Automatic De-Identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results

Montserrat Marimon[2], Aitor Gonzalez-Agirre[1,2], Ander Intxaurrondo[1,2], Heidy Rodríguez[1], Jose Antonio Lopez Martin[3], Marta Villegas[1,2], and Martin Krallinger*[1,2]

[1] Centro Nacional de Investigaciones Oncológicas (CNIO)
[2] Barcelona Supercomputing Center (BSC)
{montserrat.marimon, aitor.gonzalez, marta.villegas, martin.krallinger}@bsc.es
[3] Hospital 12 de Octubre - Madrid

Abstract. There is an increasing interest in exploiting the content of electronic health records by means of natural language processing and text-mining technologies, as they can result in resources for improving patient health/safety, aid in clinical decision making, facilitate drug repurposing or precision medicine. To share, re-distribute and make clinical narratives accessible for text mining research purposes, it is key to fulfill legal conditions and address restrictions related data protection and patient privacy. Thus, clinical records cannot be shared directly "as is". A necessary precondition for accessing clinical records outside of hospitals is their de-identification or exhaustive removal/replacement of all mentioned privacy related protected health information phrases. Providing a proper evaluation scenario for automatic anonymization tools is key for approval of data redistribution. The construction of manually de-identified medical records is currently the main rate and cost-limiting step for secondary use applications of clinical data. This paper summarizes the settings, data and results of the first shared track on anonymization of medical documents in Spanish, the MEDDOCAN (Medical Document Anonymization) track. This track relied on a carefully constructed synthetic corpus of clinical case documents, the MEDDOCAN corpus, following annotation guidelines for sensitive data based on the analysis of the EU General Data Protection Regulation. A total of 18 teams (from the 51 registrations) submitted 63 runs for first sub-track 1 and 61 systems for the second sub-track. The top scoring systems were based on sophisticated deep learning approaches, representing strategies that can significantly reduce time and costs associated to accessing textual data containing privacy-related sensitive information. The results of this track might help in lowering the clinical data access hurdle for Spanish language technology developers, showing also potentials for similar settings using data in other languages or from different domains.

Keywords: GDPR · IberLEF · de-identification · anonymization · sensitive data · data privacy · named entity recognition · deep learning · Gold Standard corpus · NLP · Plan TL · text mining · EHR.

1 Introduction

There is an increasing interest in exploiting the content of unstructured clinical narratives by means of language technologies. Therefore, and because there is clear interest in the health sector by the language technology industry, one of the flagship projects of the Spanish National Plan for the Advancement of Language Technology (Plan TL[4]) is related to the clinical and biomedical field. The Plan TL has promoted the generation of a collection of resources for Spanish biomedical NLP[5], including corpora [26], gazetteers [26], components [2, 19] and tools, as well as evaluation efforts [18, 11, 12]. Due to their central role in fostering language technology resources, the promotion of shared tasks and evaluation campaigns is of particular relevance for the Plan TL, being considered a key instrument for: (1) independent quality evaluation of components, (2) promotion of standards, interoperability and harmonization of resources, (3) generation of new systems, tools and software components, (4) promotion of confidence by end users, investors and commercial partners in language technologies, (5) promoting new start ups and innovative ideas, (6) improving access to data, (7) create collaborative research interactions and networks and (8) serve as a knowledge transfer and learning experience engaging both academia and industry. Structured clinical data, in the form of codified clinical information using controlled indexing vocabulary such as ICD10, only covers a fraction of the medically relevant information stored in electronic health records (EHRs) and clinical texts. Complex relations such as drug-related allergies, constituting a serious health risk, cannot be captured well by the coding schemes followed typically by clinical documentalists and, thus, require direct processing of clinical narrative texts.

Being able to transform automatically clinical documents into some structured representations is nonetheless needed to enable secondary use of EHRs to carry out population and epidemiological studies, to detect medication-related adverse events or for monitoring systematically treatment-related responses, just to name a few.

To be able to share, re-distribute and make clinical narratives accessible for text mining and natural language processing (NLP) purposes, it is key to fulfill legal conditions and address restrictions related data protection and patient privacy legislations [5]. Some efforts have been made to examine GDPR demands

[4] https://www.plantl.gob.es

[5] https://github.com/PlanTL-SANIDAD

for the construction of de-identified textual corpora for research purposes [15]. Thus, clinical records with protected health information (PHI) cannot be directly shared "as is", due to privacy constraints, making it particularly cumbersome to carry out NLP research in the medical domain. A necessary precondition for accessing clinical records outside of hospitals is their de-identification, i.e., the exhaustive removal (or replacement) of all mentioned PHI phrases.

Studies describing services for pseudonymization of EHRs based on standards such as the ISO/EN 13606 were previously published for data in Spanish [4], but are generally limited to the structured fields of the clinical documents, have not been evaluated against any particular Gold Standard dataset (i.e. lack proper evaluation), and, most importantly, are not accessible or released on public software repositories, making it impossible to actually carry out a proper independent benchmark study. Providing a proper evaluation scenario of automatic anonymization tools, with well-defined sensitive data types, is crucial for approval of data redistribution consents signed by ethical committees of healthcare institutions. It is important to highlight that the construction of manually de-identified medical records is currently the main rate and cost-limiting step for secondary use applications. Moreover, such settings also require very carefully designed annotation guidelines and interfaces to assure that there is no leak of sensitive information from clinical records and that the resulting de-identified datasets are compliant with all legal constraints.

The practical relevance of anonymization or de-identification of clinical texts motivated the proposal of two shared tasks, the 2006 and 2014 de-identification tracks [24, 21], organized under the umbrella of the i2b2 (i2b2.org) community evaluation effort. The i2b2 effort has deeply influenced the clinical NLP community worldwide, but was focused on documents in English and covering characteristics of US-healthcare data providers. Systems used for de-identifying English clinical texts like Carafe, based on Conditional Random Fields or MIST (the MITRE Identification Scrubber Toolkit) have benefited from i2b2 shared tasks to improve, evaluate and analyze these tools. The interest in automated de-identification and anonymization systems is not limited to data in English, and there is also a growing awareness in developing such systems for other languages, such as French [9, 7], German [22], Dutch [20], Portuguese [13], Danish [17], Swedish [1] or Norwegian [23].

In case of texts in Spanish, there has been so far a rather limited attempt in developing and characterizing automatic de-identification strategies [10, 14, 25, 6], even though some in house tools, such as the AEMPS anonymizer or a recent publication by Medina and Turmo [14] show that efforts in this direction are being made and such tools are already explored in practice. We, therefore, organized the first community challenge track specifically devoted to the anonymization of medical documents in Spanish, called the MEDDOCAN (Medical Document Anonymization) track, as part of the IberLEF evaluation initiative.

2 Methods

2.1 Track Description

The MEDDOCAN track was one of the nine challenge tracks of the Iberian Languages Evaluation Forum (IberLEF 2019)[6] evaluation campaign, which had the goal of promoting the development of language technologies for Iberian languages. MEDDOCAN was the first community challenge track specifically devoted to the anonymization of medical documents in Spanish and it evaluated the performance of the systems for identifying and classifying sensitive information in clinical case studies written in Spanish.

The evaluation of automatic predictions for this track had two different scenarios or sub-tracks:

1. *NER offset and entity type classification*: the first sub-track was focused on the identification and classification of sensitive information (e.g., patient names, telephones, addresses, etc.).
2. *Sensitive span detection*: the second sub-track was focused on the detection of sensitive text more specific to the practical scenario necessary for the release of de-identified clinical documents, where the objective is to identify and to mask confidential data, regardless of the real type of entity or the correct identification of PHI type.

