Automatic de-identification of protected health information

Jelena Jaćimović*†, Cvetana Krstev*, Drago Jelovac†

* University of Belgrade, Faculty of Philology
Studentski trg 3, 11000 Belgrade, Serbia
jjacimovic@rcub.bg.ac.rs
cvetana@matf.bg.ac.rs

†University of Belgrade, School of Dental Medicine
Dr. Subotića 8, 11000 Belgrade, Serbia
drago.jelovac@stomf.bg.ac.rs

Abstract

This paper presents an automatic de-identification system for Serbian, grounded on a rapid adaptation of the existing named entity recognition system. Based on a finite-state methodology and lexical resources, the system is designed to detect and replace all 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.

Avtomatska dezidentifikacija zaščitenih zdravstvenih podatkov

V prispevku predstavimo sistem za avtomatsko dezidentifikacijo v srbščini, ki temelji na hitri prilagoditvi obstoječega sistema za identifikacijo imenskih entitet. Sistem je zasnovan na metodologiji končnih avtomatov in jezikovnih virov ter identificira in zamenja vse eksplicitne zaščitene zdravstvene osebne podatke v medicinskih narativnih besedilih, pri čemer pa ohrani relevantne medicinske koncepte. Rezultati preliminarne evalvacije so pokazali uporabnost te metode, in sicer tako pri zaščiti osebnih podatkov pacientov kot pri interoperabilnosti dezindentificiranih dokumentov.

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.

A de-identification is focused on detecting and removing/modifying all explicit personal Protected Health Information (PHI) present in the 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 the de-identification, but the most frequently referenced regulation is the US Health Information Portability and Accountability Act (HIPAA) (HIPAA, 1996). According to the HIPAA "Safe Harbor" approach, the clinical records are considered de-identified when 18 categories 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 manual removal of PHI by medical professionals proved to be prohibitively time-consuming, tedious, costly and unreliable (Douglass et al., 2004; Neamatullah et al., 2008; Deleger et al., 2013), extracting PHI requires more reliable, faster and cheaper automatic de-identification systems based on Natural Language Processing (NLP) methods (Meystre et al., 2010).

The extraction of PHI can be viewed as a Named Entity Recognition (NER) problem applied in medical domain for the de-identification (Nadeau, 2007). However, even though both traditional NER and the de-identification involve the automatic recognition of particular phrases in text (persons, organizations, locations, dates, etc.), the de-identification differs in important ways from traditional NER (Wellner et al., 2007). In contrast to general NER focused on newspaper texts, the 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 the 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 subsequent automatic processes performed on the de-identified documents.

In this paper we introduce our automatic clinical narrative text de-identification system, based on a rapid

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

adaptation of the existing 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 the 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. The extensive review of recent research in the automatic de-identification of narrative medical texts is given in (Meystre et al., 2010). 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 (Gardner, 2008; Neamatullah et al., 2008; Uzuner et al., 2008; Gardner et al., 2010), while only a few of them were evaluated on a larger scale, with a more heterogeneous document corpus (Sweeney, 1996; Taira, 2002; Ruch at al., 2000; Ferrández et al., 2013). The selection of targeted PHI varied from patient names only (Taira, 2002) to all 17 textual HIPAA PHI categories (Aramaki et al., 2006; Neamatullah et al., 2008; Wellner et al., 2007), or even everything but valid medical concepts (Berman, 2003; Morrison et al., 2009).

The de-identification approaches applied in medical domain are mostly classified into the rule-based or machine learning methods, while some hybrid approaches (Ferrández et al., 2013) efficiently take advantage of both previous methods. The rule-based methods (Neamatullah et al., 2008; Morrison et al., 2009) 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 (Meystre et al., 2014). On the other hand, the machine-learning methods (Aramaki et al, 2006; Wellner et al., 2007; Gardner, 2008; Uzuner et al., 2008; Aberdeen et al., 2010), 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 often unpredictable effects of a change. 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 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 (Uzuner et al., 2007).

Aside from systems specifically designed for the de-identification purpose, some NER tools trained on newspaper texts also obtained respectable performance with certain PHI categories (Benton et al. 2011, Wellner et al., 2007).

3. Materials 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.

The characteristic of medical narratives confirmed in our corpus are 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 are consisting of lexical resources and local grammars developed using the finite-state methodology as described in (Courtois and Silberztein, 1990; Gross, 1989). For development and application of these resources the Unitex corpus processing system is used (Paumier, 2011). 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 processed texts (Krstev et al., 2013). 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. The Serbian NER system is organized as a cascade of FSTs – CasSys (Maurel et al., 2011), 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:s1v}, 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):

(1) {Dom zdravlja "\{Milutin Ivković\, \.NE\+persName \+full\\:s1v\}",.NE+org+:1sq:2sq:7sq:3sq:4sq:5sq:6sq}

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 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 Bramslav Dimitnjević, specijalista za stomatološku protetiku i ortopediju 'Prof. PhD Bramslav Dimitnjević, a specialist for Prosthetic Dentistry and Orthodontics'*

The finite-state transducers used in the NER cascade use beside general and specific e-dictionaries, as explained before, local grammars that model various triggers and NEs context, 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 if grammatical information were 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)).

