Medical Document Anonymization with a Semantic Lexicon

Patrick Ruch, Robert H. Baud, Anne-Marie Rassinoux; Pierrette Bouillon, Gilbert Robert
Medical Informatics Division, University Hospital of Geneva; ISSCO, University of Geneva

We present an original system for locating and removing personally-identifying information in patient records. In this experiment, anonymization is seen as a particular case of knowledge extraction. We use natural language processing tools provided by the MEDTAG framework: a semantic lexicon specialized in medicine, and a toolkit for word-sense and morpho-syntactic tagging. The system finds 98-99% of all personally-identifying information.

INTRODUCTION AND BACKGROUND

For centuries, care providers have been taking written notes about their patients. But, with the age of the electronic patient record (EPR), the very confidential relationship between the clinician and the patient may reache its limits, if hundreds of professionals have access to this information; even if such access is legitimate: as for example the retrieval of similar cases in clinical document warehouses[1]. Research on corpora, as it needs both important amounts of textual data and frequent cooperation with more specialized groups, tends to extend this problem out of the medical sphere.

If we refer to some of the rare[2] works [1] studying document anonymization, our system is a data de-identifier -a 'scrubbing' system- likely to remove explicit identifiers[3] such as name, address, phone number, and date of birth. However such operation does not guarantee that the result is universally anonymous, i.e. that nobody will ever be likely to infer some information concerning some patients, for example by linking the document to other sources. Therefore the tool should be used together will more classical and legal procedural barriers.

From a technical point of view, the system is based on the MEDTAG lexicon [2] for the lexical resources, and on an original rule-based word-sense (WS) and morpho-syntactic (MS) tagger for the disambiguation task. The basic idea is to rely on markers to locate identifiers, but some markers may be ambiguous (i.e. they may not be markers in some contexts). Taggers are used to solve such ambiguities. Like the taggers described earlier [2], the toolkit we used is of Markovian type, i.e. it uses local grammar [3], but the first system was data-driven, while the present one is rule-based: disambiguation rules are written manually, as in the FACILE project[4]. A third component using formal recursive transition networks (RTN) [5] for extraction and removal of the confidential items has been specially applied.

OVERVIEW OF THE PROBLEM

The anonymization is seen as a particular case of knowledge extraction, as before removing the specific information, it is necessary to localize it. To show the complexity of the task, here are some excerpts[4]:

Miss Maria Christina GOMEZ DA LOVIS (0), born 3.3.1956 (1), without medical antecedents, but a caesarian section by Pfannentiel (2)
However, doctors Robert de Baud (3) and Anton Gicebuhler (4) as they know her well [...]
One stitches up by Donati (5), with Maxon (6) 4.0.
Doctors of the Geneva University Hospital (7)[...]
In the night, she decided to phone the EMS (8) [...]
Alan River, MD (9)

Table 1: Some examples of identifiers

2 and 5 are technical names. 0 is the full name of a patient, 1 is a date of birth. 3, 4 and 9 are physician's names; they are composed of several items. Let us note that the last name in 3 is starting with a lower case letter, while in 9, the last name is also a common noun. 6 is a medical device. 7 and 8 are health care institutions. In conclusion, in the sample only 0, 1, 3, 4 and 9 must be anonymized, whereas 2, 5, 6, 7 and 8 must be left intact. The replacement operation we have designed is very simple, it replaces each character of any confidential items by an 'x', and it respects the case (capital letters are replaced by an 'X'); punctuation (hyphen, dot...) occurring within confidential items is kept (see Fig. 2).

METHODS

Hopefully, among so many capitalized words, one very productive pattern emerges: an IDentity Marker (IDM, such as markers of politeness: *Sir*, *Mrs.*, *Miss.*,

[1] This tool is part of the University Hospital EPR.

[2] If medical papers are rare, the message understanding conferences (MUC) provide an interesting framework for name entity recognition.

[3] In the following, identifier refers to explicit personally-identifying items.

