ЛЕНИВЦЕВА Ю.Д.,
Национальный Центр Когнитивных Разработок, Университет ИТМО, г. Санкт-Петербург, Россия,
e-mail: lenivezzki@gmail.com

КОПАНИЦА Г.Д.,
к.т.н., Национальный Центр Когнитивных Разработок, Университет ИТМО, г. Санкт-Петербург, Россия,
e-mail: georgy.kopanitsa@gmail.com

ОПРЕДЕЛЕНИЕ ТИПА АЛЛЕРГИИ НА ОСНОВАНИИ НЕСТРУКТУРИРОВАННЫХ МЕДИЦИНСКИХ ЗАПИСЕЙ

DOI: 10.25881/ITP.2021.44.51.002

Аннотация.
Использование разных форматов данных затрудняет стандартизацию и обмен медицинских данных. Более того, большая часть медицинских данных хранится в виде неструктурированных медицинских записей, что затрудняет их обработку. В данной работе мы решаем задачу категоризации неструктурированных аллергологических анамнезов по категориям, предоставленным в стандарте обмена FHIR. Была разработана двухэтапная модель классификации на основе размеченных вручную медицинских записей. На первом этапе модель фильтрует записи с информацией об аллергии, а на втором этапе классифицирует каждую запись. Модель показала высокую точность. Развитие предложенного подхода обеспечит вторичное использование и обмен данными.

Ключевые слова: *стандартизация медицинских данных, FHIR, аллергия и непереносимость, обработка естественного языка, интероперабельность.*

Для цитирования*: Ленивцева Ю.Д., Копаница Г.Д. Определение типа аллергии на основании неструктурированных медицинских записей. Врач и информационные технологии. 2021; 1: 18–24. doi: 10.25881/ITP.2021.44.51.002.*

LENIVTCEVA I.D.,
National Center for Cognitive Technologies, ITMO University, Saint-Petersburg, Russia,
e-mail: lenivezzki@gmail.com

KOPANITSA G.D.,
PhD, National Center for Cognitive Technologies, ITMO University, Saint-Petersburg, Russia,
e-mail: georgy.kopanitsa@gmail.com

AUTOMATIC ALLERGY CLASSIFICATION BASED ON RUSSIAN MEDICAL FREE TEXTS

DOI: 10.25881/ITP.2021.44.51.002

Abstract.
Different data formats are challenging for the standardization and exchange of medical data. In addition, most medical data in medical information systems (MIS) or databases is stored in an unstructured way, causing difficulties in processing the data. The article proposes an approach for categorizing unstructured medical records of patients with allergies into the categories provided in the FHIR exchange standard. We developed a two-stage classification model based on manually labelled medical records. The method is based on machine learning algorithms, as well as international standards for the exchange of medical data. The model has shown high accuracy. The development of the presented approach for structuring medical texts will ensure the reuse and interoperability of medical data.

Keywords: *medical data structuring, allergy, machine learning, unstructured texts analysis, interoperability.*

How to cite: *Lenivtceva ID, Kopanitsa GD. Automatic allergy classification based on Russian medical free texts. Medical doctor and information technology. 2021; 1: 18–24. (In Russ.). doi: 10.25881/ITP.2021.44.51.002.*

ВВЕДЕНИЕ

Преемственность медицинской помощи требует связи и обмена данными для обеспечения высококачественного медицинского обслуживания [1]. Основная проблема возникает из-за использования разных форматов данных, когда возникает необходимость в обмене медицинскими данными между несколькими агентами, предоставляющими услуги одному и тому же пациенту. Международные терминологические стандарты, такие как SNOMED CT [2] и LOINC [3], логические модели данных, такие как openEHR [4], ISO13606 [5], стандарты HL7 [6] и подробные клинические модели, такие как ISO 13972 [7], были разработаны для решения проблемы интероперабельности медицинских данных. Одним из наиболее перспективных стандартов обмена данными является HL7 FHIR [8].

