Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2020) 28: 3285 – 3303
© TÜBİTAK
doi:10.3906/elk-2002-9

Automated labeling of terms in medical reports in Serbian

Aldina AVDIĆ[1,*], **Ulfeta MAROVAC**[1], **Dragan JANKOVIĆ**[2]
[1]Department of Technical Sciences, State University of Novi Pazar, Novi Pazar, Serbia
[2]Department of Computer Science, Faculty of Electronic Engineering, University of Nis, Nis, Serbia

Received: 02.02.2020 • **Accepted/Published Online:** 28.07.2020 • **Final Version:** 30.11.2020

Abstract: Nowadays, many electronic health reports (EHRs) are stored daily. They consist of the structured part and of an unstructured section written in natural language. Due to the limited time for medical examination, EHRs are short reports which often contain errors and abbreviations. Therefore it is a challenge to process an EHR and extract knowledge from this part of the text for different purposes. This paper compares the results of three proposed methods for automatic labeling of medical terms in unstructured parts of EHRs. All words are categorized as words within the medical domain (symptoms, diagnoses, therapies, anatomy, specialties etc.) and those beyond the medical domain (numbers, places, stop words etc.). The first method is based on dictionaries of medical terms, the second on the training set, and the third on the training set and rules. The results of application of different methodologies to reduce a word to its basic form (pure, prefix, stem) are given for each of the methods. The paper shows that in labeling medical terms, the methods based on medical dictionaries (diagnosis, symptoms, medications etc.) do not produce best results, therefore it is better to use manually annotated part of the data set as a model. A significant number of words (17.36%) in medical reports are abbreviations and errors, so for better results, we should focus on creating rules to solve this problem. Better results are obtained for supervised methods compared to the dictionary-based method (with relative improvement of 42.82%). The inclusion of the algorithm for processing errors and abbreviations increased the results (with a relative improvement of 4.21%) and gave the largest F1 measure (0.9082). The advantage of the proposed method is that the use of rules for processing errors and abbreviations provides good results regardless of how the word is reduced to its basic form.

Key words: Automatic annotation, normalization, electronic health record, natural language processing, medical terms

1. Introduction

The electronic medical report (EHR) [1] is a patient report that stores data on health status, diagnoses and therapies. EHRs are created by a software found in healthcare institutions. EHRs contain a patient's private information (name, surname, personal ID number, date of birth, card number, insurance number, address, etc.), the physician's notes, diagnoses, laboratory reports, therapies, etc. The data included in the EHR are structured (name, surname, age, diagnosis, etc.) and unstructured (description of a patient's condition that cannot be expressed through the structured fields offered, like symptoms, history etc.). In addition to records, all the information is used for various purposes aimed at scientific research, review of the situation or improvement of the processes in health care institutions. Very often these analyses include different methods and tools such as statistical methods, business intelligence, artificial intelligence methods and others. It is known that it is easier

*Correspondence: aplaskovic@np.ac.rs

This work is licensed under a Creative Commons Attribution 4.0 International License.

to search for and extract knowledge from structured data in comparison to the unstructured. For example, if there is a code for every symptom a user may have, it will be easier to search for symptoms by code. However, if they are written in natural language, then synonyms and abbreviations are possible, including different presentations of the same symptom. The unstructured part also carries significant information, therefore it is extremely important to enable their analysis. As a rule, considerable effort is required to reach them, and manual word marking must be usually done.

The simplest and the most demanding way to solve this problem in terms of time spent (although practically useless in the case of massive data) is to have experts who read and assign meaning to any medical document [2]. It is impossible to use this approach in real time, given the amount of documents being uploaded to the EHR database.

To solve this problem, it is necessary to use techniques of medical text mining aimed at extracting knowledge from the text. For extracting knowledge from an unstructured piece of data, natural language processing (NLP) methods are needed. NLP software is designed to convert free text into machine-readable, structured data and it contains methods which infer the sentiment of the text and the latent characteristics of the natural language [3]. The aim of this research is to develop methods for labeling terms within medical reports with the best accuracy. Using these methods, all words contained in the section of medical reports are categorized as terms which belong to the medical domain (symptoms, symptom descriptions, diagnoses, biochemical analyses, Latin words, anatomic names of organs, therapies, and other medical terms) and those out of the medical domain (numbers, negation symbols, stop words and other words without significance for the patient's health condition). The labeling of symptoms, diagnoses, and other medical terms within the unstructured part of the EHR can contribute to the extraction of nonspecific symptoms, which can accelerate diagnostics in times of epidemics. There are no adequate publicly available lexical resources or medical corpora for the Serbian language which can be used for labeling medical terms. Therefore, a set of dictionaries with medical (diagnoses, medications, Latin terms, etc.) and nonmedical terms (stop words and some proper nouns) were prepared for this research. Manual labeling of medical terms in unstructured section of 4112 medical reports was applied. Three methods were proposed: a method based on thus obtained dictionaries of medical terms, and two supervised methods, one of which being hybrid and applying rules to process abbreviations and errors. The paper presents the results for all methods together with their variations when stemmer and cutting off to n-size prefix are applied to reduce the word to its basic form.

Due to the specificity of the language used in writing medical reports (use of abbreviations, presence of errors, Latin expressions, etc.), this paper shows that the dictionary-based method does not give the best results. The results are better if automatic annotation is done based on manually labeled corpus and the rules for processing errors and abbreviations. It has been shown that the hybrid method based on training set and rules produces the best results of all the proposed methods without reducing the word to its basic form.

The proposed supervised method uses rules to correct deficiencies in the anamnesis of EHRs (errors and abbreviations) and allows the labeling of medical and nonmedical terms in the EHR, so that they can be more easily automated and analyzed. This is the main contribution of the paper. Due to the importance of accuracy in medical decision making, pure machine learning methods are not applicable in practice, but greater importance is attached to rule-based methods. Sometimes incorrect terms can be legalized with the use of machine learning, which can later lead to incorrect conclusions. The side contribution of the paper is achieved in preparation of data for system of quick diagnostics based on symptoms, which in times of epidemics can speed up diagnostics and/or can be used as a dashboard in service of epidemic control intended for citizens of smart cities. Actual

COVID-19 epidemic is a good example of the importance of creating tools for rapid diagnostics. This system could be implemented as a public service to help citizens describe the symptoms themselves, see how likely they are to be infected and whether they should go to healthcare facilities or stay home and seek advice online only. The paper is organized as follows: The second section gives an overview of papers dealing with similar problems. The third section illustrates the description of the resources that were needed to implement the methods. This is followed by a description of the methods and experiment together with the results of the experiment. In the penultimate section, the results are discussed and the observations and encountered problems are presented. Finally, the conclusion and directions for further work are given.

