Named Entity Recognition for De-identifying Real-World Health Records in Spanish

Guillermo López-García[1(✉)], Francisco J. Moreno-Barea[1], Héctor Mesa[1], José M. Jerez[1], Nuria Ribelles[2], Emilio Alba[2], and Francisco J. Veredas[1,3]

[1] Departamento de Lenguajes y Ciencias de la Computación, Escuela Técnica Superior de Ingeniería Informática, Universidad de Málaga, Málaga, Spain
guilopgar@uma.es
[2] Unidad de Gestión Clínica Intercentros de Oncología, Instituto de Investigación Biomédica de Málaga (IBIMA), Hospitales Universitarios Regional y Virgen de la Victoria, Málaga, Spain
[3] Research Institute of Multilingual Language Technologies, Universidad de Málaga, Málaga, Spain

Abstract. A growing and renewed interest has emerged in Electronic Health Records (EHRs) as a source of information for decision-making in clinical practice. In this context, the automatic de-identification of EHRs constitutes an essential task, since their dissociation from personal data is a mandatory first step before their distribution. However, the majority of previous studies on this subject have been conducted on English EHRs, due to the limited availability of annotated corpora in other languages, such as Spanish. In this study, we addressed the automatic de-identification of medical documents in Spanish. A private corpus of 599 real-world clinical cases have been annotated with 8 different protected health information categories. We have tackled the predictive problem as a named entity recognition task, developing two different deep learning-based methodologies, namely a first strategy based on recurrent neural networks (RNN) and an end-to-end approach based on transformers. Additionally, we have developed a data augmentation procedure to increase the number of texts used to train the models. The results obtained show that transformers outperform RNN on the de-identification of Spanish clinical data. In particular, the best performance was obtained by the XLM-RoBERTa large transformer, with a strict-match micro-averaged value of 0.946 for precision, 0.954 for recall and 0.95 for F1-score, when trained on the augmented version of the corpus. The performance achieved by transformers in this study proves the viability of applying these state-of-the-art models in real-world clinical scenarios.

Keywords: Named Entity Recognition · Natural Language Processing · Electronic Health Records · De-Identification · Spanish

J. Mikyška et al. (Eds.): ICCS 2023, LNCS 14075, pp. 228–242, 2023.
https://doi.org/10.1007/978-3-031-36024-4_17

1 Introduction

The adoption of electronic health records (EHR) [12] is a key component for health systems and medical professionals, as well as representing an important source of information to advance medical research and improve healthcare-related services. However, for their widespread use in medical research, it is necessary to remove identifiable information to protect patients' data privacy. EHRs store information in a wide variety of formats that contain information related to clinical diagnoses, treatments, procedures, and especially, the privacy of patients and medical professionals. However, the unstructured nature of the textual fields makes the task of automatically extracting the relevant concepts from them especially difficult, but the manual extraction of concepts is non-reusable, time-consuming and costly [7].

The automatic extraction and masking of the concepts related to individually identifiable data thus becomes the primary task to treat the information contained in the EHR in other medical-analytical processes. This task, called de-identification, is not only an ethical prerequisite, but also a legal requirement imposed by data privacy legislation. In the United States (US), the Health Insurance Portability and Accountability Act (HIPAA) requires the deletion of 18 categories of protected health information (PHI) [29]. Similarly, the General Data Protection Regulation (GDPR) of the European Union (EU) [4], and the Ley Orgánica Española de Protección de Datos Personales y Garantía de Derechos Digitales (LOPD-GDD) of Spain [3] in particular, prohibit the processing of personal data unless identifiable information is masked. In this paper, we focus on the de-identification of Spanish EHRs for compliance with the LOPD-GDD.

De-identifying clinical texts is a named entity recognition (NER) task from the standpoint of natural language processing (NLP). NER is the process of identifying sections of text that reference rigid designators belonging to predefined semantic types, such as person, organisation, location, etc. The term Named Entity (NE) was first used at the 6th Message Understanding Conference (MUC6) [8], a scientific event designed to promote and evaluate research in information extraction. In the case of the de-identification process in EHR, the PHI categories are treated as NEs.

In the last decades, text de-identification has been addressed following three different approaches: rule-based methods, machine learning (ML) systems and deep learning (DL) models. Early de-identification systems were generally rule-based. However, considering that these methods are not reproducible for different domains, researchers began designing ML algorithms, especially motivated by the organisation of various NLP de-identification challenges. ML algorithms used for this task include decision tree, hidden Markov model, support vector machines and conditional random field (CRF) [14]. The three main systems in the 2014 i2b2 de-identification challenge [27] were based on CRF, since it was the state-of-the-art (SOTA) method at the time the shared task was held.