[4] Any print of real names is fortuitous.

or titles: *Dr.*, *doctors*) precedes -or follows as *MD* in case 9- each occurrence of patient and doctors identifiers. Notice that it is also true for 4, which is related to the marker *Doctors* by the coordination *and*. Unfortunately, if each occurrence of identifiers is preceded -or followed- by an IDM, the reverse is not true: we may have IDM, which are not followed - or preceded- by identifiers. As in 7, where *Doctors* refers to some humans, without indicating their names, or in 8, where *phone* is not followed by a digit but by the name of a medical institution. Here is the starting point of the study, and the task-orientated semantic separation between strict IDM (*Ms.*, *Dr.*...), which are always directly followed -or preceded- by identifiers, and tokens likely to refer to general persons (*doctors*, *professors*...), which are not necessary followed -or preceded by names. Again, this rule has never been violated within our corpora, instead, we found a lot of misspelled occurrences (*prfessor* for *professor*), and a lot of errors on mixing up small letters and capitalized letters. As for example in *Miss Jane Dermott* and *Mr. Lawrence van Belleghem*, which were incorrectly written *jane Dermott* and *Lawrence Van Belleghem*.

Concerning phone numbers and dates, together with explicit markers -similar to IDM- such as *born* (cf. case 1), we can also rely on very well defined patterns, which can be exhaustively listed. For example a token with 3 digits, followed by 4 digits, with a dash in-between, is considered as a phone numbers.

Methodological hypothesis

From an epistemic point of view, three main hypotheses are guiding the experiment:

a. syntax can help to distinguish meanings of words having different syntactic categories;
b. syntactic and semantic ambiguities can be solved using simple taggers
c. information extraction can also be tag-assisted

We earlier verified hypotheses a and b [2]. This experiment focuses particularly on hypothesis c, and reports on our attempt to apply the MEDTAG tag-like framework to the task of anonymization. These hypotheses have been tested in the following way: texts are first annotated with the MS tagger, in order to make explicit the part-of-speech (POS). This stage provides the first disambiguation filter at the MS level. Then, the WS tagger provides the WS tag. Finally, this information serves to extract and annotate the identifiers, via a formal RTN algorithm.

Choice of a corpus

Let us note that French is the regular working language at the Geneva University Hospital. However a small amount of documents is written in English (less than 1%), and an even smaller amount is written in German. This is particularly true with discharge summaries, mainly for patients leaving abroad, due to the international status of the city and the different languages spoken across the country. Therefore the system we designed is able to work in such multilingual environment. In the test corpus only two documents were written in English, but the examples are provided in English for sake of clarity.

<Header>

<mpi[5]> -423751	<name> Dr. P. GALLAWAY	
	<address> 56 Montaigne Av.	
	<city code> 1211 GENEVA	
<department> Surgery		<unit> Digestive Surgery
	<phone> Phone: 345-7343	
<place, date> Geneva, the 5th of December, 1999		

<Body>

Dear Dr. *Gallaway*,
Your patient, Mr. *A.-M. COGER*, born *11.8.1959*, stayed in our service, from *05/05/99* to *05/06/99*.

DIAGNOSIS :
Acute pancreatitis of unknown origin.
Comorbidities :
Gastritis.
Hypertension.
Esophagitis with backflow.

Mr. *Coger*, 72 year-old[6], has been admitted in emergency, after [...]. However tests for the cytomegalovirus and the EBV were negative. Therefore we perform an abdominal CT-scan [...] M. *Coger* will be followed in ambulatory by Drs. *David Ducruet* and *Robert van Belleghem* [...].

Figure 1: Example of a complete medical document

Standard documents within the University Hospital of Geneva include a header and a body (see Fig. 1). The header, where only structured data occurs, can be easily handled as we can process individually each field. So the document header has been automatically discarded in this experiment. We finally picked 1000 documents for a total of 80784 tokens: 600 post-operative reports, 200 laboratory and test results, and 200 discharge summaries. Two sets of documents have been extracted from this ad hoc corpus. The first set (set A, a representative sample with about 20% of the corpus, i.e. 16456 tokens) helped set up the system, whereas the second (set B, about 80% of the corpus, i.e. 64328 tokens) was used for assessment purposes. In the first set, we counted 124 identifiers.

[5] MPI stands for Master Patient Index.

[6] The age of the patient is considered as a clinical data, and therefore is not to be anonymized.

Adapting taggers and lexicons

The MEDTAG tagset (see Tab. 3) based on the UMLS Metathesaurus is seen as a basic ontology of the domain. With less than 40 tags, it aims at describing major semantic features related to the medical domain. The main target of the MEDTAG taggers is the word-sense disambiguation. Considering an ambiguous token, both taggers attempt to provide the right tag (first MS, then WS), as for example a token like *miss*, is ambiguous out of its contexts between an action (*to fail*) and a person (*a young lady*), but may become unambiguous in a particular context. However, applying both taggers to the anonymization task has implied some refinements, mainly on the WS tagger.