2. Related work

The related research on this topic describes methods for normalizing and extracting knowledge from medical records not directly related to the application in Serbian. Most of the research refers to English-speaking corpora and lexical resources which are publicly available. One of the best-known clinical corpora for the English language is the informatics for integrating biology and the bedside (i2b2) [4]. There are also multiparameter intelligent monitoring in intensive care (MIMIC II) [5] as well as a corpus of biomedical texts annotated for uncertainty, negation and their scopes the BioScope [6]. The corpus for medical domain named Health Bank — Swedish Health Record Research Bank [7] is available in Swedish. For the Bulgarian language, there are results of extracting information from a large corpus of unstructured data and their obtaining in the structured form [8].

In [9], the differences between medical and standard texts are presented and the problems that may arise when extracting information from medical texts are pointed out. The characteristics of clinical reports from the corpus of medical reports in Sweden taken in 2014–2015, are presented in [10]. Normalization is the first step used in the classification and labeling of the medical terms. The methods for normalization of electronic medical data are shown in the review papers [11]. Many systems that have been built to process free texts in medical domain use NLP methods for their further application in the health care system. Medical report processing consists of several steps such as data refinement, integration, transformation, reduction and ultimately, data protection. The main purpose is to translate semistructured and unstructured medical reports into computer-readable information using NLP methods. The key methods in this process are named-entity recognition (NER) and relation extraction (RE) [11]. By clinical relation extraction from medical reports, the relationships between drug references and their attributes can be identified [12]. A review of clinical information delivery systems [7] shows that 60% of commercial systems use rule-based methods, while scientific research rather suggests machine learning methods. According to [13], better results in extracting text in medical reports are obtained by using rule-based methods. The most popular systems for labeling medical text are CTAKES and CLAMP systems. The complete architecture of the CTAKES system is described in [14, 15], and a similar NLP-based CLAMP system in the paper [16]. The identification of medical terms in patient-written texts using crowdsourcing was addressed by the authors in the paper [17]. In [18], one approach to correct errors in the free text of medical reports is demonstrated.

Serbian has a complex grammar and is close to other Slavic languages, so the similar approach to text processing can be applied to other languages belonging to this group. The automatic labeling of diagnoses in an unstructured medical text with appropriate lexical resources for medical terms in Bulgarian was made in [19]. In the languages of former Yugoslavia, no papers were found dealing with the free text process in medical reports and publicly available lexical resources in the medical domain. Wordnet for biomedical sciences [20] for the

Serbian language contains sets of synonymous words or more exactly different parts of speech (PoS) with a new concept for six ontological categories (genetics, virus, bacterium, cell, science fields and microorganism). There are no publicly available lexical resources of classified medical terms (diagnoses, medications, symptoms etc.). Regardless of the medical domain, the paper [21] describes a method for labeling semantic roles in Slovenian and Croatian. A complex grammar and two alphabets (Latin and Cyrillic) make the normalization of documents in the Serbian language more complicated. An algorithm for normalization of medical documents in Serbian is presented in [22]. Of the papers dealing with this problem, there are no papers directly related to the medical field. The recognition of named entities in the Serbian language was dealt with by authors in the paper [23].

A bottom-up approach of natural language processing based on taggers and machine learning methods applied to texts in Serbian is shown in papers [24, 25]. Taggers can be used to classify terms into groups with different tags. These papers describe publicly available tools, the Stanford PoS Tagger [1] [26] and TreeTagger [2] [27]. Therefore, we used them for comparison with proposed methods, as shown in the section with experiments and results.

3. Lexical resources

The access to medical information is limited and it is difficult to find it in a format suitable for use in automatic medical text processing. Specialized medical resources containing medical terms and expressions are also lacking. For the purposes of this research, a set of lexical resources in Serbian that are specific to medical text processing have been created. It can be divided in four groups: diagnoses, therapies, symptoms and diagnostics, specialties. Some general lexical resources (personal names and settlement names, stop words and symbol of negation) have also been used to remove words that are not medically relevant. All presented medical resources are obtained by the automatic processing of medical documents that can be found in any medical information system or are publicly available. The significance of this approach is in the language independence of the process for obtaining such resources and the ability to easily adapt to another language.

The goal of creating these medical resources is to label terms in electronic medical reports. The unstructured part of the data contains descriptive information obtained by the physician. Symptoms, diagnostics, diagnosis and prescribed therapy are emphasized as important medical information. The suggested lexical resources can be used to identify the stated medical terms at the word level. By using the same resources it is possible to identify named entities, which is not covered by this research and will be the subject of our interest in the future.

3.1. Nonmedical resources

Natural language text processing requires resources to extract essential words from irrelevant facts. All personal data related to patients and doctors must be extracted from the available data, for which available resources exist [28, 29]. Often, medical reports include names and location of health centers and cities; therefore a resource with the names of settlements in Serbia was created. The assumption is that nouns, adjectives and verbs carry the contents of the text, and that other types of words are less important for informative value of the text. Such words are called stop words. This set contains: words with high frequency in most documents, conjunctions, exclamations, suggestions, etc. This set was collected for the purpose of normalization of documents in the Serbian language [30].

[1]The Stanford NLP Group (2020). Stanford Log-linear Part-Of-Speech Tagger [online]. Website https://nlp.stanford.edu/software/tagger.shtml [accessed 16 August 2020].

[2]University of Stuttgart Institute for Computational Linguistics (2020). TreeTagger [online]. Website https://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/ [accessed 16 August 2020].

Labeling the range of negation is of importance in medical records. For the purpose of denoting the negation symbol, a set of negation symbols together with a set of rules proposed in a given paper was used [31]. The purpose of the paper was not to denote the scope of the negation; but by using tokenization to denote the symbol of negation, tokenization is adapted to the rules of the negation. Thus, the range of action of the corresponding symbol of negation is also indicated. It is important to note that in medical reports, only one or two forms of negation that are easily processed are most commonly reported.

3.2. The processing of medical lexical resources

There are no publicly available specialized medical lexical resources for the Serbian language. Therefore, the available documents in different formats were used to process medical domain-specific terms. The textual data in its original form must undergo preprocessing in order to be used.

3.2.1. Normalization

The process of preparing medical resources includes removal of irrelevant information. Another specificity of the Serbian language is the use of two alphabets (Latin and Cyrillic). Therefore, it was necessary to transfer all the data to a single alphabet to allow for further uniform processing. Serbian alphabet is different from English since there are letters with diacritical symbols (ć, č, đ, ž, š and dž). Serbian script does not use the letters x, y, w and q.

The process of normalization consists of the following steps: (1) removal of all redundant parts from the text and preparation for tokenization (division of text into words): (1.1) conversion from Cyrillic to Latin alphabet, (1.2) changing letters with a diacritical symbol to a combination of letters (ć, č, ž, š, đ)⇒ (cx, cw, zx, sx, dx); this mapping is correct because the letters x, y, w do not belong to Serbian alphabet, (1.3) marking the separators of sentences (’.’, ’,’, ’;’ , ’-’, ’!’, ’?’, ’(‘, ’)’, ‘‘, ’’), (1.4) removal of unnecessary spaces, all other symbols except letters, numbers and separators of the sentence; (2) division of original text into sentence units; (3) division of the sentence units into words while maintaining the connection with the sentence unit; (4) labeling the stop words; (5) labeling the symbols of negation; and (6) labeling the numeric values.