In recent years, deep neural networks, which have the ability to automatically learn effective features from large-scale datasets, have been extensively applied in different NPL tasks. Architectures based on feedforward neural net-

works and Recurrent Neural Networks (RNN), as well as other modified and combined deep neural networks, show impressive results in the NER task. Especially successful are the Long Short-Term Memory networks (LSTM) [10] and its variations, e.g., bidirectional LSTM (BiLSTM) and BiLSTM-CRF. Currently, large pretrained language models based on the multi-head self-attention mechanism [30], specifically the Bidirectional Encoder Representations from Transformers (BERT) model [6], outperform other ML systems for the task of NER, particularly in the biomedical domain [16].

In this paper, we have addressed the problem of automatic detection of personally identifiable information in Real-World Data (RWD) from EHR written in Spanish. For this purpose, we have used several models based on RNN (BiLSTM, 2-BiLSTM and BiLSTM-CRF) and the Transformer architecture (XLM-RoBERTa [2] and RoBERTa-BNE [9]). The models studied herein represent the SOTA in various NER tasks [5]. However, to the best of our knowledge, this is the first study that analyses the application of these models to the problem of identifying PHI using real-world medical texts in Spanish. Through this study, the transformers analysed in this work achieve a higher performance in this task with respect to the RNNs applied, demonstrating why they currently represent the SOTA in many NER tasks.

2 Related Works

With the organisation of different NER challenges and NLP shared tasks [27], and the increased adoption of EHRs in healthcare systems worldwide, text de-identification studies proliferated. In particular, great progress has been made in the development of de-identification systems based on the CRF model [32]. Recently, the rise of DL has increased the number of proposed architectures based on RNNs. LSTM networks [10] and their combinations with CRF (LSTM-CRF) [15], showed better performance on the 2016 CEGS N-GRID de-identification task than CRF models. On the other hand, in [5], the authors developed a BiLSTM model to tackle the de-identification problem, which achieved SOTA in the 2014 i2b2/UTHealth dataset. This model implements a layer of BiLSTM units at a character level input to obtain character embeddings that are concatenated with pretrained token embeddings. The enhanced embeddings are returned to the BiLSTM units and the sequence of probabilities is adjusted with the CRF sequence optimiser to produce the system output. In this way, with the superior performance demonstrated by the LSTM-CRF architectures, most of the previous works addressing the de-identification of medical texts have implemented these models [13].

Although the majority of developed techniques for de-identifying clinical cases have focused on English, national laws around the world vary, thus language-specific methodologies are needed. Consequently, automatic de-identification strategies have been proposed for documents in other languages [11,26]. Being the second most spoken language in the world in terms of the number of native speakers [31], there is a pressing need to develop medical NLP

methodologies focused on Spanish. In fact, some competitions and projects have been organised to exploit the content of unstructured medical records in Spanish, although due to the limited availability of annotated corpus with clinical-entity information, the task remains a challenge. One example of this type of initiatives is Cantemist (Cancer Text Mining SharedTask), a shared task focused on the recognition of NEs of a critical type of concept related to cancer and tumor morphology in Spanish medical records [19]. Another example is MEDDOCAN (Medical Document Anonymization), a shared task organised in 2019 that focused on the de-identification of clinical EHRs in Spanish [21]. As in the most recent NER and de-identification competitions worldwide, the NLP models that showed the best performance in MEDDOCAN used methodologies based on DL [22].

Even though MEDDOCAN represents the first shared task specifically devoted to the de-identification of medical texts in Spanish, the corpus derived from the competition does not closely resemble the documents found in clinical practice. In this way, the organisers created a synthetic collection of curated and well-structured clinical texts enriched with PHI expressions [21], in contrast with the unstructured and complex nature of the real-world clinical cases. In this study, we have addressed the de-identification of EHRs in Spanish written by physicians during clinical practice. For this purpose, we have systematically analysed the performance of RNN and transformers based models as automatic systems to detect PHI contained in medical texts. Our results show that transformer models outperform RNNs in de-identifying Spanish medical cases, demonstrating the ability of these models to be successfully applied not only in general and biomedical domains in Spanish [19,22,23], but also in real-world clinical scenarios.

Finally, for reproducibility purposes, all the code needed to replicate our work is publicly available at https://github.com/guilopgar/DeIdentSpanishEHR.

