Anonymization of Clinical Reports in Spanish: a Hybrid Method Based on Machine Learning and Rules.

Pilar López-Úbeda, Manuel C. Díaz-Galiano, L. Alfonso Ureña-López, and M. Teresa Martín-Valdivia

Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{plubeda, mcdiaz, laurena, maite}@ujaen.es

Abstract. Biomedicine is an ideal environment for the use of Natural Language Processing, due to the huge amount of information processed and stored in electronic format. This information cannot be shared with confidential patient data. In order to achieve this task, the Medical Document Anonymization workshop has been created.
In this paper, we present an automated anonymization system for clinical reports written in Spanish. Three different methods are evaluated and compared. The first method is rule-based, the second method uses machine learning and the third is a hybrid method between the first two. The evaluation showed that the use of the hybrid method obtained the best results. The results are as expected, we obtained 90% in measure F1 in sub-task 1 and 95% in sub-task 2.

Keywords: Anonymization · Named Entities Recognition · CRF · Machine Learning · Regular Expressions

1 Introduction

Named Entity Recognition (NER) in a text is an important key in many natural language applications such as the anonymization of clinical records. This task is crucial because a hospital cannot freely publish information related to a patient [11].

NER consists in automatic identification of fragments of texts called entities which refer to information units such as persons, geographical locations, sex, names of organizations, dates, occupation or references to documents [2].

The *OTG de Sanidad* of the Plan TL in collaboration with the National Center for Oncological Research and the *Hospital 12 de Octubre* in Madrid have organized a task to contribute to these studies mentioned above. This task is part of the IberLEF (Iberian Languages Evaluation Forum) in SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural) 2019.

Medical Document Anonymization (MEDDOCAN) task is the first community challenge task specifically devoted to the anonymization of medical documents in Spanish [7]. The MEDDOCAN task is structured into two sub-tasks: NER offset and entity type classification and Sensitive span detection.

The first sub-track wants to match exactly the beginning and end locations of each Protected Health Information (PHI) entity tag, as well as detecting correctly the annotation type. In the order hand, the second sub-track is more specific to the practical scenario needed for releasing de-identified clinical documents, where the main goal is to identify and be able to obfuscate or mask sensitive data, regardless the actual type of entity or the correct offset identification of multi-token sensitive phrase mentions.

2 Dataset

This MEDDOCAN corpus was selected manually by a practicing physician and augmented with PHI phrases by health documentalists, adding PHI information from discharge summaries and medical genetics clinical records. Usually, the clinical records that we will treat are structured, most of them follow a common format.

The organizers provided us with the following datasets:

- The training set consists of 500 documents
- The validation set consists of 250 documents
- The test set contain 3751 documents

The MEDDOCAN annotation scheme defines a total of 29 entity types [1] and the official annotation guidelines used to annotate the MEDDOCAN data sets is available [2].

3 Strategies

In this section we will describe the methods and strategies followed to achieve the tasks. These methods are used for both subtasks: NER offset and entity type classification and Sensitive token detection.

3.1 Rule-based method

An initial survey using the training dataset showed that the majority of the values for a given field were recorded in certain patterns. Some patterns we can find are shown in Table 1. In this table we can see that it can be easy to identify some named entities. Therefore, we designed rules by incorporating regular expression (REs or regex) for each field according to the description types of the fields [1].

There have been many studies on the use of regular expressions in different areas of medicine [8, 5]. Some of them very current so that it is still a method that is applied in the area of Natural Language Processing (NLP).

[1] http://temu.bsc.es/meddocan/index.php/annotation-guidelines/

[2] 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

Table 1. Examples PHI phrases in text.

Text	PHI phrases	Annotation
Nombre: M. del Mar	M. del Mar	NOMBRE_SUJETO_ASISTENCIA
Edad: 68	68	EDAD_SUJETO_ASISTENCIA
Sexo: Hombre	Hombre	SEXO_SUJETO_ASISTENCIA
Médico: Josep Rubio Palau	Josep Rubio Palau	NOMBRE_PERSONAL_SANITARIO
Domicilio: Av. Melchor Fernández 59. 4,2	Av. Melchor Fernández 59. 4,2	CALLE
CP: E-28015	E-28015	TERRITORIO
Matricula: 5478GDV	5478GDV	IDENTIF_VEHÍCULOS_NRSERIE_PLACAS

The first step to elaborate this experiment was to define the rules to extract the required entities. These rules were taken from the annotation guide provided by the task organizers. The text is converted to lower case to have it homogeneous. A total of 25 rules were defined in this method. Some of these rules are shown in Table 2.