3.3. Medical resources

A lexical resource specific to a certain medical domain is a set of distinct terms extracted by processing appropriate medical documents, as described in the previous section.

Diagnosis

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 medical and administrative purposes. It is publicly available and translated into several languages [3]. It contains disease codes, a description of the disease, symptoms and signs, social circumstances and external causes of the disease, and more. The initial classification contains about 14000 diagnoses. Extended versions and national editions of this classification contain multiple diagnoses.

The initial set is the set of diagnoses given in the [32], which contains diagnoses translated into Serbian and includes 14194 diagnoses. By preprocessing the names of the diagnoses, the set of 72652 words was extracted, after which it was reduced to 7942 different words. It should be emphasized that the name of the diagnosis also

[3]ICD10Data.com (2020). International Statistical Classification of Diseases and Related Health Problems [online]. Website https://www.icd10data.com [accessed 16 August 2020].

includes symptoms, anatomical parts, causes of disease and many medical and nonmedical terms. However, this set also contains many medical terms that often occur, so it is difficult to distinguish nonmedical terms which are not significant for the labeling. Additional resources are needed to label these medical terms. In addition to the name of the diagnosis, there is a resource with ICD-10 [4] diagnosis codes that are also used when writing medical reports. A set of 14194 codes was extracted. In medical reports, the use of Latin is common, so a resource with Latin names of diagnoses [5] was created for those with a Latin name (3794 diagnoses). This resource also includes some Latin names for anatomical organs, symptoms, and more. This postprocessing set (tokenization without excluding Serbian nonmedical terms) contains 2844 Latin terms.

Therapies

Therapy or medical treatment is an attempt to repair the health condition after diagnosis. One type of therapy includes prescribing medications and ampoules. Medications are described by the factory name, generic name and anatomical therapeutic chemical (ATC) classification system [6], [7]. Therapy resources are made of 2232 drugs and 1317 ampoules. After processing 3.2.1 these baseline data, 2255 different terms were formed to constitute the therapy resource.

Symptoms and diagnostics

Many of the symptoms of disease are already covered by a set of diagnoses. For the purposes of this research, the documents describing the symptoms of measles (B05) were collected [33, 34] due to the specificity of the data set described in Section 4.1 The initial set consists of 28 separate symptom descriptions that were reduced by processing to 95 different words. In addition to symptoms, biochemical results were used to determine the diagnosis. A resource with different types of biochemical analyses available in the observed information system [35] was processed and 63 terms were obtained from the domain of blood biochemistry.

Specialties

Medical reports often include the reports from previous treatments or from other medical institutions, as well as recommendations to other specialists for further treatments at other specialized departments. To identify this type of word, the names of available specializations as well as health services of various specialties that were taken from the medical information system were processed. The obtained resources contained 41 terms referring to specialization and 41 terms associated with the health service of the respective specialty.

4. Dataset, proposed methods and used metrics

This section describes the dataset that was labeled, as well as a description of the proposed methods and metrics that we used to compare them.

4.1. Dataset

There is neither publicly available corpus of medical reports in the Serbian language, nor a paper dealing with the automatic extraction of information from medical texts in Serbian. Two datasets are considered from 32 medical centers in the City of Nis, Serbia. These medical reports were collected from the MEDIS.NET [35] information system in the period 2012–2018. The primary dataset contains 2212 medical reports written by

[4]ICD10Data.com (2020). International Statistical Classification of Diseases and Related Health Problems [online]. Website https://www.icd10data.com [accessed 16 August 2020].

[5]ICD10Data.com (2020). International Statistical Classification of Diseases and Related Health Problems [online]. Website https://www.icd10data.com [accessed 16 August 2020].

[6]WHO Collaborating Centre for Drug Statistics Methodology (2020). ATC [online]. Website https://www.whocc.no/atc_ddd_index/ [accessed 16 August 2020].

[7]Mediately (2020). ATC klasifikacija [online]. Website https://mediately.co/rs/atcs [accessed 16 August 2020].

169 doctors. In one part of this period there was a measles epidemic in the City of Nis, so this dataset contains EHRs for this diagnosis. The processing of this dataset is important for the analysis of epidemic. The control dataset consists of 2000 medical reports with ten distinct types of diagnoses used for validation of the results of supervised methods. The corpora are created in accordance with ethical standards, with the deidentification of the patient and the medical staff. The corpora are manually labeled by four independent annotators, who are medical experts. Interrater agreement is 93.3% calculated by Felis's measuring nominal scale for agreement among many raters [36]. Each EHR is assigned a diagnosis code, as shown in Table 1. In the primary set, all diagnoses are assigned code B05, while the second set is composed of EHRs that are joined by 10 different diagnoses (B00, B01, B02, H10, H650, H66, J11, J18, N390, S60) evenly distributed in the set (10% for each class). In addition to the diagnoses associated with EHR in the unstructured part of EHR, there are diagnoses of the disease that occur as complications of the corresponding disease, as well as diagnoses of accompanying diseases of the patient. Thus, in the primary set, although the records are associated with the same diagnosis code B05, the terms (related to diagnoses 29, symptoms 116, therapies 158 and specialties 11) appear in different forms (code, name, synonym, abbreviation).

The primary dataset

Table 1 shows an example of a medical report to be labeled. There, a structured part can be identified, including the date of provided service, service name, diagnosis, diagnosis code, organizational unit and location of the provided service. An unstructured part consists of free-text history. It is more difficult to extract relevant data from this unstructured part because it must be transformed into a standardized format, suitable for further processing.

Table 1. An example of the used medical record.

Date of the service: 23-03-18	Name of the service: Reexamination of adults
Anamnesis: Pacijent dobio sinoc osip po koži. – Eng. The patient received skin rash last night.	
Diagnosis: Morbilli–measles	Diagnosis code: B05
The organizational unit of the service: General medicine	Location of the service: Central building

Prior to labeling, it was necessary for the anamnesis to perform the normalization indicated in step (explained in Subsection 3.2.1). In this way, 25425 words were extracted, so the average number of words by history was 11.49. All words contained in the unstructured section of medical reports were categorized as words within the medical domain (symptoms, symptom descriptions, diagnoses, biochemical analyses, Latin words, anatomic names of organs, therapies, and other medical terms) and those beyond the medical domain (numbers, negation symbols, stop words and other words without significance for the patient's health condition). By manual annotation, a small number (2.74%) of words were not classified into any group, either errors or abbreviations.

When marking the corpus, it was noted that abbreviations and errors were quite present in the text (12.9% of abbreviations and 4.46% of errors, making the total of 17.36%). Most common abbreviations were standard (often for laboratory analyzes, e.g., wbc–white blood cells), but there were also a lot of personal abbreviations (e.g., izv (izveštaj, Eng. report), temperature temp, etc.). Different types of abbreviations (e.g., temperature) were used for the same term. Errors that occurred were classified into 11 types to help us correct the errors later. The types of error that occurred were: double letter omitted (T1), replacement of letter positions (T2), additional letters (T3), missing letters (T4), substitution with a similar word (T5), merged words without space

(T6), merged words with random letter instead of space (T7), diacritical symbol omitted (e.g., c instead č) (T8), incorrect letter (T9), use of letters that do not belong to Serbian alphabet (X instead ks) (T10) and multiple errors in one word (T11).

