UNIVERSIDAD REY JUAN CARLOS

MÁSTER EN DATA SCIENCE

Trabajo Fin de Máster
Curso 2018-2019

Anonimización de Informes Médicos

David Córdoba Ruiz

Tutora: Soto Montalvo Herranz

Resumen

Compartir datos en forma de texto es importante para una amplia gama de actividades, pero también genera preocupación sobre privacidad al compartir datos que podrían ser confidenciales. En particular, los registros clínicos con información de salud protegida (*Protected Health Information* o PHI) no se pueden compartir directamente no solo por motivos humanos sino porque el acceso a la historia clínica con fines judiciales, epidemiológicos, de salud pública, de investigación o de docencia, se rige en España por lo dispuesto en la Ley Orgánica 3/2018, de Protección de Datos de Carácter Personal (LOPD), y en la Ley 14/1986, General de Sanidad [1].

Una condición previa necesaria para acceder a los registros clínicos fuera de los hospitales es su anonimización o disociación, es decir, la eliminación exhaustiva o el reemplazo de toda la información que pueda asociarse a persona identificada o identificable.

Una solución para eliminar toda la información confindencial es anonimizar el texto de forma automática. Sin embargo, esta es una tarea difícil debido a la forma no estructurada de datos textuales y la ambigüedad del lenguaje natural.

En este trabajo se ha hecho un estudio previo de los métodos de automatización que han usado diversos autores para proponer uno que se ajuste a los informes médicos cedidos en la tarea de MEDDOCAN: Medical Document Anonymization task. Esta tarea forma parte de la iniciativa IberLEF 2019, donde han organizado la primera tarea de desafío comunitario dedicada específicamente al anonimato de documentos médicos en español (más información en: `http://temu.bsc.es/meddocan/`). Se ha concluido que el mejor enfoque es el que usa los Campos Aleatorios Condicionales (*Conditional Random Fields* o CRFs) generalmente añadiendo algún procesamiento complementario como el uso de expresiones regulares o diccionarios, entre otros.

Finalmente, el presente trabajo propone un modelo supervisado que usa los CRFs junto con indicaciones específicas, expresiones regulares y diccionarios para la identificación de PHI en los informes de la tarea MEDDOCAN.

Índice general

Capítulo 1

Introducción

En este capítulo se pretende explicar la motivación del presente trabajo, los problemas que pueden surgir en su desarrollo y el objetivo del mismo. La finalidad del trabajo es presentar un sistema capaz de detectar información personal o sensible de personas físicas en documentos clínicos. Este propósito viene motivado por el creciente aumento de textos informatizados y la preocupación por proteger la privacidad de las personas y respetar las leyes existentes.

1.1. Motivación

El verbo anonimizar es de reciente aceptación en el idioma español. Tal es así que en el anterior diccionario de la Real Academia Española publicado (22a edición, año 2001) no lo definía. Tuvo que esperarse a la 23a edición (2014) para añadirlo. La R.A.E. define anonimizar como "expresar un dato relativo a entidades o personas, eliminando la referencia a su identidad" [2]. La ley 14/2007 del Reino de España que regula la investigación biomédica, define en su artículo tercero la anonimización como el "proceso por el cual deja de ser posible establecer por medios razonables el nexo entre un dato y el sujeto al que se refiere" [3].

Existen además dos niveles de anonimización. La *anonimización* propiamente dicha y la *deidentificacion* o *disociación*. La primera implica eliminar de forma irreversible toda información que permita identificar a un individuo u organización. La segunda en cambio añade la posibilidad de que se guarde algún registro referencial que permita a una entidad autorizada o de confianza acceder a los datos personales eliminados. La ley española también da dos definiciones en relación a los datos que son anonimizados o deidentificados respectivamente:

- Se define como *dato anonimizado* o *irreversiblemente disociado*, a aquel dato que no puede asociarse a una persona identificada o identificable por haberse destruido el nexo con toda información que identifique al sujeto, o porque dicha asociación exige

un esfuerzo no razonable, entendiendo por tal el empleo de una cantidad de tiempo, gastos y trabajo desproporcionados.

- Se entiende como dato *codificado* o *reversiblemente disociado* aquel dato no asociado a una persona identificada o identificable por haberse sustituido o desligado la información que identifica a esa persona, utilizando un código que permita la operación inversa.

Existen diversos ámbitos de aplicación de la anonimización como herramienta en la gestión documental, siendo de interés para este trabajo el ámbito médico.

En el entorno médico se generan gran cantidad de documentos con información relevante para consulta profesional. El ejemplo más notable es el de las historias clínicas. Estos documentos guardan información médica de un paciente, que puede ser utilizada para realizar un mejor diagnóstico o tratamiento de otros pacientes con síntomas similares.

