clinical text data,
data de-identification,
natural text processing

Małgorzata MARCINIAK[1], Agnieszka MYKOWIECKA[1], Piotr RYCHLIK[1]

MEDICAL TEXT DATA ANONYMIZATION

The paper discusses a program for removing patient identification information from hospital discharge documents in order to make them available for scientific research e.g. information extraction system designing. The presented method allows de-anonymization of documents using a key-code file that is created on the basis of a patient's surname, forename and date of birth. Problems of normalization of crucial data used in the key-code file creation are presented.

1. INTRODUCTION

Clinical information describing patients health conditions can be expressed by various types of data. Most examinations results are given by means of numbers and images but crucial information is also given by means of free natural language text. Text form is used for describing patient's medical history and results of clinical examinations. Examination results which are given in a form of images are usually accompanied with text interpretations. All these medical texts include knowledge which can be used not only to help a particular patient but also for statistical and scientific purposes. For statistical analysis it can be sufficient to make appropriate aggregations and combine them with those coming from other sources, but for scientific research it would be much more desirable to have detailed data available. This approach would give an opportunity to analyze data in various, not foreseen aspects. The problem is that clinical information is hardly available for anyone except a physician who cares about a particular patient. Making clinical data available for research purposes is limited not only by the lack of will to share these data but, what is more important, by privacy issues. Clinical data include information serving for identification purposes which have to be eliminated before making the entire dataset available. However, this seems to be a quite straightforward task, in practice there are many decisions which have to be undertaken during this process. Their importance differs according to the type of data and the planned scope of data availability. The most important issues are:

- deciding what types of information have to be removed (names, dates, …),
- deciding what should they be replaced by,
- providing means for identification documents concerning the same person (important especially for chronic diseases).

Although it is relatively easy to point out particular information items which can identify a person before making the data public, one also should take into account combinations of data which make identification possible. Personal identification number identifies a person unambiguously, but one has to have access to the appropriate database; forename and surname are less precise but very discriminating and when combined with the address or date of birth are almost always unique. Generally, the more data we have, it is more likely that combinations of several items may make identification possible. This is especially true for patients with rare diseases. However, removing selected data should be done with care as they can be important for further analysis.

In this paper we present the problems we encountered while preparing data for information extraction task. We decided to implement temporal de-identification of patients (not permanent anonymization) as the extracted data might be designated for the further use in the hospital which is the source of the data. Identification of a person is also important if we want to check whether two documents concern the same patient, or if we want to eliminate inclusion of the same or slightly modified documents in the dataset.

[1] Instytut Podstaw Informatyki PAN, ul. J.K. Ordona 21, 01-237 Warszawa

The medical data anonymization problem has already been noted by authorities (e.g. the report written for the Access to Information and Privacy Division of Health Canada, [1], or the report to US Congressional Requesters [8]) and addressed by researchers (a short description of several works can be found in [7]). Most systems were implemented for English texts and use permanent anonymization which means that personal data is completely removed. Scrub system [5] used a set of a word-based templates and lexicons. Ruch [4] presented a system which used some NLP tools such as a medical semantic lexicon and a word-sense tagger. In [6] a maximum entropy classifier to indentify personal data was trained on the basis of a manually tagged data set. In [3] there was presented a Perl-based de-identification software package using look-up tables, regular expressions and heuristics. The system achieved 0.96 precision and 0.749 recall on a set of English nursing notes containing 1779 personal health information occurrences. For Polish data, there is a system for removing personal information from a dataset consisting of security text messages [2].

Below we present a rule based de-identification program. We describe our formulation of the anonymization task, a data set we worked on, the main encountered problems as well as accepted solutions.

2. TASK DESCRIPTION

We present a task which consists in pre-processing of a set of discharge summaries from one of the Warsaw hospitals so that they will be available for research purposes. The specific goal of this work was to prepare a data set which can be used to create and test information extraction system which is planned to be used to process both historical and current discharge records. The system will select from documents the most important medical information about patient's illnesses, results of examinations, and treatment; and insert them into a hospital database. Like many other Polish healthcare institutions this hospital does not use any computer system for storing medical documents. Most of the documentation is still in a paper form but discharge summaries are created in an electronic form and can be further processed. Unfortunately, although it is generally agreed what should be placed within a discharge record, there is no standard form of these documents – each ward or even each physician has its own style of writing them. The use of different word processors is also a common practice. The task which will be described below consists in temporal de-identification of data.