3 Materials

3.1 Galén Texts Annotation

In this study, we have used a private collection of 599 clinical cases retrieved from the Galén Oncology Information System [25,28], which collects information on more than 62,250 cancer patients from the *Hospital Regional Universitario* and the *Hospital Universitario Virgen de la Victoria* in Málaga, Spain. In total, Galén stores 600,000 documents corresponding to clinical episodes as well as a significant number of structured fields (which are completed both in real time, during clinical care activity, and later by specific personnel in charge of this supervised task).

Once the medical documents were obtained, we proceeded, as it is mandatory according to the LOPD-GDD [3], with the de-identification of the medical records to guarantee their correct anonymization and dissociation of the personal information of the individuals involved in the health system. In order to label each of the 599 available records for the subsequent training and evaluation

of automatic de-identification algorithms, a manual supervised annotation was performed exclusively by authorised clinical personnel of the Intercentre Clinical Management Unit of Oncology (UGCOI). Thus, the collection of clinical records was processed and labelled by our authorised staff, also replacing sensitive information with standardised labels in a non-reversible way.

Finally, with the aim of evaluating the quality of the annotations, we measured the inter-annotator agreement (IAA) on a subset of 100 documents—which corresponds to the test subset (see Sect. 3.2). For this purpose, we computed a widely used metric to quantify the IAA in NER tasks [17], namely the F1-score, comparing the labelling produced by two different authorised annotators. The IAA value was a micro-averaged F1-score of 0.9633, which indicates a significant agreement between the labelling performed by both annotators.

3.2 Named Entities

In this section, we provide a detailed description of the distinct PHI categories considered to perform the annotation process. We examined the presence of the PHI categories defined by the HIPAA from the US. A reliable interpretation of the HIPAA guidelines was made, adapting some PHI entities to fit the reality of health records in Spain. The Spanish legal system does not provide specific guidance on what information must be removed to de-identify medical texts, but the annotation guidelines made by the Spanish National Plan for the Advancement of Language Technology (Plan TL) for the MEDDOCAN task [21] were taken into account. The task was also carried out from a position of "risk aversion", due to the great variability of users who would later view and interpret the information based on the de-identification.

A sub-selection of 8 PHI categories was made after manually reviewing the data, with the aim of adapting the guidelines mentioned above to the specificities of our real-world clinical corpus. The chosen NEs were:

- CENTRO (healthcare centre): includes any reference to names of clinical centres, general hospitals, institutions or health centres.
- CONTACTO (contact): includes any form of contact with a patient, doctor, nurse or health centre, such as telephone numbers and email addresses.
- DIRECCION (address): includes the appearance of a physical address, such as streets, avenues or buildings.
- HISTORIA (EHR number): includes the identifiers used to control hospital medical records (NHC) and Andalusian health history (NUHSA).
- IDENT (identifier): includes personal identifiers such as the national identity document (DNI), the social security number (NSS), identifiers associated with insurers, the personal numerical code in the Andalusian Health Service (CNP), or any other type of unique identifier.
- PERSONA (person): includes names and surnames of people, as well as initials.
- UBICACION (location): includes references to the location of a person or centre, without this representing a specific physical address, but rather locations relative to the name of a city/town, region or country.

– REFERENCIA (reference): includes identifiers related to medical tests performed, such as tests, biopsies, scans or x-rays.

Table 1 shows the distribution of NEs in the selected documents from the Galén corpus. The collection of 599 clinical cases was randomly split into 3 subsets: the training set (399 documents), the validation set (100 documents) and the test set (100 documents). In Table 1, for the different training, validation and test sets (including DA sets, see Sect. 3.3), the columns show the absolute number (abs) of NEs for each of the 8 classes considered as well as their relative frequency (%). The majority of NEs in the Galén corpus are related to Centre, Person and Location, with approximately a presence greater than 20% in all the considered sets. Meanwhile, the entities related to the Address are almost non-existent in non-DA sets (2 annotations in train, 1 in val and 1 in test sets).

Table 1. Description of the number of annotations per corpus subset: the training and validation sets with and without DA, and the test set.

