Original research paper
UDC 614.2:004.738.5(497.11)
DOI 10.7251/IJEEC2001018M

Creating Resources for Marking Diagnoses in Electronic Health Reports in Serbian

Ulfeta Marovac[1], Aldina Avdić[1], Dragan Janković[2], Sead Marovac[3]

[1]Department of Technical Sciences, State University of Novi Pazar, Novi Pazar, Serbia
[2]Faculty of Electronic Engineering, University of Nis, Nis, Serbia
[3]Department of General Surgery, General Hospital of Novi Pazar, Novi Pazar, Serbia

E-mail address: umarovac@np.ac.rs, apljaskovic@np.ac.rs, dragan.jankovic@elfak.ni.ac.rs, smarovac@yahoo.com

Abstract— **Thanks to medical information systems, many medical reports are collected in an electronic form daily. Apart from the fields with allowed values for input (the structural part), one part of these reports consists of the free, non-structural text. It contains a more detailed description of the patient's condition, which could not be described using the structural part. Symptoms, results of laboratory analyses, accompanying diagnoses, etc. can often be found in it. Due to a lack of time, doctors often write these descriptions in non-standard ways, using their abbreviations and synonyms, and they often contain typos. All this makes it difficult to extract information in documents specific to the medical domain. This paper presents the creation of medical lexical resources for the automatic labeling of terms from diagnoses in medical reports. In order to perform the automatic marking of the free text, methods of the computer processing of natural languages are needed, as well as appropriate lexical resources. As there are no publicly available medical lexical resources for the Serbian language, as well as a corpus with medical reports, the contribution of this paper is the construction of such resources for needs of automatic marking of diagnoses. Using the proposed resources, diagnosis codes, Latin and Serbian terms specific to certain ICD-10 can be mapped with precision of 83.47%, 86.86% and 78.29%, respectively.**

Keywords- medical reports; diagnoses; automatic marking; computer text processing; lexical resources

I. INTRODUCTION

The widespread use of information systems has led to the accumulation of a large amount of textual data in an electronic form. With the development of the computer field that deals with the natural language processing (NLP), the possibilities of extracting information not only from structured data but also the free text written in different natural languages have opened up. The use of information systems in medical institutions has integrated the work of different parts of the medical care system. A large amount of data in medical information systems is structured and it carries information about the patient's personal data, information about employees, medications, specializations, establishments, diagnoses, etc. In addition to structured data, an indispensable part of each examination is a textual description of the patient's condition (an anamnesis). As the limitations of the system often do not allow essential facts to be expressed as structured data, this free text carries a lot of information that can be important for various analyses of both the medical condition of an individual patient and the condition of a part of the population. Extracting information from medical reports requires the existence of appropriate special lexical resources that will separate important facts from irrelevant ones.

The processing of medical reports consists of several steps such as the data cleaning, the integration, the transformation, the reduction and finally, the privacy protection [1]. The result is processed data suitable for the knowledge extraction. In order 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 electronic health records (EHRs). To minimize information loss, only stop words are removed, as well as information used to identify patient data.

This paper presents the construction of resources for marking diagnoses written in medical reports as the free text. Also, their evaluation is presented, using standard methods for the information extraction.

This paper is a revised and expanded version of the paper presented at the XIX International Symposium INFOTEH-JAHORINA 2020

This paper is partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia under projects III44007 and ON 174026.

The paper is organized into five sections. The second section presents an overview of research related to the topic of the paper. The third section shows the construction of the dictionary of diagnoses. The testing of created dictionaries over a set of marked anamneses is given in the fourth section. The last section contains the conclusion and directions for further research.

II. Related Work

Electronic health records (EHR) carry a lot of important information such as the patient's condition on admission to the hospital, the process of his recovery, and the state of health at discharge. This information is the easiest to express in natural language, which makes the extracting of the information more difficult [2]. The specific medical terminology is defined by different standards and classification systems. The classification and descriptions of diseases, treatments and drugs control the vocabulary used in medical reports and administration and reduce ambiguity to a degree. The international classification of diseases (ICD - *International Statistical Classification of Diseases and Related Health Problems*) has been used since the 18th century with constant revisions and additions. It is available in several languages and it is under the jurisdiction of the World Health Organization [3]. The tenth revision of this ICD-10 classification has been translated into Serbian and used in medical institutions [4].

