Компьютерный анализ текстов

Методы извлечения биомедицинской информации из патентов и научных публикаций (на примере химических соединений)

Н.А. Колпаков[I], А.И. Молодченков[II,III], А.В. Лукин[III]

[I] Московский физико-технический институт, г. Москва, Россия
[II] Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва, Россия
[III] Российский университет дружбы народов, г. Москва, Россия

Аннотация. В данной статье предложен алгоритм для решения задачи извлечения информации из биомедицинских патентов и научных публикаций. Предложенный алгоритм основан на методах машинного обучения. Были проведены эксперименты на патентах из базы USPTO. Эксперименты показали, что лучшее качество извлечения показала модель, построенная на основе BioBERT.
Ключевые слова: *машинное обучение, обработка естественного языка, извлечение именованных сущностей, обработка биомедицинских текстов.*
DOI: 10.14357/20790279230118

Введение

С каждым годом число биомедицинских патентов и научных публикаций значительно увеличивается. Зачастую эти тексты не содержат какие-то описательные метаданные, а это, в свою очередь, приводит к большому объёму неструктурированных данных. Следовательно, увеличивается потребность в инструментах, которые бы могли точно извлекать требуемую информацию из таких текстов.

Для извлечения информации из текстов для дальнейшей её обработки можно использовать как подходы машинного обучения, так и алгоритмы, основанные на регулярных выражениях. В работах [1, 2] ключевую роль играют регулярные выражения, и, напротив, в [3, 4] используются достижения области глубокого машинного обучения, в частности модель условных случайных полей. А в [5] используется нейросетевая модель-трансформер, которая при правильной настройке параметров может достаточно неплохо извлекать биомедицинских данные.

Хотя, были созданы инструменты для анализа и взаимодействия с неструктурированными данными, зачастую эти решения основаны на правилах, которые применимы к конкретным обрабатываемым данным. В этой работе мы предлагаем решение для задачи извлечения биомедицинской информации из патентов с помощью регулярных выражений. Таким образом, полученная структурированная информация может быть использована для обучения сложных нейросетевых моделей, которые позволят корректно извлекать информацию из большего числа текстов.

1. Обзор релевантных работ

Существует не так много решений, которые решают поставленную задачу. Зачастую, существующие алгоритмы разработаны для решения

большого спектра задач, поэтому они показывают недостаточно высокие результаты для задачи извлечения определений из биомедицинских патентов и научных публикаций.

Например, Jinhyuk Lee и его коллеги представили BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) [5] – нейросетевую модель-трансформер [6], разработанную для автоматической обработки языка биомедицинской области, которая предварительно обучена на больших биомедицинских текстах. Данная модель способна извлекать биомедицинские именованные сущности, биомедицинские отношения в тексте, а также может выдавать ответы на биомедицинские вопросы. BioBERT инициализируется со значениями весовых функций, которые были получены для BERT [7] (данная модель предварительно обучена на текстах из английской Википедии и BooksCorpus), после чего BioBERT был дообучен на биомедицинских текстах (сюда входят аннотации с PubMed и полнотекстовые статьи PMC).

В статье [3] представлен другой подход для решения задач NLP из области биомедицины. CLAMP (Clinical Language Annotation, Modeling, and Processing) для извлечения информации использует как методы, основанные на машинном обучении, так и методы, основанные на правилах. Данный инструментарий позволяет извлекать именованные сущности, разбивать текст на токены, и многое другое. В своей программе авторы используют 3 типа токенизаторов (на выбор):

1) OpenNLP токенизатор [8] на основе машинного обучения,
2) токенизатор на основе разделения слов по заданным символам,
3) токенизатор на основе правил с различными параметрами конфигурации.

А для задачи извлечения именованных сущностей авторы предлагают использовать:

1) алгоритм условных случайных полей (conditional random fields – CRF) [9],
2) алгоритм на основе словаря с большим количеством биомедицинской лексики, собранной из разных ресурсов, таких как UMLS,
3) алгоритм на основе регулярных выражений для объектов с общими шаблонами.