Hoy día es habitual que los centros asistenciales cuenten con algún sistema de gestión médica informatizado, que contenga alguna base de datos documental con las historias clínicas. Sin embargo, existen restricciones a la hora de hacer pública esa información incluso para consulta de otros profesionales de la medicina, ya que las historias contienen información de salud protegida (*Protected Health Information* o PHI), resguardada por leyes de protección de datos personales. Por tal motivo, el uso de algoritmos capaces de anonimizar los sistemas de gestión documental orientados al ámbito médico es de gran utilidad.

1.2. Problemas

A la hora de enfrentarse a una tarea de anonimización de documentos, surgen diversos problemas que deben tenerse en cuenta, entre los que caben destacar los siguientes:

1. Acceso a documentos para la creación de un algoritmo de anonimización. Debido a que la divulgación de documentos que contienen información personal está altamente protegido, el uso de estos documentos para estudiar un sistema que los anonimice requiere de una serie de implicaciones legales que la persona implicada debe conocer y aceptar.

2. Procesamiento de texto plano. La gran mayoría de las historias clínicas vienen en texto plano. El inconveniente de este es que posee información no estructurada, lo que requiere un preprocesamiento del texto para transformarlo en información útil. En este punto hay que saber cuál es la información útil que se quiere obtener. Además, se debe tener en cuenta que no todos los textos pueden estar estructurados de igual forma entre sí, pueden contener erratas o incluso errores de formato.

3. Uso de librerías. Es habitual apoyarse en librerías existentes para preprocesar un texto y facilitar la obtención de diversas características del mismo. Sin embargo, en ocasiones los métodos de las librerías no son impecables y pueden generar errores o crear ruido que se va arrastrando. Esto no se puede evitar por lo que hay que aceptarlo y tenerlo en cuenta.

4. Forma de los PHIs. Algunos PHIs están compuestos por varias palabras. Por ejemplo: Doctor Juan García, Dr Juan García y Dr J. García se refieren a la misma persona pero no es fácil para una máquina relacionar ambas. Además, plantea el dilema de si es conveniente estudiarlo como conjunto o separarlo por palabras. Si se separan se tiene al menos el apellido. Por contra, Dr Juan García y Dr Jose García no se refieren a la misma persona. Si bien, en un mismo informe médico la probabilidad de que sucedan ambas es baja, es un ejemplo extensible a otros nombres y otras categorías.

5. Escasos PHI anotados. Para generalizar lo máximo posible un modelo para futuros documentos, es deseable tener un gran número de PHI anotados. Además, esto es imprescindible si se quiere utilizar un modelo de aprendizaje automático y que sea capaz de aprender medianamente bien. Sin embargo, es frecuente encontrar insuficientes PHI para algunas categorías.

1.3. Objetivo

Aunque existen numerosos algoritmos de anonimización la mayoría de ellos se centran en textos en inglés. Por tanto, el objetivo de este trabajo es estudiar técnicas utilizadas para la detección de información sensible en textos en inglés y proponer un modelo que consiga funcionar de manera razonable para historias clínicas en español. Para ello, se ha usado el corpus compuesto por informes médicos en este idioma que ofrece la tarea de MEDDOCAN (*Medical Document Anonymization task*). Esta tarea forma parte de la iniciativa IberLEF 2019, donde han organizado la primera tarea de desafío comunitario dedicada específicamente al anonimato de documentos médicos en español [4].

En esta memoria se detalla cómo se han ido realizando todas las partes para lograr el objetivo propuesto y los elementos que se han tenido en cuenta para su realización.

La organización del presente trabajo viene estructurado como sigue:

El segundo capítulo está dedicado al estado del arte. Aquí se exponen distintas propuestas seguidas por diversos autores de trabajos similares al que se está exponiendo que llevaron a cabo para la disociación en textos médicos. En el capítulo tercero se detallan las tecnologías utilizadas para la realización del presente trabajo. En el cuarto capítulo se presenta el sistema de anonimización de informes clínicos llevado a cabo en este trabajo y las distintas partes de dicho sistema. El capítulo cinco muestra la evaluación del sistema

propuesto y el resultado final junto con los obtenidos en las pruebas previas realizadas. Por último, se expresan las conclusiones obtenidas del proyecto y posibles trabajos futuros que se podrían realizar para la mejora del sistema o planteamientos distintos que se podrían haber realizado.

Capítulo 2

Estado del arte

En este capítulo se presentan algunos sistemas automáticos para anonimización de textos clínicos que se han considerado más representativos. La particularidad de todos ellos es que se desarrollaron para la detección de información sensible de textos en inglés.

Todos ellos utilizan sistemas de anonimización basados en Reconocimiento de Entidades Nombradas (*Named Entity Recognition*, NER). Esto no es de extrañar puesto que el objetivo fundamental del NER es identificar personas, organizaciones y localizaciones, que coincide con gran parte de la información que se desea anonimizar.

