pHealth 2020
B. Blobel et al. (Eds.)

doi:10.3233/SHTI200634

Category of Allergy Identification from Free-Text Medical Records for Data Interoperability

Iuliia LENIVTCEVA[a,1], Mariya KASHINA[a], Georgy KOPANITSA[a]
[a] *ITMO University, Saint Petersburg, Russian Federation*

Abstract. The use of different data formats complicates the standardization and exchange of valuable medical data. Moreover, a big part of medical data is stored as unstructured medical records that are complicated to process. In this work we solve the task of unstructured allergy anamnesis categorization according to categories provided by FHIR. We applied two stage classification model with manually labeled records. On the first stage the model filters records with information about allergies and on the second stage it categorizes each record. The model showed high performance. The development of this approach will ensure secondary use of data and interoperability.

Keywords. Medical data standardization, FHIR, allergy and intolerance, NLP, interoperability

Introduction

Integrated care requires to enable high level communication and data exchange to ensure a high-quality medical care [1]. The main challenge occurs when there is a need to exchange medical data between multiple agents providing services to the same patient due to the use of different data formats. Many international standards for terminology such as SNOMED CT [2] and LOINC [3]; logical data models such as openEHR [4], ISO13606 [5], HL7 standards [6] and detailed clinical models such as ISO 13972 [7] were developed to overcome this problem and ensure interoperability. One of the most prospective standards for data exchange is HL7 FHIR [8].

It is widely accepted that about 80% of medical data are stored as unstructured medical notes which are complicated to process compared to structured information [9]. However, these notes contain useful information for modelling and research [10]. Manual information filtering and extraction is time-consuming. Thus, this task requires the use of Natural Language Processing (NLP) techniques.

Information extraction (IE) and free-text classification are language and domain specific tasks. Neural networks (NN) show high performance for medical text classification. Dudchenko et al [11] used deep classifiers to discover diagnosis from free-text medical notes in Russian and German and achieved over 95% accuracy. The main limitation in NN applications is the need for a big dataset. Graph-based classification by

[1] Corresponding Author, Iuliia Lenivtceva, ITMO University, 49 Kronverkskiy prospect, 197101 Saint Petersburg, Russian Federation; E-mail: lenivezzki@gmail.com.

Shanavas et al [12] showed 0.86 F-score and almost 0.87 Precision and recall. Shallow classifiers also perform well for text classification. Oleynik et al [13] reported 0.80 F-score by Logistic Regression (LR) and 0.81 by Support Vector Machines (SVM) in patient-phenotyping. Weng et al [14] SVM showed 0.93 F-score in subdomain medical classification. Tafti [15] reported 0.82 Precision and Recall of LR in biomedical sentence classification.

The aim of this work is to develop a method for allergy category identification from Russian free-text allergy anamnesis to facilitate medical data standardization and interoperability.

1. Methods

Free-text allergy anamnesis can be mapped to AllergyIntolerance which is one of the FHIR Summary resources. It includes information about undesired reactions on different substances. The task of this work is to identify the category of allergy from the record. Figure 1 represents four categories of allergies introduced in FHIR.

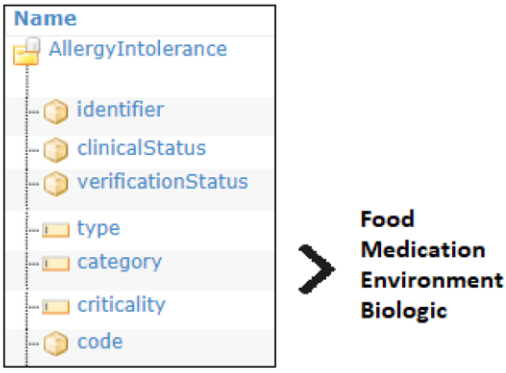

Figure 1. Categories of allergy in FHIR

Biologic allergy is not represented in the dataset; thus, the study is limited to food, medication and environment categories of allergy.

Russian medical records of more than 250 thousand patients were provided by the Almazov National Medical Research Centre (St. Petersburg, Russia). The personal information of patients was discarded. The records contain medical history fragments and anamnesis of life including allergy anamnesis. Table 1 shows the examples of records and labelling.

