Вагун Илья Владимирович
магистрант, обучающийся по направлению подготовки:
09.04.03 Прикладная информатика,
АНО ВО «Российский новый университет»
Vagun Ilya Vladimirovich,
ANO VO Russian New University

ИССЛЕДОВАТЕЛЬСКИЙ ФРЕЙМВОРК ДЛЯ ОБНАРУЖЕНИЯ НОВОЙ МЕДИЦИНСКОЙ ТЕРМИНОЛОГИИ В НАУЧНЫХ ПУБЛИКАЦИЯХ
A RESEARCH FRAMEWORK FOR DETECTING NOVEL MEDICAL TERMINOLOGY IN SCIENTIFIC PUBLICATIONS

Аннотация. В данной статье представлен исследовательский фреймворк для обнаружения новой медицинской терминологии в научных публикациях на основе обработки естественного языка. Его применение продемонстрировано на корпусе публикаций PubMed.

Abstract. This article presents a research framework for detecting new medical terminology in scientific publications using natural language processing techniques. The framework is evaluated on a corpus of PubMed publications.

Ключевые слова: Медицинская терминология, обработка естественного языка, извлечение именованных сущностей, PubMed, исследовательский фреймворк.

Keywords: Medical terminology, natural language processing, named entity extraction, PubMed, research framework.

Введение. Современная медицина стремительно развивается и это порождает огромный объём публикаций и постоянный процесс появления новых медицинских терминов [1]. С таким большим объемом данных традиционные лингвистические методы справить не могут, и необходима автоматизация этого процесса на основе современных IT-решений.

Целью данной работы является разработка и апробация исследовательского фреймворка для проведения серии экспериментов по обнаружению новой медицинской терминологии в научных публикациях и анализу ее динамики во времени.

Объект и предмет исследования. Объект исследования – медицинская научная литература, собранная по базам PubMed и PubMed Central [2]. Предмет исследования – процессы формирования и развития новой медицинской терминологии.

Материалы и методы. Процесс поиска новых терминов можно представить как последовательный запуск отдельных этапов:

1. Импорт статей из базы данных PubMed или PubMed Central.
2. Извлечения именованных сущностей.
3. Поиск в словаре для фильтрация уже существующих терминов.
4. Обработка и вывод результатов.

Все эти этапы составляют «рабочий процесс» обработки текстов («NLP workflow»), который запрограммирован на Python и является основной частью данного исследовательского фреймворка [3]. Последовательность запуска и параметры этапов задаются в управляющем файле. Статистическая обработка и вывод результатов происходят автоматически.

Ключевые концепты фреймворка. Большинство представленных в открытом доступе аналогичных решений имеет жесткую структуру. Это ограничивает возможности изменения параметров эксперимента без изменения исходного кода. А для реальной задачи поиска новой терминологии необходимо провести множество таких экспериментов. Предлагаемый фреймворк свободен от таких ограничений и позволяет строить собственные эксперименты.

Модульная архитектура фреймворка дает возможность его расширять за счет написания новых модулей. Список существующих модулей по типам:

1. «cleaner» – очистка базы данных от результатов прошлого запуска.
2. «fetcher» – импорт статей, 2 модуля: PubMed и PubMed Central.
3. «dictionary» – поиск терминов в словарях: MeSH, SNOMED CT, CUI, DrugBank, GO, HPO, ICD10, NCI, WHO. Если термин найден, то сохраняется его код из словаря.
4. «output» – расчет и вывод результатов. Реализованы 2 модуля: вывод данных в Excel, и вывод в виде диаграмм.

Гибкая настройка эксперимента через YAML-файлы. Возможно прописать все необходимые параметры (например, строку поиска по PubMed) в этом файле без необходимости это делать в коде. Все параметры задокументированы в этом же файле в виде комментариев. Для отладки возможно отключать некоторые этапы (например, импорт статей) простым комментированием соответствующего блока YAML.

Разработка через тестирование. Код фреймворка написан по методологии test-driven development (TDD). С одной стороны это дало уверенность в корректности его работы, а с другой стороны рекомендуется придерживаться этой методологии при написании собственных модулей.

Автоматический вывод результатов в нескольких форматах. Этот принцип позволяет автоматизировать проведение экспериментов и концентрироваться именно на исследовательской работе, доверив расчеты и построение диаграмм фреймворку.

Возможность расшаривания для обеспечения воспроизводимости эксперимента. Отдельного механизма обмена конфигурациями между пользователями нет. Для простоты YAML-файлы и результаты работы передаются между исследователями по email.

Результаты и обсуждение. В рамках данной работы фреймворк был апробирован на корпусе публикаций по тематике рака и кальцификации молочной железы. Эксперимент запускался с параметром поиска « (((calcification [Abstract]) AND cancer) AND breast) AND 2005:2025 [DP]». По полученным результатам построен график количества терминов по годам и их покрытие словарями (рис. 1).

