SCIENTIFIC PUBLICATIONS OF THE STATE UNIVERSITY OF NOVI PAZAR
SER. A: APPL. MATH. INFORM. AND MECH. vol. 14 (2) 2022, 65-72

Automatic Labeling of Diagnosis in Medical Reports in Serbian

A. R. Avdić, U. A. Marovac, D. S. Janković, S. S. Marovac

Abstract: A large number of patient health data is collected daily in medical information systems. This data contains a non-structural part written in natural language that contains the physician's notes on specific characteristics of the patient's medical condition. This section may contain symptoms, diagnoses, therapies, specialties, Latin terms, and other words specific to the medical domain. Useful information suitable for various analyzes could be extracted by processing this section of the text. There are no electronic lexical resources in the Serbian language that are suitable for normalizing and extracting knowledge from medical texts, as well as methods for marking terms in this domain. One reason is that, before any method is applied, the de-identification of patients and staff must be ensured. Also, the evaluation of the results requires manually marked corpora of medical reports in the Serbian language. This paper proposes a method for identifying words belonging to diagnoses in medical texts written in Serbian using natural language processing (NLP) techniques. The proposed method is based on the use of lexical resources, and two set of 1000 medical reports are manually marked for research purposes. In the experimental part, the results of automatic labeling of diagnoses on the marked corpus using the proposed method are presented.

Keywords: Electronic health records, natural language processing, diagnosis labeling

1 Introduction

The development of information and communication technology has contributed to digitization in various fields, including healthcare. The existence of medical information systems that enable the entering and storing of data on the health status of patients, in addition to administrative facilities, also enables the constant storage of data that can be analyzed to obtain knowledge. Saved patient health reports in electronic form are called electronic health reports (EHRs) [1]. EHRs are made mainly according to the internal needs of the hospital. They are needed by various actors such as: medical staff, patients whose health is documented, clinical research (medical researchers, pharmacists, epidemiologists, etc.), hospital management to monitor finances and inventory planning, budget, etc. EHRs may contain

Manuscript received June 12, 2022; accepted

A. R. Avdić is with the State University of Novi Pazar, Novi Pazar, Serbia; U. M. Marovac is with State University of Novi Pazar, Novi Pazar, Serbia; D. S. Janković is with Faculty of Electronic Engeneering, Niš; S. S. Marovac is with General Hospital, Novi Pazar, Serbia

numerical and textual information. Medical data consist of structured, semi-structured and unstructured data. The structural part is entered into the EHR within the marked field, and it can be e.g. name and surname of the patient, health insurance number, age of the patient, etc. The non-structural part is usually entered in EHRs through a text field in which there can be a more detailed description of the patient's health condition, the circumstances of the disease, the doctor's remarks. Often in this part you can find the results of laboratory analyzes of the patient, accompanying diagnoses and other things of importance that cannot be expressed in the offered structural fields. Semi-structural fields combine the properties of structural and non-structural data. As the non-structural part contains data on the health status of patients written by a doctor or nurse, this part requires more complex processing, which usually involves the existence of appropriate lexical sources.

The set of artificial intelligence methods that deal with the processing of natural languages, i.e. texts written in natural languages are NLP Natural Language Processing methods [2, 3], and they represent a subset of methods for text mining. The use of these methods is necessary for information retrieval from unstructured part, free text in EHRs.

The motivation for this research is to create a basis for the use of data collected daily in medical information systems. This means creating free text processing methods in EHRs, which would provide a large number of opportunities that would contribute to better management of EHRs and gaining knowledge from them. Some of these possibilities are design and implementation of software for labeling EHRs, creation of tools for error detection and correction in EHRs, smart city services based on data from EHRs such as e.g. epidemic control service etc. Our contribution in this paper is to create a method for labeling diagnoses in the free text in EHRs, based on lexical resources. This paper describes the procedure of necessary steps for free text processing in electronic medical documents to mark diagnoses in them, as well as evaluation of methods on the data sets consisting of 1000 EHRs originating from 2018, collected during the measles epidemic in Nis, using the MEDIS information system [4].

The paper is organized as follows. The second chapter provides an overview of related works that has had an impact on our research. This is followed by a description of the medical lexical resources necessary for the implementation of the proposed methods, as well as a description of the data set over which the methods were performed. The following is an overview of NLP techniques for marking diagnoses in the nonstructural part of EHRs. Then, the results of the experiment in which the proposed methods were applied to the mentioned data set are presented. Finally, conclusions and directions for further research are given.

