ПРЕДСТАВЛЕНИЕ РЕЗУЛЬТАТОВ НАУЧНЫХ ИССЛЕДОВАНИЙ В ОБЛАСТИ БИОМЕДИЦИНЫ В МАШИНОЧИТАЕМЫХ ФОРМАТАХ

Пашук А. В., Гуринович А. Б., Кузнецов А. П.
Кафедра систем управления, кафедра вычислительных методов и программирования, кафедра систем управления, Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: {pashuk, gurinovich, kuznap}@bsuir.by

В последнее время все более остро встает проблема воспроизводимости результатов научных исследований в различных областях науки. Все чаще можно услышать, что научные исследования, в частности в области биомедицины, которые публиковались в известных научных изданиях, оказываются невоспроизводимыми или изначально подстроенными, основанными на ложных данных. Это приводит к замедлению развития медицины, появлению лекарств, не оказывающих никакого эффекта или ухудшающих состояние больных, а также к огромным тратам средств, выделяемых ученым в виде грантов и в качестве финансирования дальнейших исследований.

Введение

Под воспроизводимостью результатов понимается возможность любого заинтересованного лица проверить положения, предложенные в научной статье (которые включают в себя графическую, числовую или текстовую информацию), восстановив их из исходных данных. Данный термин подразумевает, что авторы статьи должны предоставить доступ к исходным данным исследования (насколько это возможно, ввиду ограничений законодательства или организации, финансирующей исследования), а также к программному коду, с помощью которого были получены результаты.

I. Проблема воспроизводимости результатов исследований

Полностью решить проблему воспроизводимости научных результатов можно только комплексным подходом к каждому этапу исследования.

- Формат исходных данных и данных, полученных в ходе проведения исследования должен быть общепринятым, не требующим дополнительных действий для получения доступа к данным. В качестве такого формата обычно используют CSV, который имеет хорошую степень сжатия. В случае большого объема информации, ее обычно разбивают на несколько CSV файлов;
- Программный код, позволяющий полностью повторить ход исследования и восстановить результаты, описанные в статье. Для этих целей все чаще используются Jupyter Notebooks [1], позволяющие объединить код, написанный на Python или R, в один файл, состоящий из набора ячеек с фрагментами кода.
- Формат научных статей. Важное значение имеет формат публикации. Обычно научные публикации предоставляется в одном из трех форматов: LateX, DOCX или PDF, однако ни один из этих форматов не позволяет добавить дополнительную информацию о содержимом статьи. Стоит отметить, что информация в таких форматах сложна для автоматизированной обработки.

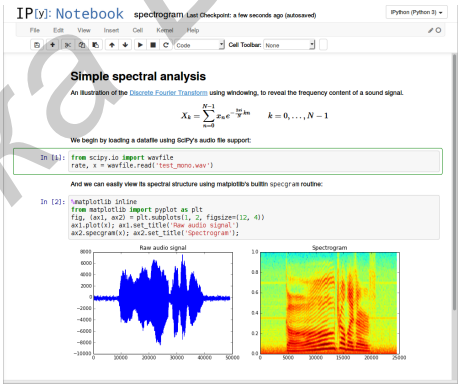

Рис. 1 – Интерфейс Jupyter Notebooks

II. Формат научных статей

Кроме описанных выше существует несколько открытых форматов, позволяющих представить информацию, содержащуюся в научной статье в структурированном виде. Достоинством таких форматов является возможность дальнейшей автоматизированной обработки информации. В ходе исследования было разработано веб-приложение, позволяющее загрузить текст научной статьи в текстовом формате, добавить дополнительную информацию о терминах и их взаимосвязях в тексте статьи и затем экспортировать в один из следующих форматов:

- JATS - Journal Article Tag Suite [2] - XML-based формат, разработанный сообществом ученых как альтернатива неструктурированным форматам. Особенностью данного формата является простота извлечения информации при автоматизированной обработке, а также наличие документации, поз-

воляющей добавлять собственные теги при необходимости. Так, в рамках исследования, JATS был расширен набором тегов, необходимых для разметки терминов в научных статьях.

– HTML with microdata [3] - Расширение HTML разметки, позволяющее добавить дополнительную информацию о содержимом веб-страницы, например, информацию об авторах, дате публикации и др.