Lexical refinements

Some lexical information had to be added, but we decided not to modify the MEDTAG tagset, and we worked with lists of particular cases, in order to provide the anonymization-specific tags. Thus, we created two tags. First *idm*, which is attributed to lexemes such as *Dr.*, second, *id*, which is given to human proper nouns (*David*, *Louise*...). Few proper nouns are present in the lexicon, and this tag is mainly used for recognizing identifiers (cf. the distribution of these tags, together with the MEDTAG tagset in Tab. 3). It was also necessary to improve the coverage of the MEDTAG lexicon, which contains now 5131 entries. Thus, it was necessary to list more tokens likely to occur with a capital letter. As it is clearly impossible to put in the lexicon all proper names (tagged *id*), we decided to work on lexemes written with a capital letter, but which are not proper names. We focused on tokens referring to medical institutions (hospitals, clinics, ... tagged *hcorg* or *spec*). We also linked partially the MEDTAG lexicon together with the Swiss Compendium, which describes all the drugs used within the country. Finally, most of the medical devices (tagged *mdev*, such as CT scan, Doppler, Macron...) were also added to the MEDTAG lexicon.

Strategy

Basically, it is necessary to recognize if a given token is an IDM. However, for some of these markers, to recognize the occurrences is not enough, as they may appear without being followed by confidential items, as for example *Doctors* in 7. Therefore the class *pers* has been split into two tags: *idm* and *pers*. For example *Dr.* is unambiguously an *idm*, i.e. it is necessary followed by an name identifier, while *Doctors*, may be both an *idm*, if it is followed by an identifier, or a *pers*, if it is followed by anything else. A last group gathers tokens such as *miss*. *Miss* is ambiguous at an MS level between a common noun and a verb, before being ambiguous at a WS level, between an action and an IDM.

Hopefully, such ambiguities are rare: IDM are generally clearly defined at the MS level (they are common noun, tagged *nc*), as for example *Dr.*, *Mr.*, *Prof.*, *Professor*. And only some of them are ambiguous at the WS level (*Professor, Doctors*...) between and *idm* and a *pers*.

Disambiguation: MS and WS tagging

The output of the disambiguation task is a 3 level stream (see Tab. 2): the token, the MS tag, and the WS tag. Tab. 2 shows the MS tagging process: column 1 provides the token, as it appears in the corpus, column 2 provides the lexical tag-like morpho-syntactic information. Column 3 picks up the preferred MS tags, which contains the POS. Using the POS, a lexical access returns the hypothetical WS tags. Lexical ambiguities are separated by a '/'. Finally, column 5 provides the preferred WS tags.

Tokens	Lex. MS	MS	Lex. WS	WS
Miss	v/nc[s]	**nc[s]**	idm/pers	**idm**
L.	x	**x**	x	**id**
Mitchell	x	**x**	x	**id**
phones	v/nc[p]	**v**	act	**act**
the	det	**det**	def	**def**
Hospital	nc	**nc[s]**	hcorg	**hcorg**
[...]				
Doctors	nc[p]	**nc[p]**	idm/pers	**pers**
of	sp	**sp**	rel	**rel**
the	det	**det**	def	**def**
[...]				

Table 2: morpho-syntactic and word-sense disambiguation

MS tags: The symbol before the bracket means the part-of-speech (verb, noun, adjective...), while the symbol between the brackets provides optional information about the morpho-syntactic features (s for singular, p for plural) [7]
nc is the MS tag for common nouns,
v is the MS tag for verbs
det is the MS tag for determiners
sp is the MS tag for prepositions
x is the MS tag for unknown tokens

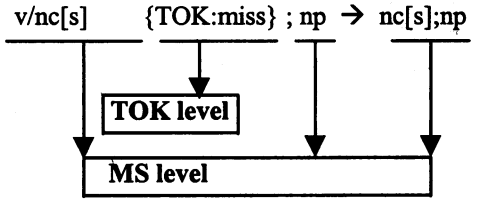
WS tags: see Tab. 3, in appendix.

