УНИВЕРСИТЕТ ИТМО

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
май–июнь 2021 Том 21 № 3 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
May–June 2021 Vol. 21 No 3 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК УНИВЕРСИТЕТ ИТМО
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

КРАТКИЕ СООБЩЕНИЯ

BRIEF PAPERS

doi: 10.17586/2226-1494-2021-21-3-433-436
УДК 004.912

Автоматическое определение типа аллергии из неструктурированных медицинских текстов на русском языке

Юлия Дмитриевна Ленивцева[1]✉, Георгий Дмитриевич Копаница[2]

[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] lenivezzki@gmail.com✉, https://orcid.org/0000-0002-5572-5151
[2] Georgy.kopanitsa@gmail.com, https://orcid.org/0000-0002-6231-8036

Аннотация

Большая часть медицинских данных в базах медицинских информационных систем хранится в неструктурированном виде. Методы обработки неструктурированных записей широко представлены в литературе для английского языка. В работе предложен метод интеллектуального анализа неструктурированных аллергологических анамнезов на русском языке с целью определения наличия и типа аллергии и непереносимости у пациента. В основе метода лежат алгоритмы машинного обучения, а также используются международные стандарты обмена медицинскими данными, такие как FHIR и SNOMED CT. В результате эксперимента обработано около 12 тысяч медицинских записей. Значение F-меры для разработанных моделей классификации составило от 0,93 до 0,96. Полученные модели показали высокие значения метрик оценки эффективности моделей. В дальнейшем структурированные данные могут быть использованы в моделях предсказания медицинских рисков. Развитие методов структурирования медицинских текстов обеспечит интероперабельность медицинских данных.

Ключевые слова

структурирование медицинских данных, аллергия, непереносимость, машинное обучение, анализ неструктурированных текстов, интероперабельность

Ссылка для цитирования: Ленивцева Ю.Д., Копаница Г.Д. Автоматическое определение типа аллергии из неструктурированных медицинских текстов на русском языке // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21, № 3. С. 433–436. doi: 10.17586/2226-1494-2021-21-3-433-436

Automatic allergy classification based on Russian unstructured medical texts

Iuliia D. Lenivtceva[1]✉, Georgy D. Kopanitsa[2]

[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] lenivezzki@gmail.com✉, https://orcid.org/0000-0002-5572-5151
[2] Georgy.kopanitsa@gmail.com, https://orcid.org/0000-0002-6231-8036

Abstract

Most of the medical data in hospital information systems databases are stored in an unstructured form. Techniques for processing unstructured records are widely presented in scientific papers focused on English data. This paper proposes a method for intellectual analysis of unstructured allergy anamnesis in Russian in order to identify the presence and type of allergy and intolerance of a patient. The method is based on machine learning algorithms and uses international standards for the exchange of medical data and terminology standards, such as FHIR and SNOMED CT. As a result of the experiment, about 12 thousand medical records were processed. F-measure for the developed classification models ranged from 0.93 to 0.96. The models showed high values of metrics for evaluating the effectiveness of the models. In the future, structured data can be used in models for predicting medical risks. Further development of methods for structuring medical texts will ensure the interoperability of medical data.

Keywords

medical data structuring, allergy, intolerance, machine learning, unstructured text analysis, interoperability

For citation: Lenivtceva Iu.D., Kopanitsa G.D. Automatic allergy classification based on Russian unstructured medical texts. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2021, vol. 21, no. 3, pp. 433–436 (in Russian). doi: 10.17586/2226-1494-2021-21-3-433-436

Большая часть медицинских данных хранится в виде неструктурированного текста, что вызывает трудности в процессе обработки [1]. Стандартизация медицинских данных — важная задача для эффективного обмена данными и интеграции. Для обеспечения интероперабельности медицинских данных используются международные терминологические стандарты (SNOMED CT [2], LOINC [3], МКБ-10[1]) и стандарты обмена данными (openEHR [4], ISO 13606 [5], стандарты HL7 [6]). Одним из перспективных стандартов на Российском рынке является HL7 FHIR [7].