Принято считать, что около 80% медицинских данных хранятся в виде неструктурированных медицинских записей, которые сложнее обрабатывать по сравнению со структурированной информацией [9]. Однако эти записи содержат полезную информацию для моделирования и исследований [10]. Фильтрация записей вручную, а также обработка и извлечение информации требует много времени и не исключают влияние человеческого фактора на точность. Таким образом, эта задача требует использования методов обработки естественного языка и машинного обучения.

Извлечение информации и классификация текста — это задачи, специфичные для языка и предметной области. Нейронные сети показывают высокую производительность для классификации медицинских текстов. А. Дудченко и др. [11] использовали глубокие классификаторы для выявления диагноза по медицинским записям в произвольном порядке на русском и немецком языках и достигли точности более 95%. Основное ограничение при использовании нейронных сетей заключается в необходимости иметь большой размеченный набор данных. Классификация на основе графовых моделей, выполненная Н. Шанавасом и др. [12], показала 0,86 F-меру и почти 0,87 точность. Простые классификаторы также хорошо подходят для классификации текста. М. Олейник и др. [13] описали в работе классификаторы на основе логистической регрессии со значением F-меры 0,80, а также на основе метода опорных векторов с F-мерой 0,81 в задаче фенотипирования пациентов. В.-Х. Вэнг и др. [14] получили значение F-меры 0,93 в задаче классификации медицинских субдоменов. А.Р. Тафти [15] сообщил о точности 0,82 логистической регрессии в классификации предложений биомедицинской тематики.

Целью данной работы является разработка метода определения категории аллергии на основе русскоязычного неструктурированного текста аллергологических анамнезов для стандартизации медицинских данных.

МЕТОДЫ

Текстовый анамнез аллергии и непереносимости можно сопоставить с ресурсом AllergyIntolerance стандарта FHIR. Он включает информацию о нежелательных реакциях на различные вещества. Задача данной работы выявить категорию аллергии на основе текста медицинской записи согласно атрибутам, выделенным в FHIR. На рисунке 1 представлены четыре категории аллергии в FHIR. Биологическая аллергия не представлена в наборе данных. Таким образом, исследование ограничено категориями аллергии на продукты питания, лекарства и окружающую среду.

Российские медицинские карты более 250 тысяч пациентов предоставлены Национальным медицинским исследовательским центром им. А.А. Алмазова (Санкт-Петербург, Россия). Личная информация пациентов была удалена. Записи содержат фрагменты истории

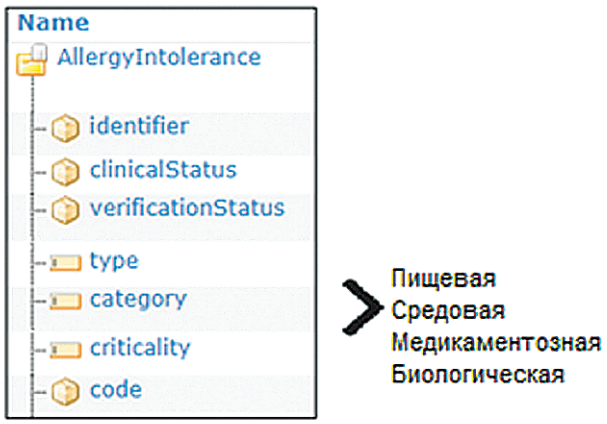

Рисунок 1 — Категории аллергий.

Таблица 1 — Записи и разметка

Запись	Аллергия	Пищевая	Средовая	Медикамент
Аллергический анамнез. Аллергической реакции не отмечено	✗	—	—	—
Аллергия на препараты пенициллина — крапивница (шоколад, яйца)	✓	✓	✗	✓
Аллергическая реауция на пыль и пыльцу, сезонная чувствительность	✓	✗	✓	✗
Аллергическая бронхиальная астма неустановленного генеза	✓	✗	✗	✗
Непереносимость спиртных напитков — аллергические высыпания на коже и отеки	✓	✓	✗	✗

болезни и анамнез жизни, включая аллергологический анамнез. В таблице 1 приведены примеры аллергоанамнезов и их разметки согласно выделенным категориям. Чтобы получить необходимые записи:

- Мы отфильтровали записи пациентов с аллергией и непереносимостью, используя ключевые слова и регулярные выражения («аллергия», «(не)переносимость»).
- Обрезали записи до одного предложения, включающего ключевое слово, чтобы уменьшить информационный шум.
- Удалили полные дубликаты и похожие шаблоны в записях.