2.2 Track data

For this track, we prepared a synthetic corpus of clinical cases enriched with PHI expressions, named the MEDDOCAN corpus. The MEDDOCAN corpus, of 1,000 clinical case studies, 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.

To carry out the manual annotation, we constructed the first public guidelines for PHI in Spanish [16], following the specifications derived from the General Data Protection Regulation (GDPR) of the EU, as well as the annotation guidelines and types defined by the i2b2 de-identification tracks, based on the US Health Insurance Portability and Accountability Act (HIPAA). The construction of these annotation guidelines involved active feedback over a six-month period from a hybrid team of nine persons with expertise in both healthcare and NLP, resulting in a 28-page document that has been distributed along with the corpus. Along with the annotation rules, illustrative examples were provided to make the interpretation and use of the guidelines as easy as possible.

The MEDDOCAN corpus was randomly sampled into three subset: the train set, which contained 500 clinical cases, and the development and test sets of 250 clinical cases each. These clinical cases were manually annotated using a customized version of AnnotateIt. Then, the BRAT annotation toolkit was used to

[6] http://hitz.eus/sepln2019/?q=node/21

correct errors and add missing annotations, achieving an inter-annotator agreement (IAA) of 98% (calculated with 50 documents). Together with the test set, we released an additional collection of 3,501 documents (background set[7]) to make sure that participating teams were not able to do manual corrections and also to promote that these systems would potentially be able to scale to larger data collections.

The MEDDOCAN annotation guidelines defined a total of 29 entity types. Table 1 summarizes the list of sensitive entity types defined for the MEDDOCAN track and the number of occurrences among the training, development and test sets.

Table 1. Entity type distribution among the data sets.

Type	Train	Dev	Test	Total
TERRITORIO	1875	987	956	3818
FECHAS	1231	724	611	2566
EDAD_SUJETO_ASISTENCIA	1035	521	518	2074
NOMBRE_SUJETO_ASISTENCIA	1009	503	502	2014
NOMBRE_PERSONAL_SANITARIO	1000	497	501	1998
SEXO_SUJETO_ASISTENCIA	925	455	461	1841
CALLE	862	434	413	1709
PAIS	713	347	363	1423
ID_SUJETO_ASISTENCIA	567	292	283	1142
CORREO_ELECTRONICO	469	241	249	959
ID_TITULACION_PERSONAL_SANITARIO	471	226	234	931
ID_ASEGURAMIENTO	391	194	198	783
HOSPITAL	255	140	130	525
FAMILIARES_SUJETO_ASISTENCIA	243	92	81	416
INSTITUCION	98	72	67	237
ID_CONTACTO_ASISTENCIAL	77	32	39	148
NUMERO_TELEFONO	58	25	26	109
PROFESION	24	4	9	37
NUMERO_FAX	15	6	7	28
OTROS_SUJETO_ASISTENCIA	9	6	7	22
CENTRO_SALUD	6	2	6	14
ID_EMPLEO_PERSONAL_SANITARIO	0	1	0	1
IDENTIF VEHICULOS NRSERIE PLACAS	0	0	0	0
IDENTIF DISPOSITIVOS NRSERIE	0	0	0	0
NUMERO BENEF PLAN SALUD	0	0	0	0
URL WEB	0	0	0	0
DIREC PROT INTERNET	0	0	0	0
IDENTF BIOMETRICOS	0	0	0	0
OTRO NUMERO IDENTIF	0	0	0	0

The MEDDOCAN corpus was distributed in plain text in UTF-8 encoding, where each clinical case was stored as a single file, while PHI annotations were released in the BRAT format, which makes visualization of results straightforward, as you can see in Fig. 1 For this track, we also prepared a conversion script[8] between the BRAT annotation format and the annotation format used by the

[7] The background set included the train, development and test sets, and an additional collection of 2,751 clinical cases (totalling 3,751 clinical cases).

[8] https://github.com/PlanTL-SANIDAD/MEDDOCAN-Format-Converter-Script

previous i2b2 effort, to make comparison and adaptation of previous systems used for English texts easier.

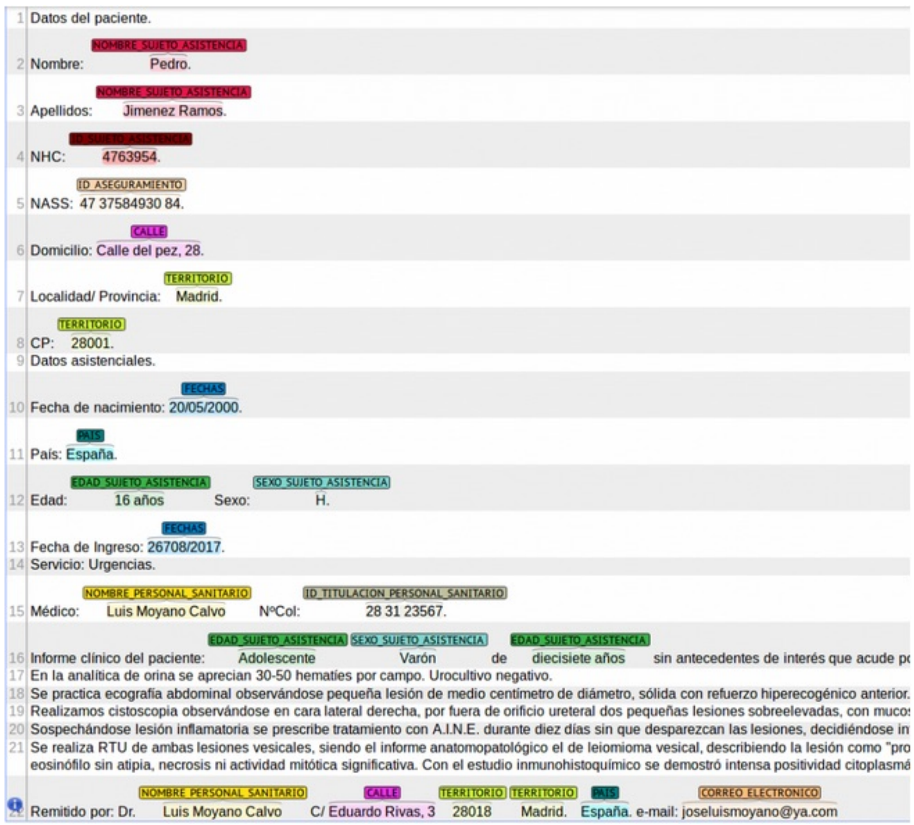

Fig. 1. An example of MEDDOCAN annotation visualized using the BRAT annotation interface..

2.3 Evaluation metrics

We developed an evaluation script that supported the evaluation of the predictions of the participating teams. For both sub-tracks the primary evaluation metrics used consisted of standard measures from the NLP community, namely micro-averaged precision, recall, and balanced F-score, being the last one the only official evaluation measure of both sub-tracks:

$$\text{Precision: } P = \frac{TP}{TP+FP}$$

$$\text{Recall: } R = \frac{TP}{TP+FN}$$

$$\text{F-score: } F1 = 2 * \frac{(P*R)}{(P+R)}$$

where TP = true positives, FP = false positives and FN = false negatives.

In addition, in case of the first sub-track, the leak scores; i.e., #false negatives/#sentences present, previously proposed for the i2b2 challenges, were also computed. In the case of the second sub-track, we also additionally computed another evaluation where we merged the spans of PHI connected by non-alphanumerical characters.

Teams could submit up to five prediction files (runs) in a predefined prediction format (BRAT o i2b2).

