Resource-Based Anonymization for Spanish Clinical Cases

Fernando Sánchez-León

unaffiliated
f.sanchez.lcmcvp@gmail.com

Abstract. This implementation notes describe a system for anonymization of clinical case texts written in Spanish based on careful collection of gazetteers for several entity types. The entities gathered are improved converting their strings to regular expressions covering commonest text variations. A simple grammar formalism is implemented to *sugar* the creation of context-dependent regular expressions. These expressions are applied in order (organized according to reliability) and context specified by rules may refer to the character/word/regular expressions and/or annotations previously introduced in the running text by already applied rules. With this simple approach, the system obtained an F1-score of 0.9595 in the NER offset and entity type classification (subtask 1) and 0.96409 in the sensitive token detection (subtask 2).

Keywords: Anonymization · de-identification · resource-based processing · Electronic Health Records · Protected Health Information · clinical cases · MEDDOCAN.

1 Introduction

The anonymization of Electronic Health Records (EHRs) opens a new range of uses for these otherwise private data, being just a few of them the application of big data techniques to gain feedback on treatment effectiveness, the estimation of survival of a patient that has overcome a serious illness, and, generally, the effective easiness of the study of human diseases in broad sense. In this new (at least for Spanish) avenue of research activities, IberEval 2019 has launched evaluation activity of Medical Document Anonymization (henceforth, MEDDOCAN). This working paper accounts for our participation in the referred evaluation track.[1]

Section 2 describes the resources we have collected and enriched for our the task at hand. Some general issues on the development of our system are presented in section 3. System results are presented in section 4 along with both a discussion on errors by the system and some comments of the track objectives and the golden dataset. Finally, section 5 includes some concluding remarks.

[1] http://temu.bsc.es/meddocan/

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). IberLEF 2019, 24 September 2019, Bilbao, Spain.

CEUR-WS.org/Vol-2421/MEDDOCAN_paper_9.pdf

2 Resource building

This section describes the resources built from scratch (or reused, adapted and further developed) specifically for the MEDDOCAN competition. Since the organizers released gazetteers for some of the annotation classes (henceforth, simply *classes*, we have decided to explore this avenue.

From previous work, we had lists of first names and last names (c. 80,000 elements) that we have reused as such. Other gazetteers have been complemented using different information sources, mainly c. 175,000 articles in the health domain downloaded from SciELO electronic library[2]; but also Wikipedia[3], GeoNames[4] and certain web searches and sites have been visited.

For instance, the gazetteer for hospitals released by MEDDOCAN organizers, which is basically the list published by the Ministerio de Sanidad, Consumo y Bienestar Social[5], has been cleaned for orthographic errors (lack of stresses: `HOSPITAL RIO CARRIÓN`, `HOSPITAL QUIRONSALUD MIGUEL DOMINGUEZ`, missing final dots: `AVANCES MÉDICOS S.A`). This list has been complemented with all hospital and clinical centers extracted in the SciELO collection. Some other webs have been used to download hospital names. The list used in the competition has around 2,000 unique hospital names.

As regards, *known* addresses we have used exclusively those included in the gazetteer released by the organizers. Again, these addresses have been cleaned and canonicalized (more on this below), and duplicate addresses have been removed.

We have also gathered gazetteers for services, units, laboratories (with cue words in Spanish, English and Portuguese), as well as clinical specialties. These class is not to be annotated within MEDDOCAN, but its identification will be of help for the matching of positive classes.

Institutions have been collected from a number of web pages, most notably from the Ranking Web of Universities[6]. These have been complemented (in another gazetteer) with company names, specially companies and holdings within the business area of health.

As for location names, we have collected names of countries, autonomous communities, provinces and cities in Spain (from Wikipedia), as well as a list of provinces and largest cities in countries around the world. Besides this list of location names, we use another gazetteer with all cities with population greater than 500, downloaded from GeoNames database[7]. These GeoNames are filtered using five heuristics: 1) delete all name shorter than four characters long; 2)

[2] `https://www.scielo.org`. The collection used was downloaded during February 2019.
[3] `https://es.wikipedia.org`
[4] `https://www.geonames.org`
[5] `https://www.mscbs.gob.es/ciudadanos/prestaciones/centrosServiciosSNS/hospitales/docs/2018_CNH.pdf`
[6] `http://www.webometrics.info/en`
[7] `http://downloads.geonames.org/export/dump/cities500.zip`

remove all names ambiguous with first names; 3) delete those ambiguous with last names; 4) discard single- or multi-word names all whose parts are lexical (i.e., found in a full form lexicon); 5) finally, forms ambiguous with any other entity in the rest of gazetteers are also removed (*Calcium*). As a final pruning of the set of GeoNames, all those names seen in train+dev datasets as part of the construction `signo|fórmula|pasta|asa|drenaje|escala|... de <name>` have been blocked.