Table 1. Records examples and labeling

Record	Allergy	Food	Environment	Medication
Allergy anamnesis. No allergic reaction noted.	×	–	–	–
Allergy to medications penicillin – urticaria; chocolate, eggs.	✓	✓	×	✓
Dust and weed pollen allergy reaction, seasonal sensitivity.	✓	×	✓	×
Allergic bronchial asthma of unknown origin.	✓	×	×	×
Intolerance to alcoholic drinks with allergic skin rush and edema.	✓	✓	×	×

To get relevant records:

- We filtered patients' records with allergy and intolerances using keywords and regular expressions («allergy», «(in)tolerance»)
- Cut the records with a one-sentence window from a keyword to reduce noise
- Removed full duplicates and similar patterns in records.

After these steps we obtained 12590 medical records. All these records were labeled manually by two experts. In case of disagreement the decision was made by consensus.

Preprocessing:

- Clean the records from extra symbols and extra spaces.
- Correct syntactic, case and spaces errors using regular expressions where possible
- Correct space and spelling errors using «symspellpy» (dictionary based)
- Tokenize and normalize words with «nltk» and «pymorphy2»
- Represent text as Bag of Words (BOW).

The approach on allergy category identification consists of two stages.

- Binary classifier identifying if a record is related to allergy or intolerance.
- Three binary classifiers identifying if a record is related to one of three allergy categories.

For both classifiers we used LR with C=3, penalty='l2', solver='saga', max_iter=4000, multi_class='ovr' from «scikit-learn» implementation. F-Score, Precision and Recall were used to evaluate classifiers.

2. Results

Figure 2a illustrates the number of records per classes in a labeled dataset. On the second stage each record can be labeled with several categories. Some records do not report the allergen nature and have no category. We removed records with no category and, thus, the dataset for allergy categorization contains 9140 records. Figure 2b illustrates the distribution of categories number per record. A patient is reported to have all three types of allergy if a record is assigned with three categories. For instance, 7741 records in the

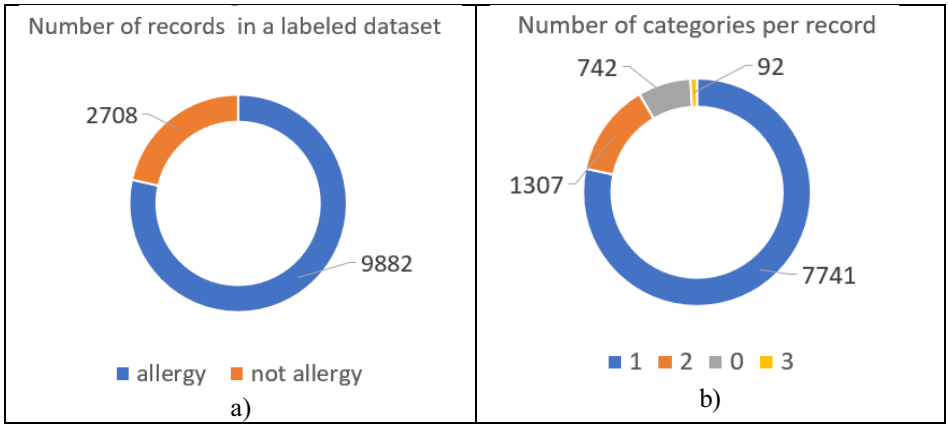

Figure 2. Data distribution in the labeled dataset a) number of records in a labeled dataset, b) number of categories labeled per record (from 0 to 3)

dataset are labelled with one category and 1307 records contain information about two different allergy categories (food and medication or food and environment).

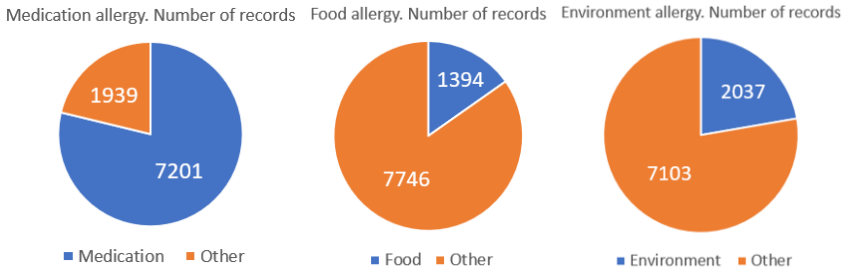

Figure 3. Number of records in categories

Table 2 represents the performance of the applied classifiers.