A freely written section of a medical report often includes numbers indicating the values of laboratory analyzes, therapies, and descriptions of some patient specificities, such as details of personal history of previous treatment, family history, employment, etc. There are also many postrecovery reports where sick leave is concluded, and patient remittances made.

4.2. Proposed methods

After labeling, the dataset is split in two groups. Two thirds of the anamneses are used as a model. The last third of the data is used as a test and the task is to label this group using the three proposed methods. The first method (method M0) is based on dictionaries of medical and non-medical terms. The second method (method M1) labels terms using training set, and the third method (method M2) is hybrid, using both training set and rules for labeling terms. Methods M1 and M2 do not require dictionaries of medical terms.

For each method its variations are created. Differences in variations are in the way of reduction of a word to its basic form (a–cutting-off to n-size prefix, b–using a stemmer). The reason why this is done is due to the existence of declensions (change of noun words by case) and conjugations (change of verb forms by tenses, gender and number) in the Serbian language, whereby variable types of words can be found in many forms. In the proposed methods, n-size prefix is array of n first letters of a word. By cutting to n-size prefix, the long word suffix is removed and then searched for in the resources as the beginning of a word. The words shorter than n letters are also required as the beginning of words. Thus 4 is used for value of n, since a great number of prefixes in Serbian consist of three letters. The stemming is a language-dependent method used to identify and replace the suffixes (inflection form) with the appropriate inflection for the basic word form (stem). The stemming process is performed by using stemmer for Serbian [37].

Method M0

In this method, the text is labeled by using the proposed medical and nonmedical resources described in Section 3. The free text from the test data should be automatically labeled. The free text from anamnesis is taken, and tokenization and normalization described in Section 3.2.1 are performed on the anamnesis. This part is common to both M1 and M2 methods and their variations. After tokenization, each word is searched for in dictionaries. As soon as it is found in one of the resources, the search ends and the following words are processed. Method M0a is a variation of this method where the original word is cut off after first four letters and searched as the beginning of word in dictionaries. Method M0b is a variation of the method M0, where the word is replaced with its stem and searched in dictionaries. The procedure for performance of methods M0, M0a and M0b is shown in detail in Figure 1.

Method M1

In the M1 method, after the normalization the word is labeled by using a data model. Labeling of word w with label l is done by considering the rule extracted from the model with the greatest confidence ($Conf$), presented in Formula 1. If a word is labeled as an error in the model, it is not considered, i.e. the words in the test are labeled only when based on the correct words in the model. The M1 method does not consider personal abbreviations, as they do not belong to the domain of correctly spelled words. In the M1 method, the word in its original form is searched; in the M1a method, the 4-size prefix form of word is searched, while in the M1b method, the stem is searched in labeled model. The procedure for executing methods M1, M1a and

M1b is shown in detail in Figure 2.

$$Conf(w \rightarrow l) = \frac{number\_of\_words\_w\_labeled\_with\_l\_in\_data\_model}{number\_of\_words\_w\_in\_data\_model} \quad (1)$$

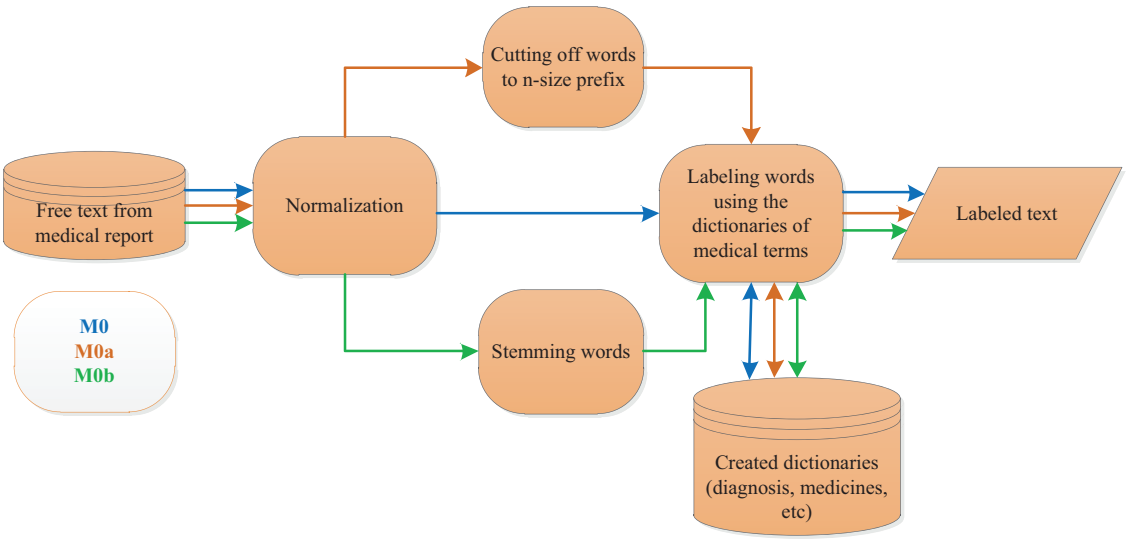

Figure 1. Processing steps in the method M0.

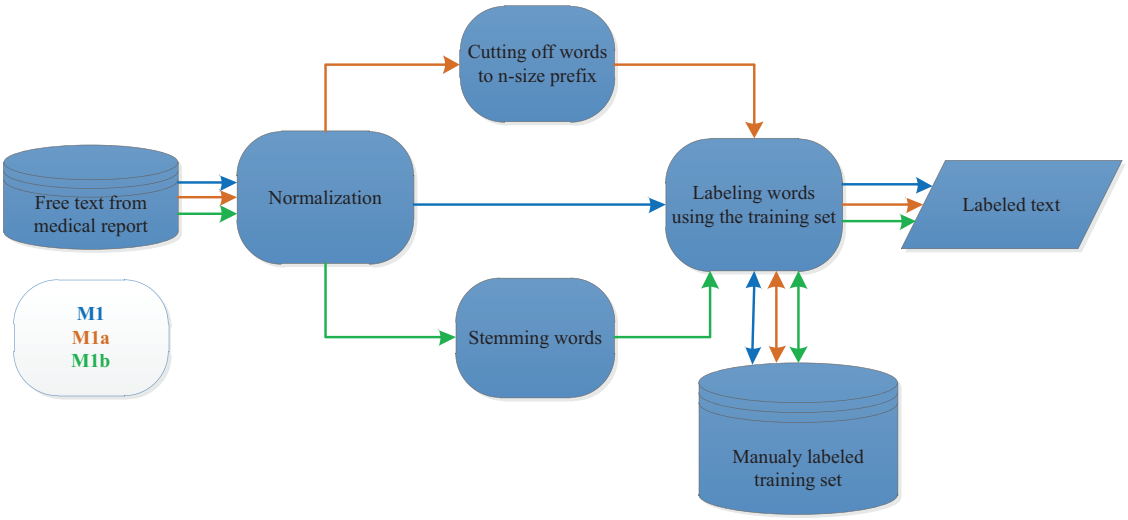

Figure 2. Processing steps in the method M1.