NE	Train		Train + DA		Val		Val + DA		Test	
	abs	%	abs	%	abs	%	abs	%	abs	%
CENTRO	468	.3145	5148	.3153	114	.2953	1254	.2960	91	.2600
CONTACTO	41	.0276	451	.0276	14	.0363	154	.0364	17	.0486
DIRECCION	2	.0013	22	.0013	1	.0026	11	.0026	1	.0029
HISTORIA	18	.0121	198	.0121	14	.0363	154	.0364	15	.0429
IDENT	63	.0423	682	.0418	9	.0233	99	.0234	8	.0229
PERSONA	439	.2950	4828	.2957	147	.3808	1617	.3817	130	.3714
REFERENCIA	115	.0773	1235	.0756	13	.0337	133	.0314	22	.0629
UBICACION	342	.2298	3762	.2304	74	.1917	814	.1922	66	.1886
Total	1488		16326		386		4236		350	

3.3 Data Augmentation

The nature of the NER task assumes the need of recognising words with a high probability of being out of vocabulary. In order to mitigate the lack of context from a small corpus like Galén, a method for augmenting the amount of documents has been applied. For each document in the training and validation dataset, 10 different synthetic documents are generated by text surrogation of entities susceptible to natural replacement. Entities which present a numerical pattern (e.g., telephone numbers, numerical identifiers) are replaced by perturbing digits randomly. Alpha-numeric entities (e.g., DNI, NSS) are replaced by perturbing digits and characters. The surrogation of the entities made primarily of proper nouns—such as people, locations, countries and centres—is a dictionary-based replacement using data from the Spanish National Statistics Institute (INE)[1].

[1] https://www.ine.es/inebmenu/indiceAZ.htm.

Table 1 shows the distribution of NEs in documents created by this augmentation process from the Galén corpus. Although the number of NEs present in the training and validation sets increases, the proportion with respect to the rest of the entities remains essentially the same.

4 Methods

This section presents the two distinct NLP methodologies developed in this study to tackle the de-identification of real-world EHRs in Spanish. The first methodology addresses the problem using an approach based on RNN models, while the second methodology applies transformer-based models. In both cases, the de-identification problem is tackled as a sequence labelling NER task, using the IOB2 tagging scheme [24].

4.1 Recurrent Neural Models

Features. Due to the small order of magnitude of the training corpus, several decisions were made in the tokenization stage to reduce the size of the vocabulary and maximise the inference of contextual relationships. The tokenization was not case sensitive and the easily detectable expressions with a low degree of error based on alpha-numerical sequences (e.g., DNI, NSS, dates, telephone numbers) were replaced by special tokens. To keep information at the character level, a series of descriptors are added to indicate the presence of initial case, uppercase, lowercase, and digits. In addition, a flag is added to tokens that have been detected as a new line, a numeric expression, or if the token is a separator character. Finally, a fastText skip-gram model [1] was used as embedding model trained with the Gálen corpus (training documents and stored clinical documents not used for de-identification).

Thus, we have three types of entry descriptors: indicator descriptors character ranges in the word, descriptors of detected expressions and vector from word embedding. These descriptors are joined in an embedding layer and constitute the input fed to RNN models, as seen in Fig. 1.

Recurrent Neural Networks. RNNs are a generalization of feed-forward neural networks specially designed to deal with temporal problems. In particular, the recurrent SOTA models are based on the LSTM network [10]. An LSTM network has three different gates: the input gate infers the values used to modify the block memory, the forget gate infers features to be discarded and the output gate determines the output using both input and block memory.

BiLSTM consists of two LSTM networks which learns each token in the sequence based on its future and past context [15]. The sequence is processed from left-to-right by one LSTM to learn the past context (current step and previous unit state), while the other LSTM network processes it from right-to-left to learn the future context (current step and subsequent unit state). The designed model includes two time distributed layers, in order to process each

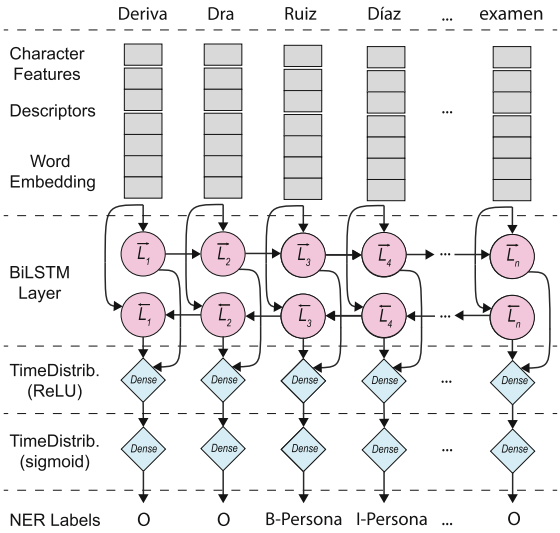

Fig. 1. The BiLSTM model structure for the NER system.

of the time steps and infer the label for each input sequence. This RNN model is shown in Fig. 1. Additionally, 2 BiLSTM layers can be linked to produce a more context-aware algorithm. In this case, the hidden unit forward layer of the second BiLSTM layer receives the output of the previous state and the output of the hidden unit forward layer of the first BiLSTM layer. In contrast, the hidden unit backward layer is computed based on the hidden unit backward layer of the first BiLSTM and the future hidden state.

