THE USE OF ASR TO MAKE CLINICAL DOCUMENTATION IN SERBIAN

Aldina Avdić[1]*,
Ulfeta Marovac[1],
Dragan Janković[2]

[1]State University of Novi Pazar,
Novi Pazar, Serbia

[2]Faculty of Electronic Engineering,
University of Nis,
Niš, Serbia

Abstract:

Medical information systems are used to manage electronic medical records (EHRs) which store information about patients' health and medical treatments. These records store data daily in a structured, semi-structured, and unstructured form. As unstructured data preserves details about the health of patients written in natural language, artificial intelligence techniques, such as natural language processing (NLP) techniques, can be applied to this part of medical reports. To obtain as useful knowledge as possible from patients' data, before data processing, it is necessary to make adequate preparation. Due to the limited duration of the examination, physicians often make typos when writing clinical documentation. The processing of misspellings that occurred during the writing of the electronic medical records is one of the steps in data preparation. If the ASR (automatic speech recognition) is used when creating a medical report, some common typing errors can be avoided. In this paper, a set of electronic medical data written in Serbian is read through using the ASR, and differences in the distribution of misspellings is analyzed compared to manually entered anamnesis. The high-level architecture of the healthcare knowledge extraction system has been proposed, which would serve to take a patient's data using ASR, and then further processing of the NLP to correct errors and classify the text.

Keywords:

Electronic health records, automatic speech recognition, Serbian, misspellings processing, natural language processing.

INTRODUCTION

Medical Information Systems (MIS) [1] plays an important role in modern health systems. In addition to keeping patient health data and being centralized, easier search and overall management are enabled, and space for analysis of this data is opened. Also, these systems can keep track of medical personnel, the state of necessary resources, reduce administration costs, so these systems have many advantages and have quickly become an indispensable part of health care systems.

Electronic medical reports Electronic Health Record (EHRs) [2] stores data on patients' health and is usually written by doctors. EHRs can have structured, semi-structured, and unstructured parts. The structured part is created by entering text in marked text fields and their

Correspondence:
Aldina Avdić

e-mail:
apljaskovic@np.ac.rs

DOI: 10.15308/Sinteza-2022-373-378

structure is fully known (e.g., examination date, personal number of insurers, diagnosis, etc.). The semi-structured part has a partially known structure, unlike the unstructured part consisting of free text, in which a physician gives additional observations about the health of patients that cannot be expressed through previous data entry fields. Most commonly in this part are notes on the results of laboratory tests, previous or related diseases, symptoms, diagnoses, therapies, or other data of importance. Due to the limited duration of the medical examination, usually, this part of the medical report is susceptible to typos. Errors in medical reports can be fatal, for example, may lead to the wrong therapy if the name of the drug is misspelled. The existence of errors in medical reports makes its analysis more difficult, as the gain of new knowledge too. Therefore, adequate preparation of these records before processing is the main motivation of this research. There are numerous examples of the application of free text analysis from medical reports, but one of the current ones would be to conclude on changing symptoms over time for a suitable disease, given the current situation with the corona virus pandemic and the emergence of various strains that carry different symptoms.

In this paper, natural language processing techniques were applied over a set of Serbian-language electronic medical reports collected by the information system MEDIS.NET [3], to detect and correct errors in the free text of medical reports. Data entered by typing and entered using ASR was analyzed. This research aims to analyze errors that occur in free text in Serbian-language medical reports to form rules for their autocorrect. The main contribution of the paper is the classification of the types of errors and displaying of their distribution in analyzed medical reports for each method of input. Therefore, the pros and cons of both text input methods will be given. The architecture of a knowledge extraction system with EHRs entry using ASR has also been proposed, with the error detection and correction based on specialized dictionaries, natural language processing, a training set consisting of hand-marked reports, machine learning, and rules.

The paper is organized as follows. The second chapter gave an overview of works dealing with a similar theme. The third chapter describes the data and methods used for analysis. The following is a view of the classification and distribution of found errors in the analyzed set. A proposal has also been made to the architecture of the system for entering patient data with ASR possibility and for detecting and correcting errors in medical reports, and further knowledge extraction in Serbian. In the end, the conclusion and direction of further research were given.