3 Participation and Results

3.1 Participation

To participate in the MEDDOCAN track it was necessary to register both on the official website[9] and in the CodaLab competition[10]. Training and development sets were made available for download on the official website[11], and the evaluation script was uploaded to GitHub[12], to ensure a transparent evaluation.

Submissions had to be provided in a predefined prediction format (BRAT or i2b2). The participants had a period of almost two months to develop their system. In the middle of this period, the text and background sets were released with the 3,751 documents that the participants had to process and label, although the final evaluation was done on the 250 documents of the test set. As we have mentioned, the participants could submit a maximum of 5 system runs, and, once the submission deadline expired, we published the Gold Standard annotations of the test set, in order to ensure a transparent evaluation process.

A total of 18 teams participated in the track, submitting a total of 63 systems for sub-track 1 and 61 systems for sub-track 2. Teams from eight different nationalities participated in the track: ten from Spain, two from the United States, and one from Argentina, China, Germany, Italy, Japan, and Russia. Among all the participants, only one belonged to an institution of a commercial nature. Table 2 summarizes the most relevant information about the participants.

3.2 Baseline system

We produced a baseline system using a vocabulary transfer approach. Each annotation from the train and development datasets was transferred to the test dataset using strict string matching. For those cases where the text was the same, but the entity type was different, we decided to annotate all entity types that matched that text.

[9] http://temu.bsc.es/meddocan/
[10] https://competitions.codalab.org/competitions/22643
[11] http://temu.bsc.es/meddocan/index.php/data/
[12] https://github.com/PlanTL-SANIDAD/MEDDOCAN-CODALAB-Evaluation-Script

Table 2. Overview of Team Participation in the MEDDOCAN track.

Username	Organization/Institution/Company	Members	Country	Comm.
Aspie96	University of Turin	1	Italy	No
ccolon	Carlos III University of Madrid	3	Spain	No
Fadi	Universitat Rovira i Virgili, CRISES group	6	Spain	No
FSL	Unaffiliated	1	Spain	No
gauku	University of Pennsylvania	2	USA	No
jiangdehuan	Harbin Institute of Technology	9	China	No
jimblair	University of Maryland	2	USA	No
Jordi	Centro de Estudios de la Real Academia Espaola	1	Spain	No
lsi_uned	National Distance Education University	4	Spain	No
lsi2_uned	National Distance Education University	2	Spain	No
lukas.lange	Bosch Center for Artificial Intelligence	3	Germany	Yes
m.domrachev	Unaffiliated	3	Russia	No
mhjabreel	Universitat Rovira i Virgili, iTAKA Research Group	5	Spain	No
nperez	Vicomtech	4	Spain	No
plubeda	Advanced Studies Center in ICT, SINAI	4	Spain	No
sohrab	National Institute of Advanced Industrial Science and Technology	3	Japan	No
vcotik	Universidad de Buenos Aires	3	Argentina	No
VSP	Carlos III University of Madrid	1	Spain	No

3.3 Results

Table 3 shows the results for sub-track 1 (*NER offset and entity type classification*), ordered by team performance (first column), then system performance (second column). Note that almost all of the systems were well above the baseline, which would rank 18.

The top scoring system was submitted by *lukas.lange*, with an F-score of 0.96961, being relatively close to the next two participants: *Fadi*, ranked 2nd with a F-score of 0.96327, and *nperez*, ranked 3rd with a F-score of 0.96018. If we focus our attention on the recall (which is a crucial metric for de-identification) obtained by the systems, we see that best performing systems were *lukas.lange*, with a recall of 0.96944, FSL, with a recall of 0.96043, and *mhjabreel*, with a recall of 0.95707.

Tables 6 and 7 show the results for sub-track 2A (*Sensitive token detection with strict spans*) and sub-track 2B (*Sensitive token detection with merged spans*), respectively, ordered by team performance (first column), then system performance (second column). As in sub-track 1, almost all of the systems were well above the baseline.

The top scoring system for sub-track 2A was submitted by *lukas.lange*, with a F-score of 0.97491. The second team was *Fadi*, with a F-score of 0.96861, and the third team was *nperez*, with a F-score of 0.96799. The best results in terms of recall were obtained by *lukas.lange*, with a recall of 0.97474, *mhjabreel*, with a recall of 0.96591, and, *FSL*, with a recall of 0.96520.

The results for sub-track 2B were quite surprising. The top scoring systems was submitted by *lukas.lange*, with a F-score of 0.98530, but the second team for this sub-track was *jiangdehuan*, with a F-score of 0.98184, very close to the best team. Note that *jiangdehuan* ranked 7th for sub-tracks 1 and 2A (their best system ranked 25th). This boost in performance was quite surprising and probably need further analysis. The third team was *nperez*, with a F-score of 0.97593. Finally, the best results in terms of recall were obtained by *jiangdehuan*,

Table 3. Results for sub-track 1: *NER offset and entity type classification*.

Team Rank	System Rank	User	Leak	Precision	Recall	F1
1	1	lukas.lange	0.02299	0.96978	0.96944	0.96961
	2		0.02378	0.97078	0.96838	0.96958
	3		0.02365	0.97044	0.96856	0.96950
	4		0.02432	0.96956	0.96767	0.96861
	5		0.02724	0.96720	0.96379	0.96549
2	6	Fadi	0.03255	0.96991	0.95672	0.96327
	7		0.03388	0.97160	0.95495	0.96321
	8		0.03508	0.97191	0.95337	0.96255
	9		0.03322	0.96867	0.95584	0.96221
	10		0.03402	0.96933	0.95478	0.96200
3	11	nperez	0.03282	0.96403	0.95637	0.96018
	15		0.03946	0.96823	0.94754	0.95777
	19		0.03946	0.96492	0.94754	0.95615
	20		0.04146	0.96570	0.94489	0.95518
	21		0.04770	0.97124	0.93658	0.95360
4	12	FSL	0.02976	0.95857	0.96043	0.95950
	16		0.03096	0.95597	0.95884	0.95740
	18		0.03096	0.95547	0.95884	0.95715
5	13	mhjabreel	0.03242	0.95978	0.95690	0.95834
	14		0.03282	0.95976	0.95637	0.95806
	17		0.03229	0.95741	0.95707	0.95724
	22		0.03734	0.95610	0.95036	0.95322
	24		0.04783	0.94779	0.93641	0.94207
6	23	lsi_uned	0.05381	0.95877	0.92846	0.94337
7	25	jiangdehuan	0.03574	0.92806	0.95248	0.94011
	26		0.03681	0.92892	0.95107	0.93986
	28		0.04106	0.92868	0.94542	0.93697
	30		0.03747	0.92217	0.95019	0.93597
	58		0.16835	0.91580	0.77619	0.84023
8	27	jimblair	0.06617	0.96451	0.91203	0.93753
	29		0.06604	0.96164	0.91221	0.93627
	33		0.05395	0.93306	0.92828	0.93067
	35		0.05567	0.93125	0.92598	0.92861
	36		0.05594	0.92547	0.92563	0.92555
9	31	ccolon	0.05421	0.93653	0.92793	0.93221
	34		0.05195	0.92700	0.93093	0.92896
10	32	sohrab	0.07002	0.95676	0.90691	0.93117
	39		0.08026	0.94119	0.89331	0.91662
	40		0.07348	0.92553	0.90231	0.91377
	41		0.06325	0.90997	0.91592	0.91293
	42		0.08570	0.93252	0.88606	0.90870
11	37	Jordi	0.07095	0.93150	0.90567	0.91841
	38		0.06218	0.91912	0.91733	0.91822
	57		0.12091	0.86571	0.83925	0.85227
12	43	plubeda	0.08491	0.92113	0.88712	0.90381
	52		0.11998	0.89369	0.84049	0.86627
	62		0.34600	0.66457	0.54001	0.59585
13	44	m.domrachev	0.08318	0.91098	0.88942	0.90007
	47		0.07813	0.89313	0.89613	0.89463
	48		0.08225	0.87824	0.89066	0.88441
14	45	lsi2_uned	0.12052	0.96902	0.83978	0.89978
	59		0.18164	0.91929	0.75852	0.83120
15	46	vcotik	0.09022	0.91413	0.88006	0.89677
	49		0.07308	0.86568	0.90284	0.88387
	50		0.07308	0.86568	0.90284	0.88387
	51		0.07308	0.86568	0.90284	0.88387
	60		0.13540	0.76223	0.82000	0.79006
16	53	VSP	0.10165	0.85535	0.86486	0.86008
	54		0.10165	0.85535	0.86486	0.86008
	55		0.10058	0.84639	0.86628	0.85622
	56		0.10058	0.84639	0.86628	0.85622
17	61	gauku	0.31464	0.90841	0.58170	0.70924
-	-	**Baseline-VT**	*0.37351*	*0.37023*	*0.50344*	*0.42668*
18	63	Aspie96	0.35384	0.18829	0.52959	0.27781