Many systems have been built to use natural language analysis (NLP) methods to process medical language that can be found as the free text with the aim of its further application in the health care system. Spyns [5] has written an overview of the system for the application of NLP techniques in medicine in 1996. Even then, there were ideas about creating NLP systems that are multilingual, which is of the special importance in medicine, given the primary use of Latin as a part of medical reports in any other natural language. The main purpose is to convert semi-structured 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) [1]. The name entity recognition method refers to the process of identifying a specific symbol or a type of name in documents. In medical reports, this method is used to identify medical subjects that have specific significance for treatment, such as disease names, symptoms, and drug names. It consists of two steps: finding the entity boundary and determining the entity class. The specificity of written medical reports such as the large number of errors, abbreviations, personal styles and physician labels complicate this process [6].

Methods of extracting information from medical reports that are based on rules and dictionaries require the assistance of relevant experts in the field of medicine for the establishment of the vocabulary and rules. In [7], the dependence of the separation of named entities on the use of the different corpora is shown. Savova et al. presented cTAKES an open solution for processing medical reports by NLP methods [8]. Methods based only on dictionaries, rules and the machine learning can be used to extract medical expressions from the free text [9, 10].

The process of extracting information is preceded by the process of normalization and it is specific to the appropriate language but also the type of documents being processed. The normalization of medical reports in the Serbian language is presented in [11]. The authors deal with the recognition of named entities in the Serbian language [12].

There are no publicly available medical lexical resources or works dealing with the automated extraction of information from medical resources in the Serbian language. The paper [13] describes the challenges of extracting information from medical reports in German with an emphasis on the problem of lack of adequate lexical resources for this specific domain. Here are suggested steps necessary to overcome the difficulties in processing clinical texts. As far as Slavic languages are concerned, there are similar researches for the Bulgarian language [14]. This paper presents software modules that support the automatic extraction of diagnosis. The paper describes relationships between OWL ontologies and the domain model for medical information extraction system on the medical reports written in Polish language [15]. Balabaeva et al. described an approach to develop a spell checker module for clinical text in Russian [16]. This approach combines string distance measure algorithms with technics of machine learning embedding methods.

III. The Dictionary of Diagnoses

In order to label the words that participate in the description of the diagnosis, there must be a dictionary with the terms that denote diagnoses. In medical reports, diagnoses are written both in the mother tongue and Latin; names are often used, as well as internationally accepted names. Abbreviations are often used as they are officially established, but also personally made. The heterogeneous writing of diagnoses in medical reports makes the process of labeling them more difficult.

In order to obtain a resource with terms related to diagnoses, we started from a structured set of data that exists and refers to the ICD-10 classification of diagnoses in the Serbian language [4]. This classification contains disease codes, the name and the description of the disease (symptoms and signs, social circumstances and external causes of the disease, etc.). The initial classification contains about 14,000 diagnoses. Extended versions and national editions of this classification contain some additional diagnoses (Serbian version 14194). An alphanumeric string of maximum length 4 consisting of one letter and up to three numbers is used to encode each diagnosis. 25 letters of the English alphabet were used (letter 'U' was not used, which was left for additional changes). In the free text, these diagnoses can be found in different forms: with a different number of digits, with a dot in front of the last digit, etc. Examples of diagnosis codes are: 'A000','B05.8' and 'B05'.

The diagnosis data consists of (Table I):

- The diagnosis codes,
- The description and the name of the diagnosis in Serbian,
- The description and the name of the diagnosis in Latin.

TABLE I. Examples of diagnoses data

Diagnosis code	Serbian name of diagnosis	Latin name of diagnosis
A00	Kolera	Cholera
A000	Kolera, uzročnik Vibrio cholerae 01, biotip cholera...	Cholera classica
A001	Kolera, uzročnik Vibrio cholerae 01, biotip El Tor	Cholera El Tor
A009	Kolera, neozačena	Cholera, non specificata

In order to successfully mark the terms in the medical reports that are a part of the diagnosis, we must convert the set of diagnoses of the ICD-10 classification into the set of tokens of which the diagnoses are composed.