2. RELATED WORK

Data mining and text mining [4] differ in terms of the type of data they process. While data mining processes structured data (e.g., databases), text mining deals with unstructured text data (e.g., social media posts) [5] [6]. Both use a wide range of features to convert available data into knowledge. Data mining combines disciplines that include statistics, artificial intelligence, and machine learning over structured data. Text mining requires an additional step in retaining the same goal as data research. Text mining deals with unstructured data, so before any data modeling or pattern recognition feature is applied, unstructured data must be organized and structured in a way that enables their modeling and analysis. This process is usually associated with an artificial intelligence technique called NLP – Natural Language Processing [7] and enables the system to understand the meaning of data in human language. The NLP's goal is to read, decrypt, understand, and find meaningfulness in a natural language. Most NLP techniques rely on machine learning to determine the meaning of data in natural languages.

Authors from various speaking areas dealt with the detection and correction of errors in medical reports. In a review paper for misspelling processing techniques [8] [9], three issues have been identified: detecting non-word errors, correcting isolated word errors, and correcting errors depending on context. Non-word error detection techniques fall into two categories. In n-gram analysis [10], which is mainly used in optical character recognition systems, unusual character sequences are error recognition indicators. In the paper [11] n-gram analysis is used to correct errors in the medical domain in Persian. Often, error correction systems use dictionaries: any word that is not in the dictionary is probably misspelled. Most systems for isolated word error correction use some form of minimal distance to edit or rank suggestions. In the paper [12] states that over 80% of spelling errors consist of one of the following operations: an inserted letter, a deleted letter, a letter replaced by another letter, or two transposed or replaced letters. The DL distance represents the number of operations it takes to transform one word into another and in the paper [13] is used for the Russian language.

Correcting errors depending on the context is used when a spelled word is replaced with another. These techniques use statistical language models to detect poorly formatted sequences of words. The paper [14] provided a way to detect and correct errors used by Name Entity Recognition (NER) [15], the NLP methods, and Shannon's model for noise in communication channels.

The most popular software using NLP techniques for knowledge extraction from EHRs in English are Apache CTAKES [16] and CLAMP [17]. These NER tools provide automated labeling of clinical documentation.

ASR in healthcare is detailly described in the review paper [18]. It increases medical staff productivity, facilities completeness of medical documentation, and inspires patient management.

Regardless of the medical domain, the detection and correction of errors in the Serbian language are discussed in the dissertation [19].

3. MATERIALS AND METHODS

A corpus consisting of 100 EHRs was used for this research. These reports were written in Serbian from health care institutions belonging to the Health Care Center of the city of Nis and were collected by the information system MEDIS.NET [3]. This corpus is built according to all ethical standards, with the removal of the identities of patients and medical staff.

The following scientific methods were used in this paper: description, content analysis, experimental and comparative methods. The description was applied to existing methods for detecting and correcting errors in medical reports, while the described data set was first analyzed for content, to find the types of errors that occur in the corps and create methods to correct them. Also, the same set of anamneses was read using the ASR system (Google Text to Speech API) [20], and the types of errors that appear here were also analyzed. The types of errors that occur during anamnesis entry in one way and in another are discussed. NLP methods have been proposed to detect and correct errors in free (unstructured) part of EHRs in Serbian.

4. TYPES OF MISSPELLINGS IN EHRS

By analyzing the contents of the described data set, eleven types of errors occur in electronic medical reports manually entered, while only one error type occurs when EHR is entered using an ASR system (type 5 – replacing with a similar word), which is expected since ASR systems already have built-in NLP functions. Table 1 lists these types of errors and an example for each of them from the analyzed data set.

Misspelling's Type	Description	Example	Correct word
Type 1	omitted double letter	Poštreno	pooštreno
Type 2	Replacement of letters	Malakslaost	malaksalost
Type 3	additional letters	Trupzu	trupu
Type 4	missing letters	Temeratura	temperatura
Type 5	replacing with a similar word	Zrelo	ždrelo
Type 6	joined words without spaces	Thbrufen	th. brufen
Type 7	conjoined words with a random letter instead of a space	Pomtelu	po telu
Type 8	omitted (replaced) diacritic symbol (e.g. "c" instead of "č" or "ć")	Kozi	koži
Type 9	incorrect letter	Uirus	virus
Type 10	use of letters that do not belong to the Serbian alphabet ("x" instead of "ks")	Extremitetima	ekstremitetima
Type 11	multiple errors in one word	Makolozma	makulozna

