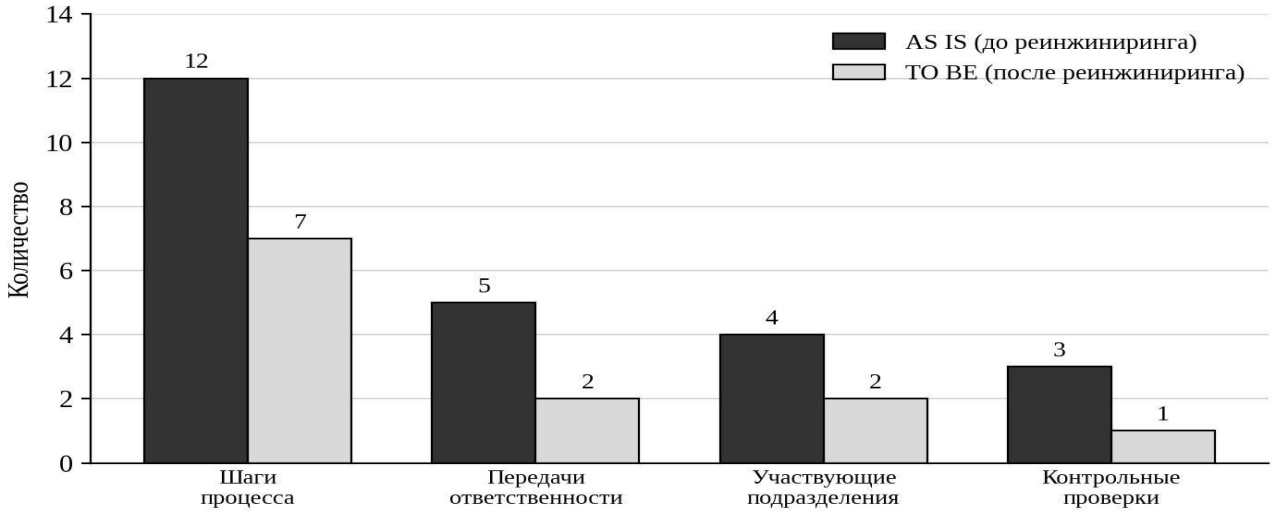

Рисунок 2. Структурные показатели процесса до и после реинжиниринга

Заключение

В работе построена и исследована система извлечения информации о лекарственных препаратах и заболеваниях из медицинских статей на основе дообученной модели RoBERTa. Дообучение повысило F1-меру с 0,67 до 0,85: модель приближается к ансамблевым решениям при меньших требованиях к ресурсам, чем у LLM, и заложена в основу реинжиниренного процесса актуализации базы знаний медицинской организации. Перспективы развития включают извлечение отношений между сущностями («препарат – показание – побочный эффект»), многоязычную обработку и интеграцию в МИС.

Список литературы:

1. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proc. of NAACL-HLT. – 2019. – P. 4171–4186.
2. Liu Y., Ott M., Goyal N. et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach // arXiv:1907.11692. – 2019.
3. Lee J., Yoon W., Kim S. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining // Bioinformatics. – 2020. – Vol. 36, № 4. – P. 1234–1240.
4. Alsentzer E., Murphy J., Boag W. et al. Publicly Available Clinical BERT Embeddings // Proc. of the 2nd Clinical NLP Workshop. – 2019. – P. 72–78.
5. Lample G., Ballesteros M., Subramanian S. et al. Neural Architectures for Named Entity Recognition // Proc. of NAACL-HLT. – 2016. – P. 260–270.
6. Li J., Sun Y., Johnson R.J. et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction // Database. – 2016. – Article baw068.
7. Dogan R.I., Leaman R., Lu Z. NCBI disease corpus: A resource for disease name recognition and concept normalization // Journal of Biomedical Informatics. – 2014. – Vol. 47. – P. 1–10.
8. Romero P., Han L., Nenadic G. INSIGHTBUDDY-AI: Medication Extraction and Entity Linking using Large Language Models and Ensemble Learning // arXiv:2409.19467. – 2024.
9. Obeidat M.S., Al Nahian M.S., Kavuluru R. Do LLMs Surpass Encoders for Biomedical NER? // Proc. of IEEE ICHI. – 2025.