2 Related Research

This section provides an overview of the corpses for medical domain and the most popular software solutions for labeling clinical text.

Most of the known medical corpora were created for the English-speaking countries. Such corpses are Informatics for Integrating Biology & the Bedside (i2b2) [5] and Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC II) [6] as well as a corpus of biomedical texts annotated for uncertainty, negation, and their scopes The BioScope [7]. The corpus for the medical domain available in Swedish is HEALTHBANK — Swedish Health Record Research Bank [8].

WordNet is a lexical database of semantic relations between words in more than 200 languages, and it is available in Serbian. Its extension for biomedical sciences [9] is the only resource available for the Serbian language, but it is not suitable and sufficient for the extraction of certain medical terms, since some of them are in Latin or have informal synonyms.

The most popular software solutions for labeling medical text are systems cTAKES and CLAMP. cTAKES represents a natural language processing system for the extraction of information from electronic medical record clinical free-text and its complete architecture is described in [10, 11]. A similar NLP-based CLAMP system that enables recognition and automatic encoding of clinical information in narrative patient reports is described in the paper [12]. There is also the software system MedaCy [13], which is a medical text mining framework built over spaCy [14] (Industrial-Strength Natural Language Processing Tool) to facilitate the engineering, training, and application of machine learning models for medical information extraction.

3 Materials and Methods

The processing of medical reports consists of several steps such as data cleaning, integration, transformation, reduction, and finally, privacy protection [15].

The first step is to remove noise from the original data, then inconsistencies, and supplement incomplete data with default values. The second step is to unify the data coming from different sources. The data reduction implies its compression, to achieve greater efficiency in their processing. The data transformation is the conversion of data into a unique format suitable for data mining. Since the EHR contains various sensitive data about the patient and his health condition, about the staff, irreparable damage would occur if they came into the possession of a third party. Therefore, it is necessary to apply some of the methods that enable the protection of this data (encryption, access control, protection protocols, etc.), which is the task of the last step in the processing, and that is the privacy protection step.

The result is processed data suitable for knowledge extraction. To achieve the transformation of data into a suitable form for the further processing and extraction of knowledge, it is necessary to reduce medical terms to a standardized format, i.e. to be able to indicate diagnoses, symptoms, drugs, laboratory analyzes within the report. Therefore, for these purposes, resources (diagnoses and diagnosis codes) have been collected and adapted to facilitate the process of marking diagnoses in EHRs.

Medical lexical resources. ICD-10 (International Statistical Classification of Diseases and Related Health Problems) is a standard for the classification and coding of diseases and medical problems and is used for both clinical and administrative purposes. It is publicly available and translated into several languages. It contains disease codes, its description, symptoms and signs, social circumstances and external causes of the disease, and more. The initial classification contains about 14000 diagnoses [16]. Extended versions and national editions of this classification contain multiple diagnoses. Creating the resource of diagnosis in Serbian is presented in detail in the previous research paper [17].

The data-set. The data-set consists of 1000 EHRs from 31 clinical centers in the City of Nis, Serbia. These medical reports were collected from the MEDIS.NET information system in the period 2012-2018. In one part of this period, there was a measles epidemic in the City of Nis, so one data-set contains EHRs for this diagnosis. The example of one EHR is given in Table 3.

Table 1. The example of EHR from MEDIS informaton system

Date of the service	02-03-18
Name of the service	First examination of adults
Anamnesis	povisena t 38.5, hiperemija grla (en. febrile, throat hyperemia)
Diagnosis	Measles
Diagnosis' code	B05
Organizational unit	General medicine
Location of the service	Central building

Natural Language Processing (NLP) techniques are required to extract information in the free text of the EHRs. The steps necessary for the extraction of information are:

- Reduction to the one alphabet, abbreviation processing, and tokenization. As these EHRs are written in Serbian, the free text can be found in Latin and Cyrillic. To transform the data into a standard format, in this step the text is translated into Latin, with special regard to letters with diacritical marks. After that, the abbreviations are marked, and then the sentence is divided into tokens.

- Deleting stop words - this step eliminates words that do not carry meaning.

- Determining negation - only a few negation symbols are used in medical reports, so the negation mark is attached to the close word, to indicate the possible absence of symptoms;

- Reduction on the basis - since the Serbian language has a rich grammar, words can be found in various forms, it is necessary to reduce them on the basis. In the absence

of a morphological dictionary, as well as for faster results, the base may be a prefix of length n or stem [18];

- Classification (labeling) - After preprocessing the text, classification can be performed. Classification can be done using machine learning methods, taggers, but also rule-based methods, if diagnoses are marked in the data model.