OSCAR4 (Open-Source Chemistry Analysis Routines) [2] – это открытая система для автоматического извлечения химических терминов из научных статей. В основе данной работы лежит распознавание химических веществ на основе регулярных выражений и распознавание на основе словаря заранее заданных слов. Но для распознавания сложных химических соединений (которые состоят из нескольких токенов) – используется модель Маркова максимальной энтропии.

Ещё, есть работа [1], где авторы используют морфологию для извлечения биомедицинских слов. Система распознавания химических объектов состоит из двух подсистем. Первая извлекает химические объекты и помечает их в нормализованном входном документе с использованием словаря заранее заданных слов и морфологического подхода. Основанный на морфологии подход идентифицирует различные элементы в химическом соединении и объединяет их для создания конечного соединения.

Вторая подсистема – извлекает дополнительные химические элементы и распределяет все распознанные объекты по классам соединений, а также имеет такие возможности как расшифровка аббревиатур и исправление орфографических ошибок. Для того, чтобы определить является ли определённая сущность “химической”, авторы собрали статистическую информацию для каждого уникального объекта. Данная информацию используется как последний этап извлечения именованных сущностей и предназначен для классификации извлеченного объекта (либо объект является химическим, либо нет).

Приведённые методы извлекают информацию из биомедицинских текстов в целом – они не направлены на извлечение структур Маркуша [10] (см. Рис. 1).

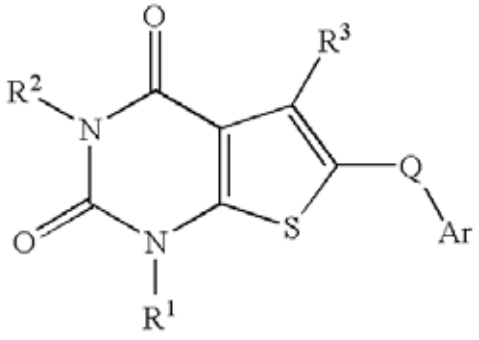
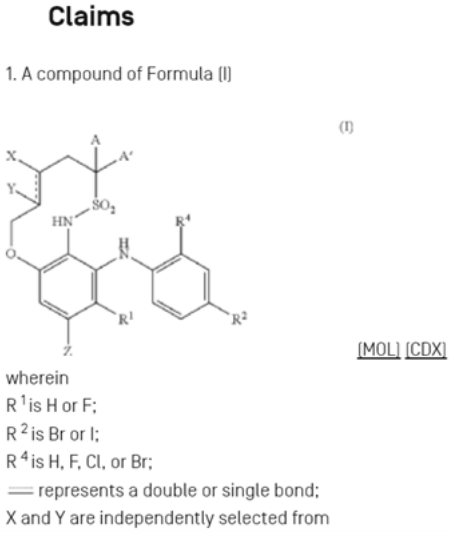


Рис. 1. Пример структуры Маркуша, взят из US Patent 20040171623

2. Постановка задачи

Данные, касающиеся различных биомедицинских патентов, находятся в открытом доступе в различных патентных ведомствах. Патенты обычно имеют четкую структуру, которая включает в себя: название патента, аннотацию, описание, формулы изобретения (Claims) и библиографическую информацию (дата, номер патента, авторы).

Интересующий нас раздел – Claims (см. Рис. 2), содержит описание химических соединений, которые заявлены авторами патента. На это как раз направлена правовая охрана, предоставляемая патентом. Раздел Claims может содержать внутри себя несколько подразделов, которые содержат информацию по разным химическим цепочкам.


Claims

1. A compound of Formula (I)

(I)

[MOL] [CDX]

wherein

R^1 is H or F;

R^2 is Br or I;

R^4 is H, F, Cl, or Br;

═ represents a double or single bond;

X and Y are independently selected from

Рис. 2. Пример данных, содержащихся в разделе Claims, взят из US Patent 20120208859

Соединения, представленные в разделе Claims, могут быть описаны с помощью структуры Маркуша [10] (см. Рис. 1). Для того, чтобы найти патенты, у которых структура Маркуша либо такая же, либо схожая, – нужно сравнить эти структуры. Так как структура Маркуша – сетевая модель, то сравнение напрямую таких моделей – очень ресурсоёмкий процесс. Поэтому, зачастую используют так называемые fingerprints, которые отражают в себе информацию, представленную в структурах Маркуша. Но перед этим нужно извлечь информацию, которая входит в такие структуры, на что и направлена данная работа.