Estos sistemas de disociación pueden basarse en reglas (diccionarios, expresiones regulares), modelos de aprendizaje automático (ML) o híbridos, que combinan reglas y modelos de aprendizaje.

2.1. Sistemas basados en patrones y diccionarios

Los primeros sistemas para encontrar información sensible se basaban en el uso de expresiones regulares y consultas a diccionarios. Uno de ellos fue planteado por Sweeney [11] en 1996 al que denominó *Scrub*.

El sistema *Scrub* tiene múltiples algoritmos en paralelo para detectar 25 clases. Usa diccionarios para clasificar los nombres propios (personas, países, localizaciones, etc.) mientras que para las entidades alfanuméricas utiliza expresiones regulares. Para comprobar su eficacia, utilizaron un conjunto de historias clínicas pediátricas de pacientes y cartas a médicos. Comprobaron que el sistema podía detectar el 99 % de las entidades dentro de un documento.

En 2003, Bernam desarrolló el sistema *Concept-Match* [12] basado en diccionarios. Realizó un módulo de preprocesamiento donde eliminaba todas las stopwords. Posteriormente, buscaba las palabras restantes en vocabularios biomédicos y de salud del UMLS Metathesaurus (libro nacional de medicina de EEUU). Se reportó que este sistema obtuvo un alto recall y una baja precisión debido a los falsos-positivos.

2.2. Sistemas basados en Modelos de Aprendizaje Automático

El sistema de Wellner llamado *MITRE* [13] estuvo basado en métodos NER y dos clasificadores, uno basado en CRF y otro en un Modelo Oculto de Márkov (HMM)[1]. Además, crea un módulo de postprocesamiento donde utiliza expresiones regulares. Durante la evaluación de la tarea en la que colaboraron resultó un *F1-score* de 0.983.

Szarvas presentó un sistema basado en NER usando Boosting y árboles de decisión [14]. Se basa en la pregunta planteada por Kearns y Valiant: "¿Puede un conjunto de aprendices débiles crear un único alumno fuerte?". La herramienta NER se adaptó para detectar PHI usando frecuencia y co-ocurrencias de las palabras, diccionarios e información contextual. Sin embargo, no usó POS-tagging. Se entrenaron tres clasificadores diferentes y se ejecutaron en paralelo para estimar si un token dado pertenecía a una PHI. Un token se aceptaba como PHI si, al menos, dos clasificadores votaron positivamente. Durante la evaluación, este sistema obtuvo un *F1-score* superior a 0.960 para todas las NEs en la tarea que participaron.

Arakami desarrolló un sistema basado en CRF [15]. Detectaba entidades usando características locales (morfológicas, POS, palabras circundantes), características de la frase (posición dentro de la frase y del documento) y características externas (diccionarios de nombres y expresiones regulares para entidades numéricas). El autor indica que las características de la frase fueron la mayor contribución para detectar algunas entidades como números de identificación, fechas y nombres propios.

2.3. Sistemas Híbridos

Aunque se ha visto que los modelos descritos con anterioridad funcionan bien, en general, son los sistemas híbridos los que mejor resultados dan. Esto se debe principalmente a que el uso de reglas o modelos de aprendizaje no resultan tan efectivos para la detección de algunos tipos de PHI. Sin embargo, combinados pueden llegar a abarcar más categorías de PHI distintas.

Hara describió un sistema basado en Máquinas de Vector Soporte (SVM) y expresiones regulares [16]. Este incluía un proceso para detectar entidades numéricas mediante expresiones regulares, un clasificador SVM para detectar entidades textuales dentro de una frase y otro clasificador de código abierto basado también en SVM para determinar la clases de

[1]Un modelo oculto de Márkov o HMM (por sus siglas del inglés, *Hidden Markov Model*) es un modelo estadístico que, a diferencia de los CRFs, es generativo, es decir, basado en un modelo de distribución conjunta $P(X, Y)$ en el que se asume que el sistema a modelar es un proceso de Márkov de parámetros desconocidos. El objetivo es determinar los parámetros desconocidos (u ocultos, de ahí el nombre) de dicha cadena a partir de los parámetros observables.

las entidades. Este clasificador usaba POS y características morfológicas.

El *Health Information DE-Identification* (HIDE) fue desarrollado por Gardner [17] para la de-identificación de PHI basado en CRF y un modelo de privacidad k-anonimato [2]. El sistema consistía en tres componentes: el primero para extracción de entidades usando un clasificador CRF, el segundo componente enlazaba entidades que se habían encontrado previamente y el último llevaba a cabo la anonimización de los datos.

