Normalization of Medical Records Written in Serbian

Aldina Avdić*, Ulfeta Marovac*, Dragan Janković**, Dženan Avdić*
* State University of Novi Pazar / Department of Technical Sciences, Novi Pazar, Serbia
** University of Niš / Faculty of Electronic Engineering, Niš, Serbia
apljaskovic@np.ac.rs, umarovac@np.ac.rs, dragan.jankovic@elfak.ni.ac.rs, dzavdic@gmail.com

***Abstract*— A bunch of patient data in hospital information systems is kept every day. A part of the data that contains the symptoms, anamnesis, and observations of a doctor can be used for further processing in order to improve public health. This paper describes the algorithm for normalizing medical textual data in order to prepare them for further processing, and within the framework of smart healthcare services. The result of its application is normalized data but also a corpus of stop words specific to the medical domain.**

I. INTRODUCTION

Every day, a large amount of data is stored in systems for storing medical reports. The purpose of health information systems is not only the keeping of standard records (the number of patients examined per doctor, the consumption of materials, the monitoring of issued recipes, etc.), but they contain medical reports with information such as symptoms, anamnesis and diagnosis of patients. The large amount of data collected at the daily basis should be used for analyses and predictions in order to improve the health system. In order for these data to be used for the abovementioned purposes, it is necessary to process these data in an adequate way, and this problem precisely is the subject of this paper. These data can be used to advance medical information systems in the direction of creating smart health services [1], which promotes the smart health as one of the smart-city components [2].

A smart city is a place where traditional services become more flexible and more efficient by using information, digital and telecommunication technologies [2]. The use of digital technologies provides better public services for residents and better use of resources. One of the definitions of the smart city is the following: a city that connects physical infrastructure, information technology infrastructure, social infrastructure, and business infrastructure to foster the collective intelligence of the city.

The smart city infrastructure includes physical, ICT and services. Physical infrastructure is the real physical or structural part of the smart city including facilities, roads, railways and a water supply system. ICT infrastructure is the basic intelligent component of a smart city that keeps together all other components. Service infrastructure is based on physical infrastructure and can have some ICT components [1]. The ICT infrastructure of smart cities is based on two closely related and new technologies, the Internet of Things (IoT) and Big Data [3-4]. IoT consists of: things (a sensor, a device connected to a network with the ability to interact with a user or manage another device), a local area network, the Internet, and a cloud. IoT application is found in smart healthcare, traffic and energy efficiency in smart cities. Big Data consists of large amounts of data collected from sensors, crowdsourcing, etc. and whose storage and processing is carried out in a more complex way than usual.

The most important components of smart cities are the smart transport, the smart healthcare, the energy efficiency, the smart technology and infrastructure, the smart education, the smart management and smart people. The smart healthcare includes e-Health and its aspects to improve the public healthcare services in smart cities [1].

The main idea of the smart healthcare is using technological innovations as much as possible in the health care system. It is questionable whether a part of the data from medical reports could be processed and used for the purposes of the smart health and its services, such as an epidemic control, the visualization of vaccination data, disease prevention, a self-diagnosis, and the like.

The motivation for our work is to improve the quality of life of citizens through the concept of smart cities and smart healthcare services. The medical reports we dealt with in this paper are written in Serbian language. In this paper, the aim was to create an algorithm that would delete the non-relevant data from the input set of medical data, and those which have been relevant were being prepared for further processing. They are prepared by purifying from excessive words and punctuation marks that did not carry any informative value. These redundant words are remembered, in order to take advantage of the normalization of new data. The data we used was about 5000 parts of medical reports collected from 32 of 63 ambulances belonging to the Health Center Nis (DZ Niš), by the use of the MEDIS.NET system.

The work is organized in the following way. The second chapter describes similar researches. The third and fourth chapter provides the description of our approach of normalization medical data written in Serbian. Then the results of application of our algorithm on medical records are given. Finally, conclusions and directions for further research are given.

II. RELATED WORK

The papers related to our research described the possible application of medical textual data, and the ways for them to be processed, without including the ones in the Serbian language [5-10]. Also, in the papers that had described the normalization of documents in the Serbian language, there were no specially processed medical

documents, which had their own domain-specific terms [11-14].

The difference between clinical and ordinary texts and problems with which we can encounter when obtaining information from medical texts is described in [5]. In [6, 7], a description of the methods used in the normalization of electronic medical data is given, but these methods do not include the specificities of the Serbian language, and their application to medical data in our language is not effective. In the paper [8], one way of classifying medical data and the application of neural networks in solving this problem is described. Medical data in this case are texts of patients from the Internet, and their descriptions of the state of the health and symptoms, and the like. Paper [9] describes the significance of Big Data in bioinformatics, i.e. how to get medical data and how to analyze them. A complete system that normalizes and extracts information from medical records is described in [10], and its architecture is presented here. The mentioned methods and systems do not take into account the specificities of the Serbian language. On the other hand, currently available papers that describe normalization of the text written in Serbian are not specifically dealt with and analyzed medical texts.