Table 4. Results by label for sub-track 1: *NER offset and entity type classification.*

Category	Sub-category	Best Team(s)	Leak	Precision	Recall	F1
AGE	EDAD_SUJETO_ASISTENCIA	jiangdehuan	0.0004	0.9828	0.9942	0.9885
CONTACT	CORREO_ELECTRONICO	lukas.lange nperez	0.0001	0.9920	0.9960	0.9940
	NUMERO_FAX	jimblair jiangdehuan lsi_uned	0.0000	1.0000	1.0000	1.0000
	NUMERO_TELEFONO	jiangdehuan	0.0000	1.0000	1.0000	1.0000
DATE	FECHAS	jiangdehuan lukas.lange	0.0004	0.9935	0.9951	0.9943
ID	ID_ASEGURAMIENTO	FSL jiangdehuan jimblair lsi_uned lukas.lange m.domrachev mhjabreel nperez sohrab	0.0001	1.0000	0.9950	0.9975
	ID_CONTACTO_ASISTENCIAL	lsi2_uned lukas.lange mhjabreel nperez sohrab vcotik	0.0000	1.0000	1.0000	1.0000
	ID_SUJETO_ASISTENCIA	jiangdehuan	0.0001	0.9758	0.9965	0.9860
	ID_TITULACION_PERSONAL_SANITARIO	jiangdehuan jimblair lsi_uned lsi2_uned lukas.lange mhjabreel nperez sohrab	0.0000	0.9957	1.0000	0.9979
LOCATION	CALLE	lukas.lange	0.0031	0.9353	0.9443	0.9398
	CENTRO_SALUD	FSL jiangdehuan lsi2_uned lukas.lange mhjabreel	0.0001	1.0000	0.8333	0.9091
	HOSPITAL	FSL	0.0016	0.9672	0.9077	0.9365
	INSTITUCION	jiangdehuan	0.0036	0.6061	0.5970	0.6015
	PAIS	jiangdehuan	0.0004	0.9890	0.9917	0.9904
	TERRITORIO	lukas.lange	0.0035	0.9759	0.9728	0.9743
NAME	NOMBRE_PERSONAL_SANITARIO	lukas.lange	0.0003	0.9960	0.9960	0.9960
	NOMBRE_SUJETO_ASISTENCIA	jiangdehuan	0.0000	1.0000	1.0000	1.0000
OTHER	FAMILIARES_SUJETO_ASISTENCIA	lukas.lange	0.0017	0.8293	0.8395	0.8344
	OTROS_SUJETO_ASISTENCIA	nperez	0.0008	1.0000	0.1429	0.2500
	SEXO_SUJETO_ASISTENCIA	FSL	0.0004	0.9892	0.9935	0.9913
PROFESSION	PROFESION	lukas.lange	0.0004	1.0000	0.6667	0.8000

with a recall of 0.98335, *lukas.lange*, with a recall of 0.98264, and, *mhjabreel*, with a recall of 0.97471.

An analysis of errors showed that some of the annotations in the Gold Standard (GS) corpus were not detected by any of the systems (at least not exactly). Some of them are listed here:

- HOSPITAL: *Hospital General de Agudos P. Piñero*
- FAMILIARES_SUJETO_ASISTENCIA: *tres hermanos varones sordomudos y otro con baja visión*
- OTROS_SUJETO_ASISTENCIA: *estudiante de administración de empresas*

On the contrary, some systems annotated entities that were not in the GS but probably should be. For instance, "*ex-operario de la industria textil*" was annotated as *PROFESION* by *jiangdehuan*, *jimblair*, and *Jordi*, but this annotation was not in the GS.

Table 5. Statistics by track.

Track	Measure	Leak	Precision	Recall	F1
1	Min	0.02299	0.18829	0.52959	0.27781
	Mean	0.07594	0.90219	0.89327	0.89410
	Median	0.05567	0.93252	0.92598	0.93117
	Max	0.35384	0.97191	0.96944	0.96961
	Std	0.06857	0.10736	0.09116	0.10223
2A	Min	-	0.19771	0.55609	0.29171
	Mean	-	.92907	0.91058	0.91724
	Median	-	0.95965	0.92616	0.94118
	Maxi	-	0.97747	0.97474	0.97491
	Std	-	0.10200	0.08190	0.09535
2B	Min	-	0.19780	0.55626	0.29183
	Mean	-	0.94661	0.92494	0.93320
	Median	-	0.97180	0.95001	0.95774
	Maxi	-	0.98749	0.98335	0.98530
	Std	-	0.10260	0.08247	0.09624

3.4 Combination of systems

One of the primary goals of this track was to develop systems capable of completely de-identifying sensitive information from clinical documents. However, none of submitted systems managed to obfuscate all the sensitive information. In this section, we present two experiments we performed that evaluated the performance of combined systems to de-identify the test dataset without leaks. The first experiment was based on a joint system, the second experiment, on a voting system.

Joint system The goal of this experiment was to find the combination of individual systems that achieved the best possible performance. For this, first, we ranked all the systems by F-score, and then we joined the annotations of the two best system. If the performance of the Joint system improved, we continued with the next best system, if not, we kept the previous system (or the previous joint system). We repeated this until no systems were left. We measured the performance of the joint system using three metrics:

1. Best F1: If the F-score of the joint system improved when we added the annotations from the next system, we updated the joint system with the new one. If the F-score did not improve, but it was maintained and the recall was better, we also updated the joint system with the new one (same F-score, better recall, worse precision).
2. Best Recall: If the recall of the joint system improved, we updated the joint system, regardless of the drop in the F-score. It tried to maximize the chances of completely de-identifying the documents.
3. Balanced: If the recall of the joint system improved, we updated the joint system only if the decrease of the F-score was at much four times the increase

