ReCRF: Spanish Medical Document Anonymization using Automatically-crafted Rules and CRF

Fadi Hassan[1], **Mohammed Jabreel**[2], **Najlaa Maaroof**[2],
David Sánchez[1], and **Josep Domingo-Ferrer**[1] and **Antonio Moreno**[2]

[1]CYBERCAT-Center for Cybersecurity Research of Catalonia. UNESCO Chair in Data Privacy.
[2] iTAKA: Intelligent Technologies for Advanced Knowledge Acquisition.
Department of Computer Science and Mathematics
Universitat Rovira i Virgili, Av. Països Catalans 26, E-43007 Tarragona, Catalonia
{fadi.hassan, mohammed.jabreel, najlaa.maaroof, david.sanchez,
josep.domingo, antonio.moreno}@urv.cat

Abstract. This paper describes ReCRF, a named-entity recognition system submitted to the Medical Document Anonymization (MEDDOCAN) challenge in the IberLEF 2019 Workshop. We propose a general method based on a data-driven rule generator and Conditional Random Fields (CRFs) to automatically detect protected health information (PHI) in Spanish medical documents. The reported experiments show that our system achieves a micro-F1 of 96.33% on the test dataset for the first sub-task and a micro-F1 of 96.86% and 97.50% on strict and merged metrics, respectively, on the test dataset for the second sub-task.

Keywords: Anonymization · CRF · Medical Documents.

1 Introduction

Medical documents containing detailed patients' data are of utmost importance for research. When healthcare data are associated to individuals, they are considered protected health information (PHI). The new European General Data Protection Regulation (GDPR) [6], states that explicit consent from the affected individuals is needed to use personally identifiable information (PII), and PHI in particular, for secondary purposes. That means, the data collector should strive to gather such consent. To avoid the need for consent, data used for secondary purposes should no longer be personally identifiable. Document anonymization provides a way to turn PII into information that cannot be linked to a specific identified individual any more, so that it is not subject to privacy regulations anymore.

In 2006 and 2014, i2b2 organized two shared tasks on document anonymization [2]. The i2b2 effort had a significant impact on the medical natural language processing (NLP) community, but that effort was focused on English documents only. IberLEF 2019 organizes the first community challenge task specifically devoted to the anonymization of Spanish medical documents, called the MEDDOCAN task [5]. The purpose of the MEDDOCAN task is to detect and remove PHI from Spanish plain text medical records. The task is structured into two sub-tasks: ”NER offset and entity type classification” and ”sensitive token detection”. The first sub-task aims at detecting entity types and locations in the text and the second sub-task aims at detecting just entity locations.

The remainder of the paper is organized as follows. In Section 2 we briefly describe the data to be anonymized. Section 3 describes the methodology we propose. Results and discussions are presented in Section 4. Section 5 presents the conclusions and depicts some lines of future work.

2 Data Description

The MEDDOCAN challenge task aims at identifying and extracting several types of PHI categories from plain text medical documents. The PHI categories are grouped into eight main categories with 22 sub-categories. The corpora released for the tasks consists of 1000 documents, divided into: 500 as training data, 250 as development data and 250 as test data. The distributions of PHI categories and sub-categories in the training, development and test data are shown in Table 1.

Table 1: Distributions of PHI categories in the training, development and test corpora

PHI Category	**Sub-Category**	**Training Data**	**Development Data**	**Test Data**
AGE	EDAD_SUJETO_ASISTENCIA	1035	521	518
CONTACT	CORREO_ELECTRONICO	469	241	249
	NUMERO_FAX	15	6	7
	NUMERO_TELEFONO	58	25	26
DATE	FECHAS	1231	724	611
ID	ID_ASEGURAMIENTO	391	194	198
	ID_CONTACTO_ASISTENCIAL	77	32	39
	ID_EMPLEO_PERSONAL_SANITARIO	0	1	0
	ID_SUJETO_ASISTENCIA	567	292	283
	ID_TITULACION_PERSONAL_SANITARIO	471	226	234
LOCATION	CALLE	862	434	413
	CENTRO_SALUD	6	2	6
	HOSPITAL	255	140	130
	INSTITUCION	98	72	67
	PAIS	713	347	363
	TERRITORIO	1875	987	956
NAME	NOMBRE_PERSONAL_SANITARIO	1000	497	501
	NOMBRE_SUJETO_ASISTENCIA	1009	503	502
OTHER	FAMILIARES_SUJETO_ASISTENCIA	243	92	81
	OTROS_SUJETO_ASISTENCIA	9	6	7
	SEXO_SUJETO_ASISTENCIA	925	455	461
PROFESSION	PROFESION	24	4	9
Total:		11 333	5801	5661