Табл. 1.

Примеры химических соединений

Название соединения	Молекулярная формула
nitrogen monoxide	NO
glucose	$C_6H_{12}O_6$
copper (II) sulfate	$CuSO_4$
carbon dioxide	CO_2
dichlorine heptoxide	Cl_2O_7

Задача состоит в извлечении из раздела Claims химических соединений (Табл. 1), названий переменных (вместо которых могут быть подставлены различные значения), химических элементов, формул и InChI кодов [11] (Рис. 3) с целью преобразования данной текстовой информации в некоторую структуру формального представления.

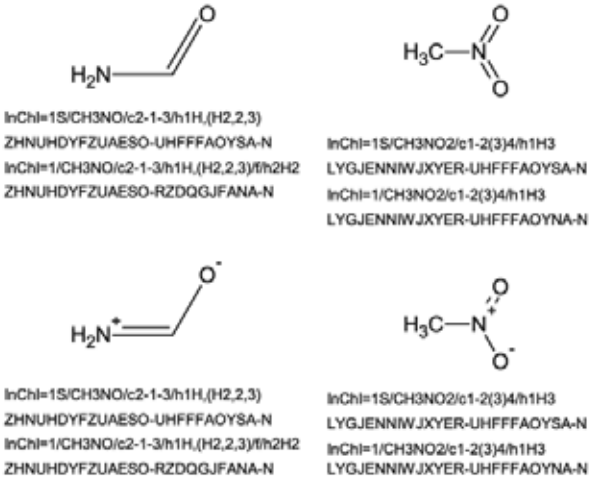


Рис. 3. Примеры InChI кодов [11]

В теоретико-множественной аннотации задачу можно сформулировать следующим образом: имеются патенты и научные публикации X, где каждый элемент $x \in X$ представлен в виде $x = x_1, \ldots, x_n$ ($x_1, \ldots, x_n$ – последовательность слов (токенов)), и задано множество классов $Y = (y_1, \ldots, y_5)$, где:

y_1 – номер Claim,

y_2 – переменная, к которой ищем описание,

y_3 – описание переменной,

y_4 – ссылка на другой Claim,

y_5 – в случаях, если токен не соответствует $y_1, \ldots, y_4$.

Необходимо построить отображение F, которое бы сопоставляло каждому элементу $x \in X$ – соответствующий элемент $y \in Y$.

Задача извлечения информации из текста представляет собой поиск и классификацию именованных сущностей (Named Entity Recognition), имеющихся в неструктурированном тексте, по заранее заданным категориям. Именованная сущность – n-грамма в тексте, для которой определена категория (класс, метка).

3. Описание метода

Алгоритм извлечения информации из текстов можно разбить на следующие шаги:

1) составление набора данных,
2) предварительная обработка входных данных,
3) векторизация данных и извлечение признаков,
4) обучение моделей для извлечения необходимой информации из текстов.

Составление набора данных включает также включает в себя автоматизированную разметку то-

кенов. В предварительную обработку данных входит нормализация и токенизация входных данных. Схема алгоритма приведена на Рис. 4.

Рис. 4. Схема предложенного алгоритма извлечения информации из текстов

3.1. Составление набора данных

Данные, с которыми мы работали, взяты из базы USPTO [12]. Все данные изначально представлены в XML файлах, которые содержат структурированную информацию о патентах: описание, аннотацию, библиографические данные и Claims. Для разработки алгоритма из файлов берется только раздел Claims.

3.2. Предварительная обработка данных

Первым этапом обработки данных является извлечение из имеющихся данных – раздела Claims. Так как такие данные имеют схожее оформление, то извлечение выполняются с помощью регулярных выражений.