The papers [11-14] describe the process of normalization of documents in the Serbian language with the aim of faster search. Input data that are normalized in these works are standard documents written in an informal language. Some steps can be used to normalize medical data, but most of them need to be adapted so that their application will improve the availability of usable information from medical data.

III. The Process of Normalization of Medical Data

Medical data preserved in an electronic form contain forms that are incomplete, linguistically uneven, informal and non-standard abbreviations which makes them difficult for computer processing and analysis. Therefore, prior to the analysis, data preprocessing should be performed to bring it into a standardized form. Like all other textual data, these go through some of the standard text pre-processing phases, such as data cleansing, transforming data into a standard form, and reducing data volumes. Medical data is of a mixed type (structured, semi-structured and unstructured) and therefore requires a more complex processing involving the existence of appropriate specialized lexical resources. The ultimate analysis of medical texts usually requires separation of relations, so good normalization raises the possibility of finding them.

A structured data set consists of patient data, such as the name, the surname, the date of birth, the health card number, the address, etc. The semi-structured part of the set consists of data such as the temperature, the pressure, the laboratory analysis. The unstructured part is the free text given by the doctor and consists of symptoms, anamnesis, observations, and conclusions. These texts contain a large number of information that is more difficult to handle because they do not have the same structure. Very often they abound in printing, grammatical errors as well as features of a local communication.

The specificity of the language on which the data is written, in this case of the Serbian language, specifies the steps to be undertaken in normalization. The standard normalization process consists of the following steps of data cleansing, data integration, transformation etc. Data cleansing is to remove data that is not relevant. A part of the data that is not complete can be manually supplemented or automated by the average value by the method of the nearest neighbors or you can simply ignore if the missing value has a major impact on the data analysis.

Semi-structured and structured data may contain values that are not valid and can be automatically excluded from the further testing (for example, non-volatile values of body temperature, weight ...). It is also often necessary to reduce data to the same measurement units and measurements, so that the analysis of semi-structured data is more efficient. The text mining is reduced to analyzing an unstructured text and finding the appropriate relationships that are hidden in the text. Analyzing the medical text of anyone and anyone else is based on: the information gathering, the transformation of information, the separation of relations from the text, the discovery and the knowledge transfer.

In the field of medicine, the recognition of the named entities encounters many obstacles, such as the writing of doctors (writing errors and grammar), various forms of writing, medical terms (such as an epilepsy and an atrophy related to the same disease), and the ambiguity of an abbreviation (for example, a PC, which can be related to prostate cancer, phosphatidylcholine, or a personal computer).

IV. Description of the Applied Normalization Method

In the introductory part we mentioned the number of medical records over which we performed normalization during this research, as well as their origins.

TABLE I.
An example of the used medical record

Date of Birth	12-05-87
Date of the service	23-03-18
Name of the service	Re-examination of adults
Anamnesis	Pacijent dobio sinoc osip po koži. Makulopapulozna ospa po kozi iza ušiju, čela i spušta se na trup. Vezikularni disajni šum (en. The patient received a skin rash last night. Maculopapular rash on the skin behind the ears and forehead and going down to the hull. Vesicular breathing noise)
Diagnosis	Morbili - smallpox
Diagnosis' code	B05
Organizational unit of the service	General medicine
Location of the service	Central building

The Table 1 shows an example of a medical record from our database, and a portion of it is marked as anamnesis, and its normalization is carried out. The normalization method that is performed in this paper on

the above data includes six steps: the tokenization, stopping words, cutting words to *n*-grams, the counting *n*-grams, the *n*-gram classification and the associating *n*-grams with synonyms. This is shown in Fig. 1.

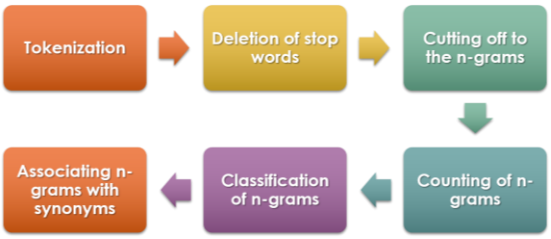

Figure 1. The steps of proposed normalization method

Tokenization – In this step, the processing of diacritic symbols is carried out, and the deletion of punctuation marks, numbers and special characters. In this case, the text in the Serbian language is translated into ASCII code, in order to clear the text so that the output format contains all ASCII codes for Serbian specific characters such as č, č, š, ž and đ. Also, text is purified by numbers and special characters by using regular expressions, after which the text is ready for processing.