3 Methodology

We developed an automatic system to detect PHI categories from Spanish medical documents. The next subsections describe the steps followed to train and use the system.

3.1 Text Tokenization

In this step, we tokenize the text at two levels: sentence-level and word-level. First, a sentence tokenizer takes a single document as input and produces list of sentences. Afterwards, we split each single sentence into a list of tokens. The sentence tokenizer is based on newline delimiter whereas a manually-crafted regular expression based tokenizer and a spaCy pre-trained model for Spanish[1] are used sequentially to perform the word-level tokenization.

3.2 Rules Generation

In this step, we developed a data-driven regular expression generator so that we avoid implementing hand crafted regular expression rules. This generator analyses all the appearances of the PHI categories in the training data set and, from that, it generates rules to detect those categories. These rules are later used to extract sudo-labelled tokens that are used to guide the CRF tagger in taking the final decision.

3.3 Feature Extraction

We extract a wide variety of linguistic features, similarly to previous studies [9, 8]. These features characterize the semantics of PHI terms. The main types of features are:

- **Lexical Features**: they include the target word itself, its prefix and suffix, word lemma, and Part-of-Speech (POS) tag.
- **Orthographic Features**: they detail word form information, e.g. target word length, word shape (CAPITALIZED, ALL_UPPER, ALL_ LOWER, MIX), ends with s, contains alpha and contains number.
- **RegEx features**: a RegEx model is used as first-pass recognizer for the PHI entities in the text. We use the output of the RegEx model to detect the location of the token, either at the beginning, middle, end or outside of PHI entity.
- **External Resource Features**: we also consider if a token appears into one or several external resources, which include lists of English and Spanish names of countries and cities, names and abbreviations of time expressions (e.g. 'año', 'mes'), or names and abbreviations of places (e.g. 'plaza', 'av.'). Additional resources include lists of Spanish last names, Spanish first names, addresses, hospitals, cities and towns, professions and autonomous communities, and provinces.

Extracting these features from just the target word does not consider the context in which the word appears, which may lead to misclassifying tokens due to language ambiguity. To tackle this, we consider a window of 5 words centered at the target word (i.e., the two words on the left and the two words on the the right).

3.4 Training the system

We used both a set of automatically-crafted rules (RegEx model) and Conditional Random Fields[4] (CRF model) to identify PHIs in medical documents. The system is implemented using Python 3.7 with sklearn-crfsuite package [3] and spaCy package [1] for the tokenization. We also use the BIO tagging scheme to set the labels of the tokens [7]. Each word token in the document is labeled using one of three possible tags: *B*, *I*, or *O*, which indicate if the word is at the beginning, middle, or outside of a PHI entity.

