A Rule-Based System for Automatic De-identification of Medical Narrative Texts

Jelena Jaćimović[1,2], Cvetana Krstev[1] and Drago Jelovac[2]
[1]University of Belgrade, Faculty of Philology, Studentski trg 3, 11000 Belgrade, Serbia
[2]University of Belgrade, School of Dental Medicine, Dr. Subotića 8, 11000 Belgrade, Serbia
E-mail: jjacimovic@rcub.bg.ac.rs, cvetana@matf.bg.ac.rs, drago.jelovac@stomf.bg.ac.rs

Keywords: named entity recognition, finite-state transducers, rule-based system, de-identification, medical narrative texts, Serbian

Received: May 7, 2014

This paper presents an automatic de-identification system for Serbian, based on the adaptation of the existing rule-based named entity recognition system. Built on a finite-state methodology and lexical resources, the system is designed to detect and replace all the explicit personal protected health information present in the medical narrative texts, while still preserving all the relevant medical concepts. The results of a preliminary evaluation demonstrate the usefulness of this method both in preserving patient privacy and the de-identified document interoperability.

Povzetek: Razvit je nov sistem za de-identifikacijo besedil v srbskem jeziku.

1 Introduction

Current advances in health information technology enable health care providers and organizations to automate most aspects of the patient care management, facilitating collection, storage and usage of patient information. Such information, stored in the form of electronic medical records (EMRs), represents accurate and comprehensive clinical data valuable as a vital resource for secondary uses such as quality improvement, research, and teaching. Besides the vast useful information, narrative clinical texts of the EMR also include many items of patient identifying information. For both ethical and legal reasons, when confidential clinical data are shared and used for research purposes, it is necessary to protect patient privacy and remove patient-specific identifiers through a process of the de-identification.

De-identification is focused on detecting and removing/modifying all explicit personal Protected Health Information (PHI) present in medical or other records, while still preserving all the medically relevant information about the patient. Various standards and regulations for health data protection define multiple directions to achieve de-identification, but the most frequently referenced regulation is the US Health Information Portability and Accountability Act (HIPAA) [1]. According to the HIPAA "Safe Harbor" approach, clinical records are considered de-identified when 18 categories (17 textual and one regarding images) of PHI are removed, and the remaining information cannot be used alone or in combination with other information to identify an individual. These PHI categories include names, geographic locations, elements of dates (except year), telephone and fax numbers, medical record numbers or any other unique identifying numbers, among others. Since the manual removal of PHI by medical professionals proved to be prohibitively time-consuming, tedious, costly and unreliable [2, 3, 4], extracting PHI requires more reliable, faster and cheaper automatic de-identification systems based on the Natural Language Processing (NLP) methods [5].

The extraction of PHI can be viewed as a Named Entity Recognition (NER) problem applied in the medical domain for de-identification [6]. However, even though both the traditional NER and de-identification involve automatic recognition of particular phrases in text (persons, organizations, locations, dates, etc.), de-identification differs importantly from the traditional NER [7]. In contrast to the general NER focused on newspaper texts, de-identification deals with the clinical narratives characterized by fragmented and incomplete utterances, the lack of punctuation marks and formatting, many spelling and grammatical errors, as well as domain specific terminology and abbreviations. Since de-identification is the first step towards identification and extraction of other relevant clinical information, it is extremely important to overcome the problem of significantly large number of eponyms and other non-PHI erroneously categorized as PHI. For instance, the anatomic locations, devices, diseases and procedures could be erroneously recognized as PHI and removed (e.g. *"The Zvezdara method"*[1] vs. *Clinical Center "Zvezdara"*), reducing the usability and the overall meaning of clinical notes, and thus the accuracy of

[1] The original surgical 2-step arteriovenous loop graft procedure developed in the Clinical Center "Zvezdara", Belgrade, Serbia. *Zvezdara* is a municipality of Belgrade.

subsequent automatic processes performed on the de-identified documents.

In this paper we introduce our automatic clinical narrative text de-identification system, based on the adaptation of the existing rule-based NER system for Serbian. The aim of this study is to evaluate the accuracy of PHI removal and replacement while preserving all the medically relevant information about a patient and keeping the resulting de-identified document usable for subsequent information extraction processes.

2 Related work

Over the past twenty years, various text de-identification approaches have been developed, but relatively few published reports are focused only on the unstructured medical data. Extensive review of recent research in automatic de-identification of narrative medical texts is given in [5]. However, most of them are highly specialized for specific document types or a subset of identifiers. Regarding the general nature of applied de-identification methods, the majority of the systems used only one or two specific clinical document types (pathology reports, discharge summaries or nursing progress notes) for the evaluation [3, 8, 9, 10], while only a few of them were evaluated on a larger scale, with a more heterogeneous document corpus [11, 12, 13, 14]. The selection of targeted PHI varied from patient names only [12] to all 17 textual HIPAA PHI categories [3, 7, 15, 16, 17, 18, 19], or even everything but valid medical concepts [20, 21].