Deletion of stop words - This step is done by comparing the stop words and removing them from the text. Stop words are words that are not important in the process of obtaining information from documents. Stop words are usually attachments, suggestions, questionable words, pronouns, and other words that are not relevant for determining the content of the text. The process of removing the stop word begins by creating a dictionary of stop words. The dictionary of stop words in Serbian, which is the result of our previous research [11], contains 3117 stop words. In Table 2, some words from this dictionary are given.

TABLE II.
SOME OF STOP WORDS FROM THE DICTIONARY

Stop words	Stop words
kakva (en. which)	nećete (en. wouldn't)
igde (en. where)	neću (en. won't)
iako (en. although)	će (en. will)

Cutting off to the n-grams – This step in normalization is the definition of the basis of the word. In many languages, words occur in many different forms, while retaining a common meaning. In Serbian, there are ten types of words, five of which are variable: nouns, pronouns, adjectives, numbers and verbs. The basis is the bearer of the meaning of the word, and the complex derivation in the Serbian language (the existence of prefixes, suffixes and infixes) makes it difficult to find. Lemmatization is the process of grouping different variables of the word in such a way that they can be analyzed as a unique form. In the computing, the lemmatization is an algorithmic process for determining the lemma (morphological basis of a word) for a given word. The complex grammar of the Serbian language makes this task very complicated and requires its vast knowledge and possession of appropriate lexical resources like the morphological vocabulary. Stemming is the process of reducing the different forms of words to their common ground. A stem is not necessarily identical to the lemma and it does not have to be the right word. It is sufficient that the common basis for different forms of words. Since our method of normalization independent of morphological resources such as the Dictionary of the Serbian language, instead lemmatization we used to cutting and reducing the *N*-grams of length 4. The words which may create 4-grams are taken into consideration while the rest are being ignored.

N-grams are successive sequences of length *n* characters. *N*-grams are generated by moving the frame length n along the text. Each text can be presented as a *n*-gram vector that appears in it. The text can be compared using its vector representation based on n-grams. *N*-gram analysis calculates the likelihood of occurrence of an *n*-gram, the relative frequency of occurrence of different *n*-grams or other statistical properties of the *n*-gram. By analyzing the content of the *n*-gram, one can notice the correlation between the appearance of an *n*-gram and the characteristics of the text. *N*-grams are suitable for use in the analysis of textual documents in natural languages, due to language independence, in which error tolerances are written.

Counting of n-grams – In this step, in the anamnesis reduced to 4-grams, *n*-grams are counted, their occurrences are sorted in decreasing order, and the ones most often appearing are distinguished.

Classification of n-grams - The biggest problem for the classification of documents is the large volume of texts written in natural language. In order to reduce the size of documents, each of them should be presented using key words (small word and phrase sets that describe the content of a document). Automatic keyword allocation is the way to find words that are most commonly used in a document, so the format does not affect their semantic meaning in which they appear in the document. In order to find the key words in the document, the document first goes through all the levels of normalization. In this step, the most commonly occurring *n*-grams from the previous step are classified in a group with semantically belonging (keywords or stop words).

Associating n-grams with synonyms - In this step, the words from the original anamnesis with key *n*-grams are linked. The reason for this is the formation of keyword phrases and stop words from the medical domain, which can be used for further use. The effectiveness of medical documents of normalization will be increased if it is removed from the anamnesis of not only the common stop words, but these newly obtained from the medical domain.

V. RESULTS AND DISCUSSION

The results show which *n*-grams of length of four letters appear most frequently in the anamnesis and how they are classified by meaning (KW for keywords and SW for stop words). In Table 3, the most commonly-occurring 4-grams are sorted, and they are presented with related words from anamnesis. The table shows how many times the *n*-gram appears in all anamnesis (NAA), and then the number of different anamnesis (NADA) in which it appears, since some words are repeated several times in a single anamnesis. This is because one anamnesis describes the health conditions of the one patient.

TABLE III.
MOST FREQUENT *N*-GRAMS IN ANAMNESIS

4-gram	Associated synonim	Eng. meaning	NAA	NADA	Type
infe	infekcija	infection	552	520	KW
kont	kontrola	appointment	459	447	SW
morb	morbili,	morbili	377	353	KW
pulm	pulmo	pulmo	298	295	KW
ospa	ospa	rash	291	282	KW
izve	izveštaj	report	275	262	SW
ždre	ždrelo	pharynx	261	260	KW
dana	danas	today	253	221	SW
osip	osip	rash	250	231	KW
bolo	bolovi	pains	240	222	KW
hipe	hiperemija	hyperemia	235	223	KW
telu	telo	body	230	217	KW
dozn	doznake	remittances	196	188	SW
temp	temperatura	temperature	177	176	KW
licu	lice	face	154	147	KW
uput	uput	refer	148	134	SW
kašlj	kašalj	cough	126	125	KW
nala	nalaz	finding	121	115	SW
koži	koža	skin	112	110	KW
preg	pregled	examination	112	112	SW

The results show that the most common occurring words relate to symptoms of the disease (rash, temperature, cough, etc.) and to medical terms (appointment, report) that we have declared as stop words in the medical domain because they do not indicate the patient's condition (Fig. 2).