Unfortunately, the information processed from GeoNames includes only the column for the English name, discarding other columns with the GeoNames in other languages. For this reason, we have missed some otherwise common locations like *Milán* and *Nueva York* in the test set.

Gazetteer elements are not plain strings but regular expressions. This way, we can allow for the matching of text variants of *known* class instances found in the clinical cases. However, these (possibly regular expression-form) entities are further processed as described below. Before this further processing, the system potentially uses c. 200,000 entities, plus 430,000 unique names from GeoNames pre-processed database.

Resource munging As already said, prior to munging the gazetteer data, we have canonicalized all the entries in the new gazetteer types[8]. As a result of this, the entry `DR ESQUERDO, 46` (which actually appears twice in the released gazetteer for *known addresses*) has been converted to

```
DR. ESQUERDO, 46
```

The following are all the text versions of this "canonical" address in the train+dev dataset:

```
Calle del Dr. Esquerdo, 46
c/ Doctor Esquerdo 46
c/ Doctor Esquerdo, 46
C/ Dr. esquerdo, 46
C/ Dr. Esquerdo 46
C/ Dr. Esquerdo, 46
Doctor Esquerdo, 46
Dr Esquerdo 46
```

When loading resources and depending on the class set as first line of each gazetteer file, list elements are munged by a set of methods applied to each of them. This way, after munging the former canonical address, we store the following regular expression (split in two lines for editorial purposes):

[8] Canonicalization has been done using editor search and replace commands on the dataset. A random sample representing 20% of the addresses was exhaustively checked and, when satisfied with the result, the rest was spot checked.

```
   (?:(?:calle|c[\/.]*|avenida|avda\.?|av\.?|paseo|ps?\.?)
(?: del?| de la)? ?)?(?:dr\.|dr|doctor) esquerdo(?:,?) (?: *46)?
```

We use a similar strategy also for hospital and institution names, but haven't used it (due to lack of time) to companies (that map also to `INSTITUCION` in MEDDOCAN[9]).

3 Development

Our system's approach is very simple, as it is described in this section. A **core** subset of all entities[10] is dynamically complemented with all entity candidates found in the corpus to be processed. This is achieved performing a fast light tokenization[11] that spot fragments of (possibly multi-word) names in the above mentioned three categories. All munging processes are performed also on the newly gathered entity candidates and compiled with the rest of regular expressions. A regular grammar, with the possibility to express left and right (regular) context and/or tags already inserted on both sides of the focus, due to previous applications of matching rules, has been defined. In order to access fragments already identified by a previous rule, the system uses an offset vector, where we record the span and the type of information identified[12].

The current grammar has nearly 100 rules, ordered in three sections: first section contains rules that block the further identification of (mainly) entities as any of the proposed classes, while the second and third sections identify positively text fragments to be tagged, being the latter of these of a more heuristic shape.

As an example of the type of context captured by rules in the former section, take, for instance, *escala de Glasgow*, where the name must remain untagged. In order for the reader to have a cleared insight of the rules shape, one of them is included below:

```
HOSPITAL => {@HOSPITALS}
resp_known_hospital        => '({HOSPITAL})
              (?:{SEP}* ({CITY_OR_PROVINCE}))?',
              types => [ 'HOSPITAL', 'TERRITORIO' ]
```

[9] This lack of further processing of company names accounts for some errors in our system run on the test data —*Alcon Cusí*, *Novartis*, *Allergan* are companies in our gazetteer not annotated because in the text show with a *SA* or *S.A.* post-modifier. These and other cases could have been easily captured.

[10] These entities were selected using frequency information gathered from SciELO collection processing. Currently the core has the full content of gazetteers for all classes except for locations, first names and last names.

[11] Light tokenization takes 5.01 seconds on the MEDDOCAN background corpus — 1,644,964 running words as counted with `wc`— using an Intel$^{(R)}$ Core$^{(TM)}$ i7-3770K CPU @ 3.50GHz. The system is implemented in Perl and it relies on a library developed entirely by the authors for last years BARR2 competition and extended for the MEDDOCAN and the PHARMACONER competitions this year. This library uses other libraries from the Perl ecosystem.

[12] Neither the algorithm nor the regular expressions dynamically built have been optimized.

A symbol table is declared, and some notation sugar is implemented. For instance, an `@` sign at the beginning of an identified symbol instruct the program to return the disjunction of all the gazetteer elements (possibly regular expressions) of that class. The same is done for the symbol `CITY_OR_PROVINCE`. If matched in the text, either a hospital name and optionally a city or province, they get tagged with the types given in the rule.