Table 1 - Types of errors identified in EHRs

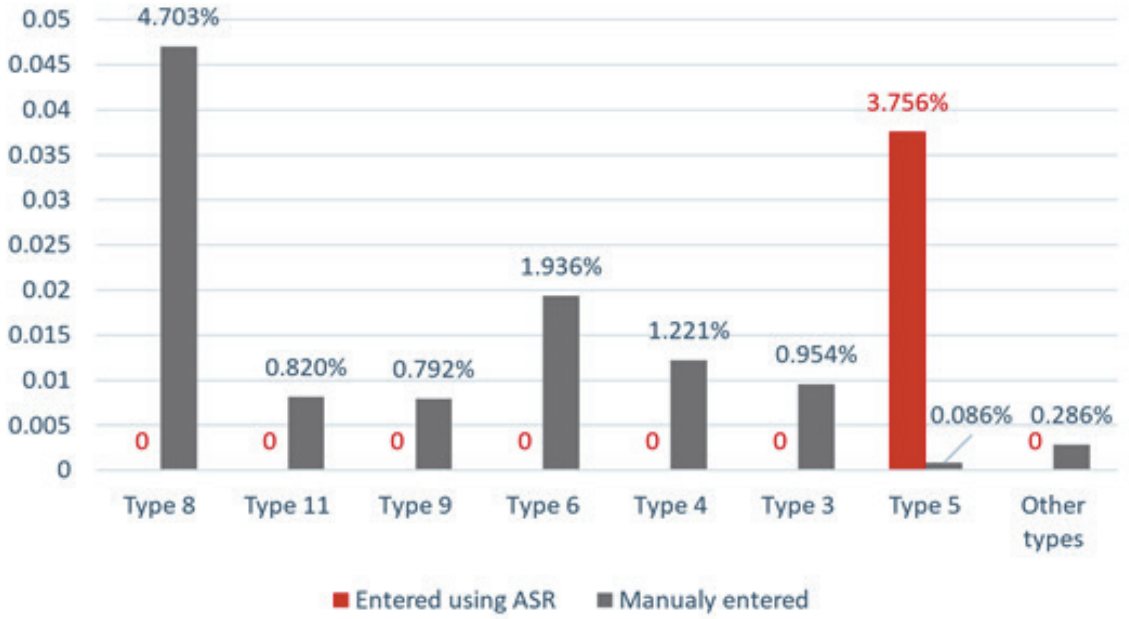

Figure 1 - Percentage of misspellings frequency in EHRs

By analyzing the data set, the percentage of errors in electronic medical reports manually entered is 10.7981% and entered using the ASR system is 3.7559%. Figure 1 shows the distribution of errors in the data set.

Based on the percentage of errors that appear in the processed data set, the types 8, 6, 4, and 3 are the most prevalent in manually entered EHRs, and only type 5 in EHRs entered through ASR, so that error correction rule formulation should resolve these errors.

5. SYSTEM FOR KNOWLEDGE EXTRACTION FROM EHRS IN SERBIAN

Although entering EHRs through ASR reduces the number of errors, the ability to enter through the keyboard must be left to provide the writer with the ability to correct the error before saving EHR entered in the system. Therefore, input through ASR can be used to reduce input time and give better results for errors that occur, but NLP methods must still be used to process the entered EHRs. Because the type of error that occurs during an ASR entry is the wrong word, which does not contain a typo, but does not belong to the syntagma by context, then techniques such as NER should be used for these purposes.

Figure 2 shows the architecture of the knowledge extraction system from EHRs in Serbian with the detection and correction capability. The first layer is the user interface, with the possibility of entering EHR manually using a keyboard or using Text to Speech API. Considering Type 5 of error which can be found after ASR entrance of EHR, before saving, a manual correction of an incorrectly recognized word is allowed. Due to time limitation and large number of examinations it is very possible that this feature would not be always used.