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

3.3. The PHI de-identification

We used our training corpus for creation and adaptation of patterns that will capture the characteristics of PHI. Through the corpus examination 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 number). 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 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 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.

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 the 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.). 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 (3).[3] The same expertise after the de-identification and tagging is given in Example (4).[4]

(3) Vaš broj Posl. Br. Ki 250/08
Naš broj 33/06
OPŠTINSKI SUD Istražni sudija G-đa Rada Anđelić-
Vašom naredbom zatražili ste od Komisije lekara veštaka Medicinskog fakulteta Univerziteta u Nišu sudsko medicinsko veštačenje u predmetu Ki 250/08 na okolnost vrste, težine i mehanizma nastanka povreda koje je dana 20.03.2008. god. zadobio oštećeni Marković Ivan iz Dragačeva.
…
PODACI
1. Pri pregledu obavljenom dana 11.08.2007. god. od strane članova Komisije lekara veštaka Medicinskog fakulteta u Nišu Ivana, Mirka, Marković navodi da je rođen 13.05.1986. god. u Dragačevu, živi u Dragačevu, ul. Dositejeva br. 27, po zanimanju elektromehaničar za teničke i rashladne uređaje. Identitet imenovanog utvrđen je na osnovu članovima komisije pokazane lične karte br. 82193. Ivan takođe navodi da je krajem marta meseca 2008. god. oko 1 h posle ponoći sa svojim drugovima sedeo u parku ispred hotela gde je u toku bilo svadbeno veselje.
…
NALAZ
1. U izveštaju doktora Opšte bolnice u Užicu na ime Markoić Ivana, broj protokola 01241, izdatom dana 21.03.2006. god. u 2,30h navedeno je sledeće:" Fractura dens 2 traumatice (dalje nečitko) upućuje se stomatologu radi daljeg lečenja i kvalifikacije povrede "
…
(4) Vaš broj <number PHI="yes">XXXX</number>
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>-
Vašom naredbom zatražili ste od
<org PHI="yes">Komisije</org>
<org PHI="yes">fakulteta</org>
<org PHI="yes">Univerziteta</org> sudsko medicinsko veštačenje u predmetu
<number PHI="yes">XXXX</number> na okolnost vrste, težine i mehanizma nastanka povreda koje je dana
<date PHI="yes">26.09.2007.</date> zadobio oštećeni
<persName.full PHI="yes">Barni Kamenko</persName.full> iz
<top.gr PHI="yes">Kamengrada</top.gr>.
…
PODACI
1.{S} Pri pregledu obavljenom dana <date PHI="yes">17.02.2008.</date> od strane članova <org PHI="yes">Komisije</org>
<org PHI="yes">fakulteta</org>
<persName.full PHI="yes">Barni Kamenko</persName.full> navodi da je rođen
<date PHI="yes">19.11.1987.</date> u
<top.gr PHI="yes">Kamengradu</top.gr>, živi u
<top.gr PHI="yes">Kamengradu</top.gr>, <address PHI="yes">ul. Kamenolomska br. 6a</address> , po zanimanju elektromehaničar za teničke i rashladne uređaje.{S} Identitet imenovanog utvrđen je na osnovu članovima komisije pokazane lične karte <number PHI="yes">XXXX</number>.{S} **Ivan** takođe navodi da je krajem <date PHI="yes">septembra 2007.</date> oko 1 h posle ponoći sa svojim drugovima sedeo u parku ispred hotela gde je u toku bilo svadbeno veselje.
NALAZ
<number PHI="yes">XXXX</number>.{S} U izveštaju doktora <org PHI="yes">bolnice</org> na ime <persName.full PHI="yes">**Vilma Kremenko**</persName.full>, broj protokola <number PHI="yes">XXXX</number>, izdatom dana <date PHI="yes">27.09.2007.</date> u 2,30h navedeno je sledeće:{S}" Fractura dens 2 traumatice (dalje nečitko) upućuje se stomatologu radi daljeg lečenja i kvalifikacije povrede "

This example demonstrates our de-identification approach. 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. These occurrences are bolded and underlined in Example (4).

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 present completely unseen material containing many occurrences of PHI. Details about the PHI distribution within the test corpus can be found in Table 1.

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

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 correctly

[3] This example looks exactly as the original – however, for the purpose of protecting the personal data we have manually replaced all of it.

[4] We wanted to avoid 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.

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, personal name tags even though the utterance belongs 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 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, no matter for what reason (including the incorrect input). With 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).

	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)

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

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
address	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
Address	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

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 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) requires inclusion of missing patterns for these categories. 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.

5. Conclusion

In this paper, we described the automatic text de-identification system for medical narrative texts, based on the rapid adaptation of the existing NER 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. 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 work, we plan to focus on improvements of 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.

6. References

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.

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.

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.

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.

Courtois, B., Silberztein, M. 1990. *Dictionnaires électroniques du français*. Larousse, Paris.

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.

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.

Ferrández, O., 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.

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.

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.

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

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

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.

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.

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.

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.

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.

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

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.

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

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.

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

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.

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.

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

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.