Method M2

In this method, the word is labeled based on the data model, but the rules are additionally made for capturing errors and abbreviations and their processing (Figure 3). This method is an extension of the M1 method. In this method, the words that are spelled correctly are labeled based on the model as in the method M1. However, all words that are unlabeled after that step go through a few more steps to find their label. Therefore, the unlabeled word goes through the following steps:

a. Correction of the T8 error: The most common type of errors is an omitted diacritical symbol. In this step, it is checked whether the unlabeled word contains one letter c, s, z, and if so, it is replaced with its diacritical pair č, č, š, ž, respectively, and the special group of letters dj is changed by the sign đ. If the word remains unlabeled after the change of letters, then the process is repeated until every potential diacritical symbol is replaced. Similar rules were applied here as in the paper [38].

b. Correction of the T10 error: Letters such as x, w, q, and y do not belong to Serbian alphabet. It is observed that there are examples of their incorrect use in Serbian words (the letter x is abbreviated to the group of letters ks, the letter y due to the specificity of the keyboard is written instead of z). In this step, if the searched word is not found as a Latin term, the letters are replaced with a group of corresponding letters or with a single letter from Serbian alphabet, and an attempt is made to label the changed word.

c. Labeling personal abbreviations: Personal abbreviations are processed by searching at the beginning of word from data model, so if such a word is found, the unlabeled word is labeled as the found word.

d. Labeling words with errors T6 and T7: In this step, two merged words are labeled, by searching for the word from data model which is a substring of this two-word compound; if found, the compound is labeled as the found word.

e. Labeling the words with other types of errors: If after all the previous steps the word is still unlabeled, in the last step we try to label the unlabeled word based on the words that sound like the unlabeled word. This is done by modifying the Soundex function adapted to the rules for the similarity of letters in the Serbian language [39]. In this way, some of the errors caused by the replacement of two letters, duplication of letters, absence or addition of one letter, or replacement with a similar word are corrected. As the word can get wrong label in this step, it is important that this step goes at the end of processing, after all other possibilities for labeling are exhausted.

The rules are given in the order of application. If more than one rule can be applied to a single word, the rules are applied as a matter of priority in order to modify the word less, i.e. to reduce the possibility of assigning the wrong label.

The M2 method also has variations M2a and M2b (Figure 3).

Examples of labeled sections of anamnesis using the M2 method are shown in Figure 4.

4.3. Used metrics

We used the metrics to compare the results of M0, M1, and M2 methods, and their variations are Precision, Recall, F1-Score, and Accuracy. They are defined by Formulas (2–5).

$$Precision = \frac{NCLT}{NLT} \tag{2}$$

$$Recall = \frac{NCLT}{NTL} \tag{3}$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

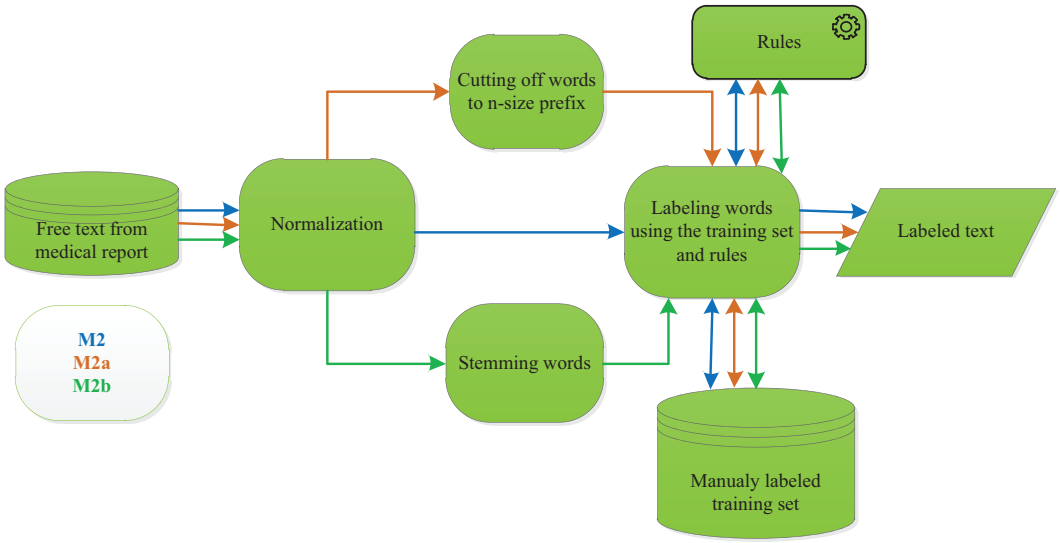

Figure 3. Processing steps in the method M2.

Original anamnesis	Labeled anamnesis
Ždrelo hiperemicno,nad plucioma grubljre disanje (The pharynx is hyperemic, over the lungs the coarser breathing.)	zxdrelo hiperemicxno nad plucxima grubljre disanje
Kaslje od juce,promukla,slina oci joj krmeljaju,temp. nije imala (She was coughing yesterday, hoarse, her saliva was steaming, her temperature was gone.)	kasxlje od jucxe promukla slina ocxi joj krmeljaju temperatura nije imala
Izvestaj dermatologa (A report from a dermatologist)	izvesxtaj dermatologa
Labels: anatomical organ, symptom, speciality, other medical term, nonmedical, without a label, underlined are corrected with M2	

Figure 4. Examples of labeled anamnesis using the method M2.

$$Accuracy = \frac{NCLT}{NT} \tag{5}$$

Here, $NCLT$ is the number of correctly labeled terms, NLT is the number of labeled terms, NTL is the number of terms to be labeled, and NT is the total number of terms. For multiclass classification, all the term categories are taken together using micro metrics (metrics are calculated globally by counting the total true positives, false negatives and false positives).

The results for medical terms only are also presented, using metrics defined by Formulas 6 and 7.

$$Precision = \frac{NCLMT}{NLMT} \tag{6}$$

$$Recall = \frac{NCLMT}{NMTL} \tag{7}$$

Here, $NCLMT$ is the number of correctly labeled medical terms, $NLMT$ is the number of terms labeled as medical and $NMTL$ is the number of medical terms to be labeled. When comparing the methods, we also

used the relative improvement measure given by Formula 8, and here, the F1-Scores of methods are compared.

$$RelativeImprovement = \frac{Compared\_Method - Baseline}{Baseline} * 100 \quad (8)$$

5. Results and discussion

The results of implementation of the methods and their variations, showing the success of the methods in labeling medical and nonmedical terms are presented in Table 2.

Table 2. Results of labeling medical and nonmedical terms.