IV. THE PROCESSING OF THE DICTIONARY OF DIAGNOSES

Text data must go through a preprocessing process before each processing. The problem of processing written data in the Serbian language lies in the complexity of the grammar of the Serbian language, so morphological processing of words is difficult and unreliable. Another problem is the existence of two alphabets that are both in official use and very often both can be found in the same document. Letters with diacritical marks also make automated processing difficult, so the Serbian alphabet will be translated into the English alphabet.

Before any text processing, the following steps are taken to solve the previously mentioned problems and prepare the text for tokenization:

a. the conversion from Cyrillic to Latin alphabet,
b. the replacement of 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,w do not belong to Serbian alphabet.
c. marking the separators of sentences (‘.’, ’,’, ’;’ , ’-’, ’!’, ’?’, ’(‘, ‘)’, ‘{‘, ‘}’),
d. the removal of unnecessary spaces and all other symbols except letters, numbers and separators of the sentence.

The next step is the tokenization. Bearing in mind that the goal of creating this resource is to mark a set of words that describe the diagnosis, in the process of tokenization, information about sentence units is stored.

The tokenization steps are following:

a. division of the original text into sentence units,
b. division of the sentence units into words while maintaining the connection with the sentence unit,

The third step is marking of words which are not significant for extraction of diagnoses:

a. adding a label to stop words,
b. labeling symbols of negation,
c. marking numeric values.

Stop words are words that have no informative value and appear with great frequency in most documents (for example: conjunctions, exclamations, suggestions, etc.). Such a set is made for the needs of normalization of documents in the Serbian language [17].

After this process, each description of the diagnosis is reduced to a set of words that carry significant information about it (stop words and numbers, marked in the third step are excluded). A representation of normalization is shown in Fig. 2.

As the expressions in Serbian and Latin are equally present in the medical reports (Fig. 1), the same procedure of normalization has been performed for the set of Latin names, but without marking the stop word.

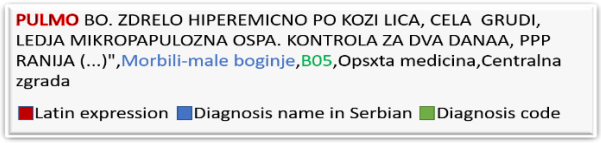

Figure 1. Example of anamnesis with different contents

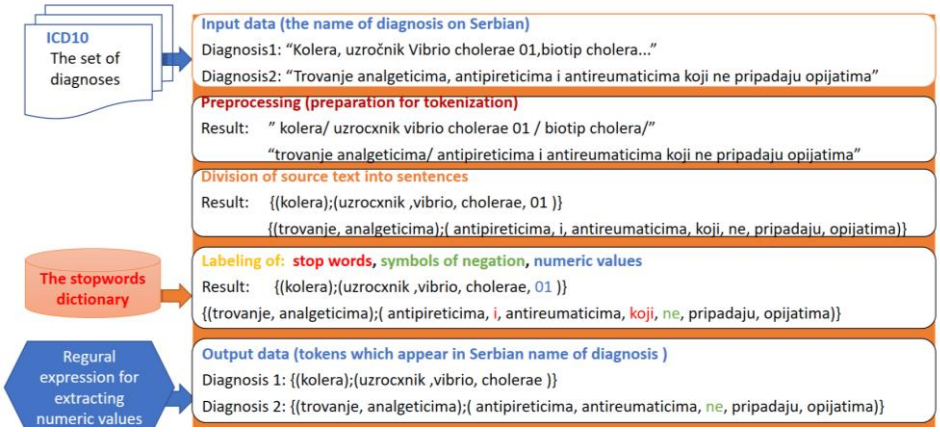

Figure 2. The process of extracting significant words from a set of diagnoses

V. ANALISIS OF DICTIONARY OF DIAGNOSES

Normalization yields a set of words, some of which are more or less important for the identification of appropriate medical terms that indicate the diagnosis. The significance in the occurrence of the word (w) in the diagnosis (d) in relation to its occurrence in the whole set of diagnoses (s) was calculated using the formulas for *tf* (1), *idf* (2) and *tf_idf* (3):

$$Tf(t,w) = \frac{Word\ w\ frequency\ in\ diagnosis\ d}{Number\ of\ words\ in\ diagnosis\ d} \quad (1)$$

$$Idf(\mathrm{w},\mathrm{s}) = \frac{Number\ of\ diagnosis\ in\ set\ s}{Number\ of\ diagnosis\ with\ word\ w} \quad (2)$$