Table 2. Performance of the classifiers

Classifier	F-score	Precision	Recall
Relation to allergy	0.945	0.923	0.945
Food category	0.953	0.932	0.953
Environment category	0.932	0.902	0.932
Medication category	0.962	0.944	0.962

After classification we obtained lists of keywords for each category of allergy. Top keywords are shown in Table 3.

Table 3. Unigrams indicating category of allergy in a record

Category	Top unigrams
Food	Strawberry, food, chocolate, lactose, citrus, product, milk, honey, fish, red, alcohol, egg, nuts
Medication	Medication, novocaine, penicillin, polyallergy, antibiotic, bicillin, iodine, drug, medicine, analgin, aspirin, diphenhydramine
Environment	House, plaster, wool, flowering, cold, dust, pollen, metals, bite, sun, paint, insect

3. Discussion

Figure 2a shows that more than 20% of records are not related to allergy after filtering by keywords and regular expressions. It causes the need in additional classifier to filter the records in the dataset with imbalanced classes. We chose F-score, Precision and Recall metrics as they are not sensitive to classes imbalance.

We developed one filtering classifier and three classifiers for free-text allergy anamnesis categorization. According to figure 2b the number of categories assigned to a record differs and depends on the number of allergen types mentioned in a record. There are records with no category. Typically, these records specify only a reaction, allergy related diagnosis or reports the unknown allergen. We did not include these records in the dataset for categorization. Figure 3 shows that most records (more than 75%) are

related to allergies on medications, only 15% are related to food allergies and 22% are related to environment allergies.

The applied models perform well, however, misclassifications take place. Misclassification is a situation when the classifier labels a record with wrong category. For instance, the record *"Pollen allergy, no medication allergy"* would be classified with no allergy tag because of negation. Many misclassifications are connected with specific sentence structure of medical records. One record can report that a patient has food allergy but does not have medication allergy. Thus, the performance of models can be improved by applying classifiers to a meaningful segment of a sentence.

Table 3 contains lists of top important keywords for each allergy category after classification. Mostly each list contains allergens specified in the records according to category. These lists are helpful for terminology mappings (SNOMED CT) and automatic codes assignment.

The performance of the approach (table 2) is close to performance of deep classifiers such as over 95% accuracy in [11]. The developed classifiers outperform most shallow classifiers. Ye et al [16] represented 0.8 Recall and close to 0.9 Precision for emergence reports classification. Weng et al in [14] represented a shallow classifier which showed 0.87 F-score which is lower than the results of the suggested approach. However, many researchers use concepts databases, such as UMLS, which improves the performance of the classification. Thus, the classifier with UMLS concepts in [14] showed 0.93 F-score. These databases have no Russian mappings and are not available for classification. However, the use of international terminologies and identifiers is the essential part of semantic interoperability.

The suggested solutions on standardizing free-text medical data should have impact in practice. To achieve full interoperability and prepare data for integration we plan to develop a model for standard terminology codes assignment such as SNOMED CT and ICD-10. As there is no Russian version of SNOMED CT this task requires its translation. Also, data extraction tools will be developed to specify substances and undesired reactions.

4. Conclusions

In this work we developed and evaluated a method for automated category of allergy identification from Russian free-text medical records. The two-stage method performed well and is comparable with state-of-the-art results.

This classification approach is a part of Russian free-text standardization module. The standardized data then can be used to construct predictive and automated therapy appointment models providing recommendations to clinicians. The development of this approach will ensure secondary use of data and interoperability of unstructured medical records.

Acknowledgments

This work financially supported by the government of the Russian Federation through the ITMO fellowship and professorship program. This work was supported by a Russian Fund for Basic research 18-37-20002. This work is financially supported by National Center for Cognitive Research of ITMO University.