The de-identification approaches applied in medical domain are mostly classified into the rule-based or machine learning methods, while some hybrid approaches [14] efficiently take advantage of both previous methods. The rule-based methods [3, 15, 17, 19, 21] make the use of dictionaries and hand-crafted rules to identify mentions of PHI, with no annotated training data. Although these systems are often characterized with the limited generalizability that depends on the quality of the patterns and rules, they can be easily and quickly modified by adding rules, dictionary terms or regular expressions in order to improve the overall performance [22]. On the other hand, the machine-learning methods [7, 8, 9, 16, 18, 23], proved to be more easily generalized, automatically learn from training examples to detect and predict PHI. However, these methods require large amounts of annotated data and the adaptation of the system might be difficult due to the often unpredictable effects of a change. Extensive review of the published strategies and techniques specifically developed for de-identification of EMRs is given in [24]. In 2006, within the Informatics for Integrating Biology and the Bedside (i2b2) project and organized de-identification challenge, a small annotated corpus of hospital discharge summaries were shared among the interested participants, providing the basis for the system development and evaluation. Detailed overview and evaluation of the state-of-the-art systems that participated in the i2b2 de-identification challenge is given in [25].

Aside from systems specifically designed for the purpose of de-identification, some NER tools trained on newspaper texts also obtained respectable performance with certain PHI categories [7, 26].

3 Data and methods

This section provides an overview of our rule-based de-identification approach for narrative medical texts.

3.1 Training and text corpus

The training corpus for our system development consisted of 200 randomly selected documents from different specialties, generated at three Serbian medical centers. They included discharge summaries (50), clinical notes (50) and medical expertise (100), with a total word count of 143,378. The discharge summaries and clinical notes are unstructured free text typed by the physicians at the conclusion of a hospital stay or series of treatments, including observations about the patient's medical history, his/her current physical state, the therapy administered, laboratory test results, the diagnostic findings, recommendations on discharge and other information about the patient state. Medical expertise documents were oversampled because of their richness in the PHI items.

Main characteristics of medical narratives were confirmed in our corpus: fragmented and incomplete utterances and lack of punctuation marks and formatting. Moreover, as these documents are usually written in a great hurry there is also an unusual number of spelling, orthographic and typographic errors, much larger than in, for instance, newspaper texts from the Web. For the moment, we have taken these documents as they are and we are not attempting to correct them. In some particular situations we are able to guess the intended meaning, as will be explained in the next section.

3.2 The NER system

The primary resources for natural language processing of Serbian consist of lexical resources and local grammars developed using the finite-state methodology as described in [27, 28]. For development and application of these resources the Unitex corpus processing system is used [29]. Among general resources used for NER task are the morphological e-dictionaries, covering both general lexica and proper names, as well as simple words and compounds, including not only entries collected from traditional sources, but also entries extracted from the processed texts [30]. Besides e-dictionaries, for the recognition and morphosyntactic tagging of open classes of simple words and compounds generally not found in dictionaries, the dictionary graphs in the form of finite-state transducers (FSTs) are used. Due to the high level of complexity and ambiguity of named entities, the additional resources for NER were developed [31]. The Serbian NER system is organized as a cascade of FSTs – CasSys [32], integrated in the Unitex corpus processor. Each FST in a cascade modifies a piece of text by replacing it with a lexical tag that can be used in subsequent FSTs. For instance, in a sequence *Dom*

zdravlja "Milutin Ivković" 'Health Center 'Milutin Ivković" first a full name 'Milutin Ivković' is recognized and tagged {Milutin Ivković,.NE+persName+full:ms1v}, and then a subsequent transducer in the cascade uses this information to appropriately recognize and tag the full organization name that can also be subsequently used (see Figure 1 and Example (1)).

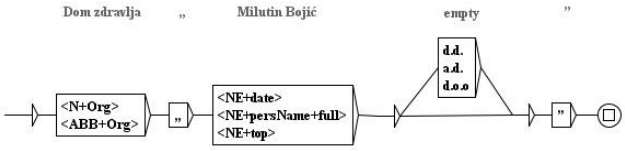

Figure 1. A path in a cascade graph that uses already recognized NEs to recognize organization names.

(1){Dom zdravlja "\{Milutin Ivković\, \.NE\+persName \+full\\:s1v\}",.NE+org:1sq:4sq}

Serbian NER system recognizes a full range of traditional named entity types:

- Amount expressions – count, percentage, measurements and currency expressions;
- Time expressions – absolute and relative dates and times of day (fixed and periods), durations and sets of recurring times;
- Personal names – full names, parts of names (first name only, last name only), roles and functions of persons;
- Geopolitical names – names of states, settlements, regions, hydronyms and oronyms;
- Urban names – at this moment only city areas and addresses are recognized.