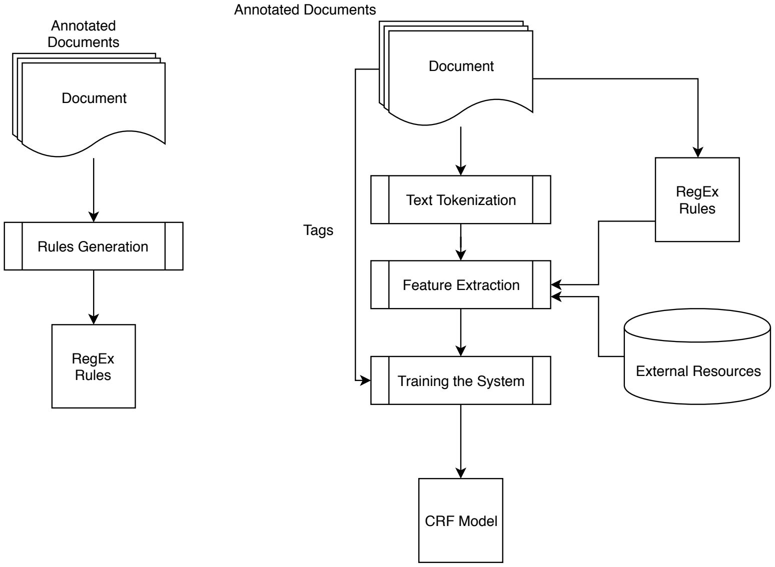

Fig. 2: Training the system

Fig 2 shows that our system has two outputs: the RegEx model and the CRF model. The RegEx model is built by the automatic rule extractor by analyzing the PHI categories that appear in well-structured contexts (e.g. Nombre: Xxxxx., Fecha de nacimiento: dd/mm/yyyy.).

The CRF model is trained by passing all the extracted features from the tokens plus the decision of the RegEx model which add extra information and make the decision easier for the CRF model.

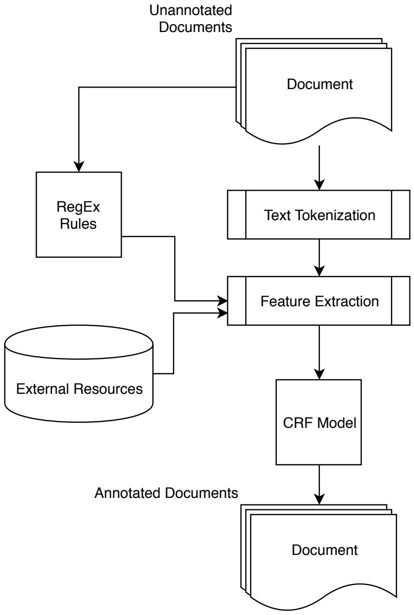

Fig. 3: Using the system

3.5 Using the system

Fig 3 shows how both RE and CRF models are used to make the annotations. Even though the RegEx model is accurate enough to detect well-structured entities, it is not effective in front of small changes in the text format. So, we decided to use the RegEx model to perform a preliminary annotation, which is then passed to the CRF model that will make the final decision.

4 Results and Discussions

The performance of the detection of PHI categories has been evaluated using Precision, Recall and F1 scores at the entity level. The results of our system on the test set for the different PHI categories are shown in Table 2; the confusion matrix is shown in Table 3. Notice that categories that have low frequency in the training dataset have less F1 score (e.g. NUMERO_FAX appears only 15 times in the training set and CENTRO_SALUD appears six times). This result is expected because the model didn't get enough examples in order to learn how to accurately detect them.

Table 2: Overall performance of PHI sub-categories detection on the first-task

PHI Category	Sub-Category	#Expected	#Predicted	#Correct	Precision	Recall	F1
AGE	ESA	514	516	501	97.09	97.47	97.28
CONTACT	CE	248	249	246	98.80	99.19	98.99
	NF	7	7	5	71.43	71.43	71.43
	NT	26	26	22	84.62	84.62	84.62
DATE	Fech	612	612	603	98.53	98.53	98.53
ID	IA	199	199	197	98.99	98.99	98.99
	ICA	38	39	38	97.44	100.00	98.70
	IEPS	0	0	0	0.00	0.00	0.00
	ISA	277	277	274	98.92	98.92	98.92
	ITPS	234	235	233	99.15	99.57	99.36
LOCATION	Call	412	406	382	94.09	92.72	93.40
	CS	4	6	2	33.33	50.00	40.00
	Hosp	129	120	109	90.83	84.50	87.55
	Inst	58	49	24	48.98	41.38	44.86
	Pais	366	354	348	98.31	95.08	96.67
	Terr	950	945	916	96.93	96.42	96.68
NAME	NPS	499	501	497	99.20	99.60	99.40
	NSA	503	502	502	100.00	99.80	99.90
OTHER	FSA	82	76	59	77.63	71.95	74.68
	OSA	6	2	0	0.00	0.00	0.00
	SSA	461	459	455	99.13	98.70	98.91
PROFESSION	Prof	6	4	3	75.00	50.00	60.00

The overall results of our system for the two sub-tracks of the competition on the development and test datasets are shown in Table 4. We see very small differences in F1 scores between the development and test datasets. This proves that our system generalizes well in front of new data.