После извлечения Claims – необходимо подготовить данные для дальнейшей работы. Для этого выполняется следующая нормализация строк:

1) Удаляются лишние пробелы в начале и конце строк.
2) Пустые строки тоже удаляются.
3) Каждые строки разбиваются таким образом, чтобы в них содержалось только одно описание переменных. Это выполняется с помощью поиска в каждой строке следующей конструкции: *… variables … definition_verb … definitions … definition_end_symbol*. При этом, учитывается ситуация, когда на этой же строке может быть приведено описание вложенных переменных. Например, “Z is OR3, wherein R3 is C1-C6 alkyl”. В этом случае строка не разбивается.
4) Если в строке не встретился *definition_end_symbol*, то строки объединяются до тех пор, пока не найдется нужный символ.
5) Если строка является начальной для Claim, но номера Claim указаны через тире, то последующее содержимое копируется для каждого номера Claim из указанного промежутка.

Затем полученные строки группируются по Claim. Все описанные выше действия по нормализации строк, тоже выполняются с помощью регулярных выражений.

Применение нормализации позволит обучить модель на небольшой выборке более качественно, а также повысит её точность. А группировка и разбиение строк упростят последующую разметку данных.

Следующим этапом является присвоение каждому токену метки из возможных:

- CLAIM – номер Claim,
- VAR – переменная, к которой ищем описание; описание к этой переменной подставляется только к последнему месту, где она была упомянута до встречи этой самой переменной,
- VAR-ALL – переменная, к которой ищем описание; описание к этой переменной подставляется во все места, где она была упомянута,
- DEF – описание переменной,
- REF – ссылка на другой Claim,
- O – в случае, если токену не присвоена ни одна из вышеперечисленных меток.

Присвоение токенам соответствующей метки выполняется с помощью средств разметки ФИЦ ИУ РАН. Результатом работы этих средств являются данные, содержащие токен, его метку, номер строки, где он был найден и уникальный номер Claim.

3.3. Векторизация данных и извлечение признаков

Так как не все модели классификации в качестве входных данных принимают строковые значения данных, то необходимо векторизовать такие признаки. К ним относятся токены и соответствующие метки.

Если каждой уникальной метке сопоставляется число, то с токенами дело обстоит совсем иначе. Для каждого токена строится вектор размерности 100 с помощью модели Word2Vec [13, 14] для получения векторных представлений слов естественного языка.

Word2Vec была обучена на собранном наборе данных. Обучение происходило 10 эпох, с размером скользящего окна равным 8.

Чтобы в дальнейшем обучить модель машинного обучения – необходимо объединить токены в списки на основе принадлежности к Claims, а затем подать эти списки на вход Word2Vec. Результатом работы такой модели будет сопоставление каждому токену его векторного представления.

Некоторые алгоритмы машинного обучения, например основанные на Conditional Random Fields, будут лучше работать с признаками, содержащими информацию о соседних токенах, относительно рассматриваемого.

Поэтому, ещё одним способом представления данных, подаваемых на вход таким моделям, – является сопоставление каждому токену набора признаков. Этими признаками являются:

- сами токены,

- последние 2–3 символа токена,
- флаг, начинается ли токен с заглавной буквы,
- флаг, является ли токен числом,
- флаг, содержит ли токен только заглавные буквы,
- информация о соседних токенах (соседний токен и 3 флага, как в предыдущих пунктах).

3.4. Обучение моделей

В качестве методов классификации, которые бы на основе размеченных и векторизованных данных присваивали бы ранее неизвестным данным метки, – использовались как стандартные методы машинного обучения (Support Vector Machine [15], Conditional Random Fields [9]), так и модели глубокого обучения (Stanford NER [16], BERT [7], BioBERT [5]), которые уже заранее предобучены.

Дообучение BioBERT, BERT и Stanford NER производилось на данных, полученных в пункте 3.2 – токенах, метках и номерах Claims. Для BioBERT и BERT, чтобы не происходило переобучение, число эпох было выбрано равным 5 (см Рис. 5 и Рис. 6).