The second layer (application logic) uses techniques for preprocessing, classifying, learning, and suggesting correct words. Precondition for using NLP techniques is normalization of EHRs (tokenization, stop words removal, processing of abbreviations, negation, reducing words to a basic form). These steps are detailly described in paper [10]. Error detection can be performed by trying to label words using the algorithms in Serbian medical texts. Methods based on dictionaries or machine learning that use a labeled set of electronic medical reports may be used for error detection [15]. If the word is not labeled, it will be considered a misspelling. Error correction can also be made using methods based on dictionaries or on learning about the training set, but these methods need to include the use of algorithms for normalization [10] based on n-gram analysis and Serbian language stemmer [21] and to use additional steps to correct errors as error correction rules [15]. To correct a misspelling, first, for each doctor's ID special vocabulary should be created and searched. The performance of labelling EHRs in Serbian using proposed methods is quite high, considering that F1-score is over 90% [15]. The goal of such EHR data structuring is to reach the structure which enables semantic search and, possibly, automated reasoning.

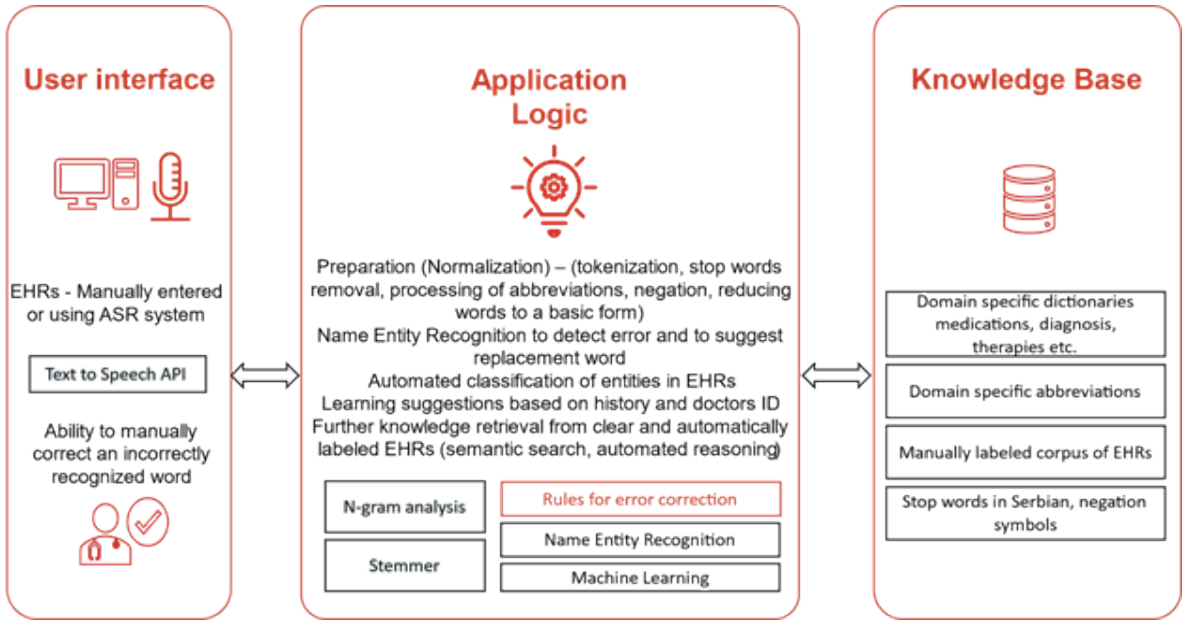

Figure 2 – The architecture of the knowledge extraction system from EHRs in Serbian with the detection and correction capability

Third layer is knowledge base which consist of training set (manually labelled EHRs) and domain specific dictionaries (therapies, diagnoses codes, diagnoses, anatomic organs, Latin word, etc.), stop words, negation symbols etc.

6. CONCLUSION

Correctly written electronic medical reports can affect the success of treating patients, and their incorrectness can have dire consequences. A set of medical reports was analyzed in this paper, and errors were found and marked in it, to create methods for their detection and correction. The main contribution of the paper is the classification of the type of errors, the display of their frequency in analyzed medical reports both manually entered and using ASR and the proposal of the architecture of the system for EHRs entry using ASR and manually, error detection and correction based on specialized dictionaries, natural language processing, training set consisting of hand-marked reports, machine learning and rules, and further processing and knowledge extraction. The subject of further research will be a quantitative analysis of the proposed methods and their experimental performance and a comparison of results with similar methods for detecting and correcting errors in similar languages.