После этих шагов мы получили 12590 медицинских записей. Все эти записи были вручную размечены двумя экспертами. В случае разногласий решение принималось на основе консенсуса.

Предварительная обработка:

- Очистка записей от лишних символов и лишних пробелов.
- Исправление синтаксических ошибок, ошибок регистра и пробелов, используя регулярные выражения.
- Исправление пробелов и орфографических ошибок с помощью питоновской библиотеки «symspellpy» (на основе словаря).
- Токенизация и нормализация слов с помощью библиотек «nltk» и «pymorphy2».
- Представление текста в виде мешка слов.

Подход к определению категории аллергии состоит из двух этапов.

- Бинарный классификатор, определяющий, связана ли запись с аллергией или непереносимостью.
- Три бинарных классификатора, определяющих, относится ли запись к одной из трех категорий аллергии.

Для обоих видов классификаторов мы использовали модель на основе логистической регрессии с C = 3, penalty = ‘l2’, solver = ‘saga’, max_iter = 4000, multi_class = ‘ovr’ из реализации «scikit-learn».

Для оценки классификаторов использовались F-мера, точность и полнота.

РЕЗУЛЬТАТЫ

На рисунке 2а представлены диаграммы с распределением размеченных записей по наличию аллергии. На втором этапе каждой записи можно присвоить несколько категорий. Некоторые записи не содержат сведений о природе аллергена и не имеют категории. Мы удалили записи без категории, и, таким образом, набор данных для категоризации аллергии содержит 9140 записей. На рисунке 2б показано распределение количества категорий, которые были упомянуты внутри одной записи. Сообщается, что у пациента есть все три типа аллергии, если записи присвоены три категории. Например, 7741 запись в наборе данных размечена одной категорией, а 1307 записей содержат информацию о двух разных категориях аллергии (продукты питания и лекарства или продукты питания и окружающая среда).

На рисунке 3 представлены диаграммы распределения количества записей по категориям аллергий.

В таблице 2 представлены характеристики примененных классификаторов. После классификации мы получили списки ключевых слов

для каждой категории аллергии. Самые распространенные ключевые слова по категориям показаны в таблице 3.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

На рисунке 2а показано, что после фильтрации по ключевым словам и регулярным выражениям более 20% записей не имеют отношения к аллергии. Это означает, что необходим дополнительный классификатор для фильтрации записей в наборе данных. Мы выбрали в качестве показателей F-меру, точность и полноту, поскольку они не чувствительны к дисбалансу классов.

Мы разработали один фильтрующий классификатор и три классификатора для категоризации неструктурированного аллергологического анамнеза. Согласно рисунку 2б количество категорий, присвоенных записи, различается и зависит от количества типов аллергенов, упомянутых в записи, также присутствуют записи без категории. Обычно в таких записях указывается только реакция, диагноз, связанный с аллергией, или аллерген неизвестен. Мы не включили такие записи в набор данных для категоризации. На рисунке 3 показано, что большинство записей (более 75%) связано с аллергией на лекарства, только 15% связаны с пищевой аллергией и 22% связаны с аллергией на объекты окружающей среды.

Разработанные модели показывают высокие значения метрик, однако также имеют место ошибки в классификации. Например, запись «Аллергия на пыльцу, аллергия на лекарства отсутствует» будет классифицироваться без тега аллергии из-за отрицания. Многие ситуации неверного присвоения тэга связаны с конкретной структурой предложения в медицинских записях. Одна и та же запись может сообщать, что у пациента есть пищевая аллергия, но нет аллергии на лекарства. Таким образом, производительность моделей можно улучшить, применив классификаторы к значимому сегменту предложения.