$$TfIdf(t,w,s) = Tf(t,w) * Idf(w,s) \quad (3)$$

The value of *tf* indicates the ratio of the frequency of occurrence of the word *r* in the diagnosis *d* in relation to the total number of words in the diagnosis. The value of *idf* indicates the ratio of the total number of diagnoses and the number of diagnoses in which the word *r* appears. The higher the *tf-idf*, the more significantly the word r appears in diagnosis d compared to other diagnoses. Thus, for example, the name and description in Serbian of the A000 diagnosis is mapped to a set of words with the associated *tf, idf* and *tf-idf* values shown in Table II.

It can be seen from the table that the words "biotip", "cholerae" have the highest *idf*, which means that they rarely appear in the entire corpus. The word "kolera" has a slightly lower *idf*, which is influenced by diagnoses from group A00, which all contain the given word. The word "uzročnik" occurs in many diagnoses, so it is not important for the identification of a given diagnosis.

TABLE II. EXAMPLE OF TF, IDF AND TF-IDF VALUES FOR DIAGNOSIS A000

Word	Number of appearances in d	Tf	Idf	tf-idf
Biotip	1	0.14	7097	1013.86
Cholerae	2	0.29	7097	2027.71
Kolera	1	0.14	3548.5	506.93
Uzrocxnik	1	0.14	84.49	12.07
Vibrio	1	0.14	4731.34	675.90

The initial set consisted of 14,194 diagnoses. By preprocessing the name of the diagnosis, 72652 words were separated, after which the set was reduced to 7942 different words. It should be emphasized that the name of the diagnosis also included symptoms, anatomical parts, causes of the disease and many medical and non-medical terms. By calculating *tf-idf* it can be noticed that there is a large set of words that contain a very low *idf* which means they often appear in terminology. As about a quarter of the extracted terms (2071 terms) have *idf* that differ from the average value by more than one standard deviation. However, there are a lot of medical terms that often appear in this set, so it is difficult to extract non-medical terms that do not constitute key words in the designation. Due to the specificity of the classification of diagnoses by the ICD-10 system, the appropriate factors should be considered at the level of the group of diagnoses, which could improve the identification. *Tf-idf* provides good heuristics for selecting candidate keywords, and its application has proven effective in many papers [18].

A resource with Latin names of diagnoses was also created for those which had a Latin name (3794 diagnoses). This resource also includes some Latin names of anatomical parts, symptoms and more. This post-processing set contains 2844 Latin terms.

In addition to the name of the diagnosis, a resource with ICD-10 diagnosis codes is also extracted, which are used when writing medical reports. This set consists of 14194 diagnosis codes.

VI. RESULTS AND DISCUSSION

The proposed dictionaries of terms from diagnoses are tested on a set of medical reports and on their unstructured part (the anamnesis). This set contains 2212 medical reports from the period (2012-2018) from 32 medical stations in Nis. These medical reports were collected by the MEDIS.NET information system [19], which is used in more than 20 health care institutions in the Republic of Serbia. The corpus was used in accordance with ethical standards, with the de-identification of the patient and the medical staff.

The same normalization process used to normalize dictionary is performed on these anamneses. After normalization, the words that form part of the description of the diagnosis, as well as other medical terms in the anamnesis are manually marked by the annotators in the field of medicine.

Using the previous steps (sections 3-5), three dictionaries were obtained:

- The diagnosis codes dictionary (DCD),
- The dictionary of terms which appear in the description and the name of the diagnosis in Serbian (DTD),
- The dictionary of Latin terms that appear in the description and the name of the diagnosis in Latin (DLTD).

The validation of the obtained dictionaries was performed on a set of annotated anamneses. Terms from the dictionary can be found in the anamnesis with a label that refers to a medical term (MT) such as diagnoses, symptoms, descriptions with symptoms, specialties, etc., or as a non-medical term (NMT) such as time, preposition, verb etc.

A search of terms from the created dictionary of ICD-10 diagnosis codes (DCD) is performed. The mapping is done on 26 different codes with a total of 121 occurrences in the resource. By comparison with manually assigned marks, it is determined that 20 errors have been made by this separation of diagnoses. For example, a doctor's error in recording the temperature t38 (temperature 38) is mapped as a diagnosis T38 (Poisoning by, adverse effect of and underdosing of hormones and their synthetic substitutes and antagonists, not elsewhere classified). The precision of mapping the diagnosis codes in the marked corpus is 83,47%.