Table 2. Rules defined for PHI phrases.

Regular expression	Annotation
r'(nass):(.*)'	ID_ASEGURAMIENTO
r'nombre:(.*)'	NOMBRE_SUJETO_ASISTENCIA
r'edad:(.*)'	EDAD_SUJETO_ASISTENCIA
r'(cipa|dni|nif|cie|pasaporte|nhc):(.*)'	ID_SUJETO_ASISTENCIA
r'(4)[a-zA-Z]3'	IDENTIF_VEHÍCULOS_NRSERIE_PLACAS
r'(bastidor):(.*)'	ID_CONTACTO_ASISTENCIAL'
r'(\ + \d{2}){0, 1}(\d){9}d'	NÚMERO_TELEFONO

In addition to obtaining greater precision in recognizing PHI information within the text. We use some resources to correctly identify territories and peoples. First, we used a list of Spanish provinces on which we relied to identify territories, and on the other hand, we use Stanford Named Entity Recognizer [6] with a pre-trained Spanish model. Spanish models on a combination of two corpora, after very heavy modifications: AnCora Spanish 3.0 corpus[3] and DEFT Spanish Treebank V2[4]. With these Stanford models, people and territories will be identified in the text.

3.2 CRF

Conditional Random Fields (CRF)[4] classifier is a stochastic model commonly used to label and segment data sequences or extract information from medical documents [3]. We used CRFsuite, the implementation provided by Okazaki [9], as it is fast and provides a simple interface for training and modifying the input features.

[3] http://clic.ub.edu/corpus/ancora
[4] https://catalog.ldc.upenn.edu/LDC2018T01

We incorporate some basic features of each word such as isLower, isUpper, isTitle, isDigit, isAlpha, isBeginOfSentence and isEndIfSentece.

Similar to most machine learning-based de-identification systems, the token-level CRF requires a tokenization module at first. The tokenizer used is WordPunctTokenizer of the NLTK[5] library in Python.

Below are a few short lines from the training file (S1130-01082009000500012-1).

```
S1130-01082009000500012-1  Médico      0
S1130-01082009000500012-1  :           0
S1130-01082009000500012-1  David       NOMBRE_PERSONAL_SANITARIO
S1130-01082009000500012-1  Hernández   NOMBRE_PERSONAL_SANITARIO
S1130-01082009000500012-1  Alcaraz     NOMBRE_PERSONAL_SANITARIO
S1130-01082009000500012-1  .           0
S1130-01082009000500012-1  NCol        0
S1130-01082009000500012-1  :           0
S1130-01082009000500012-1  29          ID_TITULACION_PERSONAL_SANITARIO
S1130-01082009000500012-1  29585       ID_TITULACION_PERSONAL_SANITARIO
```

The CRF algorithm trained with the parameters: algorithm = lbfgs, c1 = 0.1, c2 = 0.1, max_iterations = 100, all_possible_transitions = False.

The output provided by this method is shown below. As we can see, CRF returns tokens and their predicted annotation, so it is necessary to perform a treatment to join different tokens in the same concept.

```
[("Médico",0), (":",0), ("David",NOMBRE_PERSONAL_SANITARIO),
("Hernández",NOMBRE_PERSONAL_SANITARIO),
("Alcaraz",NOMBRE_PERSONAL_SANITARIO), (".",0),
("NCol",0), (":", 0),("29",ID_TITULACION_PERSONAL_SANITARIO),
("29585",ID_TITULACION_PERSONAL_SANITARIO)]
```

This treatment consisted of joining all the contiguous tokens of the same category. In this way, we create the correct output file as shown below:

```
David Hernández Alcaraz    NOMBRE_PERSONAL_SANITARIO
29 29585                   ID_TITULACION_PERSONAL_SANITARIO
```

3.3 Hybrid method

The last method applied was using the two methods described above: rule-based method and machine learning with CRF.

At the end of the machine learning method, an error analysis was carried out and we found that there were some inconsistencies according to the PHI phrases that had been annotated. This analysis was developed with the development dataset.

[5] https://www.nltk.org/

Most of the error cases that we could observe were with annotations like: *TERRITORIO*, *ID_CONTACTO_ASISTENCIAL* and *ID_ASEGURAMIENTO*.