Method	NT	NTL	NCLT	NLT	Precision	Recall	F1-Score	Accuracy
M0	10969	10883	5119	5895	0.868363	0.470367	0.610204	0.466679
M0a	10969	10883	7717	10320	0.747771	0.709088	0.727916	0.703528
M0b	10969	10883	6965	9013	0.772773	0.639989	0.700141	0.634971
M1	10969	10883	8415	8429	0.998339	0.773224	0.871479	0.767162
M1a	10969	10883	9424	10341	0.911324	0.865938	0.888051	0.859149
M1b	10969	10883	9234	9927	0.93019	0.848479	0.887458	0.841827
M2	10969	10883	9612	10285	0.934565	0.883212	0.908163	0.876288
M2a	10969	10883	9569	10656	0.897992	0.879261	0.888528	0.872368
M2b	10969	10883	9609	10522	0.913229	0.882937	0.897828	0.876014

The method M0 uses resources of medical terms that belong to sets of diagnoses, ampoules, medicines, specializations, etc. The words in these resources are spelled correctly, with diacritical symbols. The results of the experiment show that, based on such resources, a smaller number of words from the data set are labeled in comparison with other methods. The Precision of the method M0 is best when a word is searched without reducing it to its basic form. By the methods M0a and M0b, the word is reduced to a base, thus increasing the number of labeled terms, but the Precision is lower, because the word base is often classified as a similar word without the label. Still, F1-Score and Accuracy are best for M0a (the relative improvement of F1 values in relation to the M0 method is 19.29%) where a word is searched as its 4-size prefix as the beginning of words from the resource.

The reason for this is that by looking for the base of the word, both personal abbreviations and some types of errors 6 and 7 are labeled (since the error that occurs at the end of the word is cut off), although sometimes not correctly. The reason why the M0b method produces poorer results is following: the inflected form is not eliminated but replaced with the appropriate inflection to obtain the base, and the words are less commonly annotated in that basic form. The medical resources used in Method M0 also contain nonmedical terms as described in section 3. The common error type in this method is labeling nonmedical terms as medical. Most nonmedical terms coexist with other medical terms in the diagnosis description. For example, in anamnesis "Dobio boginje od majke" (Eng. Measles caught from the mother), the nonmedical term "majke" (Eng. mother) is labeled as medical because of the diagnosis O75.0 ("iscrpljenost majke tokom porođaja" – Eng. Maternal distress during labor and delivery). The resource of diagnosis contains the largest number of words, which is the reason of wrong labeling. Out of total wrong labeling, this type of mislabeling was 83.12%. The M1 method achieves better results than the M0 method (the relative improvement of the F1-Score value relative to the M0 method is 42.82%). Also, among all methods, the M1 has the best Precision because the word found in the

model in its original form is labeled properly. The Recall is also higher than in M0, which indicates that it is enough to label one part of a report for a disease, while the words within are usually repeated. By reducing the words to the basic form, the M1a and M1b methods produce better results than the M1 method. M1a shows the greatest improvement because the cutting off on n-size prefix solves part of the problem of abbreviations and errors, as explained in this variation in the M0 method. The M2 method labels a correct word in the same way as M1. However, M2 is especially concerned with labeling words that are misspelled or abbreviated. The Precision is slightly lower than in M1; however, the Recall significantly increases as well as F1-Score (relative improvement of F1-Score value compared to M1 method is 4.21%). This means that many shortened and incorrect words were corrected and, in most cases, properly labeled. M2a and M2b do not provide better results than M2, indicating that the proposed error and abbreviation rules partially cover the word reduction to the basis, so this step becomes useless and the labeling process is accelerated. For the M2 method that has the best F1-Score, 3-fold cross validation is also performed. The difference obtained by applying the method on different parts of the data set is not statistically significant ($P = 0.013185 > 0.01$, paired t-test [40]). This demonstrates that the M2 method is linguistically independent, as the marking is done based on labeled data model and rules for processing errors and abbreviations, which can be adjusted according to linguistic specifics. For similar languages, the method can be applied directly or with minor adaptations.

Table 3 shows the results of application of the methods for medical terms only. In this case, M2 provides the best results again, the F1-Score results are close to the results in the previous table. The slightly lower results for Precision and Recall for the M0 method are caused by mislabeling of nonmedical terms as medical, as described above.

Table 3. Results over medical terms.

Method	NMTL	NCLMT	NLMT	Precision	Recall	F1-Score
M0	5692	1920	2565	0.748538	0.337316	0.46506
M0a	5692	3890	5125	0.759024	0.683415	0.719238
M0b	5692	3174	4108	0.772639	0.557625	0.647755
M1	5692	3994	4007	0.996756	0.701687	0.82359
M1a	5692	4746	5075	0.935172	0.833802	0.881583
M1b	5692	4793	5262	0.91087	0.842059	0.875114
M2	5692	4731	4865	0.972456	0.831167	0.896277
M2a	5692	4787	5139	0.931504	0.841005	0.883944
M2b	5692	5013	5528	0.906838	0.88071	0.893583

Using McNemar's test [40], the statistical significance of difference in term labeling by different methods is calculated. Table 4 shows the difference in application of the method with and without reducing the word to its basic form. As can be seen from the table in case of methods M0 and M1, this difference is significant in favor of the method using n-size prefix cutoff. In the case of the M2 method, there is no statistically significant difference in the results obtained by the M2, M2a and M2b methods.

As the best F1-Score and Accuracy were obtained by the M2 method, Table 5 shows the results of statistical significance of the labeling improvement obtained by this method with respect to the M0, M0a, M1, M1a methods. In all four cases, there is statistically significant improvement of the M2 (without reducing it to its basic form method) in comparison with other methods.

The validity of the M1 and M2 methods was additionally verified on a set of medical reports in 2000, describing different diagnoses. Similar results were obtained for the F1-Score as in the primary data set (Table 6). There, the best F1-Score of the M2 was obtained, with the relative improvement of 8.34% compared to the M1. So, it can be concluded that M2 gives the best results independent of the number of diagnoses in the corpus. The results obtained are comparable to those for other languages (e.g., Bulgarian and English)[17, 19]. The results of application of the methods and the tools used to label the terms in Serbian [24, 25] are presented below, as well as the methods of supervised multiclass classification performed by the tool Weka 3.8.4 [8]. These methods were applied to our primary annotated dataset expanded with additional attributes. Table 7 shows the results of a language-dependent method, supervised classification based on POS tags and language-independent methods, for custom tagging and supervised classification based on character n-gram.

Table 4. Statistical significance of the M2 method improvement over other methods.

Baseline method	Compared method	Better method	Chi-square	P
M0	M0a	M0a	1832.094	0.000000
M0	M0b	M0b	1174.503	0.000000
M0a	M0b	M0a	248.9091	0.000000
M1	M1a	M1a	437.9769	0.000000
M1	M1b	M1b	344.4994	0.000000
M1a	M1b	M1a	31.42694	0.000000
M2	M2a	M2	2.441485	0.118170
M2	M2b	M2	0.082234	0.774293
M2a	M2b	M2b	2.410023	0.124649

Table 5. Statistical significance of differences in term designation.