By mapping words from the dictionary of terms from the description and name of diagnoses in the Serbian language (DTD), 460 different terms from the dictionary were found, which appear a total of 4750 times in the marked corpus. Of these, the term found 3434 times in the corpus was marked with a diagnosis or some other medical term that stood as a description next to the diagnosis, while 1316 occurrences in the corpus were marked as a non-medical term. The precision of this mapping of terms from diagnoses in the Serbian language in the corpus is 72.29%. Many terms could not be marked based on the dictionary in the corpus since it was in some other morphological form in the text. For example, for measles we have the following forms in the anamnesis ("mobilli", "mor", "morb", "morbila", "morbilama", "morbile" etc). In this form, these terms cannot be found from the dictionary in the resource without being reduced to a morphological form.

A set of words from the Latin names of diagnoses (DLTD) was tested in the same way. The problem with this resource is that there are Serbian terms that have the same form as a Latin term and a completely different meaning ("tel", "sa",…). From this dictionary, 30 different terms were found in the corpus, which appear a total of 236 times (205 times as medical terms and 31 times as non-medical). The precision of mapping Latin words into medical terms in the anamnesis is 86.86%.

From the previously mentioned information, we can notice that the terms from the name of the diagnosis in the Serbian language are marked with the least precision. Additional dictionary transformations (DTDs) were applied to increase precision and map as many words sets as possible.

The first transformation refers to the improvement of the mapping precision, i.e. the reduction of the number of non-medical terms in the DTD. We will achieve this with the *tf-idf* measure associated with each term. In order to avoid mapping terms from non-medical diagnoses that are not specific to it, key terms have been set aside for each individual diagnosis. Key terms are terms that occur more than expected in each diagnosis compared to the rest of the set of diagnoses. Each diagnosis is associated with two key terms as follows:

- The diagnosis goes through a process of normalization
- The *tf-idf* value is calculated for each word of the diagnosis
- Words previously marked as not significant for extraction of diagnoses (stop words, numeric values) and monosyllabic terms are rejected
- Two words with the highest *tf-idf* value are selected in the order of appearance in the diagnosis.

The dictionary of key terms (DKTD) thus obtained that appears in diagnoses contains 28227 words, of which at most two are related to one of the 14194 initial diagnoses. The DKTD dictionary contains a total of 7473 different which is less than the DTD which contains 7942 different terms. By mapping the terms from this resource over a set of anamneses, 437 different terms were found, 1130 of which appear as a non-medical term in the anamnesis and 3257 times as a medical term. This increased the precision of mapping terms from the name of diagnosis in the Serbian language to 74.24%.

Another transformation of a set of terms from the name of the diagnosis refers to their morphological form. Terms can occur in various morphological forms of diagnoses of ICD-10 classifications but also in anamnesis. Reducing different forms of the same word to one will enable better word mapping.

We used two techniques to reduce words based on:
1) the stemming
2) the reduction of the words to its first four characters.

The first method is language dependent and uses a stemmer for the Serbian language [18]. The stemmer maps the different suffixes that words can have in different forms and reduces them to a basic suffix, or they are removed. The shape obtained after stemming is called stem.

The second method is linguistically independent, and it is applied to all words longer than four characters by discarding all characters after the fourth character. The resulting four-letter word will be evaluated below to 4-gram ($n=4$ gives best results).

These two types of word combinations are based on both the anamnesis and the set of terms from the diagnosis. The set of diagnostic terms is reduced to 5232 stems or 2671 4-grams. The results of mapping the corresponding stems, 4-grams in comparison with the mapping of terms and dictionary in the basic form as well as the application of different resources are shown in Table 3.

TABLE III. RESULTS OF MAPPING TERMS FROM DICTIONARIES IN DATASET OF ANAMNESIS

Word form	Dictionary	# NMT	# MT	Precision
Term	DTD	1316	3434	72,29%
Term	DKTD	1139	3257	74,24%
Stem	DKTD	1118	4032	78,29
4-gram	DKTD	2272	7019	75.54%

It can be seen from the table that the best precision (78.29%) is achieved by applying a stemmer. The set of terms from the DKTD dictionary was found 4032 times in the anamnesis as medical, while the number of mappings in non-medical terms was reduced to 1118. As for 4-grams, the dictionary was reduced to the smallest number of words, which sped up the mapping process. A very simple algorithm made the normalization process easier. Also, with the help of this dictionary, the number of terms that have been mapped has almost doubled, and the precision is a little better than the precision obtained by mapping the original terms.