Table 6. Results for sub-track 2A: *Sensitive token detection (strict spans).*

Team Rank	System Rank	User	Precision	Recall	F1
1	1	lukas.lange	0.97508	0.97474	0.97491
	2		0.97574	0.97333	0.97453
	3		0.97540	0.97350	0.97445
	4		0.97522	0.97333	0.97427
	5		0.97217	0.96873	0.97045
2	6	Fadi	0.97529	0.96202	0.96861
	8		0.97507	0.96043	0.96770
	9		0.97556	0.95884	0.96713
	10		0.97351	0.96061	0.96701
	11		0.97569	0.95707	0.96629
3	7	nperez	0.97187	0.96414	0.96799
	15		0.97491	0.95407	0.96438
	20		0.97093	0.95001	0.96036
	21		0.96703	0.95337	0.96015
	22		0.97747	0.94259	0.95971
4	12	mhjabreel	0.96758	0.96467	0.96612
	13		0.96625	0.96591	0.96608
	14		0.96720	0.96379	0.96549
	19		0.96463	0.95884	0.96173
	23		0.95798	0.94648	0.95219
5	16	FSL	0.96315	0.96502	0.96409
	17		0.96231	0.96520	0.96375
	18		0.96180	0.96520	0.96350
6	24	lsi_uned	0.96406	0.93358	0.94858
7	25	jiangdehuan	0.93356	0.95813	0.94569
	26		0.93392	0.95619	0.94492
	30		0.92817	0.95637	0.94206
	31		0.93285	0.94966	0.94118
	57		0.91976	0.77954	0.84387
8	27	plubeda	0.96167	0.92616	0.94358
	45		0.93858	0.88271	0.90979
	59		0.86594	0.70288	0.77594
9	28	jimblair	0.96782	0.91910	0.94283
	32		0.96806	0.91539	0.94098
	33		0.96646	0.91609	0.94060
	34		0.96536	0.91556	0.93980
	36		0.95965	0.91592	0.93727
10	29	ccolon	0.94705	0.93835	0.94268
	35		0.93650	0.94047	0.93848
11	37	sohrab	0.96086	0.91079	0.93516
	40		0.93568	0.91221	0.92379
	41		0.92639	0.92033	0.92335
	43		0.94752	0.89931	0.92278
	44		0.91962	0.92563	0.92262
12	38	vcotik	0.94771	0.91238	0.92971
	50		0.87229	0.90973	0.89062
	51		0.87229	0.90973	0.89062
13	39	Jordi	0.93732	0.91132	0.92414
	42		0.92407	0.92228	0.92317
	56		0.87136	0.84473	0.85783
14	46	m.domrachev	0.91424	0.89260	0.90329
	48		0.89754	0.90055	0.89904
	49		0.88521	0.89772	0.89142
15	47	lsi2_uned	0.97187	0.84225	0.90243
	58		0.92207	0.76082	0.83372
16	52	VSP	0.86548	0.87511	0.87027
	53		0.86548	0.87511	0.87027
	54		0.85658	0.87670	0.86652
	55		0.85658	0.87670	0.86652
17	60	gauku	0.91421	0.58541	0.71376
-	-	**Baseline-VT**	*0.44174*	*0.50627*	*0.47181*
18	61	Aspie96	0.19771	0.55609	0.29171

Table 7. Results for sub-track 2B: *Sensitive token detection (merged spans).*

Team Rank	System Rank	User	Precision	Recall	F1
1	1	lukas.lange	0.98749	0.98311	0.98530
	2		0.98566	0.98264	0.98415
	3		0.98648	0.98145	0.98396
	4		0.98598	0.98162	0.98380
	7		0.98182	0.97730	0.97956
2	5	jiangdehuan	0.98033	0.98335	0.98184
	6		0.98029	0.98282	0.98155
	8		0.97496	0.98199	0.97846
	9		0.97962	0.97625	0.97793
	56		0.96913	0.80565	0.87986
3	10	nperez	0.97954	0.97235	0.97593
	20		0.97724	0.96666	0.97192
	21		0.98253	0.96136	0.97183
	22		0.98159	0.95890	0.97011
	27		0.98329	0.95001	0.96636
4	11	Fadi	0.98128	0.96886	0.97503
	14		0.98110	0.96734	0.97417
	16		0.97939	0.96750	0.97341
	17		0.98120	0.96573	0.97340
	18		0.98186	0.96419	0.97294
5	12	mhjabreel	0.97471	0.97471	0.97471
	13		0.97517	0.97350	0.97434
	15		0.97481	0.97297	0.97389
	19		0.97457	0.96957	0.97207
	28		0.97125	0.95955	0.96536
6	23	FSL	0.96694	0.96942	0.96818
	24		0.96708	0.96890	0.96799
	25		0.96645	0.96942	0.96793
7	26	m.domrachev	0.96515	0.96826	0.96670
	29		0.95890	0.96768	0.96327
	33		0.96702	0.94718	0.95700
8	30	plubeda	0.97295	0.94370	0.95810
	35		0.96825	0.93575	0.95173
	59		0.87549	0.70752	0.78259
9	31	ccolon	0.96308	0.95246	0.95774
	34		0.95648	0.95631	0.95639
10	32	lsi_uned	0.97280	0.94201	0.95716
11	36	sohrab	0.95950	0.93908	0.94918
	38		0.97695	0.92028	0.94777
	43		0.96234	0.92242	0.94196
	45		0.94907	0.92815	0.93849
	46		0.96924	0.90909	0.93820
12	37	jimblair	0.97424	0.92310	0.94798
	39		0.97505	0.91915	0.94627
	40		0.97327	0.92001	0.94589
	41		0.97180	0.92008	0.94524
	42		0.96985	0.92059	0.94458
13	44	vcotik	0.95591	0.92367	0.93951
	50		0.88734	0.92089	0.90381
	51		0.88734	0.92089	0.90381
14	47	Jordi	0.93267	0.93590	0.93428
	48		0.94357	0.92149	0.93240
	57		0.87986	0.85150	0.86545
15	49	lsi2_uned	0.98284	0.85568	0.91486
	58		0.93509	0.77562	0.84792
16	52	VSP	0.88881	0.89356	0.89118
	53		0.88881	0.89356	0.89118
	54		0.88361	0.89685	0.89018
	55		0.88361	0.89685	0.89018
17	60	gauku	0.92299	0.59848	0.72613
-	-	**Baseline-VT**	*0.50594*	*0.51363*	*0.50976*
18	61	Aspie96	0.19780	0.55626	0.29183

of the recall. That it, for every point of increase in recall, we allowed 4 point of decrease in F-score, but not more. It tried to increase the recall, but without hurting the F-Score too much.

The systems that were used to achieve the best results for these metrics were the following:

– Best F1:

```
lukas.lange/run3 improves the F-score from 0 a 0.96961.
lukas.lange/run2 improves the F-score from 0.96961 a 0.96997.
lukas.lange/run1 improves the F-score from 0.96997 a 0.97033.
```

– Recall:

```
lukas.lange/run3 improves the recall from 0 to 0.96944.
lukas.lange/run2 improves the recall from 0.96944 to 0.97209.
lukas.lange/run1 improves the recall from 0.97209 to 0.97492.
lukas.lange/run4 improves the recall from 0.97492 to 0.97562.
Fadi/15-7 improves the recall from 0.97562 to 0.97898.
Fadi/14-5 improves the recall from 0.97898 to 0.97951.
Fadi/17-3 improves the recall from 0.97951 to 0.98022.
Fadi/16-3 improves the recall from 0.98022 to 0.98039.
nperez/ncrfpp improves the recall from 0.98039 to 0.98181.
FSL/run1 improves the recall from 0.98181 to 0.98393.
FSL/run2 improves the recall from 0.98393 to 0.9841.
nperez/sp-test-03-empty improves the recall from 0.9841 to 0.98516.
mhjabreel/run3 improves the recall from 0.98516 to 0.98551.
mhjabreel/run2 improves the recall from 0.98551 to 0.98569.
jiangdehuan/run3 improves the recall from 0.98569 to 0.98693.
jiangdehuan/run2 improves the recall from 0.98693 to 0.9871.
jimblair/run2 improves the recall from 0.9871 to 0.98763.
jimblair/run3 improves the recall from 0.98763 to 0.98781.
jiangdehuan/run1 improves the recall from 0.98781 to 0.98816.
Jordi/run3 improves the recall from 0.98816 to 0.98869.
vcotik/run5 improves the recall from 0.98869 to 0.98887.
```