The BiLSTM-CRF is a type of RNN model that has been used to improve NER performance [5]. A Bidirectional LSTM and a CRF [14] are stacked together for sequence learning. CRF is an undirected discriminative probabilistic graph model, composed of a set of random variables that are used to represent probabilities about structured outputs based on a particular input sequence. The result processed by a BiLSTM layer and a time distributed layer without activation function is fed to a CRF model, which obtains the entity assigned to each word by the system.

4.2 Transformers

The second methodology developed herein to address the de-identification problem is based on transformers. The Transformer model [30] uses the self-attention mechanism to create a contextual numeric representation of each input word, as well as to increase computing efficiency through the parallelization of its network architecture. For the past three years, transformers have become one of the most successful models in multiple areas of NLP [6,22]. One of the reasons that explain their enormous popularity nowadays is that, by following a transfer learning (TL) approach, these attention models can be pretrained on general

domain corpora and further fine-tuned on a domain-specific corpus to tackle a certain NLP task [6]. SOTA results have been obtained in both biomedical and clinical domains by employing transformers in combination with different TL strategies [16,19,20].

In this work, since we are dealing with the de-identification of medical texts in Spanish, we have used two distinct transformer-based models that support the Spanish language, namely, XLM-RoBERTa (XLM-R) and RoBERTa-BNE:

- XLM-R: this multilingual version of the RoBERTa architecture [18] was pretrained on a massive 2.4TB CommonCrawl Corpus in 100 languages [2], using a large multilingual vocabulary of ∼250K subwords. We experimented with both the Base (∼277M trainable weights) and the Large (∼559M trainable parameters) versions of the model.
- RoBERTa-BNE: the Spanish version of the RoBERTa architecture [18] was pretrained on a 570GB corpus obtained from the Spanish National Library (BNE) [9]. The model employs a Spanish vocabulary of ∼50K subtokens and, again, we experimented with both the Base (∼124M trainable parameters) and the Large (∼354M trainable weights) versions of the model.

We have developed an end-to-end approach to address the de-identification problem using transformers, by fine-tuning the models on the clinical corpus obtained from Galén. Figure 2 shows a visual description of the developed methodology. In this way, on the one hand, since the transformers further segment words into a sequence of subwords, each sequence of words from the medical documents was further tokenized into a sequence of subwords. Therefore, each sequence of subwords constitutes the input to the models, without using any additional feature as input data. On the other hand, at inference time, since transformer-based models produce predictions at subword-level, the outputted labels had to be converted to word-level (see Fig. 2). For this purpose, we used the maximum probability criterion proposed in [19], which consists in, for each word, selecting the label predicted with the maximum probability across all subwords obtained from the same word. Finally, the sequence of word-level tags could be further compared with the gold standard (GS) annotations with the aim of evaluating the performance of the models.

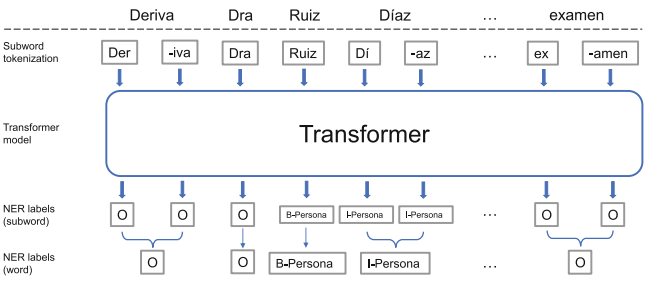

Fig. 2. Illustration of the transformer-based methodology applied to tackle the de-identification problem.

5 Experiments and Results

To experiment with both LSTM-based and transformer models, the validation set was used as an evaluation set. Hyper-parameter fine-tuning was performed by training the models and early stopping the networks using the validation set. The hyper-parameters of the networks were therefore chosen based on obtaining a higher macro averaged F1-score for the validation set. Finally, the performance evaluation of the different architectures was based on the predictions made by the models on the test set.

The standard method for performance evaluation of a NER system is to compare the GS annotations with the tagged output obtained. In this experimentation overall evaluation was considered, namely strict-match and exact-match. For each NE, the type of entity identified and its spans—start and end characters—are obtained. In the strict-match evaluation, both the entity type and spans must match with a GS annotation for a prediction to be considered as correct, while the exact-match evaluation only requires the spans to match.