Извлечение информации из неструктурированного текста — одна из распространенных задач обработки естественного языка. А. Дудченко в работе [8] использует нейронные сети для извлечения диагнозов из неструктурированных медицинских записей с F-мерой равной 0,97. Однако для эффективной работы нейронных сетей необходим большой набор данных для обучения. Логистическая регрессия и метод опорных векторов часто используются в задачах классификации текстов: М. Олейник [9] получил значение 0,8 F-меры для логистической регрессии и 0,81 для метода опорных векторов в задаче определения фенотипа пациентов.

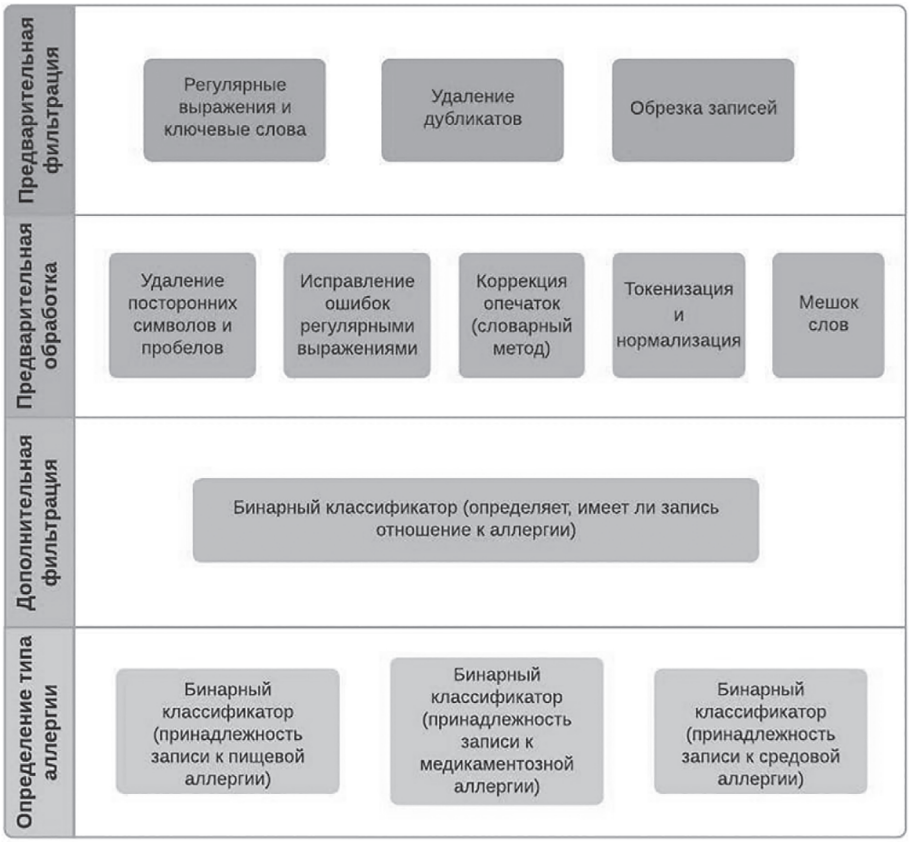
В настоящей работе решена задача автоматического определения типа аллергии из неструктурированных аллергологических анамнезов с использованием ресурса AllergyIntolerance FHIR. На основе структуры ресурса AllergyIntolerance авторы выделили три типа аллергии, описания которых присутствуют в исходном наборе данных: пищевую, медикаментозную и средовую.

Эксперимент выполнен на 12 тысячах медицинских записей более 4 тысяч пациентов, данные предоставлены ФГБУ «НМИЦ им. В.А. Алмазова» за 2017–2019 годы. Данные были размечены вручную авторами. Для разрешения спорных ситуаций привлекался третий эксперт.

Около 78 % в размеченном наборе данных содержат информацию об аллергии. Из них около 15 % записей содержат информацию о пищевой аллергии, 22 % — о средовой аллергии и 79 % — о медикаментозной аллергии.

На рисунке представлена схема и содержание этапов определения типа аллергии из неструктурированных медицинских записей.