For the purpose of PHI de-identification not all of these NEs are of interest. For instance, amount expressions should not be de-identified, and roles or functions need not be de-identified. However, we chose not to exclude them from the recognition for two reasons: first, if they are recognized correctly that may prevent some false recognition and second, even if they are not of interest for this specific task they may help in recognition of some NEs that are of interest. For instance in Example (2) a name is erroneously typed (both the first and the last name are incorrect) but due to a correct recognition of a person's function the name is also recognized.

(2) *prof. dr sci Drangan Jorvanović, specijalista za stomatološku protetiku i ortopediju* 'Prof. PhD Drangan Jorvanović, a specialist for Prosthetic Dentistry and Orthodontics'

The finite-state transducers used in the NER cascade beside general and specific e-dictionaries, as explained before, use local grammars that model various triggers and NE contexts, such as:

- The use of upper-case letters – for personal names, geopolitical names, organizations, etc.;
- The sentence boundaries – to resolve ambiguous cases where there is not enough other context;
- Trigger words – for instance, *reka* 'river', *grad* 'city' and similar can be used to recognize geopolitical names that are otherwise ambiguous;
- Other type of the context – for instance, a punctuation mark following a country name that coincides with a relational adjective[2] signals that it is more likely a country name than an adjective;
- Other NEs – for instance, an ambiguous city name can be confirmed if it occurs in a list of already recognized NEs representing cities. Also, a five digit number that precedes a name of a city (already recognized) is tagged as a postal code (as used in Serbia).
- Grammatical information – this information is used to impose the obligatory agreement in the case (sometimes also the gender and the number) between the parts of a NE. For instance, in *…istakao je gradonačelnik Londona Boris Džonson…* '…stressed Mayor of London Boris Johnson…' *Londona* can be falsely added to the person's name (because *London* is also a surname) if grammatical information was not taken into consideration (*Londona* is in the genitive case, while *Boris* and *Džonson* are in the nominative case). This is enabled by grammatical information that is part of NE lexical tags (see Example (1)).

3.3 The PHI de-identification

We used our training corpus to create and adapt patterns that will capture the characteristics of PHI. Through the corpus analysis we found that, out of 18 HIPAA PHI categories, only eight appeared in our data. Since there is no annotation standard for PHI tagging, we collapsed some of the HIPAA categories into one (telephone and fax numbers, medical record numbers or any other unique identifying numbers). In order to maximize patient confidentiality, we adopted a more conservative approach, considering countries and organizations as PHI. For the purposes of this study, we defined the resulting PHI categories as follows:

- Persons (*pers*) – refers to all personal names; includes first, middle and/or last names of patients and their relatives, doctors, judges, witnesses, etc.;
- Dates (*date*) – includes all elements of dates except year and any mention of age information for patients over 89 years of age; according to HIPAA, the age over 89 should be collected under one category 90/120;
- Geographic locations (*top*) – includes countries, cities, parts of cities (like municipalities), postal codes;
- Organizations (*org*) – hospitals and other organizations (like courts);
- Numbers (*num*) – refers to any combination of numbers, letters and special characters

[2] In Serbian many country names coincide with relational adjectives of feminine gender: *Norveška* 'Norway' and *norveška* 'Norwegian'.

representing telephone/fax numbers, medical record numbers, vehicle identifiers and serial numbers, any other unique identifying numbers;

- Addresses (*adrese*) - street addresses.

The processing usually starts with a text having undergone a sentence segmentation, tokenization, part-of-speech tagging and morphological analysis. After general-purpose lexical resources are used to tag the text with lemmas, grammatical categories and semantic features, the FST cascade is applied, recognizing persons, functions, organizations, locations, amounts, temporal expressions, etc.

Since medical narratives have specific characteristics, the primary issue of date's recognition arose and we added a small cascade of FSTs prior to detection of the sentences. For the de-identification task and the processing of medical data, we performed the adjustments of the temporal expressions FSTs, in order to cover only those temporal expressions that should be treated as PHI. We also developed new patterns for the identification of different diagnostic codes present in training documents that could be misinterpreted as an identifier and then erroneously masked. Being applied as first in the cascade, this FST produces lexical tags denoting non-PHI category of the diagnostic codes, bringing the precision and accuracy up to an acceptable level in order to prevent loss of clinical information in the de-identification process.

Lexical tags produced by FSTs (see Example (3)), even though the most convenient for the use of subsequent FSTs in the cascade, are not useful for other applications and at the end are converted to the XML tags (Example (4)).