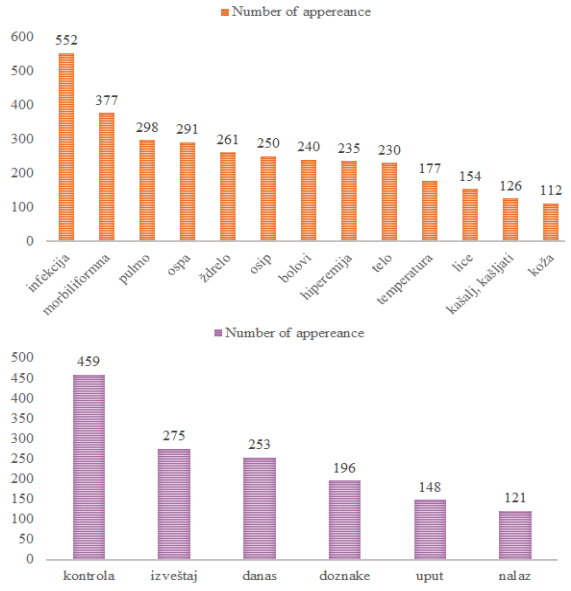

Figure 2. Most frequent medical keywords and stop words

The number of patients which have a certain symptom in their anamnesis can be extracted from these results. Therefore, this normalization can be used in the execution of statistics in the control of epidemics, which is one of the significant services in the smart health in smart cities.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we presented the problems that could arise in the analysis of medical data. A method is described that extracts relevant data from medical data that can be used for different purposes. By using this approach, new medical data used for public health and other smart healthcare services are stored in a purified form. They can be compared to each other and the keywords, symptoms and various statistics can be extracted from them. In addition, the data stored in this way can be used for further processing. The data corpus consisting of the stop words for medical domain in Serbian language is formed, which can be used for other medical documents.

ACKNOWLEDGMENT

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

REFERENCES

[1] A. Solanas, C. Patsakis, M. Conti, I. S. Vlachos, V. Ramos, F. Falcone, and A. Martinez-Balleste, "Smart health: a context-aware health paradigm within smart cities," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 74-81, 2014.

[2] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of things for smart cities," *IEEE Internet of Things journal*, vol.1, no.1, pp. 22-32, 2014.

[3] S. P. Mohanty, U. Choppali, and E. Kougianos, "Everything you wanted to know about smart cities: The internet of things is the backbone," *IEEE Consumer Electronics Magazine*, vol.5, no. 3, pp. 60-70, 2016.

[4] M. Batty, "Big data, smart cities and city planning," *Dialogues in Human Geography*, vol. 3, no.3, pp. 274-279, 2013.

[5] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle, "Extracting information from textual documents in the electronic health record: a review of recent research, "*Yearbook of medical informatics*, vol. 17, no. 1, pp. 128-144, 2008.

[6] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, "Data processing and text mining technologies on electronic medical records: a review," *Journal of healthcare engineering*, vol. 2018, pp. 1-10, 2018.

[7] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J. F., Chang, and L. Hua, "Data mining in healthcare and biomedicine: a survey of the literature," *Journal of medical systems*, vol. 36, no. 4, pp. 2431-2448, 2012.

[8] K. Lee, S. A. Hasan, O. Farri, A. Choudhary, and A. Agrawal, "Medical Concept Normalization for Online User-Generated Texts," *In 2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 462-469, 2017.

[9] J. Luo, M. Wu, D. Gopukumar, and Y. Zhao, "Big data application in biomedical research and health care: a literature review," *Biomedical informatics insights*, vol. 8, pp. BII-S31559, 2016.

[10] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "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.

[11] 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.

[12] A. Pljaković, D. Avdić, U. Marovac, A. Crnišanin, and D. Rančić, "*Pretraživanje dokumenata na srpskom jeziku za potrebe m-Uprave," ETRAN*, pp. RT4.6, 2013.

[13] P. Rajković, D. Janković, and D. Vučković, "Adaptation and Application of Daitch – Mokotoff SoundEx Algorithm on Serbian Names," *Conf. PRIM (book of abstracts)*, pp. 21, Kragujevac 2006.

[14] P. Rajković, D. Janković, and D. Vucković, "Using String Comparison Algorithms for Serbian Names," *Proceedings XLI International scientific conference on Information, communication and energy systems and technologies – ICEST*, pp. 221-224, Sofia, June 29th – July 1st, 2006.