Интерфейс приложения [3] приведен на рисунке 2.

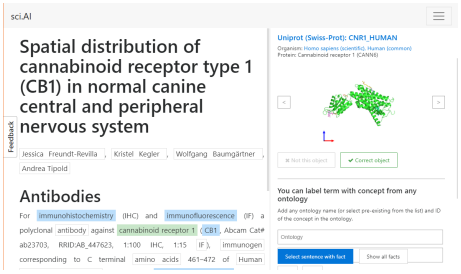

Рис. 2 – Интерфейс приложения для семантизации статей

Разработанное приложение позволяет добавить дополнительную информацию о ключевых терминах научной статьи и их взаимосвязях. Добавление такой информации несет несколько дополнительных функций:

– позволяет извлечь больше полезной информации при обработке машиной;
– позволяет улучшить качество поиска по базе научных статей, т.к. имея информацию о ключевых терминах можно лучше отфильтровать статьи по релевантности;
– позволяет рецензентам и другим ученым быстрее и проще понять предмет, о котором идет речь в научной статье.

После загрузки текста статьи в приложение происходит анализ текста и поиск в нем терминов из нескольких наиболее популярных биомедицинских онтологий (Uniprot, MeSH, ChEBI, ICD-10 и др.). Все найденные термины помечаются и требует валидации автором (рисунок 3).

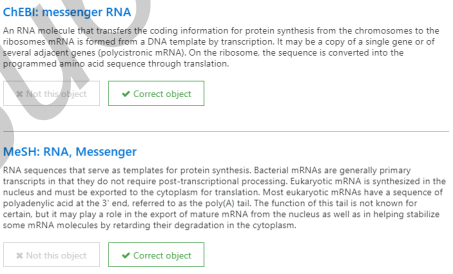

Рис. 3 – Пример термина, найденного в статье

Аналогично происходит анализ и выделение ключевых фактов в тексте статьи (рисунок 4).

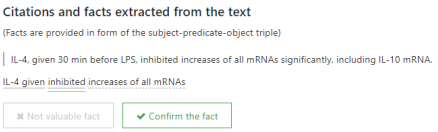

Рис. 4 – Пример термина, найденного в статье

После валидации всех найденных терминов и фактов, автор может экспортировать статью в один из форматов, описанных выше. Статья в структурированном формате при необходимости может быть преобразована в стандартный формат (PDF, DOCX и др.). Стоит отметить, что, имея информацию о ключевых терминах и фактах исследования, можно получить представление статьи в виде нанопубликации [4].

В то время как разработанное веб-приложение используется для обработки новых статей, параллельно идет автоматическая обработка текстов уже опубликованных биомедицинских статей, находящихся в открытом доступе. Качество автоматической семантизации данных статей значительно ниже, чем в автоматизированном режиме с валидацией авторами, однако позволяет увеличить объем обработанных статей и, соответственно, получить больше информации, что на данном этапе имеет большое значение.

Заключение

Преобразование статей в машиночитаемые форматы является первым шагом к улучшению воспроизводимости результатов научных исследований в области биомедицины. Существующие XML- и RDF-based форматы позволяют увеличить объем полезной информации, содержащейся в файле статьи, включить туда данные, упрощающие процесс чтения статьи, позволяющие акцентировать внимание на ключевых моментах (например, формат нанопубликаций, содержащий только ключевые факты научной статьи). Следующим шагом исследование станет построение так называемого графа знаний (knowledge graph), основанного на полученных структурированных статьях.

III. Список литературы

1. Project Jupyter [Электронный ресурс] / Режим доступа: http://jupyter.org/. – Дата доступа: 25.08.2017.
2. Journal Article Tag Suite [Электронный ресурс] / Режим доступа: https://jats.nlm.nih.gov/index.html. – Дата доступа: 26.08.2017.
3. sci.AI Papers [Электронный ресурс] / Режим доступа: http://app.sci.ai/. – Дата доступа: 26.08.2017.
4. Nanopublications [Электронный ресурс] / Режим доступа: http://nanopub.org/wordpress/. – Дата доступа: 25.08.2017.