[7] The MS tagset tends to follow the European MULTEXT standard: http://www.lpl.univaix.fr/projects/multext/LEX/LEX1.html

Writing the rules

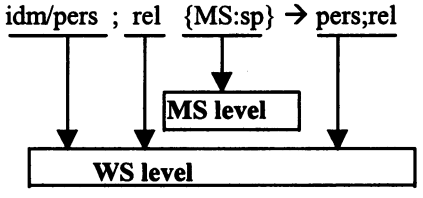
While adding more lexical information, it was necessary to write the disambiguation rules. The WS and the MS taggers use the same framework: a set of very contextual rules, applied on a short string of tokens (up to pentagrams), ranked according to their reliability. Here are two examples:

In Tab. 2, the following two-level rule is applied in order to disambiguate the token *miss*:



This rule means: 'At the MS level: A word ambiguous (/) between a *v* and a *nc[s]*, and whose the token form (between {}) is *miss*, and which is followed (;) by a *np*, must be rewritten (→) *nc[s]*'.

In Tab. 2, the following two-level rule is applied to disambiguate *Doctors*:



This rule means: 'At the WS level: A word ambiguous (/) between an *idm* and a *pers*, and which is followed (;) by a *rel*, which is also a *sp* at the MS level (between {}), must be rewritten (→) *pers*'.

Reliable and long-distance rules

Unlike most similar systems (see [6] and [7]), the output of both taggers (Tab. 2, columns 3 and 5) may still be ambiguous, if no relevant disambiguation rules can be applied. The basic idea is that we prefer not to choose, than to choose the wrong tag! When ambiguities (for example, between *idm* and *pers*) remain unsolved, the extraction module finally considers the token as being an *idm*. In this case, some tokens may be improperly removed, but as seen in the evaluation part, the system behaves very well.

Finally, the n-grams contextual rules are completed by a set of long-distance rules (coded as finite-state automata). Such rules are necessary for mastering coordination dependencies. Plural *idm* (*Drs*, *professors*...) are expected to be followed by more than one identifier, therefore at least one coordination is expected. In such case the ambiguous tag *idm/pers* is attributed to the next occurring coordination, and usual rules are applied. As in the following example, *idm/pers* is attributed to the coordination *and*:

Doctors B. Billot of the digestive surgery unit *and* G. Simali of the oncology department [...]

Extraction

The extraction module finally processes the 3-level stream returned by the tagging stage. Basically, the transition network switch on the extraction mode, when it reads a token, which is tagged *id* at the WS level, and switches off the extraction mode, when it reads a barrier, i.e. a token, which is not tagged *id*. Some nodes of the RTN are specialized for handling names with particles starting with a small letters (as *van* for Dutch names, *von* for German names, and *de* for French names).

RESULTS

More than 40 rules (1,2, and 3-level rules) were written to reach a 100% success rate for the set A. It took around three weeks to write it all.

In order to assess the performances of the anonymizer, we ran the engine on the assessment set (set B), and then we checked manually the output. We counted six types of results:

a. Identifiers in the corpus: 467 (100%)
b. Identifiers correctly removed: 452 (**96.8%**)
c. Identifiers removed with removing also irrelevant tokens: 8 (**1.7%**)
d. Identifiers incompletely removed: 3 (0.6%)
e. Identifiers left in the text: 4 (0.9%)
f. Tokens removed, which are not identifiers: 0

The first interesting result concerns the scalability of the approach: with tuning the system on 20% of the corpus we addresses correctly 96-97% of all cases. If the target is clearly reached ($98<b+c<99\%$), we believe the system is perfectible. Thus, a spelling error is responsible for one of the four errors in e, and two others errors can be solved by adding one more *idm* to the lexicon (we forgot an *idm*, similar to *MD*). In c, five errors can be mastered by adding more words to the lexicon, as unknown tokens (tagged *x*) are considered as identifiers, when they follow an identifier.

Dear Dr. *Xxxxxxx*,
Your patient, Mr. *X.-X. XXXXX*, born *xx.x.xxxx*, stayed in our service, from *xx/xx/xx* to *xx/xx/xx*.

DIAGNOSIS :
Acute pancreatitis of unknown origin.
Comorbidities :
Gastritis.
Hypertension.
Esophagitis with backflow.

Mr. *Xxxxx*, 72 year-old, has been admitted in emergency, after [...]. However tests for the cytomegalovirus and the EBV were negative. Therefore we perform an abdominal CT-scan [...] Mr. *Xxxxx* will be followed in ambulatory by Drs. *Xxxxx Xxxxxx* and *Xxxxxx xxx Xxxxxxxx* [...].