In addition, different error categories introduced at the MUC6 conference [8] are also considered in this work. For every NE type, the categories are based on the comparison between the GS annotations and the NER system output. Therefore, a predicted entity is considered as correct (cor) if it matches a GS annotation; incorrect (inc) if it does not match the annotation exactly; missed (mis) if the GS annotation is not captured; and spurious (spu) if the prediction does not match any GS annotation.

5.1 NER systems comparison

The experimentation process described above was followed, and Table 2 shows the results obtained by the two methodologies developed in this study to tackle the de-identification problem. Two main conclusion can be drawn from the results described in Table 2. On the one hand, according to the strict-match F1-score—which represents the principal metric used to evaluate automatic de-identification systems [17,21]—, all predictive models applied in this work benefit from the DA procedure. Thus, every system obtains a higher F1-score when trained on the augmented Galén de-identification corpus than when trained solely on the Galén dataset.

On the other hand, considering both strict and exact match evaluations, transformer-based models outperform RNN-based systems for the de-identification of Spanish clinical documents. Among the RNNs, the best performance is achieved by the 2-BiLSTM model when trained on the augmented Galén corpus, obtaining a strict and exact match F1-score of 0.9060 and 0.9201, respectively. In the case of transformers, the XLM-R large model achieves the highest performance obtained in this study, with a strict and exact match F1-score of 0.9502 and 0.9531, respectively, when following the DA procedure. In fact, although the RoBERTa-BNE model was pretrained on a corpus exclusively containing Spanish texts, the XLM-R model surpasses RoBERTa-BNE on the de-identification of Spanish medical cases. In this way, the base versions of XLM-R

Table 2. Micro-averaged metrics computed on Galen's test set. We report the performance of each model when trained on both the Galén de-identification corpus and its augmented version. Finally, for each evaluation strategy (strict and exact match), precision (P), recall (R) and F1-score (F1) metrics are computed.

Model	Corpora	NER (strict)			Spans (exact)		
		P	R	F1	P	R	F1
BiLSTM	Gálen	.7625	.8257	.7929	.7731	.8371	.8038
	Gálen + DA	.7855	.8686	.8250	.7933	.8771	.8331
2-BiLSTM	Gálen	.8189	.8657	.8417	.8378	.8857	.8611
	Gálen + DA	.8898	.9229	.9060	.9036	.9371	.9201
BiLSTM-CRF	Gálen	.9062	.8829	.8944	.9238	.9000	.9117
	Gálen + DA	.8840	.9143	.8989	.8895	.9200	.9045
XLM-R	Gálen	.8937	.8886	.8911	.9080	.9029	.9054
(Base)	Gálen + DA	.9218	.9429	.9322	.9302	.9514	.9407
XLM-R	Gálen	.9224	.9514	.9367	.9280	**.9571**	.9423
(Large)	Gálen + DA	**.9462**	**.9543**	**.9502**	**.9490**	**.9571**	**.9531**
RoBERTa-BNE	Gálen	.8620	.8743	.8681	.8732	.8857	.8794
(Base)	Gálen + DA	.8825	.9229	.9022	.8852	.9257	.9050
RoBERTa-BNE	Gálen	.9202	.9229	.9215	.9259	.9286	.9272
(Large)	Gálen + DA	.9405	.9486	.9445	.9405	.9486	.9445

and RoBERTa-BNE obtain a strict-match F1-score of 0.9322 and 0.9022, respectively, while the large versions of the models achieve a strict F1-score of 0.9502 and 0.9445, respectively. Hence, for this particular task, the large-scale multilingual pretraining followed by the XLM-R model has proved to be more effective than the Spanish-specific pretraining followed by the RoBERTa-BNE model.

5.2 Metrics for Each NE

With the aim of conducting a thorough analysis of the performance of the XLM-R large transformer—the system achieving the highest de-identification results (see Table 2)—, in Table 3, we show the results obtained by the model for each NE separately. Apart from the F1-score, following the "risk aversion" principle, recall is often considered as the reference metric to evaluate the performance of de-identification systems, since false negative errors have the potential to threaten the privacy of patients and medical professionals [17]. An automatic system showing a recall value over 0.95 is generally considered as reliable for de-identifying a clinical corpus [17,27]. As we can see from Table 3, the XLM-R transformer achieves a recall value over 0.95 in 6 of the 8 NEs, proving the viability of the model as a medical de-identification system.