(3) {Beogradu,.NE+top+gr:s7q}
(4) <top.gr>Beogradu</top.gr>

The de-identification can be performed in several ways: PHI that needs to be de-identified can be replaced by a tag denoting its corresponding category, with a surrogate text, or both. We have chosen the latter approach. Moreover, since we are dealing with the narrative texts as a result we want to obtain a narrative text as well. To that end, the surrogate text is chosen to agree in case, gender and number with the PHI it replaces (if applicable). Again, such a replacement is enabled by grammatical information associated with some NE types (personal names, organization names, locations, etc.). For instance, recognized geographic name (+top) of the city (+gr) in Example (3) will be replaced by the surrogate text with the same values of the grammatical categories (Example (5)).

(5) <top.gr PHI="yes">Kamengradu</top.gr>

At this moment, our system does not keep the internal structure of the numbers PHI category (*num*), and all the PHI numbers are simply replaced by placeholder characters X. Regarding the temporal information, only the month and day portion of date expressions are considered PHI. According to HIPAA, the years are excluded from this category, being important features of the clinical context. In order to preserve the existing interval in days between two events in the text or the duration of specific symptoms, all dates were replaced by a shifted date that is consistent throughout all the de-identified documents.

3.4 An example

In this subsection we will give an example taken from the part of the test corpus containing medical expertise. The part of one note is given in Example (6).[3] The same expertise after the de-identification and tagging is given in Example (7).[4]

(6)
Naš broj 23/246
OPŠTINSKI SUD - Istražni sudija G-đa Jovana Jovanović-
Vašim zahtevom u predmetu TR 123/01 od 23.07.2007. god. zatražili ste od Komisije lekara veštaka Medicinskog fakulteta Univerziteta u Kragujevcu sudskomedicinsko veštacenje o vrsti i težini telesnih povreda koje je dana 4.02. 2007. god. zadobio Petrović Dragan iz Jagodine.
…
PODACI
1. Pri pregledu na Medicinskom fakultetu u Kragujevcu, obavljenom dana 12.02.2007. god. od strane članova Komisije lekara veštaka Medicinskog fakulteta Univerziteta u Kragujevcu, Dragan, Miroslava, Petrović navodi: rođen je 14. 01. 1956. god. u Jagodini, živi u Jagodini, ul. Savska br. 7, po zanimanju pekar, broj lične karte 1234567, MUP Jagodina . Amanestički navodi operaciju kolena marta 2000. god., negira postojanje oboljenja. Dana 4,02. 2007. god. oko 12,30 h, na sportskom terenu došlo je do fizičkog obračuna između Dragana i njegovog poznanika.
…
NALAZ
1. U izveštaju Dr Petra Dragića, specijaliste za otorinolaringologiju, Zdravstvenog centra "Milutin Ivković", iz Jagodine, na ime Petrović Dragana, od 4.02.07. god., navedeno je sledeće: "Povređen u tuči od strane poznatog lica. Svest nije gubio. Dg. Fractura dentis incisiv"
…

'Our number 23/246
Municipal court - Judge Mrs Jovana Jovanović -
In your request in the case TR 123/01 from 23/07/2007 you asked for a medico-legal expertise on the type and

[3] This example looks exactly as the original – however, for the purpose of protecting the personal data we have manually replaced all the personal information with some "real world" data.

[4] We wanted to avoid the introduction of some real people names and real location names in the de-identified texts. Instead we used names: *Barni Kamenko* (Barney Rubble), *Vilma Kremenko* (Vilma Flintstone), *Kamengrad* (Bedrock), Serbian names for the characters from the sitcom *The Flinstones*, created by Hanna-Barbera Productions, Inc.

gravity of bodily injuries inflicted on Petrović Dragan from Jagodina on 4/02/2007 from the Commission of the Faculty of Medical Sciences University of Kragujevac's medical experts.
…
DATA
1. Upon the examination performed in Faculty of Medical Sciences of Kragujevac, conducted on 12/02/2007 by members of the Commission of the Faculty of Medical Sciences University of Kragujevac's medical experts, Dragan, Miroslava, Petrović states: born on 14/ 01/1956 in Jagodina, lives in Jagodina, Savska Street 7, a baker by profession, ID number 1234567, MIA Jagodina. Anamnestic states the knee surgery performed on March 2000, negates the existence of a disease. On 4/02/2007 around 12:30 PM, on the sports field it came to a physical confrontation between Dragan and his acquaintance.
…
FINDING
1. In the medical report of Zoran Dragić, MD, specialised in otorhinolaryngology, Medical Centre "Milutin Ivković", from Jagodina, on Petrović Dragan's name, from 4/02/2007, the following was stated: "He was injured in a fight by an acquaintance. He didn't lose his consciousness. Dg. Fractura dentis incisiv. "
…'