If we compare the obtained precision with the results for the Bulgarian language obtained in [14], where machine learning methods were used with the processing of abbreviations, they differ by up to 10%. It should be emphasized that the aim of the paper is to present and obtain an appropriate lexical resource, and to improve the results of mapping appropriate diagnose. In addition to the dictionaries, processing abbreviations, errors should be included and machine learning methods that would eliminate irrelevant word sets should be applied.

VII. CONCLUSION AND FUTURE WORK

This paper presents the creation of a dictionary with terms that appear in the names and descriptions of diagnoses in order to mark parts of the text in the anamnesis in which a diagnosis is described or stated. From the initial set of diagnoses listed in the international classification ICD-10, three dictionaries of terms were singled out. The first refers to ICD-10 codes of diagnoses, the second is a dictionary of terms that appear in the names of diagnoses in the Serbian language and the third is a set of terms that appear in the Latin names of diagnoses. The created dictionaries were tested on a manually marked set of anamneses. The improvement of the precision as well as the volume of the mapping is achieved by associating the key term diagnoses, as well as the reduction of the term in a unique form. The obtained precision for mapping terms from created dictionaries of codes, Serbian names and Latin names of diagnoses in the corpus is 83,47%, 78,29% i 86,86%, respectively. The obtained precision can be improved by adding special resources for abbreviations, and by marking words in dictionaries that do not carry significant information for the identification of the diagnosis. This will be the subject of our further research.

REFERENCES

[1] Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. Data processing and text mining technologies on electronic medical records: a review; Journal of healthcare engineering 2018, vol. 2018, pp. 1-10. http://downloads.hindawi.com/journals/jhe/2018/4302425.pdf

[2] Safran C, Chute C, Scherrer JR. eds. *Natural Language and Medical Concept Representation*, (Preprints of the IMIA WG6 Conference), Vevey, 1994. Also published as McCray A, Safran C, Chute

[3] International Statistical Classification of Diseases and Related Health Problems; https://www.ICD-10data.com

[4] Međunarodna statistička klasifikacija bolesti i srodnih zdravstvenih problema Deseta revizija Knjiga 1 Tabelarna lista;Institut za javno zdravlje Srbije „Dr Milan Jovanović Batut", World Health Organization, https://rfzo.rs/download/dsg/MKB102010Knjiga1%281%29.pdf

[5] Peter Spyns, Natural Language Processing in Medicine: An Overview Article in Methods of Information in Medicine · January 1997 DOI: 10.1055/s-0038-1634681 · Source: PubMed, https://www.researchgate.net/profile/Peter_Spyns/publication/14190334_Natural_Language_Processing_in_Medicine_An_Overview/links/5752da9908ae10d93371303b/Natural-Language-Processing-in-Medicine-An-Overview.pdf

[6] Dalianis H. Characteristics of Patient Records and Clinical Corpora; In: Clinical Text Mining, Springer, Cham, 2018. https://link.springer.com/content/pdf/10.1007/978-3-319-78503-5_4.pdf

[7] D. Rebholz-Schuhmann, A. Yepes, C. Li et al., "Assessment of NER solutions against the first and second CALBC silver standard corpus," Journal of Biomedical Semantics, vol. 2, article S11, Supplement 5, 2011. https://jbiomedsem.biomedcentral.com/articles/10.1186/2041-1480-2-S5-S11?optIn=false

[8] G. K. Savova, J. J. Masanz, P. V. Ogren et al., "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," Journal of the American Medical Informatics Association, vol. 17, no. 5, pp. 507–513, 2010. https://academic.oup.com/jamia/article/17/5/507/830823

[9] Jiang M[1], Chen Y, Liu M, Rosenbloom ST, Mani S, Denny JC, Xu H. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):601-6. doi: 10.1136/amiajnl-2011-000163. Epub 2011 Apr 20. https://academic.oup.com/jamia/article/18/5/601/834186