The two NEs for which the model does not reach a recall value of 0.95 are CONTACTO (a recall score of 0.9412) and UBICACION (a recall value

of 0.8788). Most of the errors the model makes for the UBICACION label are caused by the difficulty the model has in differentiating between centres and locations. In many cases, when the name of a centre is not sufficient to unambiguously identify it, its location is labelled as part of its name. For instance, in the text "Complexo Hospitalario Universitario de Lugo", the name of the Spanish city "Lugo" is labelled as part of the centre's name. However, in the text "Hospital General De Almansa de Lugo", "Lugo" is not tagged as part of the centre's name, but instead is labelled as a location. This is not only a challenge for the automatic models, but also for the human annotators.

Table 3. NER (strict-match) metrics for each NE obtained by the XLM-R (large) system on the test set when trained using Gálen and the augmented documents. Additionally, we describe the number of NEs inferred correctly (cor) and incorrectly (inc), missed (mis) or spurious (spu), and the total number of NEs in the GS set and inferred by the NER system (pred).

NE	P	R	F1	COR	INC	MIS	SPU	GS	PRED
CENTRO	.9167	.9670	.9412	88	3	0	4	91	96
CONTACTO	1.000	.9412	.9697	16	0	1	0	17	16
DIRECCION	1.000	1.000	1.000	1	0	0	0	1	1
HISTORIA	.8824	1.000	.9375	15	0	0	2	15	17
IDENT	.8889	1.000	.9412	8	0	0	1	8	9
PERSONA	.9618	.9692	.9655	126	3	1	2	130	131
REFERENCIA	.9565	1.000	.9778	22	0	0	1	22	23
UBICACION	.9667	.8788	.9206	58	5	3	1	66	60
macro-avg	.9466	.9695	.9567						
micro-avg	.9462	.9543	.9502	334	12	4	20	350	353

Finally, for illustration purposes, Fig. 3 shows the predictions made by the XLM-R large model on a sample medical document in Spanish. We elaborated a text with a similar structure to the real-world clinical cases contained in the Galén corpus. As we can see from the image, for this particular clinical document, the transformer-based model was able to correctly identify all PHI entities contained in it.

Fig. 3. De-identification performed by the XLM-R (Large) system on a sample clinical text in Spanish.

6 Conclusions

In this work, we have addressed the problem of automatic de-identification of real-world clinical documents in Spanish. For this purpose, we have produced a corpus of 599 de-identified medical cases obtained from Galén [25]. We have systematically analysed the performance of RNNs and transformer-based models when applied to this NER task. Additionally, we have developed a DA strategy to obtain a $\times 10$ augmentation of the number of documents used to train the models. The obtained results show that, on the one hand, all the models applied herein benefit from the DA procedure, since their predictive performance increases when using the augmented version of the corpus. On the other hand, transformers outperform RNNs for the de-identification of medical texts in Spanish. Among the RNN-based systems, the best performance is obtained by the 2-BiLSTM model, with strict-match micro-averaged precision, recall and F1-score of 0.8898, 0.9229 and 0.9060, respectively. For its part, the multilingual XLM-R large transformer achieves the highest performance obtained in this study, with strict-match micro-averaged precision, recall and F1-score of 0.9462, 0.9543 and 0.9502, respectively.

In future works, given the observed superiority of transformer-based models analysed in this study, we will explore how domain-specific models perform on the de-identification problem, since, as it has been shown in the literature, transformers adapted to the specificities of the medical documents obtain SOTA performance in many distinct tasks in the clinical NLP domain [16,19,20]. Finally, given the promising results obtained in this work, we will try to validate our developed methodology on external real-world corpora from other medical centres in Spain.

Acknowledgements. The authors acknowledge the support from the Ministerio de Economía y Empresa (MINECO) through grant TIN2017-88728-C2-1-R, from the Min-

isterio de Ciencia e Innovación (MICINN) under project PID2020-116898RB-I00, from the Universidad de Málaga and Junta de Andalucía through grant UMA20-FEDERJA-045, from the Malaga-Pfizer consortium for AI research in Cancer - MAPIC, and from the Instituto de Investigación Biomédica de Málaga - IBIMA (all including FEDER funds).