(7)
Naš <number PHI="yes">XXXX</number>
<org PHI="yes">SUD</org> - <pers><role>Istražni sudija gospođa</role> <persName.full PHI="yes">Vilma Kremenko</persName.full></pers>-Vašim zahtevom u predmetu <number PHI="yes">XXXX</number> od <date PHI="yes">28.12.2007.</date> zatražili ste od <org PHI="yes">Komisije</org> <org PHI="yes">fakulteta</org> <org PHI="yes">Univerziteta</org> sudskomedicinsko veštacenje o vrsti i težini telesnih povreda koje je dana <date PHI="yes">09.07.2007.</date> zadobio <persName.full PHI="yes">Barni Kamenko</persName.full> iz <top.gr PHI="yes">Kamengrada</top.gr>.
…
PODACI
1.{S} Pri pregledu na <org PHI="yes">fakultetu</org>, obavljenom dana <date PHI="yes">17.07.2007.</date> od strane članova <org PHI="yes">Komisije</org> <org PHI="yes">fakulteta</org> <org PHI="yes">Univerziteta</org>, <persName.full PHI="yes">Barni Kamenko</persName.full> navodi: rođen je <date PHI="yes">19.06.1956.</date> u <top.gr PHI="yes">Kamengradu</top.gr>, živi u <top.gr PHI="yes">Kamengradu</top.gr>, <adresa PHI="yes">ul. Kamenolomska br. 6a</adresa>, po zanimanju pekar, broj lične karte <number PHI="yes">XXXX</number>, <org>MUP<top.gr PHI="yes">Kamengrad</top.gr></org> .{S} Amanestički navodi operaciju kolena <date PHI="yes">avgusta 2000.</date>, negira postojanje oboljenja.{S} Dana <date PHI="yes">09.07.2007.</date> oko 12,30 h, na sportskom terenu došlo je do fizičkog obračuna između **Dragana** i njegovog poznanika.
…
NALAZ
1.{S} U izveštaju <pers><persName.full PHI="yes">Barnija Kamenka</persName.full><role>, specijaliste za otorinolaringologiju</role></pers>, <org PHI="yes">centra</org>, iz <top.gr PHI="yes">Kamengrada</top.gr>, na ime <persName.full><persName.full PHI="yes">**Vilma Kremenko**</persName.full></persName.full>, od <date PHI="yes">09.07.07.</date>, navedeno je sledeće:"....{S} Povređen u tuči od strane poznatog lica.{S} Svest nije gubio.{S} Dg.{S} Fractura dentis incisiv."
…

This example demonstrates our de-identification approach. Each detected PHI was enclosed in XML tags indicating its corresponding category, with the PHI attribute value set to "yes". Note that all dates were shifted into the future by the same amount. Information specific to the hospitals and other organizations was replaced by a generalized data with the same organizational hierarchy. For instance, a sequence of hierarchical organization names *Komisija lekara veštaka Medicinskog fakulteta Univerziteta u Kragujevcu* 'the Commision of medical experts of the Faculty of Medical Sciencies of the University of Kragujevac' is replaced by *Komisija fakulteta Univerziteta* 'a Commission of a faculty of a University'. Some personal data remained: the occurrence of the first name of the patient. Also, the replacement text was not always correct: the male patient's name was once replaced by the female name because the original occurrence was ambiguous and could be interpreted both as a masculine name *Petrović Dragan* (in the genitive case) and a feminine name *Petrović Dragana* (in the nominative case). Our system has randomly chosen the feminine name. These occurrences are bolded and underlined in Example (7).

4 Evaluation results

The previously described system for the automatic de-identification has been evaluated on a set of 100 randomly selected documents (total word count of 35,822), consisting of discharge summaries (60), clinical notes (27) and medical expertise (13). These chosen texts were not used in the system development and presented completely unseen material containing many occurrences of PHI. Details about the PHI distribution within the test corpus can be found in Table 1.

The performance has been evaluated with respect to recognition, bracketing and replacement of PHI. For that reason, a new attribute 'check' has been added to each XML tag. Possible values of this attribute were the following:

OK – PHI was correctly recognized, full extent was correctly determined, replacement was correctly assigned;

UOK - UOK1 (PHI type was correctly recognized, but full extent was not correctly determined, some part of PHI was revealed); UOK2 (PHI type was not

PHI/Document type	Cinical reports	Discharge summaries	Medical expertise	Total
pers	52	254	407	713
top	32	219	109	360
org	62	164	242	468
num	20	61	90	171
date	65	133	267	465
adrese	0	64	10	74
Total	231	895	1125	2251

Table 1: The PHI distribution considering document type.

correctly determined, but the full extent was correctly determined, PHI successfully masked);

NOK – an utterance tagged falsely as PHI and de-identified;

MISS – PHI was not recognized;

MISS/E – PHI was not recognized because of the incorrect input.

In some cases when it was not so easy to decide which is the most appropriate value for the 'check' attribute (e.g. personal name as a name of an organization), we always treated as correct, for example, a personal name tag even though the utterance belonged to organization category.