– Balanced:

```
lukas.lange/run3 improves the recall from 0 to 0.96944 (+0.96944)
        without losing too much F-score: 0.96961 (-0.96961).
lukas.lange/run2 improves the recall from 0.96944 to 0.97209 (+0.00265)
        without losing too much F-score: 0.96841 (0.00112).
lukas.lange/run1 improves the recall from 0.97209 to 0.97492 (+0.00283)
        without losing too much F-score: 0.96647 (0.00194).
Fadi/15-7 improves the recall from 0.97492 to 0.97863 (+0.00371)
        without losing too much F-score: 0.96181 (0.00466).
Fadi/17-3 improves the recall from 0.97863 to 0.97951 (+0.00088)
        without losing too much F-score: 0.95868 (0.00313).
nperez/ncrfpp improves the recall from 0.97951 to 0.98128 (+0.00177)
        without losing too much F-score: 0.95308 (0.00560).
FSL/run1 improves the recall from 0.98128 to 0.98375 (+0.00247)
        without losing too much F-score: 0.94342 (0.00966).
```

Table 8. Combining systems using finding the best combination (sub-track 1).

Criteria	Precision	Recall	F1
Best F1	0.96999	0.97068	0.97033
Balanced	0.90627	0.98375	0.94342
Best Recall	0.71230	0.98887	0.82811

Table 8 summarizes the results of this experiment. The joint system trying to maximize the F-score improved the result of the best system, but by a very narrow margin. The balanced systems improved the recall by 1.4 points, at the cost of decreasing the F-score by 2.6 points, being a probably desirable effect.

Voting The combination of individual systems from the previous experiment was done directly on the test set. It is very difficult for a given combination of systems to be transferable from one data set to another. Therefore, it should be taken as only an approximation of the upper bound that can be obtained by combining individual systems. In this experiment, we combined the systems using a voting scenario: we accepted as good the annotations that had predicted by N systems.

We created 50 systems for sub-track 1. The first system accepted all the annotations predicted by, at least, one of the systems, while the last one accepted only the annotations that were predicted by, at least, 50 systems. The results of this experiment is shown in Table 9. As expected, as the value of N increased (we increased the number of required votes), the recall got worse and the precision improved. The maximum value of F-score on the train and development sets was obtained combining 17 systems (F-score of 0.9942). When we used the train and development sets as train corpus to select the optimal value of N and used this value on the test set, we obtained an F-score of 0.9757. This score was lower than the best one that could be obtained (0.9768, with N = 23), but the difference was (in practice) negligible.

Comparing the results of the two experiments, we see that the voting system improved the joint system by 0.54 points. In addition, as we see in the Table 9, the values were very stable and a non-optimal choice of the value N did not vary much the result. The negative part was that the voting scenario required many systems to obtain this result (17 systems out of 63 had to agree in order to accept an annotation), while the joint system was a combination of only 3 systems. The voting system matched the performance of the joint system when N is 13, scoring 0.9701 (the joint system scored 0.9703) .

For reasons of space, we do not include the results of this experiment for sub-tracks 2A and 2B, but they showed a very similar behavior.

3.5 Performance drop

In this section we analyze the performance of the systems on the different data sets. As we have said, the background set included, the train set and the devel-

Table 9. Combining systems using a voting scheme (sub-track 1).

#	Train+Dev			Test		
	P	R	F1	P	R	F1
1	1.0000	0.2331	0.3781	0.9947	0.2084	0.3446
2	1.0000	0.7374	0.8489	0.9922	0.6054	0.7519
3	1.0000	0.8253	0.9043	0.9915	0.6789	0.8059
4	1.0000	0.8809	0.9367	0.9899	0.7575	0.8583
5	1.0000	0.9170	0.9567	0.9882	0.8477	0.9126
6	1.0000	0.9340	0.9659	0.9869	0.8739	0.9270
7	1.0000	0.9427	0.9705	0.9862	0.8989	0.9405
8	0.9997	0.9571	0.9779	0.9852	0.9170	0.9498
9	0.9995	0.9620	0.9804	0.9845	0.9244	0.9535
10	0.9994	0.9678	0.9834	0.9838	0.9349	0.9587
11	0.9992	0.9804	0.9897	0.9823	0.9483	0.9650
12	0.9989	0.9845	0.9916	0.9818	0.9530	0.9672
13	0.9985	0.9879	0.9932	0.9815	0.9591	0.9701
14	0.9982	0.9893	0.9937	0.9802	0.9652	0.9727
15	0.9974	0.9906	0.9940	0.9797	0.9699	0.9748
16	0.9966	0.9914	0.9940	0.9777	0.9731	0.9754
17	**0.9962**	**0.9922**	**0.9942**	***0.9769***	***0.9745***	***0.9757***
18	0.9953	0.9928	0.9941	0.9758	0.9768	0.9763
19	0.9946	0.9933	0.9939	0.9740	0.9791	0.9765
20	0.9938	0.9938	0.9938	0.9724	0.9802	0.9763
21	0.9931	0.9943	0.9937	0.9714	0.9818	0.9766
22	0.9925	0.9949	0.9937	0.9698	0.9837	0.9767
23	0.9918	0.9952	0.9935	0.9686	0.9851	0.9768
24	0.9913	0.9954	0.9933	0.9663	0.9863	0.9762
25	0.9906	0.9956	0.9931	0.9647	0.9879	0.9761
26	0.9898	0.9961	0.9930	0.9636	0.9884	0.9759
27	0.9892	0.9964	0.9928	0.9626	0.9891	0.9757
28	0.9883	0.9967	0.9924	0.9601	0.9896	0.9746
29	0.9877	0.9969	0.9923	0.9587	0.9905	0.9743
30	0.9865	0.9972	0.9918	0.9571	0.9912	0.9739
31	0.9855	0.9974	0.9914	0.9539	0.9917	0.9725
32	0.9846	0.9976	0.9911	0.9511	0.9917	0.9710
33	0.9833	0.9979	0.9905	0.9477	0.9919	0.9693
34	0.9821	0.9980	0.9900	0.9465	0.9922	0.9688
35	0.9806	0.9981	0.9893	0.9444	0.9924	0.9678
36	0.9788	0.9982	0.9884	0.9412	0.9927	0.9663
37	0.9767	0.9983	0.9873	0.9343	0.9934	0.9630
38	0.9743	0.9983	0.9862	0.9313	0.9938	0.9615
39	0.9715	0.9984	0.9847	0.9270	0.9941	0.9594
40	0.9674	0.9986	0.9828	0.9223	0.9947	0.9571
41	0.9632	0.9987	0.9806	0.9193	0.9950	0.9557
42	0.9568	0.9988	0.9773	0.9147	0.9952	0.9532
43	0.9529	0.9990	0.9754	0.9108	0.9952	0.9511
44	0.9493	0.9990	0.9735	0.9071	0.9955	0.9493
45	0.9449	0.9991	0.9712	0.9020	0.9957	0.9465
46	0.9411	0.9992	0.9693	0.8975	0.9959	0.9442
47	0.9378	0.9992	0.9675	0.8924	0.9959	0.9413
48	0.9338	0.9992	0.9654	0.8850	0.9960	0.9372
49	0.9286	0.9996	0.9628	0.8760	0.9962	0.9322
50	0.9214	0.9998	0.9590	0.8679	0.9964	0.9277

Table 10. Performance drop of the systems between datasets.