References

1. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Associat. Comput. Linguist. **5**, 135–146 (2017)
2. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8440–8451, Online (Jul 2020)
3. Cortes Generales de España: Ley Orgánica 3/2018, de 5 de diciembre. de Protección de Datos Personales y garantía de los derechos digitales, Boletìn Oficial del Estado (2018)
4. Council of the European Union: 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/EC (General Data Protection Regulation). Off. J. Eur. Union **119**, 1–88 (2016)
5. Dernoncourt, F., Lee, J.Y., Uzuner, O., Szolovits, P.: De-identification of patient notes with recurrent neural networks. J. Am. Med. Inform. Assoc. **24**(3), 596–606 (2017). https://doi.org/10.1093/jamia/ocw156
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171–4186 (2019)
7. Dorr, D.A., Phillips, W., Phansalkar, S., Sims, S.A., Hurdle, J.F.: Assessing the difficulty and time cost of de-identification in clinical narratives. Methods Inf. Med. **45**(03), 246–252 (2006). https://doi.org/10.1055/s-0038-1634080
8. Grishman, R., Sundheim, B.M.: Message Understanding Conference-6: A brief history. In: COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics (1996)
9. Gutiérrez-Fandiño, A., et al.: MarIA: Spanish Language Models. Procesamiento del Lenguaje Natural 68(0), 39–60 (2022). https://doi.org/10.26342/2022-68-3
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
11. Jan, T., Trienschnigg, D., Seifert, C., Hiemstra, D.: Comparing rule-based, feature-based and deep neural methods for de-identification of dutch medical records. In: ACM Health Search and Data Mining Workshop, HSDM 2020 (2020)
12. Jha, A., et al.: Use of electronic health records in US hospitals. N. Engl. J. Med. **360**(16), 1628–1638 (2009)
13. Jiang, Z., Zhao, C., He, B., Guan, Y., Jiang, J.: De-identification of medical records using conditional random fields and long short-term memory networks. J. Biomed. Inform. **75**, S43–S53 (2017)
14. Lafferty, J.D., 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, pp. 282–289 (2001)

15. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360 (2016)
16. Lee, J., et al.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
17. Liu, L., Perez-Concha, O., Nguyen, A., Bennett, V., Jorm, L.: De-identifying Australian hospital discharge summaries: An end-to-end framework using ensemble of deep learning models. J. Biomed. Inform. **135**, 104215 (2022)
18. Liu, Y., et al.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv [cs.CL] (2019)
19. López-García, G., Jerez, J.M., Ribelles, N., Alba, E., Veredas, F.J.: Detection of tumor morphology mentions in clinical reports in spanish using transformers. In: Advances in Computational Intelligence, pp. 24–35. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-85030-2_3
20. López-Garcìa, G., Jerez, J.M., Ribelles, N., Alba, E., Veredas, F.J.: Transformers for Clinical Coding in Spanish. IEEE Access **9**, 72387–72397 (2021)
21. Marimon, M., et al.: Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In: IberLEF@ SEPLN, pp. 618–638 (2019)
22. Perez, N., García-Sardiña, L., Serras, M., Del Pozo, A.: Vicomtech at MEDDOCAN: Medical Document Anonymization. In: IberLEF@ SEPLN, pp. 696–703 (2019)
23. Pérez-Díez, I., Pérez-Moraga, R., López-Cerdán, A., Salinas-Serrano, J.M., la Iglesia-Vayá, M.d.: De-identifying Spanish medical texts-named entity recognition applied to radiology reports. J. Biomed. Semant. **12**(1), 1–13 (2021)
24. Ramshaw, L.A., Marcus, M.P.: Text chunking using Transformation-Based learning. In: Natural Language Processing Using Very Large Corpora, pp. 157–176. Springer, Netherlands, Dordrecht (1999). https://doi.org/10.1007/978-94-017-2390-9_10
25. Ribelles, N., et al.: Galén: Sistema de información para la gestión y coordinación de procesos en un servicio de oncología. RevistaeSalud **6**(21), 1–12 (2010)
26. Richter-Pechanski, P., Amr, A., Katus, H.A., Dieterich, C.: Deep learning approaches outperform conventional strategies in de-identification of german medical reports. In: GMDS, pp. 101–109 (2019). https://doi.org/10.3233/SHTI190813
27. Stubbs, A., Kotfila, C.: Özlem Uzuner: Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. J. Biomed. Inform. **58**, S11–S19 (2015)
28. Urda, D., Ribelles, N., Subirats, J.L., Franco, L., Alba, E., Jerez, J.M.: Addressing critical issues in the development of an oncology information system. Int. J. Med. Informatics **82**(5), 398–407 (2013)
29. U.S. Dept. of Health & Human Services: Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Office for Civil Rights (OCR) (2012)
30. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems 30 (2017)
31. Vítores, D.F.: El español: una lengua viva. Instituto Cervantes (2019). https://www.cervantes.es/imagenes/File/espanol_lengua_viva_2019.pdf
32. Yang, H., Garibaldi, J.M.: Automatic detection of protected health information from clinic narratives. J. Biomed. Inform. **58**, S30–S38 (2015)