We report the results of the evaluation using the traditional performance measures: precision (positive predictive value), recall (sensitivity) and *F*-measure (harmonic mean of recall and precision). These measures are calculated at the phrase level, considering the entire PHI annotation as the unit of evaluation.

The harmonic mean of recall and precision is calculated in two ways, using the strict and relaxed criteria. With the strict criteria we consider as true positives only fully correctly recognized and de-identified PHI and as false negatives all PHI that were not recognized and de-identified, regardless of the reasons (including the incorrect input). With the relaxed criteria we consider as true positives all correctly recognized and de-identified PHI including partial recognition and false type attribution, and as false negatives all PHI that were not recognized and de-identified if the input was correct (see Table 2).

The overall evaluation of the system is presented in Table 3 and Table 4.

	1. Strict criteria	2. Relaxed criteria
TP	OK	OK+UOK
FP	NOK+UOK	NOK
FN	MISS+MISS/E	MISS
P	OK/(OK+NOK+UOK)	(OK+UOK)/(OK+NOK+UOK)
R	OK/(OK+MISS+MISS/E)	(OK+UOK)/(OK+UOK+MISS)

Table 2. Calculation using strict and relaxed criteria: TP (true positive), FP (false positive), FN (false negative), P (Precision), R (Recall).

PHI	OK	UOK1	UOK2	MISS	MISS/E	NOK
pers	634	12	47	15	5	30
top	337	0	0	14	9	5
org	434	0	0	28	6	6
num	132	2	0	36	1	8
date	455	4	0	1	5	7
adrese	63	0	0	1	10	2
Total	2055	18	47	95	36	58

Table 3. Evaluation data.

PHI	Precision (p1)	Recall (r1)	*F1*-measure	Precision (p2)	Recall (r2)	*F2*-measure
pers	0.88	0.97	0.92	0.96	0.98	0.97
top	0.99	0.94	0.96	0.99	0.96	0.97
org	0.99	0.93	0.96	0.99	0.94	0.96
num	0.93	0.78	0.85	0.94	0.79	0.86
date	0.98	0.99	0.98	0.98	1.00	0.99
adrese	0.97	0.85	0.91	0.97	0.98	0.98
Total	**0.94**	**0.94**	**0.94**	**0.97**	**0.96**	**0.97**

Table 4. Performance measures for PHI de-identification by applying the strict criteria (1) and the relaxed criteria. (2)

Besides the traditional *F*-measure, evaluation is also performed using the Slot Error Rate (SER) [33]. As a simple error measure, the SER equally weights different types of error directly, enabling the comparison of all systems against the fixed base. The SER is equal to the sum of the three types of errors — substitutions (UOK1, UOK2), deletions (MISS, MISS/E), and insertions (NOK) — divided by the total number of PHI in the reference corpus (Formula (1)).

(1)

$$SER = \frac{NOK + MISS + MISS/E + UOK1 + UOK2}{OK + UOK1 + UOK2 + MISS + MISS/E}$$

In measuring the accuracy of the de-identification system, the extent of PHI (UOK1), missed PHI (MISS, MISS/E) as well as entities falsely tagged as PHI (NOK), should be taken into consideration as equally weight separate error slots. In that way, unlike the more relaxed *F*-measure, the SER of 11.3% stresses out the importance

of the errors that affect revealing of PHI to a greater extent.

5 Discussion

Clinical records are considered de-identified when, after removal of PHI, the remaining information cannot be used alone or in combination with other information to identify an individual. Nevertheless, even though PHI is removed, there is a concern that de-identified medical documents could potentially be re-identified i.e. that it is difficult but still possible to reestablish the link between the individual and his/her data [24]. In the context of de-identification each PHI category is treated differently. There are obvious identifiers (e.g. name, telephone number, home address…) as well as quasi-identifiers that can play an important role in indirect re-identification, such as dates, locations, race and gender [34]. In some cases, more than one identifying variable is needed to identify an individual uniquely. For example, sex and year of birth combined with the disease name (it might be some rare disease) could be used for indirect re-identification. However, some PHI categories, such as ages over 89, geographical locations, hospitals and other organizations are most frequently ignored by the existing de-identification systems [5]. Even though according to the most frequently referenced regulation HIPAA, states and organizations are not considered as PHI, we adopted a more conservative approach, considering them as variables that could be used for re-identification.

We found that our NER system could be modified to work on medical narratives for de-identification purposes. However, certain modifications were necessary in order to preserve relevant clinical information. Previous evaluation results showed that Serbian NER system gives priority to precision over the recall [30], and the recall rate had to be improved for the de-identification task.