References

[1] Douglas HE, Georgiou A, Tariq A, Prgomet M, Warland A, Armour P, Westbrook JL. Implementing information and communication technology to support community aged care service integration: Lessons from an Australian aged care provider. Int J Integr Care . 2017 Apr 10;17(1):9. doi: 10.5334/ijic.2437.

[2] Fung KW, Xu J, Rosenbloom ST, Campbell JR. Using SNOMED CT-encoded problems to improve ICD-10-CM coding—A randomized controlled experiment. Int J Med Inform. 2019;126:19–25. doi:10.1016/j.ijmedinf.2019.03.002.

[3] Fiebeck J, Gietzelt M, Ballout S, et al. Implementing LOINC: Current status and ongoing work at the Hannover Medical School. Stud. Health Technol. Inform. 2019;267:247–248. 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, Int. J. Med. Inform. 2018;120:147–156. doi:10.1016/j.ijmedinf.2018.10.007.

[5] Santos MR, Bax MP, Kalra D. Building a logical EHR architecture based on ISO 13606 standard and semantic web technologies. Stud. Health Technol. Inform. 2010;160(Pt 1):161–165. doi:10.3233/978-1-60750-588-4-161.

[6] Ulrich H, Kock AK, Duhm-Harbeck P, HabermannJK, Ingenerf J. Metadata repository for improved data sharing and reuse based on HL7 FHIR. Stud Health Technol Inform. 2017;228:162–166. doi:10.3233/978-1-61499-678-1-162.

[7] Huff SM, R.A. Rocha RA, J.F. Coyle JF, S.P. Narus SP. Integrating detailed clinical models into application development tools. Stud Health Technol Inform. 2004;107:1058–1062. doi:10.3233/978-1-60750-949-3-1058.

[8] Hong N, Wen A, Mojarad MR, Sohn S, Liu H, Jiang G. Standardizing Heterogeneous Annotation Corpora Using HL7 FHIR for Facilitating their Reuse and Integration in Clinical NLP. AMIA Annu. Symp. Proceedings. AMIA Symp. 2018;2018:574–583.

[9] Lenivtceva ID, G. Kopanitsa G. Evaluating Manual Mappings of Russian Proprietary Formats and Terminologies to FHIR. Methods Inf Med. 2019;58:151–159. doi:10.1055/s-0040-1702154.

[10] Wang Y, Wang L, Rastegar-Mojarad M, Moon S, et al. Clinical information extraction applications: A literature review. J Biomed Inform. Jan. 2018;77:34-49. doi:10.1016/j.jbi.2017.11.011.

[11] Dudchenko A, M. Ganzinger M, G. Kopanitsa G. Diagnoses Detection in Short Snippets of Narrative Medical Texts. Procedia Comp Sci. 2019;156:150–157. doi:10.1016/j.procs.2019.08.190.

[12] Shanavas N, H. Wang H, Z. Lin Z, G. Hawe G. Ontology-based enriched concept graphs for medical document classification. Inf Sci. (Ny). 2020;525:172–181. doi:10.1016/j.ins.2020.03.006.

[13] Oleynik M, Kugic A, Kasáč Z, Kreuzthaler M. Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. J Am Med Inform Assoc. 2013;26:1247–1254. doi:https://doi.org/10.1093/jamia/ocz149.

[14] Weng W-H, K.B. Wagholikar KB, A.T. McCray AT, P. Szolovits P, H.C. Chueh HC. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Med Inform Decis Mak. 2017;17:155. doi:10.1186/s12911-017-0556-8.

[15] [15] A.P. Tafti AP, E. Behravesh E, M. Assefi M, E. Larose E, J. Badger J, J. Mayer J, A. Doan A, D. Page D, P. Peissig P. BigNN: An open-source big data toolkit focused on biomedical sentence classification. Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017, Institute of Electrical and Electronics Engineers Inc., 2017: pp. 3888–3896. doi:10.1109/BigData.2017.8258394.

[16] Ye Y, Tsui FR, Wagner M, Espino JU, Li Q. Influenza detection from emergency department reports using natural language processing and Bayesian network classifiers. J Am Med Inform Assoc. 2014;21:815–823. doi:10.1136/amiajnl-2013-001934.