The main problem we got with the *TERRITORIO* notation is that we wrote phrases together as shown below:

```
AV. San Francisco 7, 3D 50006 Zaragoza        TERRITORIO
```

And the correct annotation should be:

```
AV. San Francisco 7, 3D         TERRITORIO
50006                           TERRITORIO
Zaragoza                        TERRITORIO
```

These errors could be solved by using regular expressions that separated that annotation into different entries, in this case we used the regular expression to find the Zip Code and the list of cities in Spain, in this way, we could separate the different PHI phrases.

Other errors that we could avoid with the use of regex is the erroneous annotation *ID_CONTACTO_ASISTENCIAL* because the algorithm identified it as *ID_SUJETO_ASISTENCIA*.

4 Results and discussion

For both sub-tracks the primary de-identification metrics used will consist of standard measures from the NLP community, namely micro-averaged precision, recall, and balanced F-score.

In addition, the leak scores is also used for sub-task 1. This measure is related to the detection of leaks (non-redacted PHI remaining after de-identification), that is (#false negatives / #sentences present).

The results obtained by our team for sub-task 1 (NER offset and entity type classification) are shown in Table 3.

Table 3. Results of Task 1.

Run	Leak	Precision	Recall	F1
1	0.346	0.66457	0.54001	0.59585
2	0.11998	0.89369	0.84049	0.86627
3	**0.08491**	**0.92113**	**0.88712**	**0.90381**

This results show that we have improved our baseline (rule-based method) using machine learning algorithms. A great step that is reflected how we obtain a 0.59 in F1 score and we obtain a 0.86 with CRF. We managed to improve in all

measures with the use of some rules obtaining an precision of 0.92 and a recall of 0.88.

Table 4 and Table 5 show the evaluation of the systems for sub-task 2 (Sensitive span detection) with strict and merged spans evaluation respectively.

Table 4. Results of Task 2: Sensitive span detection and Strict span evaluation.

Run	Precision	Recall	F1
1	0.86594	0.70288	0.77594
2	0.93858	0.88271	0.90979
3	**0.96167**	**0.92616**	**0.94358**

Table 5. Results of Task 2: Sensitive span detection and Merged spans evaluation.

Run	Precision	Recall	F1
1	0.87549	0.70752	0.78259
2	0.96825	0.93575	0.95173
3	**0.97295**	**0.9437**	**0.9581**

In this second sub-task we check that we obtain values higher than the previous task, this is because the objective is only to identify confidential data. Thus, this is considered a span-based evaluation, regardless of the actual type of entity or the correct offset identification of multi-token sensitive phrase mentions.

In this sub-task we get a higher baseline than in the previous task, and also get better thanks to CRF and the rules applied. We achieve relatively equal results in both the stric and merged evaluations. The precision is almost perfect, our third method is correct in 97% of cases.

If we see a difference between the two evaluations and systems 2 and 3. In the strict evaluation system the machine learning algorithms get 90% F1, and in the merged evaluation they get 95% with the second method. We see that the second method works best when the evaluation is not strict.

Finally, it is interesting to see that between the two possible evaluations of sub-task 2 we obtain similar values with the third method proposed by our group. This means that our third experiment records almost exactly.

5 Error analysis

The main purpose of this section is to carry out an error analysis to identify the weaknesses of our best system: hybrid method (run 3). To this end, we have

obtained some basic statistics for 250 gold test files. The test files taken into account are from Subtask 1 described in Section **??**. These files contain 5661 key phrases annotated.

We have described three basic types of errors for this analysis:

1. **Does not have the same annotated label.**
   In this case, our system writes the positions of the key phrase correctly but the associated annotation is incorrect. We found a total of 221 errors in this case. The biggest confusions our system makes are described in Table 6. This table shows the annotation found in the gold test files, the annotation of our system and the number of errors found.

Table 6. Annotations incorrectly annotated by our system.

Annotation gold test	Annotation our system test	Number of errors
TERRITORIO	ID_SUJETO_ASISTENCIA	105
ID_CONTACTO_ASISTENCIAL	ID_SUJETO_ASISTENCIA	21
TERRITORIO	NOMBRE_SUJETO_ASISTENCIA	16
NOMBRE_PERSONAL_SANITARIO	NOMBRE_SUJETO_ASISTENCIA	15
ID_TITULACION_PERSONAL_SANITARIO	ID_ASEGURAMIENTO	12
ID_ASEGURAMIENTO	ID_TITULACION_PERSONAL_SANITARIO	10

   In future work, these types of errors are relatively easy to solve. The largest of the cases is found in the territory and id of the subject of assistance, this is because the zip codes are 5 digits and the IDs are usually digits as well. For the improvement of this case we should observe the context in which the digits are to annotate them correctly.