Метод опорных векторов (SVM) обучался с нуля на векторизованных данных, а метод условных случайных полей (CRF) – на данных, полученных в пункте 3.3.

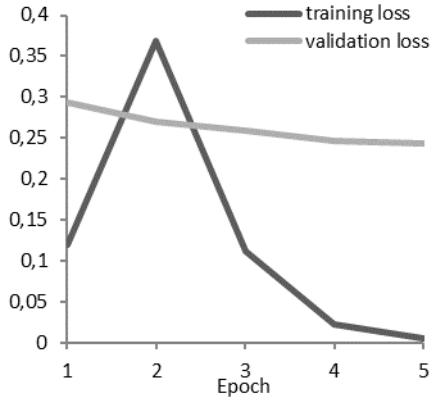


Рис. 5. График *Epoch vs Loss* для модели BioBERT.

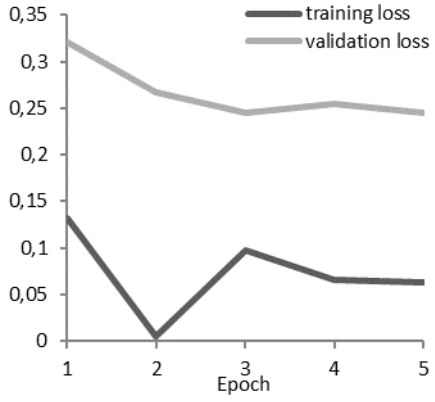


Рис. 6. График *Epoch vs Loss* для модели BERT.

4. Результаты экспериментов

В рамках данной работы была проведена серия экспериментов для решения поставленной задачи классификации. Эксперименты проводились на 100 документах с более чем 1700 Claims. Обучающая выборка состояла из 70 документов, а валидационная – из 30.

Для сравнения результатов использовались стандартные метрики качества: *precision (точность)*, *recall* (полнота) and *F1-score* [17]. Давайте, рассмотрим их более детально.

Для начала рассмотрим, что такое TP, FP и FN:

- TP – число токенов, которым классификатор присвоил правильную меток,
- FP – число токенов, которые имеют метку O, но классификатор присвоил им другую метку,
- FN – число токенов, которые имеют определённую метку (не O), но классификатор отнёс их к другой группе.

Accuracy (точность) – доля токенов, действительно принадлежащих конкретному классу, относительно всех токенов, которым классификатор присвоил такую метку класса. Данная метрика вычисляется по уравнению (1).

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

Recall (полнота) – доля токенов, которым классификатор присвоил конкретную метку класса, относительно всех токенов, имеющих эту метку. Данная метрика вычисляется по уравнению (2).

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

F1-score – среднее гармоническое значение точности и полноты. Данная метрика вычисляется по уравнению (3).

$$F1\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \quad (3)$$

Результаты классификации на тестовых данных приведены в Табл. 2.

Табл. 2

Значения метрик для различных методов классификации.

Название модели	Precision	Recall	F1-score
SVM	0.5276	0.6340	0.5675
CRF	0.6701	0.6358	0.6378
Stanford NER	0.7530	0.8488	0.7981
BERT	0.8437	0.8978	0.8699
BioBERT	**0.8467**	**0.9012**	**0.8731**

Из проведённых экспериментов видно, что классические методы машинного обучения пока-

зывают результаты намного хуже, чем предварительно обученные модели глубокого обучения, которые, в свою очередь, классифицируют токены на достаточно хорошем уровне.

Заключение

В статье описан метод решения задачи извлечения информации из биомедицинских текстов для дальнейшей ее обработки. Этот метод позволяет извлекать описание химических соединений, которые заявлены авторами патентов. Были обучены модели машинного обучения, такие как SVM, CRF, Stanford NER, BERT и BioBERT, на которых впоследствии проводились эксперименты.

В дальнейшем планируется преобразовать полученные данные в формат InChI кодов и написать fingerprints которые соответствуют структурам Маркуша, заявленным авторами патентов. Также планируется провести ещё серию экспериментов для улучшения качества извлечения информации из текстов.