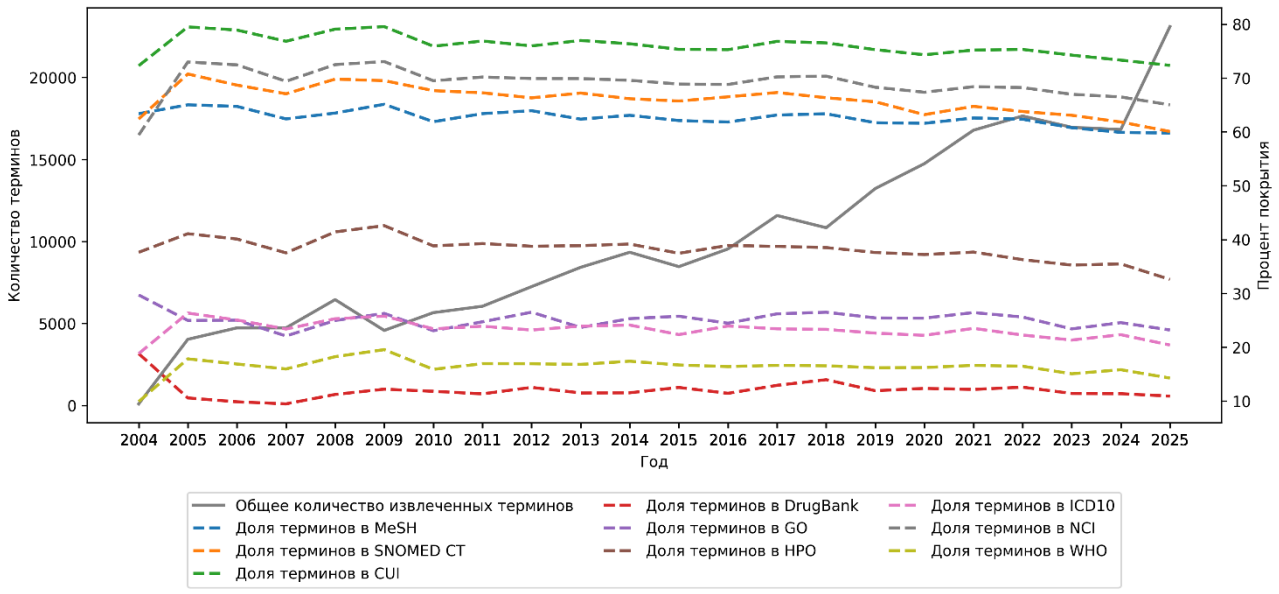

Рисунок 1. Количество извлеченных терминов из PubMed и их покрытие словарями

Из графика видно, что идет постоянное увеличение количества терминов год от года, что объясняется ростом количества публикаций. А покрытие извлеченных терминов словарями в целом имеет небольшой тренд на снижение. Это свидетельствует о появлении новой терминологии, которая еще не включена в словари.

На рис. 2 представлена динамика распределения POS-структур по годам.

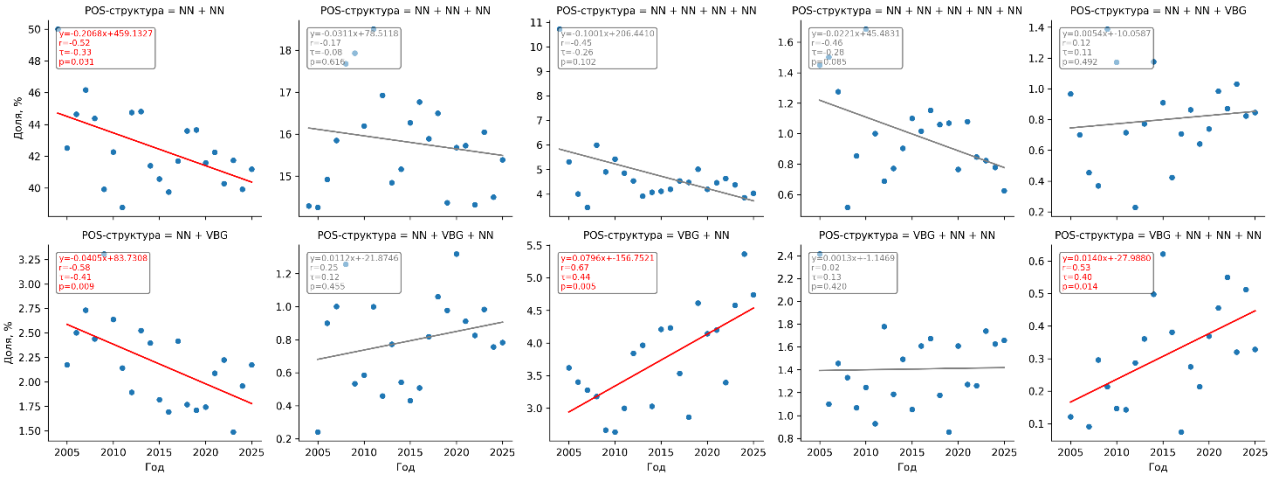

Рисунок 2. Динамика распределения POS-структур по годам, кроме униграмм

Примечательно, что POS-структуры «NN + NN» и «NN + VBG» имеют тренд на снижение, а «VBG + NN» и «VBG + NN + NN + NN» – на увеличение.

Полученные результаты подтверждают возможность данного фреймворка быть использованным для анализа процессов формирования новой медицинской терминологии.

Список литературы:

1. Золотарев О.В. и др. Извлечение терминов из биомедицинских публикаций – подход на основе N-грамм. Москва: Российский новый университет, 2023. Т. 2. С. 136–160.

2. PubMed – URL: https://pubmed.ncbi.nlm.nih.gov/ (дата обращения: 28.10.2025)

3. dbfun/NovelMedTerms – URL: https://github.com/dbfun/NovelMedTerms (дата обращения: 28.10.2025)