Original hospital documents were written in various versions of MS Word editor and in Open Office Writer editor. One document describes a visit of one patient in hospital. A document contains information that allows identification of the patient: name, address, birth date, PESEL (national identification number of citizens used in Poland). These data must be removed from documents before making them accessible even for scientific purposes. Moreover, files that contain discharge documents are usually named by the patient's name, so these filenames must also be changed. Manual anonymization of documents is labor-intensive and has to be done by a person authorized to have access to the personal data according to the Privacy Rule. Unfortunately, it is very probable that a person which makes it manually, having to open and edit hundreds of files might leave some documents unprocessed making them available in an original form, or might anonymize them partly. To overcome this problem we decided to prepare a program that automatically changes filenames and removes identification data from documents. The program converts all documents into MS Word 2007 standard and from each document prepares a text file in UTF-8 encoding in order to allow further processing of data.

A typical discharge record of a patient has a text header which contains information about a hospital (name, address, telephones), and the ward where a patient was hospitalized. It may also contain figures e.g., logo of the hospital. Sometimes, information about the hospital and the ward are enclosed in the figure containing logo on the top of the first page of a given document. In some documents a hospital ward name is mentioned only in the document header. So the name of a hospital ward included in the header should be extracted to the text file.

Identification number of the patient's visit in hospital is usually given just after the header followed by the place and date of the document issue (the date is consistent with the last day of hospital visit). Sometimes, identification number of the visit to the ward is added. All these data are left unchanged in the document.

The next section of the document contains identification information of the patient. It contains a subset of the following information: surname, forename, address, PESEL, birth date, age and dates of the patient's visit in hospital. The easiest way of anonymization is just removing the whole identification part of the document, but some of this information may be important. For example, if we want to investigate correlation between the patient's sex, age and illness, the appropriate data should be available. Usually, the patient's sex is not explicitly given in documents but it could be established on the basis of available information.

In the case of statistical analysis based on the hospital documents it is important to be able to correlate documents of the same person. It is necessary if we want to monitor treatment of a patient with a long-lasting illness like diabetes, allergy, or rehabilitation progress of cerebral palsy patients. It is indispensable for the hospital to be able to de-anonymize documents. For example, it might be necessary to contact patients who suffer from a particular disease. Their descriptions will be selected from a database created on the basis of medical data extracted from documents. Therefore, it is necessary to insert the patient's code (that substitutes identification information) into a document, and to create a file that allows to correlate the code with the identification data of the patient. The key-code file must be stored in the hospital data carrier. The same file should be used every time when new data are collected to correlate several documents of the same patient.

3. IMPLEMENTATION

The task described in the previous section is more complicated than it appears at first glance. We encountered several unpredicted difficulties during creation of the program so it had to be corrected several times before it was possible to process majority of discharge records. The program is written in Java language using standard Java 1.6 libraries. Open source library dom4j has been used to facilitate working with XML files. To convert binary Microsoft Office formats into Open XML format introduced in the MS Office 2007 we use the command-line tool called Office File Converter (OFC.EXE) which is included in the Office Migration Planning Manager available as a free download.

3.1. HOW TO CREATE A KEY-CODE FILE

The key-code file should be constructed in the way that allows to link all discharge documents of the same patient. The best method is to use PESEL that unambiguously identifies a citizen. After the inspection of a sample of discharge documents with fictitious identification data it turned out that not all documents contained PESEL. The substantial problem was that in hospitals which do not have a professional program for hospital service (where data about a patient, results of examinations, diagnosis and treatments are not transferred by the hospital intranet) discharge documents are not standardized and contain various patient's identification information edited in various ways. In the sample data it turned out that all documents contain surname, forename and address of a patient. PESEL was usually present but sometimes only the date of birth was given. As far as children hospitals with wards for newborn infants are concerned the use of PESEL as the patient identification code is irrelevant (they do not have this number assigned yet).

The method of key-code creation must be homogeneous for all documents. As often happens that the same patient obtains a discharge document with PESEL or without PESEL during two successive hospital visits we have resigned from using PESEL as the key. Finally, we decided that the method of the patient's identification by the surname, forename and date of birth is the best we can use even though, theoretically, it is not perfect.

3.2. DATA NORMALIZATION

All data that need to be removed from a given document and substituted by the patient's code are located between the title of the document e.g., Information Card, and the next section of the document which is usually called Diagnosis. The example illustrating the above is given in (1).