An error analysis shows that every correctly recognized PHI was correctly de-identified. The main source of errors were missed PHI, resulting in the information disclosure. The most missed PHI were numbers and organizations not included in our pattern rules and dictionaries, while fewer than 6% of errors resulted in the revealing of the most sensitive category i.e. person names. Another source of errors that could cause PHI exposure was wrongly determined PHI extent (4.72% of total errors). Fewer than 20% of errors were examples tagged with an incorrect PHI category which may only reduce the readability of the resulting de-identified text without exposing PHI. Since one of the main goals is to preserve medically relevant information, it is important to pay special attention to false positives, which represented 22.83% out of total errors. For *pers*, a majority of false positives were diseases and procedures names.

Our automatic de-identification system achieved very competitive precision and recall rate, showing the overall *F1*-measure of 0.94 (Table 4). High performance was achieved for most PHI types, except for numbers. The highest precision of 0.99 was reached for geographic locations and organizations, followed by dates, addresses and numbers. When partially recognized and wrongly tagged personal names are treated as true positives, the precision of their de-identification is better. With respect to the recall, the most important measure for de-identification, dates have the highest rate. Beside dates, almost all PHI categories showed high sensitivity rating from 0.99 to 0.93. The lowest recall rate for numbers (0.78) and addresses (0.85) suggests that rules for corresponding categories have to be improved. In terms of recall, especially dates and personal names, we may say that our de-identification is sufficient to guarantee high patient privacy, with achieved competitive precision and preserved document usefulness for subsequent applications.

6 Conclusion

In this paper, we presented the automatic text de-identification system for medical narrative texts, based on the adaptation of the existing rule-based NER system for Serbian. We have also produced the first versions of de-identified medical corpus that could be useful to the research community interested in both analysing different medical phenomena and producing a machine-learning automatic de-identification system for Serbian.

Even though the evaluation of the presented system is conducted on a relatively small set of documents, we have collected the heterogeneous corpus, consisting of different document types belonging to various medical specialties and institutions. The results of this preliminary evaluation are very promising, indicating that our adapted NER system can achieve high performance on the de-identification task. However, there is still much to be done.

In the future, we plan to focus on improving our strategies, such as completing the existing and adding the new patterns covering the broader formats of PHI (email addresses, URLs, IP address numbers) and the disambiguation of clinical eponyms and abbreviations. Finally, we intent to measure the impact of the de-identification through the subsequent natural language processing task of medical concepts' recognition.

References

[1] Health Insurance Portability and Accountability Act. P.L. 104-191, 42 USC. 1996.

[2] Douglass, M., Clifford, G. D., Reisner, A., Moody, G. B., Mark, R. G. 2004. Computer-assisted de-identification of free text in the MIMIC II database. Computers in Cardiology, 31:341-344.

[3] Neamatullah, I., Douglass, M. M., Lehman, L.-W. H., Reisner, A., Villarroel, M., Long, W. J., Szolovits, P., Moody, G. B., Mark, R. G. & Clifford, G. D. 2008. Automated de-identification of free-text medical records. BMC Medical Informatics and Decision Making, 8:32.

[4] Deleger, L., Molnar, K., Savova, G., Xia, F., Lingren, T., Li, Q., Marsolo, K., Jegga, A., Kaiser, M., Stoutenborough, L. & Solti, I. 2013. Large-

scale evaluation of automated clinical note de-identification and its impact on information extraction. Journal of the American Medical Informatics Association, 20:84-94.

[5] Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S. & Samore, M. H. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Medical Research Methodology, 10:70.

[6] Nadeau, D. & Sekine, S. 2007. A survey of named entity recognition and classification. Lingvisticae Investigationes, 30:3-26.

[7] Wellner, B., Huyck, M., Mardis, S., Aberdeen, J., Morgan, A., Peshkin, L., Yeh, A., Hitzeman, J. & Hirschman, L. 2007. Rapidly retargetable approaches to de-identification in medical records. Journal of the American Medical Informatics Association, 14**:**564-573.

[8] Gardner, J. & Xiong, L. 2008. HIDE: An integrated system for health information DE-identification. In: Proceedings of the 21st IEEE International Symposium on Computer-Based Medical Systems. 254-259.

[9] Uzuner, O., Sibanda, T. C., Luo, Y. & Szovits, P. 2008. A de-identifier for medical discharge summaries. Artificial Intelligence in Medicine, 42**:**13-35.

[10] Gardner, J., Xiong, L., Wang, F., Post, A., Saltz, J. & Grandison, T, 2010. An evaluation of feature sets and sampling techniques for de-identification of medical records. In: Veinot T, (ed.), Proceedings of the 1st ACM International Health Informatics Symposium. New York:ACM. 183-190.

[11] Sweeney, L. 1996. Replacing personally-identifying information in medical records, the Scrub system. Proc AMIA Annu Fall Symp, 333-337.

[12] Taira, R. K., Bui, A. T. A. & Kangarloo, H. 2002. Identification of patient name references within medical documents using semantic selectional restrictions. Proc AMIA Annu Symp, 757-761.