Литература

1. *Akhondi, S., Rey, H., Schwörer, M., Maier, M., Toomey, J., Nau, H., Ilchmann, G., Sheehan, M., Irmer, M., Bobach, C., Doornenbal, M., Gregory and M., Kors, J.* Automatic identification of relevant chemical compounds from patents. Database: the journal of biological databases and curation. 2019. Vol. 1. P. 1–14.
2. *Jessop, D., Adams, S., Willighagen, E., Hawizy, L. and Murray-Rust, P.* OSCAR4: A flexible architecture for chemical textmining. Journal of cheminformatics. 2011. Vol. 3. No. 1. P. 1–12.
3. *Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H. and Qi, W.* CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. Journal of the American Medical Informatics Association: JAMIA. 2018. Vol. 25. No. 3. P. 331–336.
4. *Swain, M. and Cole, J.* 2016. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. Journal of Chemical Information and Modeling. 2016. Vol. 56. No. 10. P. 1894–1904.
5. *Jinhyuk, L., Wonjin, Y., Sungdong, K., Donghyeon, K., Sunkyu, K., Chan, H. S. and Jaewoo, K.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2019. Vol. 36. No. 4. P. 1234–1240.
6. *Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. and Polosukhin, I.* Attention Is All You Need. Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 5998–6008.
7. *Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.* Bert: pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. Vol. 1. P. 4171–4186.
8. The OpenNLP Project. Available at: http://opennlp.apache.org (дата обращения 20.02.2022).
9. CRFsuite: a Fast Implementation of Conditional Random Fields (CRFs). Available at: http://www.chokkan.org/software/crfsuite/ (дата обращения 20.02.2022).
10. *Barnard, J.* A comparison of different approaches to Markush structure handling. Journal of Chemical Information and Computer Sciences. 1991. Vol. 31. No. 1. P. 64–68.
11. *Heller, S., McNaught, A., Pletnev, I., Stein, S. and Tchekhovskoi, D.* The IUPAC International Chemical Identifier. Journal of Cheminformatics. 2015. Vol. 7. P. 1–34.
12. USPTO. Available at: https://www.uspto.gov/patents (дата обращения 20.02.2022).
13. *Mikolov, T., Chen, K., Corrado, G. and Dean, J.* Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR. 2013. P. 1–12.
14. *Mikolov, T., Yih, W.-T. and Zweig, G.* Linguistic regularities in continuous space word representations. Proceedings of NAACL-HLT. 2013. P. 746–751.
15. *Cortes, C. and Vapnik, V.* Support-vector networks. Machine Learning. 1995. Vol. 20. No. 3. P. 273–297.
16. *Finkel, J., Grenager, T. and Manning, C.* Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005). 2005. P. 363–370.
17. *Mitchell, T.* Machine Learning. Нью-Йорк: McGraw-Hill, 1997. 432 с.

Колпаков Николай Алексеевич. Федеральное государственное автономное образовательное учреждение высшего образования "Московский физико-технический институт (национальный исследовательский университет)" (МФТИ, Физтех), г. Москва, Россия. Бакалавр. Количество печатных работ: 1. Область научных интересов: машинное обучение, глубокое обучение, извлечение именованных сущностей, обработка естественного языка. E-mail: kolpakov.na@phystech.edu

Молодченков Алексей Игоревич. Федеральное государственное учреждение "Федеральный исследовательский центр "Информатика и управление" Российской академии наук", г. Москва, Россия. Кандидат технических наук. Количество печатных работ: 96. Область научных интересов: искусственный интеллект, базы знаний, извлечение информации, медицина. E-mail: aim@tesyan.ru (Ответственный за переписку)

Лукин Антон. Федеральное государственное автономное образовательное учреждение высшего образования "Российский университет дружбы народов" (РУДН), г. Москва, Россия. Учёная степень. Количество печатных работ: 10. Область научных интересов: искусственный интеллект, анализ текстов. E-mail: antonvlukin@gmail.com

Methods of extracting biomedical information from patents and scientific publications (on the example of chemical compounds)