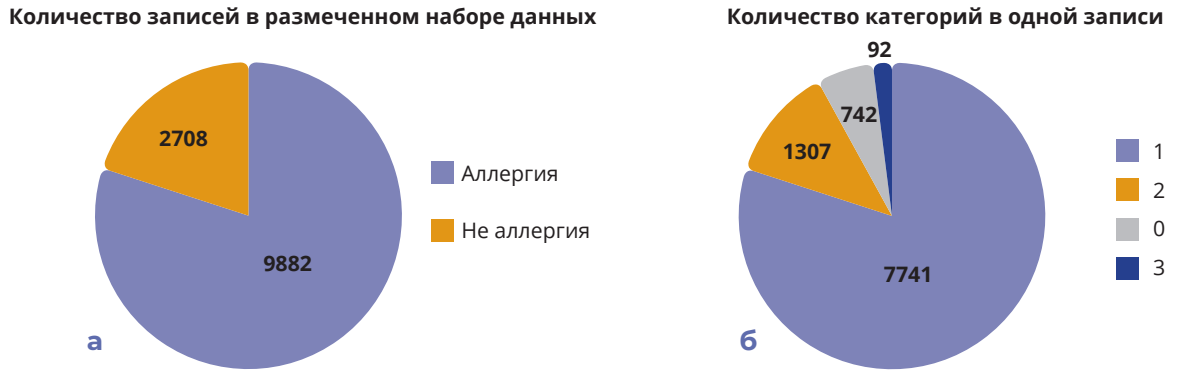

Рисунок 2 — Распределение в размеченном наборе данных: а) количество записей, б) количество категорий на запись (0-3).

Рисунок 3 — Количество записей покатегориям.

Таблица 2 — Характеристики классификаторов

Классификатор	F-мера	Точность	Полнота
Аллергия	0.945	0.923	0.945
Пищевая аллергия	0.953	0.932	0.953
Средовая аллергия	0.932	0.902	0.932
Медикаментозная аллергия	0.962	0.944	0.962

Таблица 3 — Слова, определяющие категорию аллергии в записи

Категория	Ключевые слова
Пищевая	Клубника, пищевой, шоколад, лактоза, цитрусовые, продукт, молоко, мед, рыба, красный, алкоголь, яйцо, орехи
Медикаментозная	Лекарство, новокаин, пенициллин, полиаллергия, антибиотик, бициллин, йод, лекарство, анальгин, аспирин, дифенгидрамин
Средовая	Домашний, пластик, шерсть, цветение, холод, пыль, пыльца, металлы, укус, солнце, краска, насекомое

Таблица 3 содержит списки наиболее важных ключевых слов для каждой категории аллергии, сформированных после классификации. В основном каждый список содержит аллергены, которые встречались в записях соответствующих категорий. Эти списки полезны для сопоставления терминологических кодов медицинским понятиям (SNOMED CT) в автоматическом или полуавтоматическом режиме.

Производительность подхода (таблица 2) близка к производительности глубоких классификаторов, например, точность более 95% в работе А. Дудченко [11]. Разработанные классификаторы превосходят простые классификаторы. И. Уе и др. [16] представили в своей работе значение полноты 0,8 и близкую к 0,9 точность классификации. В.Х. Вэнг и др. в [14] представили неглубокий классификатор, показавший 0,87 F-меры, что ниже результатов предлагаемого подхода. Однако многие исследователи используют англоязычные терминологические базы данных, такие как UMLS, что повышает эффективность классификации. Так, классификатор с концепциями UMLS в [14] показал 0,93 F-меру. Эти базы данных не имеют русских версий и не доступны для задач на русскоязычных текстах. Однако использование международной терминологии и идентификаторов является важной частью семантической интероперабельности.