To evaluate the classification results in the diagnosis in the experimental part, the sensitivity for binary classification was used. Sensitivity is expressed as the ratio of the number of well-marked diagnoses (TP - true positives) and the number of all diagnoses that should have been labeled (TP + FN - true positives and false negatives), or in the manner shown by the equation 1.

$$sensitivity = \frac{TP}{TP+FN} \tag{1}$$

4 Experiment Results

In the experimental part, the method of content analysis on two sets of EHRs was first applied. The first set consists of 1000 reports with a diagnosis of measles, and the second of 1000 reports with a diagnosis of five different diseases. In Table 4 and Chart 1, it is shown in which form are diagnoses in both sets. Most often, diagnoses are written in Serbian, followed by the presence of diagnoses in Latin, and the least common in the free text are diagnosis codes. Also, for both sets, more diagnoses are written with a typo than in an abbreviated form. Also, in both sets, typos and abbreviations collectively make up over 9 percent of all diagnoses.

Table 2. Types of diagnoses in the data-sets

Types of diagnoses in the dataset	EHRs (various diseases)	EHRs (morbilli)
Diagnoses (Serbian)	97.520%	65.960%
Diagnoses (Latin)	1.980%	17.020%
Diagnoses Code	0.500%	17.020%
Abbreviations	2.480%	1.420%
Typos	12.380%	8.510%

In Table 4 and Chart 2 shows the results of labeling diagnoses in both sets using medical resources. In the basic method, the word was labeled in its entirety, and in the other two methods, the word was reduced to a prefix of length of 4 letters or stem [18]. For both sets, a high degree of classification sensitivity can be achieved when the word is reduced to a base before labeling. In both sets, the reduction to the prefix of length of 4 letters (4-size prefix) showed better results, while the stemmer achieved approximate results on the set of EHRs with the same diagnosis, while in the diverse set its efficiency was lower.

Table 3. Sensitivity of labeling using different word forms

Sensitivity of labeling	EHRs (various diseases)	EHRs (morbilli)
The basic method	31.680%	58.870%
Prefix (length of 4)	95.540%	92.200%
Stem	52.480%	90.070%

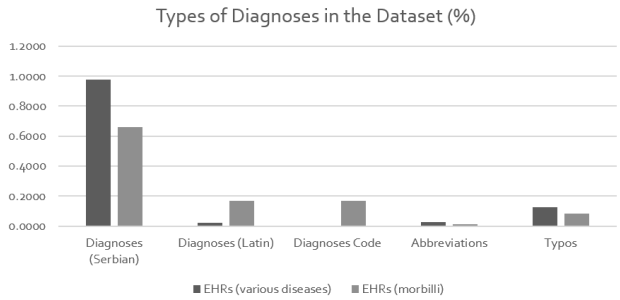

Fig. 1. Types of diagnoses in EHRs

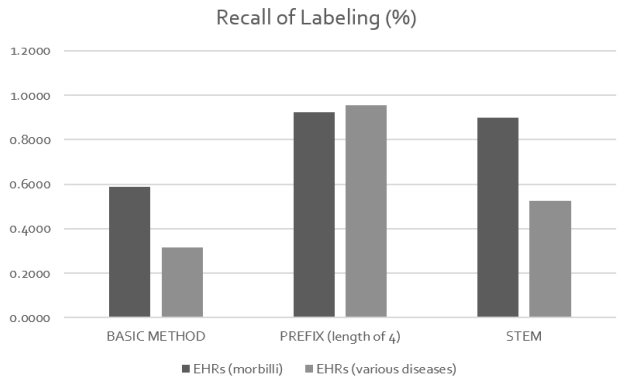

Fig. 2. Sensitivity of labeling using different methods

5 Conclusion

In addition to the structural part, medical reports include a free-form text that contains important information about the patient's health. Therefore, it is extremely important to provide the analysis for this type of data.

Serbian is very complex in grammar and therefore very challenging to analyze, so this is probably why there are neither papers related to this topic nor publicly available electronic medical dictionaries of classified medical terms in the Serbian language.

The results of the implementation of the proposed method show that better results in labeling of diagnoses are obtained when labeling is done using a 4-size prefix of word and

stemmer as word basis.