7. ACKNOWLEDGMENTS

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.

8. REFERENCES

[1] J. W. Henry and R. W. Stone, “A structural equation model of end-user satisfaction with A computer-based medical information system,” *Inf. Resour. Manag. J.*, vol. 7, no. 3, pp. 21-33, 1994.

[2] N. Menachemi and Collum, ““Benefits and drawbacks of electronic health record systems,” *Risk Manag. Healthc. Policy*, p. 47, 2011.

[3] A. M. Milenkovic, P. J. Rajkovic, T. N. Stankovic and D. S. Jankovic, "Application of medical information system MEDIS.NET in professional learning," in *19thTelecommunications Forum (TELFOR) Proceedings of Papers*, Belgrade, 2011.

[4] D. J. Hand and N. M. Adams, Data mining, Wiley StatsRef: Statistics Reference Online, 2014, pp. 1-7.

[5] M. W. Berry and J. Kogan, Text mining. Applications and Theory, West Sussex, PO19 8SQ: UK: John Wiley & Sons, 2010.

[6] G. G. Chowdhury, "Natural language processing," *Annual review of information science and technology*, vol. 37, no. 1, pp. 51-89, 2003.

[7] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1-23, 2021.

[8] K. Kukich, "Techniques for automatically correcting words in text," *Acm Computing Surveys (CSUR)*, vol. 24, no. 4, pp. 377-439, 1992.

[9] J. López-Hernández, A. Almela and R. Valencia-García, "Automatic spelling detection and correction in the medical domain: A systematic literature review.," in *In International Conference on Technologies and Innovation,* 2019, December.

[10] A. R. Avdić, U. M. Marovac and D. S. Janković, "Normalization of Health Records in the Serbian Language with the Aim of Smart Health Services Realization," *Facta Universitatis, Series: Mathematics and Informatics,* pp. 825-841, 2020.

[11] A. Yazdani, M. Ghazisaeedi, Ahmadinejad, G. M. N., H. Amjadi and A. Nahvijou, "Automated misspelling detection and correction in persian clinical tex," *Journal of digital imaging*, vol. 33, no. 33, pp. 555-562, 2020.

[12] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Communications of the ACM*, vol. 7, no. 3, pp. 171-176, 1964.

[13] K. Balabaeva, A. A. Funkner and S. V. Kovalchuk, "Automated Spelling Correction for Clinical Text Mining in Russian," 2020, June.

[14] K. H. Lai, M. Topaz, F. R. Goss and L. Zhou, "Automated misspelling detection and correction in clinical free-text records," *Journal of biomedical informatics*, vol. 55, pp. 188-195, 2015.

[15] A. Avdić, U. Marovac and D. Janković, "Automated labeling of terms in medical reports in Serbian," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 28, no. 6, pp. 3285-3303, 2020.

[16] V. Garla, V. L. Re III, Z. Dorey-Stein, F. Kidwai, M. Scotch, J. Womack, A. Justice and C. Brandt, "The Yale cTAKES extensions for document classification: architecture and application," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 614-620, 2011.

[17] E. Soysal, J. Wang, M. Jiang, Y. Wu, S. Pakhomov, H. Liu and H. Xu, "CLAMP - a toolkit for eficiently building customized clinical natural language processing pipelines," *Journal of the American Medical Informatics Association*, vol. 25, no. 3, pp. 331-336, 2018.

[18] S. Latif, J. Qadir, A. Qayyum, M. Usama and S. Younis, "Speech technology for healthcare: Opportunities, challenges, and state of the art," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 342-356, 2020.

[19] S. Ostrogonac, Modeli srpskog jezika i njihova primena u govornim i jezičkim tehnologijama, Doctoral dissertation, University of Novi Sad (Serbia), 2018.

[20] "Speech-to-text: Automatic speech recognition," [Online]. Available: https://cloud.google.com/speech-to-text. [Accessed 6 June 2021].

[21] V. Batanović, N. Ljubešić, T. Samardžić and M. M. Petrović, Otvoreni resursi i tehnologije za obradu srpskog jezika. Primena slobodnog softvera i otvorenog hardvera., 2020.