Figure 2: Example of a complete anonymization

CONCLUSIONS AND FUTURE WORK

We have designed a system for removing identifiers in medical records, with a success rate of about 99%. Unlike other systems performing similarly, this system uses natural language processing tools. One detail must be mentioned concerning the replace operation: the tractability of the anonymization is not allowed in the system, i.e. *P. Nertens* and *W. Keuster* are both replaced by *X. Xxxxxx*. This is an advantage considering the security, as reverse 'scrubbing' is forbidden, but tractability may be necessary for studies on genealogy. Although this functionality is part of the replace task, and therefore does not question the extraction task of the system, we believe that this field should be investigated in future work.

Acknowledgments

The FNRS (Swiss National Foundation) funds the MEDTAG project.

References

1. Sweeney LA, 1996, Replacing Personally-Identifying Information in Medical Records, the Scrub System. AMIA Annual Fall Symposium 1996, JJ Cimino Ed's, JAMIA, p 343-347.
2. P Ruch, J Wagner, P Bouillon, RH Baud, A-M Rassinoux, J-R Scherrer, 1999, MEDTAG: Tag-like Semantics for Medical Document Indexing. AMIA Annual Fall Symposium 1999, NM Lorenzi Ed's, JAMIA, p 137-141.
3. Gross M, 1997, The construction of Local Grammars, in Finite-State Language Processing, Roche E, Schabes Y (Eds), p. 329-354. MIT Press, Cambridge.
4. Black WJ, Gilardoni L, Dressel R, Rinaldi F, 1997, Integrated text categorization and information extraction using pattern matching and linguistic processing. Proceedings of the 5th RIAO conference, McGill University, Montreal, Canada, 25th-27th June 1997.
5. Gazdar G., and Mellish C, 1989, Natural Language Processing in Prolog: An Introduction to Computational Linguistics, Chap. 3. Eds
6. Ruch P, Bouillon P, Baud R, Robert G, 2000, *Minimal Commitment and Full Lexical Disambiguation: Balancing Rules and Hidden Markov Models*. Proceedings of the 5th CoNLL Conference (ACL-SIGNLL), Lisbon, Portugal.
7. Silberztein, M, 1997, *The Lexical Analysis of Natural Languages*, in Finite-State Language Processing, Roche E, Schabes Y (Eds), p. 176-203. MIT Press, Cambridge

Tag	Freq.(%)	Example	Label (definition)
1-qual	10.1	fat	qualifier
2-acto	9.5	leave	general act
3-loc	9.3	liver	organ, body location
4-spat	8.7	high	spatial concept
5-temp	5.3	late	temporal concept
6-mod	5.1	maybe	modal
7-quant	4.7	five	quantitative concept
8-papr	4.5	infection	pathological process
9-find	4.2	fever	signs or symptoms
10-cpt	4.1	idea	other concept
11-ther	4.0	inject	therapeutic procedure
12-mdev	3.6	scissors	medical device
13-thers	3.6	abscission	surgery procedure
14-hca	2.1	care	physician's act
15-diap	1.9	biopsy	diagnosis procedure
16-rel	1.9	same	relationships (other)
17-medi	1.7	penicillin	drugs and chemicals
18-name	1.6	Spigel	for medical techniques
19-dis	1.4	diabetes	disease or syndrome
20-bosp	1.4	vesicle	body space or junction
21-bopr	1.3	digestion	body process
22-rconj	1.2	with	conjunction relation
23-bosu	1.0	blood	body substance
24-obj	0.9	watch	general object
25-rspat	0.8	behind	spatial relation
26-actp	0.7	to suffer	patient's act
27-mpr	0.7	think	mental process
28-rtemp	0.7	during	temporal relation
29-spec	0.7	ophthalmology	medical speciality
30-occup	0.7	carpenter	occupation
31-neop	0.6	carcinoma	neoplastic process
idm	**0.6**	**Dr.**	**identity markers**
32-pers	0.5	professor	person
33-tiss	0.5	T024	tissue
34-labo	0.5	crasis	laboratory or test results
35-subst	0.4	water	substance (other)
36, 37, 38 respectively aux, def, and indef for respectively auxiliary, definite or non-definite determiner, freq << 0.1%			
id	**<<0.1**	**Louise**	**identity proper nouns**

Table 3: Distribution of the semantic tagset in the lexicon.