Nazwisko i imię: Kowalski Jan **ur. 09.01.2008r.** **PESEL: 08210999999**
'Name and forname:' ' birth date' 'PESEL'

Miejsce zamieszkania: 02-982 Warszawa, ul. Wąska 227 m 99 (1)
'Address:'

Przebywał/a w szpitalu: od 19.02.2011r. do 28.02.2011r.
'Admitted to hospital: from … to'

Data presented in (1) are substituted by (2) and in the key-code file there is information that KOWALSKI_JAN_09012008 key refers to M080000 patient's code.

Kod pacjenta: M080000 **ur.01.2008**
'Patients code:' ' birth date' (2)

Przebywał/a w szpitalu: od 19.02.2011r. do 28.02.2011r.
'Admitted to hospital: from … to'

The day from the patient's birth date has been removed to make the group of potential patients bigger. The month of birth has been left because it allows to calculate the precise age of a patient, which is important in case of small children. Address and PESEL have been removed. If a document contains PESEL but doesn't contain the date of birth, the last information is extracted from the PESEL, and inserted to the output document. The other information extracted from the PESEL is sex, which is coded as the 10th digit. Our patient's codes for males start with M, for females with K. If the PESEL is not available we try to identify the sex of a patient using the list of forenames correlated with grammatical gender. Our lists contain about 700 forenames, and do not include rare Polish names like *Protazy*, nor foreign names like *Kevin*. The next method of sex identification is looking for specific words in a document e.g.: *pacjent* 'patient$_{masc}$', *pacjentka* 'patient$_{fem}$', *chłopiec* 'boy', *dziewczynka* 'girl'. If it is impossible to determine the sex of a patient on the basis of any of the above methods, the code starts with N.

The surname and the forename are used as a part of the identification code. Sometimes they are typed in capital letters. To allow unification of strings: *Kowalski Jan* and *KOWALSKI Jan* we decided to convert them into capital letters. It is also necessary to take into account the order in which surname and forename occur. In most documents a surname precedes a forename but in some reports, they appear in the reverse order, so the program should code strings: *Jan Kowalski* and *Kowalski Jan* as the same string: KOWALSKI_JAN. We establish the order used in a given document on the basis of the opening phrase: *Surname and forename* or *Forename and surname*. The other method is to use a list of forenames and determine which string is on the list. This solution will not work out if the surname of a person can be interpreted as a forename like: *Maciej, Wacław* or *Wanda*. If above methods failed we use the same order as in the most recently processed document.

Birth date is the next crucial information that needs to be normalized by the program because it is a part of the patient's identification information. We have to identify that the following strings: *5.02.2001, 05.02.01, 05.02.2001, 05-02-2001, 05. 02 .2001* relate to the same date and are assigned to the same identification string: *05022001*. In our test data set all dates were written in the following order: day-month-year, but we also ought to take the reverse order into account. Unfortunately, sometimes there are typographical errors in dates. The most common is lack of punctuation marks, e.g. *0502.2001, 05.022001*. Usually such incorrect dates are understandable for a person but difficult to recognize by a computer

program. Dates recognition is very important during the anonymization process, so in the next version of the program we plan to implement a method of date correction.

In the cases where no date of birth is recognized in a document in any form (nor in PESEL), the document is not converted into a pure text file, and is not included into the data set available for further processing.

3.3. FILE NAMES

As it was mentioned in section 2 all file names should be changed during anonymization process since a discharge file name typically contains the surname and the forename of a patient. In the case of several visits of the same patient to the same hospital ward, the files are usually differentiated by years/month of the visit or its subsequent number. In order to make selection of all files related to the same patient easy and to differentiate names of successive visits in hospital the anonymized documents are named according to the combination of the patient's code and the date of document extracted from the header. If the date of document is not recognized (is not put in the document or because of a typographical error) we use the last day of hospital visit instead.

4. CONCLUSIONS

De-identification is a sort of an information extraction task. Our program extracts patient identification data looking up for key-phrases which accompany the important information we want to extract. This data can be given in various order. A part of identification data is used for key creation, but other are simply removed. A de-identification program should be flexible and take into account different ways in which important information is given. In case of data that is necessary for key creation we try to predict all possible formats. The program checks data consistency, e.g. PESEL contains the date of birth which usually is also given, so it is possible to compare these two dates. It checks also correctness of some data, e.g. whether PESEL contains exactly 11 digits (sometimes a digit is omitted or duplicated) and if its checksum (eleventh digit) is correct. We decided that documents in which the crucial identification data is missing are not processed. The program informs about problems encountered during the file processing. The reason for this is to identify incomplete or erroneous data.