[10] Quimbaya, AP, Múnera, AS, Rivera, RAG, et al. Named entity recognition over electronic health records through a combined dictionary-based approach. Procedia Comput Sci 2016; 100: 55–61. https://cyberleninka.org/article/n/866064.pdf

[11] Avdić A, Marovac U, Janković D, Avdić Dž**,** Normalization of Medical Records Written in Serbian , Proceedings of ICIST (2019), 9th International Conference on Information Society and Techology will be held on Kopaonik, Serbia on Mar 10-13, 2019, http://www.eventiotic.com/eventiotic/library/paper/453

[12] Krstev, C., Obradović, I., Utvić, M., & Vitas, D. (2014). A system for named entity recognition based on local grammars. Journal of Logic and Computation, 24(2), 473-489. https://ieeexplore.ieee.org/abstract/document/8200102/

[13] Starlinger, J., Kittner, M., Blankenstein, O., & Leser, U. (2017). How to improve information extraction from German medical records. *It-Information Technology*, *59*(4), 171-179. https://www.degruyter.com/view/journals/itit/59/4/article-p171.xml

[14] S Boytcheva, Automatic matching of ICD-10 codes to diagnoses in discharge letters, Proceedings of the Second Workshop on Biomedical Natural Language Processing September, 2011, Hissar, Bulgaria, Association for Computational Linguistics, pp 11–18. https://www.aclweb.org/anthology/W11-4203/

[15] Mykowiecka A., Marciniak M. (2009) Domain Model for Medical Information Extraction—The LightMedOnt Ontology. In: Marciniak M., Mykowiecka A. (eds) Aspects of Natural Language Processing. Lecture Notes in Computer Science, vol 5070. Springer, Berlin, Heidelberg

[16] Balabaeva, K., Funkner, A., & Kovalchuk, S. (2020). Automated Spelling Correction for Clinical Text Mining in Russian. *arXiv preprint arXiv:2004.04987*.

[17] U. Marovac, A. Pljaskovic, A. Crnisanin and E. Kajan, N-gram analysis of text documents in Serbian language, In Telecommunications Forum (TELFOR), pp. 1385-1388, 2012. https://ieeexplore.ieee.org/document/6419476

[18] Lott, B. (2012). Survey of keyword extraction techniques. *UNM Education*, *50*, 1-11. https://pdfs.semanticscholar.org/f9f6/8c217aef0f3f873eb602a03748ceb5806c88.pdf

[19] Milenković, A. M., Rajković, P. J., Stanković, T. N., & Janković, D. S. (2011, November). Application of medical information system MEDIS. NET in professional learning. In 2011 19thTelecommunications Forum (TELFOR) Proceedings of Papers (pp. 1474-1477). IEEE. https://www.researchgate.net/publication/261212548_Application_of_medical_information_system_MEDISNET_in_professional_learning

Milošević N. Stemmer for the Serbian language; arXiv 1209.4471, 2012. https://arxiv.org/pdf/1209.4471

Dr. Ulfeta Marovac is an assistant professor at the State University of Novi Pazar, Serbia. She holds a Ph.D. in Computer Science from the Faculty of Mathematics, University of Belgrade, Serbia. Her research interests focus on, biological sequences analysis, text mining, natural language processing, logic and its applications in computer science.

Aldina Avdić was born in Prijepolje, Serbia in 1987. She received her Dipl. Ing. degree at the University of Nis, Faculty of Electrical Engineering, 2011. She is a PhD student in computer sciences in the Faculty of Electronic Engineering at the University of Niš and also a teaching assistant at the State University of Novi Pazar, both in Serbia. Her research interests include smart cities, e-health and text mining.

Dr. Dragan S. Janković received a B.Sc., M.Sc. and a Ph.D. degree in computer science from the Faculty of Electronic Engineering, University of Niš, Serbia, in 1991, 1995 and 2001, respectively. Currently, he is a full professor at the Department for computer science, Faculty of Electronic Engineering. His research interests include logic design, design, and development of software, medical information systems and medical informatics.

Sead Marovac, general surgeon, received his medical degree from University of Belgrade Faculty of Medicine and he completed his surgical residency at the Clinical Center of Serbia and Clinic forVascular and endovascular surgery Dedinje. Employed at General hospital in Novi Pazar, Serbia.