Track	Team	Train	Dev	Test	Drop
1	lukas.lange	0.9959	0.971	0.9696	-0.0014
	Fadi	0.9977	0.964	0.9633	-0.0007
	nperez	0.9906	0.9545	0.9602	+0.0057
	FSL	0.9655	0.969	0.9595	-0.0095
	mhjabreel	0.996	0.9643	0.9583	-0.0060
	lsi_uned	0.9713	0.95	0.9434	-0.0066
	jiangdehuan	0.9625	0.9096	0.9401	+0.0305
	jimblair	1	1	0.9375	-0.0625
	ccolon	0.978	0.9356	0.9322	-0.0034
	sohrab	0.9529	0.9274	0.9312	+0.0038
	Jordi	0.9844	0.9217	0.9184	-0.0033
	plubeda	0.9808	0.8933	0.9038	+0.0105
	m.domrachev	1	1	0.9001	-0.0999
	lsi2_uned	0.9278	0.8944	0.8998	+0.0054
	vcotik	0.9689	0.8953	0.8968	+0.0015
	VSP	0.8981	0.8999	0.8601	-0.0398
	gauku	0.725	0.7108	0.7092	-0.0016
	Aspie96	0.284	0.2716	0.2778	+0.0062
2A	lukas.lange	0.9961	0.9756	0.9749	-0.0007
	Fadi	0.999	0.9681	0.9686	+0.0005
	nperez	0.9942	0.9604	0.968	+0.0076
	mhjabreel	0.9972	0.9698	0.9661	-0.0037
	FSL	0.9715	0.974	0.9641	-0.0099
	lsi_uned	0.974	0.9539	0.9486	-0.0053
	jiangdehuan	0.9638	0.9139	0.9457	+0.0318
	plubeda	0.9843	0.9327	0.9436	+0.0109
	jimblair	1	1	0.9428	-0.0572
	ccolon	0.9804	0.9427	0.9427	0.0000
	sohrab	0.9563	0.9308	0.9352	+0.0044
	vcotik	0.9719	0.9275	0.9297	+0.0022
	Jordi	0.9853	0.927	0.9241	-0.0029
	m.domrachev	1	1	0.9033	-0.0967
	lsi2_uned	0.9294	0.8977	0.9024	+0.0047
	VSP	0.9013	0.902	0.8703	-0.0317
	gauku	0.727	0.7132	0.7138	+0.0006
	Aspie96	0.2943	0.2854	0.2917	+0.0063
2B	lukas.lange	0.997	0.9805	0.9853	0.0048
	jiangdehuan	0.9934	0.9486	0.9818	+0.0332
	nperez	0.9953	0.9697	0.9759	+0.0062
	Fadi	0.999	0.9745	0.975	+0.0005
	mhjabreel	0.9986	0.981	0.9747	-0.0063
	FSL	0.9836	0.9855	0.9682	-0.0173
	m.domrachev	0.98	0.9664	0.9667	+0.0003
	plubeda	0.99	0.9485	0.9581	+0.0096
	ccolon	0.9868	0.9549	0.9577	+0.0028
	lsi_uned	0.9772	0.9617	0.9572	-0.0045
	sohrab	0.9715	0.9468	0.9492	+0.0024
	jimblair	1	1	0.948	-0.0520
	vcotik	0.9749	0.9382	0.9395	+0.0013
	Jordi	0.9878	0.9868	0.9343	-0.0525
	lsi2_uned	0.935	0.9117	0.9149	+0.0032
	VSP	0.9155	0.9165	0.8912	-0.0253
	gauku	0.7406	0.7288	0.7261	-0.0027
	Aspie96	0.2946	0.2856	0.2918	+0.0062

opment set, which allowed us to measure the F-score of all the systems on the train, development and test set, and to analyze their behavior.

All the scores of this analysis are shown in table 10, where the drop column indicates the difference of performance in the test set with respect to the development set (a negative value indicates a lower performance on the test set). There were two teams that achieved a F-score of 1.0 in both train and development set: *jimblair* (in all tracks) and *m. domrachev* (in sub-tracks 1 and 2A). The former had a performance drop of 6.25 points, and the latter of 9.99 points in the test set, probably because both systems of these competitors memorized the train and development data, obtaining a perfect score, incurring in overfitting. This also suggested that they could have used the development set to train the system, and not just to tune it.

In contrast to this, we see that *lukas.lange*, which was first team on the test set for sub-track 1, was also the first on the development set (without taking into account those who had scored 1.0), but third on the train set (without taking into account those who scored 1.0). The performance of their system only dropped 0.14 points in the test set with respect to the development set. Probably they used the train set to build the system and the development only for tuning, not incurring in overfitting. This demonstrated that the ability of the systems to generalize was very important.

Taking into account all the sub-tracks, the maximum performance drop was suffered by *m.domrachev*, losing 9.99 points in sub-track 1. Without taking into account those who had scores 1.0 on the development set, the system that lost more points was the one submitted by *Jordi*, which lost 5.25 points on track 2B (0.33 points in sub-track 1 ,and 0.29 points in sub-track 2A). The next participants with the highest loss of performance were *VSP* and *FSL*.

The maximum improvement in the test set with respect to the development set was 3.32 points, corresponding to the system submitted by *jiangdehuan*, in track 2A.

As a curiosity, *ccolon* scored exactly the same result on the development and test set. However, its performance decreased with respect to the train set (by 3.77 points).

4 Discussion

The MEDDOCAN track attracted a considerable number of teams, not only from Spain, but also from other countries, stressing the global interest in solving the clinical data access hurdles and assuring patient data privacy requirements. Compared to previous efforts for English, namely the i2b2 de-identification tracks, MEDDOCAN could even reach a higher number of participation. It is important to point out that the MEDDOCAN track benefited significantly from the experiences, setting and annotation process pioneered by the i2b2 efforts.

In case of the 2006 i2b2 shared task [24], a total of 7 teams participated in the track, providing 16 systems. The five best systems scored above 0.95 for the entity detection track and equaled or exceeded an F-score of 0.95 for the token-based evaluation. The 2014 i2b2 de-identification shared task [21] had 10 teams, submitting 22 runs. The top team reached an F-score of 0.9360 for the entity detection track, and 0.9611 for the evaluation based on tokens. It is important to mention that in case of MEDDOCAN a synthetic corpus was used so the results might not be directly comparable to i2b2. Also, it is well known that there is a considerable variability in density, distribution and characteristics of sensitive information even between different types of clinical records.

De-identification is still a very hard task, because for the special characteristics of clinical texts and the importance of recall, i.e. avoiding leakage of sensitive information. The top three teams are above 0.96 in F-score, for the track based on entity detection.

The top scoring systems make use of the most cutting-edge NLP techniques, i.e. exploiting Deep Learning. Their results are comparable to single manual anonymization done by humans. Automatic anonymization with manual revision to detect potential leakages might result in anonymized Spanish clinical records that allow data redistribution. Nevertheless, a follow up task, using real EHRs from various healthcare institutions, and assessing the practical user scenario with experts in the loop would be desirable to quantify also cost reduction and benefits of the quality of anonymization strategies assisted by automated tools.