Compared method	Baseline	Better method	Chi-square	P
M2	M0	M2	4010.383	0.000000
M2	M0a	M2	1117.108	0.000000
M2	M1	M2	1020.6	0.000000
M2	M1a	M2	23.44428	0.000001

The method (NBMT + POS tags) uses a Stanford PoS tagging tool that has a model for the Serbian language. The Stanford PoS Tagger is a tool that assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc. The annotated data set was expanded with the PoS attribute and the naïve Bayes multinomial text (NBMT) method was applied to classify the terms by the corresponding labels. The TreeTagger method for custom tagging uses the TreeTagger tool, which has the ability to train models with custom tags, with an annotated lexicon and a list of tags as input. A trained model is then obtained and can be used to label test data. The supervised classification is based on character n-gram using naïve Bayes (NB), logistic regression (LR), support vector machine (SMO) and random forest (RF) classifiers, trained for each

[8]The University of Waikato (2020). Weka [online]. Website https://www.cs.waikato.ac.nz/ml/weka/ [accessed 16 August 2020].

label in the annotated dataset with additional 3-gram (3-gram classifier) and 4-gram (4-gram classifier) features. The results of F1-Score values obtained by using these methods are presented in Table 7, and the best result is obtained with Tree Tagger. Compared to the proposed M1 and M2 methods, all methods give lower F1-Score. Presented performances for all supervised classification methods are obtained using 3-fold cross validation.

Table 8 shows labeling performances of different methods separately for each of the studied term categories. Table 7 presents two types of existing labeling methods, namely tagger-based methods, and supervised machine learning methods. Of the methods based on the taggers, the best F1-Score value is obtained for the TreeTagger method. The best F1-Score value among supervised classifiers is obtained for the RF classifier. Therefore, the results of these two methods are compared with the results obtained using the proposed M2 method (which provides the best results of all proposed methods). Table 8 shows that the M2 method gives the highest F1-Score in most term categories.

Table 6. Results of the supervised methods over the control data set.

Method	M1	M1a	M1b	M2	M2a	M2b
NLT	4478	5422	5258	5528	5671	5570
NCLT	4467	4891	4492	5332	5045	4664
Precision	0.9975	0.9021	0.8543	0.9645	0.8896	0.8373
Recall	0.764	0.8365	0.7683	0.9119	0.8628	0.7977
F1-Score	0.8653	0.868	0.809	0.9375	0.876	0.817

Table 7. Results of application some existing methods for labeling terms.

Applied classification method	Obtained F1-Score
NBMT + PoS tags	0.716255
Tree Tagger	0.80409
NB 3-gram classifier	0.749155
NB 4-gram classifier	0.749936
LR 3-gram classifier	0.773655
SMO 3-gram classifier	0.776511
SMO 4-gram classifier	0.77482
RF 3-gram classifier	0.787349
RF 4-gram classifier	0.785962

6. 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 Serbian language. The main result of the paper is the proposition of three methods for labeling text in medical reports (dictionary based, supervised method and supervised and rule-based method) and their comparison. In addition to the proposed methods, the results of our research include numerous annotated and

Table 8. F1-Scores of M2, TreeTagger and RF method for different term categories.

Term category	M2 method	TreeTagger	Random forest classifier
Symptom	0.963895	0,940355	0.933064
Biochemical analysis	0.950119	0,333333	0.625698
Diagnosis/disease	0.934211	0,918103	0.901099
Drug/therapy	0.791155	0,566327	0.771218
Anatomical organ	0.912621	0,856525	0.937468
Symbol of negation	1	0,99115	0.901763
Number	0.897315	0,892541	0.558538
Time provision	0.965347	0,903226	0.856802
Description of symptom	0.858908	0,857438	0.935256
Nonmedical term of minor significance	0.89172	0,56213	0.886491
Specialty	0.930788	0,904357	0.886364
Institution/place	0.722348	0,443946	0.944591
Term with a negative meaning	0.990476	0,898305	0.978355
Term that emphasizes meaning	1	0,891566	0.991781
Other administrative medical term	0.753986	0,455006	0.832224
Latin word	0.737589	0,268293	0.453258
Stop word	0.975724	0,9977	0.677451

normalized resources created for medical and nonmedical terms in Serbian that will be used in plenty of further research. Also, manually labeled EHR corpora in Serbian are created.

The results of implementation of the proposed methods show that better results are obtained when labeling is done based on the manually labeled dataset. Good results are already gained with use of corpora with about 2000 labeled reports. The results show that errors and abbreviations are very common in medical reports, so rules need to be created to process them. For the most successful recognition of medical terms within this text, the best method is a hybrid method (M2) that uses labeled corpus as model data and rules for correcting errors and abbreviations. The method that gives the best results does not require medical vocabularies or morphological word processing, which makes it less demanding. This proposed method can be adapted with small adjustments for use in other similar languages and those with complex grammar. The results of the methods are also comparable with the methods for labeling terms in other languages. We have shown the results of the application of standard commonly used techniques on the data set and have shown that they give poorer results compared to the proposed methods. For us, the obtained results are a stimulus to continue our work and, above all, to make additional efforts to increase the set of analyzed EHRs to obtain even better results. These good results can encourage other researchers to do similar research for their languages, which are also complex.

The nonstructural part of the report written by different doctors differs in word count and writing style. So, in future work, the possibility of associating a set of personal abbreviations to a doctor will be explored. Also, this is an introduction to creating a medical entity labeling system. Such a system will also be able to label medical reports and other texts containing medical terms. Since errors and abbreviations can be detected

and corrected using M2 method, one of directions for further work is the creation of spelling checkers for medical documents.

Acknowledgment

This paper is partially supported by the Ministry of Education, Science and Technological Development Republic of Serbia, Projects No. III44007 and ON174026.

References

[1] Rosales R. Method for automatic labeling of unstructured data fragments from electronic medical records. U.S. Patent App. 12/469,745, 2009.

[2] Buckley M, Coopey B, Sharko J, Polubriaginof F, Drohan B et al. The feasibility of using natural language processing to extract clinical information from breast pathology reports. Journal of Pathology Informatics 2012; 3: 23.

[3] Chapman W, Nadkarni M, Hirschman L, D'Avolio W, Savova K et al. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. Journal American Medical Information Association 2011; 18 (5): 540-543. doi: 10.1136/amiajnl-2011-000465

[4] Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. Journal of the American Medical Informatics Association 2013; 20 (5): 806-813. doi: 10.1136/amiajnl-2013-001628

[5] Saeed M, Villarroel M, Reisner A, Cli ord G, Lehman LW et al. Multiparameter intelligent monitoring in Intensive Care II (MIMICII): a public-access intensive care unit database. Critical Care Medicine 2011; 39 (5): 952-960. doi: 10.1097/CCM.0b013e31820a92c6

[6] Vincze V, Szarvas G, Farkas R, Móra G, Csirik J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics 2008; 9 (11): S9. doi: 10.1186/1471-2105-9-S11-S9

[7] Dalianis H, Henriksson A, Kvist M, Velupillai S, Weegar R. HEALTH BANK-a workbench for data science applications in healthcare. CAiSE Industry Track 2015; 1: 1-18.