Предлагаемые решения по стандартизации медицинских данных в виде неструктурированных текстов должны иметь практическое значение. Для достижения полной совместимости и подготовки данных для интеграции мы планируем разработать модель для присвоения стандартных терминологических кодов, таких как SNOMED CT и МКБ-10. Поскольку русскоязычной версии SNOMED CT нет, для этой задачи требуется ее перевод. Также будут разработаны инструменты извлечения данных для извлечения аллергенов и нежелательных реакций.

ЗАКЛЮЧЕНИЕ

В данной работе мы разработали и оценили метод автоматизированного определения категории аллергии из русскоязычных неструктурированных медицинских записей. Двухэтапный метод показал хорошие результаты и сопоставим с современными результатами.

Такой подход к классификации является частью модуля стандартизации русского текста. В дальнейшем стандартизованные данные можно использовать для построения прогностических и автоматизированных моделей назначения терапии, предоставляющих рекомендации для врачей. Развитие этого подхода обеспечит вторичное использование данных и функциональную совместимость неструктурированных медицинских карт.

ЛИТЕРАТУРА/REFERENCES

1. Douglas HE, et al. Implementing information and communication technology to support community aged care service integration: Lessons from an Australian aged care provider. J. Integr. Care. Igitur, Utrecht Publishing and Archiving Services. 2017; 17(1).
2. Fung KW, et al. Using SNOMED CT-encoded problems to improve ICD-10-CM coding—A randomized controlled experiment. J. Med. Inform. Elsevier Ireland Ltd. 2019; 126: 19–25.
3. Fiebeck J, et al. Implementing LOINC: Current status and ongoing work at the Hannover Medical School. Studies in Health Technology and Informatics. IOS Press. 2019; 258: 247–248.
4. Mascia C, et al. OpenEHR modeling for genomics in clinical practice. J. Med. Inform. Elsevier Ireland Ltd. 2018; 120: 147–156.
5. Santos MR, Bax MP, Kalra D. Building a logical EHR architecture based on ISO 13606 standard and semantic web technologies. Studies in Health Technology and Informatics. IOS Press. 2010; 160(1): 161–165.
6. Ulrich H, et al. Metadata repository for improved data sharing and reuse based on HL7 FHIR. Studies in Health Technology and Informatics. IOS Press. 2017; 228: 162–166.
7. Huff S.M, et al. Integrating detailed clinical models into application development tools. Stud. Health Technol. Inform. IOS Press. 2004; 107: 1058–1062.
8. Hong N, et al. Standardizing Heterogeneous Annotation Corpora Using HL7 FHIR for Facilitating their Reuse and Integration in Clinical NLP.AMIA. Annu. Symp. proceedings. 2018; 2018: 574–583.
9. Lenivtceva ID, Kopanitsa G. Evaluating Manual Mappings of Russian Proprietary Formats and Terminologies to FHIR. Methods Inf. Med. 2019; 58: 4–5.
10. Wang Y, et al. Clinical information extraction applications: A literature review. Journal of Biomedical Informatics. 2018; 77: 34–49.
11. Dudchenko A, Ganzinger M, Kopanitsa G. Diagnoses Detection in Short Snippets of Narrative Medical Texts. Procedia Computer Science. 2019; 156: 150–157.
12. Shanavas N, et al. Ontology-based enriched concept graphs for medical document classification. Inf. Sci. (Ny). 2020; 525: 172–181.
13. Oleynik Michel , et al. Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. Journal of the American Medical Informatics Association. Oxford Academic. 2013; 26(11): 1247–1254.
14. Weng W-H, et al. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Med. Inform. Decis. Mak. 2017; 17(1): 155.
15. Tafti AP, et al. BigNN: An open-source big data toolkit focused on biomedical sentence classification. Proceedings–2017 IEEE International Conference on Big Data. Institute of Electrical and Electronics Engineers Inc. 2017; 2018: 3888–3896.
16. Ye Y, et al. Influenza detection from emergency department reports using natural language processing and Bayesian network classifiers. J. Am. Med. Informatics Assoc. BMJ Publishing Group. 2014; 21(5): 815–823.