N.A. Kolpakov[I], A.I. Molodchenkov[II,III], A.V. Lukin[III]

[I] Moscow Institute of Physics and Technology, Moscow, Russia

[II] Federal research center "Computer science and control" of Russian Academy of Sciences, Moscow, Russia

[III] Peoples' Friendship University of Russia, Moscow, Russia

Abstract. This article proposes an algorithm for solving the problem of extracting information from biomedical patents and scientific publications. The introduced algorithm is based on machine learning methods. Experiments were carried out on patents from the USPTO database. Experiments have shown that the best extraction quality was achieved by a model based on BioBERT.

Keywords: *machine learning, natural language processing, named entity recognition, biomedical texts processing*.

DOI: 10.14357/20790279230118

References

1. *Akhondi, S., Rey, H., Schwörer, M., Maier, M., Toomey, J., Nau, H., Ilchmann, G., Sheehan, M., Irmer, M., Bobach, C., Doornenbal, M., Gregory and M., Kors, J.* 2019. Automatic identification of relevant chemical compounds from patents. Database: the journal of biological databases and curation, vol. 1, pp. 1–14.
2. *Jessop, D., Adams, S., Willighagen, E., Hawizy, L. and Murray-Rust, P.* 2011. OSCAR4: A flexible architecture for chemical textmining. Journal of cheminformatics, vol. 3, no. 1, pp. 1–12.
3. *Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H. and Qi, W.* 2018. CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. Journal of the American Medical Informatics Association: JAMIA, vol. 25, no. 3, pp. 331–336.
4. *Swain, M. and Cole, J.* 2016. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. Journal of Chemical Information and Modeling, vol. 56, no. 10, pp. 1894–1904.
5. *Jinhyuk, L., Wonjin, Y., Sungdong, K., Donghyeon, K., Sunkyu, K., Chan, H. S. and Jaewoo, K.* 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, vol. 36, no. 4, pp. 1234–1240.
6. *Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. and Polosukhin, I.* 2017. Attention Is All You Need. Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008.
7. *Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.* 2019. Bert: pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171–4186.
8. The OpenNLP Project. Available at: http://opennlp.apache.org (accessed February 20, 2022).

9. CRFsuite: a Fast Implementation of Conditional Random Fields (CRFs). Available at: http://www.chokkan.org/software/crfsuite/ (accessed February 20, 2022).
10. *Barnard, J.* 1991. A comparison of different approaches to Markush structure handling. Journal of Chemical Information and Computer Sciences, vol. 31, no. 1, pp. 64–68.
11. *Heller, S., McNaught, A., Pletnev, I., Stein, S. and Tchekhovskoi, D.* 2015. The IUPAC International Chemical Identifier. Journal of Cheminformatics, vol. 7, pp. 1–34.
12. USPTO. Available at: https://www.uspto.gov/patents (accessed February 20, 2022).
13. *Mikolov, T., Chen, K., Corrado, G. and Dean, J.* 2013. Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR, pp. 1–12.
14. *Mikolov, T., Yih, W.-T. and Zweig, G.* 2013. Linguistic regularities in continuous space word representations. Proceedings of NAACL-HLT, pp. 746–751.
15. *Cortes, C. and Vapnik, V.* 1995. Support-vector networks. Machine Learning, vol. 20, no. 3, pp. 273–297.
16. *Finkel, J., Grenager, T. and Manning, C.* 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363–370.
17. *Mitchell, T.* 1997. Machine Learning. New York: McGraw-Hill. 432 p.

Kolpakov N.A. Moscow Institute of Physics and Technology, 1A, building 1, Kerch str., Moscow, 117303, Russia, e-mail: kolpakov.na@phystech.edu

Molodchenkov A.I. Federal Research Center “Computer Science and Control” of Sciences, 44/2 Vavilova str., Moscow, 119333, Russia, e-mail: aim@tesyan.ru

Lukin A.V. Peoples’ Friendship University of Russia, 6, Miklukho-Maklaya str., Moscow, 117198, Russia, e-mail: antonvlukin@gmail.com