5 Conclusion and Future Work

We presented a hybrid system that automatically detects PHI entities from plain text medical documents. The system consists of an automatically constructed RegEx model and a trained CRF model. The design of the system, which includes using a variety of linguistic and semantic features to increase the accuracy, ensures that it generalizes well in front of new data.

Finally, because of the rules that we get from the automatic RegEx generator are not fully generalized, in future work, we plan to implement an automatic optimizer to get a better result.

Table 3: Confusion matrix of our system on the test dataset

	Output																							
Key	**Call**	**CS**	**CE**	**ESA**	**FSA**	**Fech**	**Hosp**	**IA**	**ICA**	**IEPS**	**ISA**	**ITPS**	**Inst**	**NPS**	**NSA**	**NF**	**NT**	**OSA**	**Pais**	**Prof**	**SSA**	**Terr**	**Miss**	**Total**
Call	382		1																			1	28	412
CS		2																					2	4
CE			246																				2	248
ESA				501	1	1																	11	514
FSA				3	59													1					19	82
Fech						603			1								1					1	6	612
Hosp		1					109																19	129
IA								197				1											1	199
ICA									38															38
IEPS																								
ISA								1			274												2	277
ITPS												233											1	234
Inst		1					1						24										32	58
NPS														497									2	499
NSA					1										502									503
NF																5							2	7
NT																1	22						3	26
OSA											1												5	6
Pais																			348			5	13	366
Prof																				3			3	6
SSA					1													1			455		4	461
Terr						1							2						1			916	30	950
Spur	24	2	2	12	14	7	10	1			2	1	23	4		1	3		5	1	4	22		138
Total	406	6	249	516	76	612	120	199	39		277	235	49	501	502	7	26	2	354	4	459	945	185	5769

Spur=Spurious, Miss=Missing

Table 4: Micro-averaged results on the development and test datasets

Sub-Track	**Development Data**			**Test Data**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Sub-Track 1 (NER)	97.36	95.45	96.40	96.99	95.67	96.33
Sub-Track 2 (Spans strict)	97.78	95.86	96.81	97.53	96.20	96.86
Sub-Track 2 (Spans merged)	98.33	96.58	97.45	98.13	96.89	97.50

Acknowledgments

This work was partly supported by the European Commission (project H2020-700540 ”CANVAS”), the Government of Catalonia (ICREA Acadèmia Prize to J. Domingo-Ferrer and grant 2017 SGR 705) and the Spanish Government (projects RTI2018-095094-B-C2 “CONSENT” and TIN2016-80250-R “Sec-MCloud”). While the authors are with the UNESCO Chair in Data Privacy, the opinions expressed in this paper are the authors’ own and do not necessarily reflect the views of UNESCO.

References

1. Honnibal, M., Montani, I.: spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear (2017)
2. I2B2: i2b2: Informatics for integrating biology the bedside. https://www.i2b2.org/, last accessed: 25-June-2019
3. Korobov, M.: sklearn-crfsuite. https://sklearn-crfsuite.readthedocs.io/en/latest/, last accessed: 31-May-2019
4. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, ICML’01
5. Marimon, M., Gonzalez-Agirre, A., Intxaurrondo, A., Rodrguez, 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
6. Regulation, G.D.P.: Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46. Official Journal of the European Union (OJ) **59**(1-88), 294 (2016)
7. Sang, E.F., Veenstra, J.: Representing text chunks. In: Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics. pp. 173–179. Association for Computational Linguistics (1999)
8. Stubbs, A., Kotfila, C., Uzuner, Ö.: Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. Journal of biomedical informatics **58**, S11–S19 (2015)
9. Yang, H., Garibaldi, J.M.: Automatic detection of protected health information from clinic narratives. Journal of biomedical informatics **58**, S30–S38 (2015)