[13] Ruch, P., Baud, R. H., Rassinoux, A. M., Bouillon, P. & Robert, G. 2000. Medical document anonymization with a semantic lexicon. Proc AMIA Symp, 729-733.

[14] Ferrández, Ó., South, B. R., Shen, S., Friedlin, F. J., Samore, M. H. & Meystre, S. M. 2013. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. Journal of the American Medical Informatics Association, 20**:**77-83.

[15] Gupta, D., M. Saul, and J. Gilbertson. 2004. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. American Journal of Clinical Pathology 121 (2):176-186.

[16] Aramaki, E., Imai, T., Miyo, K., Ohe, K. Automatic deidentification by using sentence features and label consistency. In: Workshop on challenges in natural language I2b2 processing for clinical data. Washington, DC; 2006.

[17] Beckwith, B. A., R. Mahaadevan, U. J. Balis, and F. Kuo. 2006. Development and evaluation of an open source software tool for deidentification of pathology reports. BMC Medical Informatics and Decision Making 6.

[18] Guo, Y., Gaizauskas, R., Roberts, I., Demetriou, G., Hepple, M. 2006. Identifying personal health information using support vector machines. In: Workshop on challenges in natural language I2b2 processing for clinical data. Washington, DC; 2006.

[19] Friedlin, F. Jeff, and Clement J. McDonald. 2008. A software tool for removing patient identifying information from clinical documents. Journal of the American Medical Informatics Association 15 (5):601-610.

[20] Berman, J. J. 2003. Concept-match medical data scrubbing - How pathology text can be used in research. Archives of Pathology & Laboratory Medicine, 127**:**680-686.

[21] Morrison, F. P., Lai, A. M. & Hripcsak, G. 2009. Repurposing the Clinical Record: Can an Existing Natural Language Processing System De-identify Clinical Notes? Journal of the American Medical Informatics Association, 16**:**37-39.

[22] Meystre, S. M., Ferrández, Ó., Friedlin, F. J., South, B. R., Shen, S. & Samore, M. H. 2014. Text de-identification for privacy protection: A study of its impact on clinical text information content. Journal of Biomedical Informatics. doi: 10.1016/j.jbi.2014.01.011.

[23] Aberdeen, J., Bayer, S., Yeniterzi, R., Wellner, B., Clark, C., Hanauer, D., Malin, B. & Hirschman, L. 2010. The MITRE Identification Scrubber Toolkit: Design, training, and assessment. International Journal of Medical Informatics, 79**:**849-859.

[24] Kushida, C. A., Nichols, D. A., Jadrnicek, R., Miller, R., Walsh, J. K., & Griffin, K. 2012. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. Medical care, 50(Suppl 1):S82-S101.

[25] Uzuner, O., Luo, Y. & Szolovits, P. 2007. Evaluating the state-of-the-art in automatic de-identification. Journal of the American Medical Informatics Association, 14:550-563.

[26] Benton, A., Hill, S., Ungar, L., Chung, A., Leonard, C., Freeman, C. & Holmes, J. H. 2011. A system for de-identifying medical message board text. BMC Bioinformatics, 12(Suppl 3):S2.

[27] Courtois, B., Silberztein, M. 1990. Dictionnaires électroniques du français. Larousse, Paris.

[28] Gross, M. 1989. The use of finite automata in the lexical representation of natural language. Lecture Notes in Computer Science, 377:34-50.

[29] Paumier, S. 2011. Unitex 3.0 User manual. http:// http://www-igm.univ-mlv.fr/~unitex/UnitexManual3.0.pdf.

[30] Krstev, C. Processing of Serbian – Automata, Texts and Electronic dictionaries. Faculty of Philology, University of Belgrade, Belgrade, 2008.

[31] Krstev, C., Obradović, I., Utvić, M. & Vitas, D. 2014. A system for named entity recognition based

on local grammars. Journal of Logic and Computation, 24:473-489.

[32] Maurel, D., Friburger, N., Antoine, J. Y., Eshkol-Taravella, I. & Nouvel, D. 2011. Transducer cascades surrounding the recognition of named entities. Cascades de transducteurs autour de la reconnaissance des entités nommées, 52:69-96.

[33] Makhoul, J., Kubala, F., Schwartz, R., & Weischedel, R. 1999. Performance measures for information extraction. In: Proceedings of DARPA broadcast news workshop. 249-252.

[34] El Emam, K., Dankar, F. K., Issa, R., Jonker, E., Amyot, D., Coco, E., Corriveau, J.P., Walker, M., Chowdhury, S., Vaillancourt, R., Roffey, T., & Bottomley, J. 2009. A globally optimal k-anonymity method for the de-identification of health data. Journal of the American Medical Informatics Association, 16:670-82.