[1] ICD 10 — International classification of diseases 10th revision (translated) [Электронный ресурс]. URL: https://mkb-10.com/ (дата обращения: 19.04.2021).



*Рисуно*к. Этапы метода определения типа аллергии

*Fig. S*teps of allergy identification

Задачи фильтрации и категоризации аллергоанамнезов решались с помощью алгоритмов машинного обучения. Лучший результат получен с использованием модели логистической регрессии (пакет «scikit-learn») с параметрами C=3, penalty='l2', solver='saga', max_iter=4000, multi_class='ovr'. Помимо модели логистической регрессии в эксперименте участвовали: метод опорных векторов, метод *k*-ближайших соседей, наивный байесовский классификатор, деревья принятия решений.

Для оценки работы классификаторов использована F-мера.

Классификатор, определяющий отношение записи к аллергии, показал значение F-меры, равное 0,95. F-мера для классификаторов, определяющих тип аллергена, составила 0,95 для пищевой аллергии; 0,93 для средовой аллергии; 0,96 для медикаментозной аллергии.

На основе коэффициентов логистической регрессии составлены списки ключевых слов для каждого типа аллергии:

— пищевая аллергия: пищевой продукт, лактоза, шоколад, цитрусовый;
— средовая аллергия: домашний, цветение, пыль, металл, укус;
— медикаментозная аллергия: медикамент, лекарство, антибиотик, йод, поливалентный.

Каждый список ключевых слов преимущественно содержит наименования аллергенов, которые указаны в записях согласно категории аллергии. Эти списки полезны для сопоставления терминологии (SNOMED CT) и разработки алгоритма автоматической идентификации международных терминологических кодов на основании неструктурированного текста.

Настоящая работа часть программного модуля стандартизации неструктурированных медицинских данных. Разработанный метод позволяет использовать данные, которые ранее были недоступны, в предсказательном моделировании и тем самым повысить точность предсказательных моделей. Развитие методов структурирования медицинских текстов обеспечивает повторное использование и интероперабельность медицинских данных.

Литература

1. Lenivtceva I.D., Kopanitsa G. Evaluating manual mappings of Russian proprietary formats and terminologies to FHIR // Methods of Information in Medicine. 2019. V. 58. N 4-5. P. 151–159. doi: 10.1055/s-0040-1702154
2. Fung K.W., Xu J., Rosenbloom S.T., Campbell J.R. Using SNOMED CT-encoded problems to improve ICD-10-CM coding—A randomized controlled experiment // International Journal of Medical Informatics. 2019. V. 126. P. 19–25. doi: 10.1016/j.ijmedinf.2019.03.002
3. Fiebeck J., Gietzelt M., Ballout S., Christmann M., Fradziak M., Laser H., Ruppel J., Schönfeld N., Teppner S., Gerbel S. Implementing LOINC: Current status and ongoing work at the Hannover Medical School // Studies in Health Technology and Informatics. 2019. V. 258. P. 247–248. doi: 10.3233/978-1-61499-959-1-247
4. Mascia C., Uva P., Leo S., Zanetti G. OpenEHR modeling for genomics in clinical practice // International Journal of Medical Informatics. 2018. V. 120. P. 147–156. doi: 10.1016/j.ijmedinf.2018.10.007
5. Santos M.R., Bax M.P., Kalra D. Building a logical EHR architecture based on ISO 13606 standard and semantic web technologies // Studies in Health Technology and Informatics. 2010. V. 160. N 1. P. 161–165. doi: 10.3233/978-1-60750-588-4-161
6. Ulrich H., Kock A.-K., Duhm-Harbeck P., Habermann J.K., Ingenerf J. Metadata repository for improved data sharing and reuse based on HL7 FHIR // Studies in Health Technology and Informatics. 2017. V. 228. P. 162–166. doi: 10.3233/978-1-61499-678-1-162
7. Hong N., Wen A., Mojarad M.R., Sohn S., Liu H., Jiang G. Standardizing heterogeneous annotation corpora using HL7 FHIR for facilitating their reuse and integration in clinical NLP // AMIA Annual Symposium Proceedings. American Medical Informatics Association, 2018. V. 2018. P. 574–583.
8. Dudchenko A., Dudchenko P., Ganzinger M., Kopanitsa G. Extraction from medical records // Studies in Health Technology and Informatics. 2019. V. 261. P. 62–67. doi: 10.3233/978-1-61499-975-1-62
9. Oleynik M., Kugic A., Kasáč Z., Kreuzthaler M. Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification // Journal of the American Medical Informatics Association. 2019. V. 26. N 11. P. 1247–1254. doi: 10.1093/jamia/ocz149