2. **Incorrect positions.**
   In this case, our system incorrectly marks the start or end position of the key phrase. To make this situation clearer, some examples are shown below: Example number 1, the first frame shows the correct annotations. In this frame we can see that the name of the health employee and the street are noted separately, we see that the positions are consecutive (340 - 360 and 361 - 374) but our system (second frame) takes everything as a full name with positions from 340 to 374.

```
CALLE 361 374                        Paseo Calanda
NOMBRE_PERSONAL_SANITARIO 340 360    Raquel Ridruejo Sáez
```

```
NOMBRE_PERSONAL_SANITARIO 340 374 Raquel Ridruejo Sáez Paseo Calanda
```

   The following example shows a similar error, the first frame shows the correct output and the next frame shows the output of our system. In them we can see how our system annotates everything as a street when it should

separately record institution and street.

```
CALLE 1945 1974          Gran Via Corts Catalanes, 111
INSTITUCION 1923 1944    Ciutat de la Justícia
```

```
CALLE 1923 1974  Ciutat de la Justícia Gran Via Corts Catalanes, 111
```

The number of errors of this type in our system are 215.

3. **Not found.**
   Finally, we found 203 cases that are scored in the gold test but our system does not write them down.

The total number of errors found are 639, which is 11% of the test gold annotations. It is a small number that could possibly be improved for later systems.

6 Conclusions

We have observed an increase in the number of studies associated with the identification of PHI phrases in electronic medical records. Statistical analyses or machine learning, followed by NLP techniques, are gaining popularity over the years in comparison with rule-based systems [10]. In this study we try to verify that traditional and automatic methods can still coexist and that they can be of great help if we use them together.

The SINAI group presents its first participation in this type of tasks where the main objective is to find PHI information in clinical records. The results are really good. In both subtasks we have reached more than 90% F1 in our best method. And in precision we got more than 97% in sub-task 2 with sensitive evaluation.

7 Acknowledgements

This work has been partially supported by Fondo Europeo de Desarrollo Regional (FEDER), LIVING-LANG project (RTI2018-094653-B-C21) and REDES project (TIN2015-65136-C2-1-R) from the Spanish Government.

References

1. Chen, L., Song, L., Shao, Y., Li, D., Ding, K.: Using natural language processing to extract clinically useful information from chinese electronic medical records. International journal of medical informatics **124**, 6–12 (2019)
2. Graliński, F., Jassem, K., Marcińczuk, M., Wawrzyniak, P.: Named entity recognition in machine anonymization. Recent Advances in Intelligent Information Systems pp. 247–260 (2009)

3. He, Y., Kayaalp, M.: Biological entity recognition with conditional random fields. In: AMIA Annual Symposium Proceedings. vol. 2008, p. 293. American Medical Informatics Association (2008)
4. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
5. Liang, Z., Chen, J., Xu, Z., Chen, Y., Hao, T.: A pattern-based method for medical entity recognition from chinese diagnostic imaging text. Frontiers in Artificial Intelligence **2**, 1 (2019)
6. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. pp. 55–60 (2014)
7. Marimon, M., Gonzalez-Agirre, A., Intxaurrondo, A., Rodríguez, H., Lopez Martin, J.A., Villegas, M., Krallinger, M.: Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). vol. TBA, p. TBA. CEUR Workshop Proceedings (CEUR-WS.org), Bilbao, Spain (Sep 2019), TBA
8. Nguyen, A.N., Lawley, M.J., Hansen, D.P., Bowman, R.V., Clarke, B.E., Duhig, E.E., Colquist, S.: Symbolic rule-based classification of lung cancer stages from free-text pathology reports. Journal of the American Medical Informatics Association **17**(4), 440–445 (2010)
9. Okazaki, N.: Crfsuite: a fast implementation of conditional random fields (crfs) (2007)
10. Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P.J., Elhadad, N., Johnson, S.B., Lai, A.M.: A review of approaches to identifying patient phenotype cohorts using electronic health records. Journal of the American Medical Informatics Association **21**(2), 221–230 (2013)
11. Szarvas, G., Farkas, R., Busa-Fekete, R.: State-of-the-art anonymization of medical records using an iterative machine learning framework. Journal of the American Medical Informatics Association **14**(5), 574–580 (2007)