5 Conclusions

The results of the MEDDOCAN shared task and evaluation effort on automatic de-identification of sensitive information from texts in Spanish show that advanced deep learning approaches in combination with rule based systems and gazetteer resources can provide very competitive results when a high quality manually labeled dataset is available. The construction of Gold Standard corpora is key and require very detailed annotation guidelines and a carefully designed corpus generation process with involvement of clinical domain experts. We expect that such a corpus and evaluation will also be carried out for data in other languages and that automatic anonymization and de-identification systems will be beneficial beyond EHRs, such as medical surveys [8] or legal-financial documents [3]. In order to improve the impact of future shared tasks on anonymization, the involvement should not be limited to academic groups on language technologies, but also directly data providers (health institutions), legal experts and national and European institutions. For instance, the European Medicines Agency (EMA) has launched a Technical Anonymisation Group (TAG) consisting of a group of experts in data anonymisation to help further develop best practices for the anonymisation of clinical reports. Moreover, we also would like to stress the key importance of making the systems code or developed participant tools accessible/available and the need to explore strategies to promote start-ups and commercialization of solutions resulting from shared tasks and evaluation campaigns.

Acknowledgements

We acknowledge the Encargo of Plan TL (SEAD) to CNIO and BSC for funding, and the scientific committee for their valuable comments and guidance. We would also like to thank Siamak Barzegar for his help in setting up MEDDOCAN at CodaLab, and Felipe Soares for input in preparing the manuscript and task.

References

1. Alfalahi, A., Brissman, S., Dalianis, H.: Pseudonymisation of personal names and other phis in an annotated clinical swedish corpus. In: Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) Held in Conjunction with LREC. pp. 49–54 (2012)
2. Amengol-Estapé, J., Soares, F., Marimon, M., Krallinger, M.: Pharmaconer tagger: a deep learning-based tool for automatically finding chemicals and drugs in spanish medical texts. Genomics & Informatics **17**(2) (2019)
3. Bick, E., Barreiro, A.: Automatic anonymisation of a new portuguese-english parallel corpus in the legal-financial domain. Oslo Studies in Language **7**(1) (2015)
4. Cristóbal, R.S., Carrero, A.M., Carrasco, M.P., Rodríguez, M.C., Méndez, J.F., de Mingo, M.G., Tello, J.C., de Madariaga, R.S., Serrano, A.C., Aza, I.V., et al.: Sistema anonimizador conforme a la norma une-en iso 13606 (2012)
5. Fernández-Alemán, J.L., Señor, I.C., Lozoya, P.Á.O., Toval, A.: Security and privacy in electronic health records: A systematic literature review. Journal of biomedical informatics **46**(3), 541–562 (2013)
6. García Sardiña, L.: Automating the anonymisation of textual corpora (2018)
7. Gaudet-Blavignac, C., Foufi, V., Wehrli, E., Lovis, C.: De-identification of french medical narratives. Swiss Medical Informatics **34**(00) (2018)
8. Gentili, M., Hajian, S., Castillo, C.: A case study of anonymization of medical surveys. In: Proceedings of the 2017 International Conference on Digital Health. pp. 77–81. ACM (2017)
9. Grouin, C., Névéol, A.: De-identification of clinical notes in french: towards a protocol for reference corpus development. Journal of biomedical informatics **50**, 151–161 (2014)
10. Hassan, F., Domingo-Ferrer, J., Soria-Comas, J.: Anonimizacin de datos no estructurados a travs del reconocimiento de entidades nominadas. In: Actas de la XV Reunin Espaola sobre Criptologa y Seguridad de la Informacin - RECSI 2018. pp. 102–106 (2018)
11. Intxaurrondo, A., Marimon, M., Gonzalez-Agirre, A., Lopez-Martin, J.A., Rodriguez, H., Santamaria, J., Villegas, M., Krallinger, M.: Finding mentions of abbreviations and their definitions in spanish clinical cases: The barr2 shared task evaluation results. In: IberEval@ SEPLN. pp. 280–289 (2018)
12. Intxaurrondo, A., Pérez-Pérez, M., Pérez-Rodríguez, G., López-Martín, J.A., Santamaria, J., de la Pena, S., Villegas, M., Akhondi, S.A., Valencia, A., Lourenço, A., Kralllinger, M.: The biomedical abbreviation recognition and resolution (barr) track: benchmarking, evaluation and importance of abbreviation recognition systems applied to spanish biomedical abstracts. SEPLN (2017)
13. Mamede, N., Baptista, J., Dias, F.: Automated anonymization of text documents. In: 2016 IEEE Congress on Evolutionary Computation (CEC). pp. 1287–1294. IEEE (2016)

14. Medina, S., Turmo, J.: Building a spanish/catalan health records corpus with very sparse protected information labelled. In: LREC 2018: Workshop MultilingualBIO: Multilingual Biomedical Text Processing: proceedings. pp. 1–7 (2018)
15. Megyesi, B., Granstedt, L., Johansson, S., Prentice, J., Rosén, D., Schenström, C.J., Sundberg, G., Wirén, M., Volodina, E.: Learner corpus anonymization in the age of gdpr: Insights from the creation of a learner corpus of swedish. In: Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning. pp. 47–56 (2018)
16. Mota, E., Martín, N., Moreno, A., Ferrete, E., Santamaría, J., Marimon, M., Intxaurrondo, A., Gonzalez-Agirre, A., Villegas, M., Krallinger, M.: Guías de anotación de información de salud protegida (Oct 2018), http://temu.bsc.es/meddocan/wp-content/uploads/2019/02/guías-de-anotación-de-información-de-salud-protegida.pdf
17. Pantazos, K., Lauesen, S., Lippert, S.: Preserving medical correctness, readability and consistency in de-identified health records. Health informatics journal **23**(4), 291–303 (2017)
18. Pérez-Pérez, M., Pérez-Rodríguez, G., Blanco-Míguez, A., Fdez-Riverola, F., Valencia, A., Krallinger, M., Lourenço, A.: Next generation community assessment of biomedical entity recognition web servers: metrics, performance, interoperability aspects of becalm. Journal of Cheminformatics **11**(1), 42 (2019)
19. Santamaría, J., Krallinger, M.: Construcción de recursos terminológicos médicos para el español: el sistema de extracción de términos cutext y los repositorios de términos biomédicos. Procesamiento del Lenguaje Natural **61** (2018)
20. Scheurwegs, E., Luyckx, K., Van der Schueren, F., Van den Bulcke, T.: Deidentification of clinical free text in dutch with limited training data: a case study. In: Proceedings of the Workshop on NLP for Medicine and Biology associated with RANLP 2013. pp. 18–23 (2013)
21. 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 Suppl**, S11–9 (2015)
22. Tomanek, K., Daumke, P., Enders, F., Huber, J., Theres, K., Müller, M.: An interactive de-identifica-tion-system. In: Proceedings of SMBM 2012-The 5th International Symposium on Semantic Mining in Biomedicine. pp. 82–86 (2012)
23. Tveit, A., Edsberg, O., Rost, T., Faxvaag, A., Nytro, O., Nordgard, T., Ranang, M.T., Grimsmo, A.: Anonymization of general practioner medical records. In: second HelsIT Conference (2004)
24. Uzuner, Ö., Luo, Y., Szolovits, P.: Evaluating the State-of-the-Art in Automatic De-identification. Journal of the American Medical Informatics Association **14**(5), 550–563 (09 2007). https://doi.org/10.1197/jamia.M2444, https://doi.org/10.1197/jamia.M2444
25. Vico, H., et al.: Definición de una arquitectura de referencia para anonimizar documentos (2013)
26. Villegas, M., Intxaurrondo, A., Gonzalez-Agirre, A., Marimon, M., Krallinger, M.: The mespen resource for english-spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations. In: Proceedings of the LREC 2018 Workshop MultilingualBIO: Multilingual Biomedical Text Processing, Paris, France. European Language Resources Association (ELRA) (2018)