As another example, this time from the third section of the grammar, the following, more heuristic rule, we try to identify an unknown city or province (defined via a simple regular expression) if that offset span is not already tagged and there is a specific set of entities to the left (`lc` stand for *left context*).

```
city_heur_1      => '{SEP}+ ({CITY_OR_PROVINCE_HEUR}) {SEP}+',
                 types => [ 'TERRITORIO' ],
                 lc => [ 'TERRITORIO|NOMBRE|HOSPITAL' ]
```

Among the information types for which a gazetteer does not exist in the system are addresses (`CALLE`). For these, we build a regular expression with typical cue elements (`Av., C/?.?, Pg, Paseo,...`) and common cues to express street number and floor (or kilometer point or no number —`s/n`—)[13].

```
address_heur_last_1 => '(?:{HEALTH_RECORD_RESP_CUE}.*?) {TSEP}
                        ({ADDRESS_CUE}.*?{ADDRESS_POST}) {SEP}',
                    types => [ 'CALLE' ],
                    rc    => [ 'TERRITORIO' ]
```

Finally, in order to match addresses with no cues, we have used a third scanning over the document string, with no particularly good results (see section 4).

4 Results and discussion

This section presents the results obtained by the system developed and also discusses some weak points discovered after sending our datasets to the organizers. Finally, we also express some insights from the overall task and the quality of the golden datasets.

Results We sent three runs. The three of them use the approach described in these notes. However, runs 2 and 3 (`B` and `C` in the table below) make use of the annotations gathered from train+dev datasets released by the organizers. This information (obviously pruned of certain *classes* for which propagation across documents is nonsense) was used prior to annotation by the system. Finally, run 1 (`A`) is the result of system annotation alone. MEDDOCAN is structured in two subtasks —the former, aiming to identify named entities and other sensible patient information and to provide their offsets; the latter, simply extracting the offsets of sensible text fragments. Results, as provided by the organizers, are shown in the following tables.

[13] Note that these examples are a small fragment, for illustrative purposes, of a very complex set of cues including Portuguese, Catalan, Italian and English, apart from Spanish, of cue text fragments.

	Precision	Recall	F1-score
A	**0.95857**	**0.96043**	**0.9595**
B	0.95597	0.95884	0.9574
C	0.95547	0.95884	0.95715

Table 1. Results on MEDDOCAN test dataset for subtask 1.

	Precision	Recall	F1-score
A	**0.96315**	0.96502	**0.96409**
B	0.96231	**0.9652**	0.96375
C	0.9618	**0.9652**	0.9635
A	0.96694	**0.96942**	**0.96818**
B	**0.96708**	0.9689	0.96799
C	0.96645	**0.96942**	0.96793

Table 2. Results on MEDDOCAN test dataset for subtask 2 (strict and merged).

Discussion No comparative comments can be included in this working paper since, as it was the case in last year's competition on abbreviation resolution (BARR2), the organizers don't publish results of the participants prior to paper writing. This lack of information on other participants systems performance is somewhat surprising in an applied research activity that is defined as a competition.

For the sake of saying something, not precisely positive, about the system described here, we included a third non-regular grammar-based scan on the input text in order to improve recall on *unknown* addresses not showing any cue. Due to an inadvertent bug, we applied this strategy also to addresses with a known cue and, more critically, to *known* addresses, thus, wrongly extending in some cases these *known* addresses to convert them in false positives. We haven't quantified the number of errors introduced by this bug, but we are sure that has had a significant impact on the F1-score.

Besides, some common expressions —like *3 días de nacido*— were not included in system resources. This and other new expressions can boost system performance devoting a few minutes to resources update.

Annotation and anonymization The ultimate goal of tagging certain information bits to protect patient's health information (PHI) is to anonymize a text in a way that no person or program can achieve the re-identification of the patient identity behind the data. With this primary goal in mind, and assuming a clinical case structure like that found in the datasets released by the organizers of MEDDOCAN, there are simpler ways to perform the task, at least as regards subtask 2 of the competition. Let's expand the point.

Every clinical case has three parts —a patient profile, the case description and the health professional data. The first and last parts of the clinical case can

be trivially trimmed, although preserving relevant information on the specialty of the health professional signing the case —information which is included in the department, service, unit,... sometimes explicitly stated in the health professional data part. Accepting this premise, the task at hand is that of thoroughly work the case description.