References

1. Lenivtceva I.D., Kopanitsa G. Evaluating manual mappings of Russian proprietary formats and terminologies to FHIR. *Methods of Information in Medicine*, 2019, vol. 58, no. 4-5, pp. 151–159. doi: 10.1055/s-0040-1702154
2. Fung K.W., Xu J., Rosenbloom S.T., Campbell J.R. Using SNOMED CT-encoded problems to improve ICD-10-CM coding—A randomized controlled experiment. *International Journal of Medical Informatics*, 2019, vol. 126, pp. 19–25. doi: 10.1016/j.ijmedinf.2019.03.002
3. Fiebeck J., Gietzelt M., Ballout S., Christmann M., Fradziak M., Laser H., Ruppel J., Schönfeld N., Teppner S., Gerbel S. Implementing LOINC: Current status and ongoing work at the Hannover Medical School. *Studies in Health Technology and Informatics*, 2019, vol. 258, pp. 247–248. doi: 10.3233/978-1-61499-959-1-247
4. Mascia C., Uva P., Leo S., Zanetti G. OpenEHR modeling for genomics in clinical practice. *International Journal of Medical Informatics*, 2018, vol. 120, pp. 147–156. doi: 10.1016/j.ijmedinf.2018.10.007
5. Santos M.R., Bax M.P., Kalra D. Building a logical EHR architecture based on ISO 13606 standard and semantic web technologies. *Studies in Health Technology and Informatics*, 2010, vol. 160, no. 1, pp. 161–165. doi: 10.3233/978-1-60750-588-4-161
6. Ulrich H., Kock A.-K., Duhm-Harbeck P., Habermann J.K., Ingenerf J. Metadata repository for improved data sharing and reuse based on HL7 FHIR. *Studies in Health Technology and Informatics*, 2017, vol. 228, pp. 162–166. doi: 10.3233/978-1-61499-678-1-162
7. Hong N., Wen A., Mojarad M.R., Sohn S., Liu H., Jiang G. Standardizing heterogeneous annotation corpora using HL7 FHIR for facilitating their reuse and integration in clinical NLP. *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, 2018, vol. 2018, pp. 574–583.
8. Dudchenko A., Dudchenko P., Ganzinger M., Kopanitsa G. Extraction from medical records. *Studies in Health Technology and Informatics*, 2019, vol. 261, pp. 62–67. doi: 10.3233/978-1-61499-975-1-62
9. Oleynik M., Kugic A., Kasáč Z., Kreuzthaler M. Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. *Journal of the American Medical Informatics Association*, 2019, vol. 26, no. 11, pp. 1247–1254. doi: 10.1093/jamia/ocz149

Авторы

Ленивцева Юлия Дмитриевна — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57216567381, https://orcid.org/0000-0002-5572-5151, lenivezzki@gmail.com
Копаница Георгий Дмитриевич — кандидат технических наук, ведущий научный сотрудник, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 55326019500, https://orcid.org/0000-0002-6231-8036, Georgy.kopanitsa@gmail.com

Authors

Iuliia D. Lenivtceva — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57216567381, https://orcid.org/0000-0002-5572-5151, lenivezzki@gmail.com
Georgy D. Kopanitsa — PhD, Leading Researcher, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 55326019500, https://orcid.org/0000-0002-6231-8036, Georgy.kopanitsa@gmail.com

Статья поступила в редакцию 25.03.2021
Одобрена после рецензирования 06.04.2021
Принята к печати 11.05.2021

Received 25.03.2021
Approved after reviewing 06.04.2021
Accepted 11.05.2021