We tested the program on 300 documents from 6 wards, results are given in Table 1. From each ward 50 documents were taken. We prepared the program analysing about a dozen of documents from each ward.

Table 1. Number of processed documents.

	Ward 1	Ward 2	Ward 3	Ward 4	Ward 5	Ward 6
anonymized documents	50	43	49	48	49	49
unpredicted format of data	0	3	1	1	0	0
irrelevant or lack of information	0	3	0	1	0	1
duplicated documents	0	1	0	0	1	0

96% of documents was successfully processed and all patient identification data were removed. Five documents were not processed because we did not predict all data formats, e.g. our program did not accept months as roman numerals: *02.II.2005*. Three documents from the Ward 2 were not patients discharge reports, another two documents did not contain crucial data necessary for the key-code file. Two documents were not processed because they were duplicates of already processed documents.

Table 2 shows how many different patient codes were represented in our data. Our program was tested on fragments of catalogs with hospital discharged documents. The documents were in alphabetic order, so the probability that they contain documents of the same patient was high.

Table 2. Number of processed documents.

	Ward 1	Ward 2	Ward 3	Ward 4	Ward 5	Ward 6
documents	50	43	49	48	49	49
different codes	48	32	43	43	37	41

During the test phase, we encountered that in documents from the Ward 5 patient codes were incorrectly constructed, a forename was the first element of a code in 31 cases. It turned out that in these documents, the opening phrase was: *Imię i nazwisko* 'Forename and surname' but the order of data was opposite. The same problem was in two documents from another ward. In 12 documents from the Ward 3, the opening phrase was *Nazwisko i nazwisko* 'Surname and surname' which also may cause similar problems. So we decided to change the algorithm that recognizes names. First, we try to recognize which string is a forename and a surname on the basis of the lists of forenames. The opening phrase is taken into account if that method failed.

The sex was not identified in 2 documents where all methods of sex identification failed. In one case a forename was typed with an error. In the second one a newborn child had "S" instead of a name, which indicates a son, so a male code should be assigned.

Automatic removing patient identification data is a very important task solving of which will make easier clinical data sharing. Unfortunately, natural language texts diversity makes it impossible to come up with a universal solution to this problem. The program described in this paper can be relatively easy adapted to process free text discharge records from other hospitals. De-identification of medical texts of a different kind would need more changes. To make sure that the de-identification procedure is reliable, for every new type of data the dedicated tests are required.

BIBLIOGRAPHY

[1] EL EMAM, K. Data Anonymization Practices in Clinical Research. A Descriptive Study, University of Ottawa, 2006.
[2] GRALIŃSKI, F., JASSEM K., MARCIŃCZUK M., WAWRZYNIAK P., Named Entity Recognition in Machine Anonymization, *Recent Advances in Intelligent Information Systems*, M. A. Kłopotek, A. Przepiórkowski, S. T. Wierzchoń, K. Trojanowski (Eds.), EXIT, Warszawa, 2009.
[3] NEAMATULLAH, I., DOUGLAS M., LEHMAN L., REISNER A., VILLAROEL M., LONG W., SZOLOVITIS P., MOODY G., MARK R., CLIFFORD G., Automated de-identification of free-text medical records, BMC Medical Informatics and Decision Making, 2008, pp. 8-32.
[4] RUCH, P., BAUD R., RASSINOUX A-M., BOUILLON P., ROBERT G., Medical Document Anonymization with a Semantic Lexicon, Proc. AMIA Symp., 2000.
[5] SWEENEY, L. Replacing Personally-Identifying Information in Medical Reports, the Scrub System. Proc. AMIA, 1996.
[6] TAIRA, R., BUI A., KANGARLOO H., Identification of patient name references within medical documents using semantic selectional restrictions, Proc AMIA Symp. 2002, pp. 757–761.
[7] TVEIT, A., EDSBERG O., RØST T., FAXVAAG A., NYTRØ Ø., NORDGÅRD T., RANANG M., GRIMSMO A., Anonymization of General practitioner medical Records, HelsIT Workshop, Trondheim, 2004.
[8] US GENERAL ACCOUNTING OFFICE, Medical Records Privacy, GAO/HEHS-99-55, 1999.