The test dataset contains 105,062 running words (as counted by `wc`) and 5,661 annotations in the `.ann` files, but only 769 annotations are in the body of the text. However, we believe that this part of the clinical case makes a difference in anonymization systems development for the reasons just exposed. Moreover, it is precisely in body text where a human annotator can have more problems to hear the bells of information that has to be protected. In order to prove this assumption, we have read the body of all the cases in the test dataset and checked the expert annotations. The results of this exercise are shown in table 3.

TAG	True Positives	True Negatives	False Positives
EDAD_SUJETO_ASISTENCIA	242	2	0
SEXO_SUJETO_ASISTENCIA	194	0	0
FECHAS	109	0	0
FAMILIARES_SUJETO_ASISTENCIA	77	3	3
INSTITUCION	39	0	7
TERRITORIO	26	0	4
PAIS	27	0	0
HOSPITAL	13	0	2
ID_SUJETO_ASISTENCIA	13	1	0
PROFESION	4	5	3
OTROS_SUJETO_ASISTENCIA	3	0	0
NOMBRE_SUJETO_ASISTENCIA	2	0	0
NUMERO_TELEFONO	1	0	0

Table 3. A closer look at the test dataset annotations in the body of clinical cases.

In the preceding table, we have computed span errors as false positives. Besides, *Vancouver* in the entity name *Escala de Cicatrización de Vancouver* and similar cases have been also considered false positives. We also judge *hospital militar afgano* a false positive, as it does not meet the criteria of having a unique referent, as it should be the case for proper names. In the same vein, *trabajador en canteras* and similar cases count as false positives given that the annotation guidelines explicitly state that the word *trabajador* must not be annotated as part of a `PROFESION`[14]. Finally, 2 out of 3 cases marked for `OTROS_SUJETO_ASISTENCIA` are debatable.

The number of false positives in the `INSTITUCION` class is consistent with the scarce number of examples in the annotation guidelines for this particular

[14] See page 15 of the annotation guidelines, available at `http://temu.bsc.es/meddocan/wp-content/uploads/2019/02/gu%C3%ADas-de-anotaci%C3%B3n-de-informaci%C3%B3n-de-salud-protegida.pdf`.

class[15]. Actually, there is no indication in the guidelines as to what to do when the institution names refer to medical products and, hence, are not part of the PHI. Take, for instance, the following text fragment:

> Se procedió a estudio inmunohistoquímico de cortes representativos mediante el método avidina-biotina peroxidasa, utilizando anticuerpos primarios anti antígeno de membrana epitelial EMA, (Dako M613, USA, 10/500), proteína S-100 (Dako, L1845, USA, prediluída), neurofilamentos (Biogenex 6670-0154, USA), enolasa neuroespecífica NSE, (Biogenex MU055-VC, USA, 10/1000), CD57 (Becton-Dickinson 7660, USA, 10/500), CD34 (Becton-Dickinson 7660, USA, 10/500), a-actina de músculo liso (Dako MO851, USA, 10/200), desmina (Dako M760, USA, 10/500), y vimentina (Shandon 402255, USA, pred.).

In this fragment every single occurrence of Dako, Biogenex, ..., and USA is tagged. This can be important for a NER task but is irrelevant for anonymization of PHI. Re-identification is not only possible for the set of antibodies mentioned but also of paramount importance for some human disease studies. In fact, in other similar contexts these entities remain untagged[16], a fact that mislead the participant.

Finally, some words remain untagged (and they have not been reflected in the table above), albeit carrying personal information —this is the case of ***primigesta*** *de 32 años*. However, the most argueable case for untagged information is that of gender when denoting sex of patient. There are, of course, medical specialties only for men or for women, but, for the sake of complicating re-identification of PHI, occurrences of ***la*** *paciente* —and the more difficult to capture cases of feminine agreement of adjectives, participles and the like— need to be anonymized, in our humble opinion.

The error rate found in test set for body text is relatively high, standing out that maybe some of the details in the guidelines are not clearly exemplified and have been applied inconsistently. This has a significant impact on system evaluation, especially when system scores are in the 90s.

5 Conclusions

This paper has presented a resource-based anonymization system for Spanish clinical cases. Its F1-score for the MEDDOCAN dataset is 0.9595 for NER offset and entity type classification (subtask 1) and 0.96409 in the sensitive token detection (subtask 2). We believe that, after fixing a bug in the program that garbles *known* addresses and devoting a couple of hours to expand system's regular grammar, performance can be improved significantly.

[15] See page 16 of the guidelines.

[16] For example, in the text fragment: "Como antecedentes destaca estar en tratamiento hematológico con hidroxyurea (Hydrea®, 500 mg, Bristol-Myers, Squibb) y ácido acetilsalicílico (Adiro®, 300 mg, Bayer)."