[8] Boytcheva S, Angelova G, Angelov Z, Tcharaktchiev D. Text mining and big data analytics for retrospective analysis of clinical texts from outpatient care. Cybernetics and Information Technologies 2015; 15(4): 58-77.

[9] Meystre M, Savova K, Kipper-Schuler C, Hurdle F. Extracting information from textual documents in the electronic health record: a review of recent research. Yearbook of Medical Informatics 2008; 17 (1): 128-144.

[10] Dalianis H. Characteristics of patient records and Clinical Corpora. In: Dalianis H (editor). Clinical Text Mining. Cham, Switzerland: Springer, 2018, pp. 1-20.

[11] Sun W, Cai Z, Li Y, Liu F, Fang S et al. Data processing and text mining technologies on electronic medical records: a review. Journal of Healthcare Engineering 2018; 2018: 1-10.

[12] Alimova I, Tutubalina E. Multiple features for clinical relation extraction: a machine learning approach. Journal of Biomedical Informatics 2020; 103: 103382.

[13] Gorinski PJ, Wu H, Grover C, Tobin R, Talbot C et al. Named entity recognition for electronic health records: a comparison of rule-based and machine learning approaches. arXiv 2019; arXiv:1903.03985 [cs.CL].

[14] Savova K, Masanz J, Ogren V, Zheng J, Sohn S et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association 2010; 17 (5): 507-513.

[15] Garla V, Re III VL, Dorey-Stein Z, Kidwai F, Scotch M et al. The Yale cTAKES extensions for document classification: architecture and application. Journal of the American Medical Informatics Association 2011; 18 (5): 614-620.

[16] Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S et al. CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines. Journal of the American Medical Informatics Association 2017; 25 (3): 331-336.

[17] MacLean DL, Jeffrey H. Identifying medical terms in patient-authored text: a crowdsourcing-based approach. Journal of the American Medical Informatics Association 2013; 20 (6): 1120-1127.

[18] Lai H, Topaz M, Goss R, Zhou L. Automated misspelling detection and correction in clinical free-text records. Journal of Biomedical Informatics 2015; 55: 188-195.

[19] Boytcheva S. Automatic matching of ICD-10 codes to diagnoses in discharge letters. In: Proceedings of the Second Workshop on Biomedical Natural Language Processing; Hissar, Bulgaria; 2011. pp. 11–18.

[20] Antonic S, Krstev C. Serbian Wordnet for biomedical sciences. In: INFORUM; Prague, Czech Republic; 2008. pp. 28-30.

[21] Gantar P, Štrkalj D, Krek S, Ljubešić N. Towards semantic role labeling in Slovene and Croatian. In: Proceedings of the Conference on Language Technologies Digital Humanities; Ljubljana, Slovenia; 2018. pp. 92-98.

[22] Avdić A, Marovac U, Janković D, Avdić D. Normalization of medical records written in Serbian. In: ICIST 2019 Conference; Kopaonik, Serbia; 2019; 1: 72-75.

[23] Krstev C, Obradović I, Utvić M, Vitas D. A system for named entity recognition based on local grammars. Journal of Logic and Computation 2014; 24 (2): 473-489.

[24] Popović Z. Taggers applied on texts in Serbian. INFOtheca - Journal of Informatics & Librarianship 2010; 11 (2): 1-20.

[25] Šandrih B, Krstev C, Stankovic R. Development and evaluation of three named entity recognition systems for Serbian-the case of personal names. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing RANLP; Varna, Bulgaria; 2019. pp. 1060-1068.

[26] Toutanova K, Klein D, Manning C, Singer Y. Feature-rich Part-of-Speech tagging with a cyclic dependency network. In: Proceedings of HLT-NAACL; Edmonton, Canada; 2003. pp. 252-259.

[27] Schmid H. Improvements in part-of-speech tagging with an application to German. In: Proceedings of the ACL SIGDAT-Workshop; Dublin, Ireland; 1995. pp. 1-20.

[28] Krstev C, Vitas D, Gucul S. Recognition of personal names in Serbian texts. In: International Conference Recent Advances in Natural Language Processing; Borovets, Bulgaria; 2005. pp. 288-292.

[29] Rajković P, Janković D, Vucković D. Using string comparison algorithms for Serbian names. In: Proceedings XLI International Scientific Conference on Information, Communication and Energy Systems and Technologies – ICEST; Sofia, Bulgaria; 2006. pp. 221-224.

[30] Marovac U, Pljaskovic A, Crnisanin A, Kajan E. N-gram analysis of text documents in Serbian language. In: Telecommunications Forum (TELFOR); Belgrade, Serbia; 2012. pp. 1385-1388.

[31] Ljajić A, Marovac U. Improving sentiment analysis for Twitter data by handling negation rules in the Serbian language. Computer Science and Information Systems 2019; 16 (1): 289-311. doi: 10.2298/CSIS180122013L

[32] World Health Organization. Međunarodna statistička klasifikacija bolesti i srodnih zdravstvenih problema – Deseta revizija Knjiga 1 Tabelarna lista. Belgrade, Serbia: Institute of Public Health of Serbia "Dr Milan Jovanović Batut", 2013 (in Serbian).

[33] Đurić-Petković D, Ristanović E, Kuljić-Kapulica N. Virus malih boginja. MD-Medical Data 2017; 9(3): 181-184.

[34] Krstić SS, Miljković MN, Janković IA. Kliničke karakteristike malih boginja kod dece lečene u službi za pedijatriju opšte bolnice leskovac. Apollinem Medicum et Aesculapium 2012; 10 (4): 9-12 (in Serbian with an abstract in English).

[35] Milenković A, Rajković P, Stanković T, Janković D. Application of medical information system MEDIS.NET in professional learning. In: 19th Telecommunications Forum (TELFOR) Proceedings of Papers IEEE; Belgrade, Serbia; 2011. pp. 1474-1477.

[36] Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin 1971; 76 (5): 378.

[37] Milošević N. Stemmer for the Serbian language. arXiv 2012; arXiv:1209.4471 [cs.CL].

[38] Krstev C, Stanković R, Vitas D. Knowledge and rule-based diacritic restoration in Serbian. In: Proceedings of the Third International Conference Computational Linguistics; Sofia, Bulgaria; 2018. pp. 41-51.

[39] Rajkovic P, Jankovic D, Vuckovic D. Adaptation and application of Daitch – Mokotoff SoundEx algorithm on Serbian names. In: Conference PRIM (book of abstracts); Kragujevac, Serbia; 2006. pp. 21.

[40] Rice JA. Mathematical Statistics and Data Analysis. 3rd ed. Pacific Grove, CA, USA: Duxbury Press, 2006.