Acknowledgment

This work was partially funded by the Ministry of Education and Science of the Republic of Serbia after the project III-44 007.

References

[1] R. ROSALES, *Method for Automatic Labeling of Unstructured Data Fragments From Electronic Medical Records.*, U.S. Patent Application, No. 12/469,745, 2009.

[2] M. BUCKLEY, B. COOPEY, J. SHARKO, F. POLUBRIAGNOF, B. DROHAN, K. BELLI, ..., and S. SPECHT, *The feasibility of using natural language processing to extract clinical information from breast pathology reports*, Journal of pathology informatics. Vol. 3, 1 (2012), 23.

[3] W. CHAPMAN, M. NADKARNI, L. HIRSCHMAN, W. D'AVOLIO, K. SAVOVA and O. UZUNER, *Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions*, J Am Med Inform Assoc. Vol. 18, 5 (2011), 540-3.

[4] A. MILENKOVIĆ, P. RAJKOVIĆ, T. STANKOVIĆ and D. JANKOVIĆ, *Application of medical information system MEDIS. NET in professional learning*, In 2011 19th Telecommunications Forum (TELFOR) Proceedings of Papers IEEE, Belgrade, 2011, pp. 1474–1477.

[5] W. SUN, A. RUMSHISKY and O. UZUNER, *Evaluating Temporal Relations in Clinical Text: 2012 i2b2 Challenge*, Journal of American Medical Informatics Association. Vol. 20 (2013), 806-813.

[6] M. MSAEED, M. VILLARROEL, A. REISNER, G. CLIFFORD, L. W. LEHMAN, ..., and R. MARK, *Multiparameter Intelligent Monitoring in Intensive Care II (MIMICII): A public-access intensive care unit database*, Published in final edited form as: Crit Care Med. Vol. 39, 5 (2011), 952–960.

[7] V. VINCZE, G. SZARVAS, R. FARKAS, G. MÓRA and J. CSIRIK, *The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics*, BMC Bioinformatics. Vol. 9, 11 (2008), S9.

[8] H. DALIANIS, A. HENRIKSSON, M. KVIST, S. VELUPILLAI and R. WEEGAR, *HEALTH BANK-A Workbench for Data Science Applications in Healthcare*, In CAiSE Industry Track. (2015), 1–18.

[9] S. ANTONIC and C. KRSTEV, *Serbian Wordnet for biomedical sciences*, In INFORUM, 2008, pp. 28–30.

[10] K. SAVOVA, J. MASANZ, V. OGREN, J. ZHENG, S. SOHN and C. KIPPER-SCHULER, *Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications*, Journal of the American Medical Informatics Association. Vol. 17, 5 (2010), 507-513.

[11] V. GARLA, L. RE III, Z. DOREY-STEIN, F. KIDWAI, M. SCOTCH, ... and C. BRANDT, *The Yale cTAKES extensions for document classification: architecture and application*, Journal of the American Medical Informatics Association. Vol. 18, 5 (2011), 614-620.

[12] E. SOYSAL, J. WANG, M. JIANG, Y. WU, S. PAKHOMOV, H. LIU and H. XU, *CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines*, Journal of the American Medical Informatics Association. Vol. 25, 3 (2011), 331-336.

[13] A. MULYAR, D. MAHENDRAN, L. MAFFEY, A. OLEX, G. MATTEO, N. DILL and B. MCINNES, *TAC SRIE 2018: Extracting Systematic Review Information with MedaCy*, Strain 372 (2018), 338.

[14] B. SRINIVASA-DESIKAN, *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*, Packt Publishing Ltd, 2018.

[15] W. SUN, Z. CAI, Y. LI, F. LIU, S. FANG and G. WANG, *Data processing and text mining technologies on electronic medical records: a review*, Journal of healthcare engineering. Vol. 2018 (2018), 1-10.

[16] INTERNATIONAL STATISTICAL CLASSIFICATION OF DISEASES AND RELATED HEALTH PROBLEMS, *https://www.icd10data.com*.

[17] U. MAROVAC, A. AVDIĆ, D. JANKOVIĆ and S. MAROVAC, *Creating Resources for Marking Diagnoses in Electronic Health Reports in Serbian*, International Journal of Electrical Engineering and Computing. Vol. 4, 1 (2020), 18-23.

[18] N. MILOŠEVIĆ, *Stemmer for the Serbian language*, arXiv 1209.4471, 2012.