Por último, presentar el modelo de Yang y Garibaldi [6] de la Universidad de Nottingham. El artículo donde se detalla el procedimiento de su modelo es el que ha servido de guía para la realización del presente trabajo por su sencillez y, al mismo tiempo, por los buenos resultados que obtuvieron en la tarea que participaron en 2014: *i2b2 De-identification Challenge*. Los autores, propusieron un sistema híbrido usando un clasificador CRF, diccionarios, expresiones regulares y un postprocesamiento. El sistema usaba un conjunto de tokens con características contextuales: palabra, lema, POS, forma de la palabra, mayúsculas, etc. Las expresiones regulares y diccionarios se usaban para detectar PHIs poco frecuentes. Por último, realizaron un postprocesamiento donde infirieron sobre posibles PHI que podrían aparecer simultáneamente en el mismo documento para ser detectados en una segunda ejecución del sistema. Estos PHI se denominan *PHI de confianza*.

Tras la revisión del estado del arte, se ha decidido implementar un sistema híbrido que use como clasificador un CRF y complementarlo con expresiones regulares, diccionarios y la idea de PHI de confianza que se ha presentado en el modelo de Yang y Garibaldi.

[2]Estos modelos generan datos anonimizados que cumplen la propiedad k-anonimato. Se dice que un conjunto de datos publicados tiene la propiedad de k-anonimato (o es k-anónimo) si la información de todas y cada una de las personas contenidas en ese conjunto es idéntica, al menos, con otras k-1 personas que también aparecen en dicho conjunto.

Capítulo 3

Tecnologías empleadas

Para la realización del presente trabajo se ha utilizado el lenguaje de programación *Python 3.7* ya que presenta librerías útiles tanto para procesamiento de lenguaje natural como para el entrenamiento de CRFs.

Entre las diversas librerías que facilitan el PLN cabe destacar *nltk* [8] y *spaCy* [9]. Estas librerías tienen implementados métodos para realizar las técnicas de PLN como detección y separación del texto en frases o palabras, detección de stopwords (palabras vacías), etiquetación de palabras en función de su categoría gramatical, etc. Además, ambas están bastante desarrolladas para el tratamiento de texto en distintos idiomas, en particular el español.

Tras evaluar las dos librerías, se ha decidido usar la librería *spaCy* debido a su sencillez, la claridad de la documentación publicada y el hecho de que usa texto de Wikipedia para entrenar los modelos, por tanto, es factible que funcione en muchos géneros, incluyendo el lenguaje médico. Además, una de las características que presenta *spaCy* frente a *nltk* es el cálculo de las denominadas Entidades Nombradas sin necesidad de realizar un preprocesamiento previo del texto. Por contra, se ha detectado que no ha funcionado bien a la hora de identificar si una palabra es inicio o fin de oración.

En este trabajo, se ha implementado un modelo de aprendizaje basado en CRF. La librería que se ha usado la cual tiene implementados métodos para entrenar CRFs de manera eficiente ha sido *CRFsuite* [10]. Además, esta librería facilita funciones muy prácticas para la representación de los resultados obtenidos y para la evaluación del modelo de entrenamiento.

Capítulo 4

Propuesta de sistema

En esta sección se detalla el sistema que se ha llevado a cabo en el trabajo. Además, se explican las características del corpus usado, las diferentes partes de las que está compuesto el sistema y porqué se ha realizado tal división, además de otros detalles que se han tenido en cuenta en el desarrollo del algoritmo utilizado.

4.1. Arquitectura del Sistema

Se ha decidido realizar un sistema híbrido para aprovechar las ventajas que ofrece el uso de reglas y de aprendizaje automático conjuntamente. Como algoritmo de aprendizaje supervisado se ha utilizado un CRF.

En la Figura 4.1 se puede ver un gráfico conceptual que detalla la estructura del sistema propuesto.

El sistema se compone de tres partes:

La primera parte se trata del preprocesamiento. Recibe como entrada los informes médicos de la tarea MEDDOCAN, separa por su estructura (cabecera, cuerpo y pie) y usa diferentes técnicas de procesamiento de texto para preparar la información que se va a utilizar en el siguiente módulo. La separación por cabecera, cuerpo y pie es debido a que cada una, a su vez, tiene una organización distinta entre sí pero todos los documentos clínicos de la colección con la que se trabaja presentan la misma estructura general.

El segundo módulo se dedica a la detección de PHI. Se han creado distintos métodos de detección según la parte del informe de la que se trate. Para la cabecera, se ha aprovechado su estructura bien definida para extraer de forma *directa* los PHI. Para la obtención de los PHI que se encontraban en el cuerpo se ha entrenado un modelo CRF con todas las categorías PHI que podían encontrarse en esa parte del documento. Para la parte del pie se observó que no tenía una estructura lo suficientemente adecuada para usar otro CRF

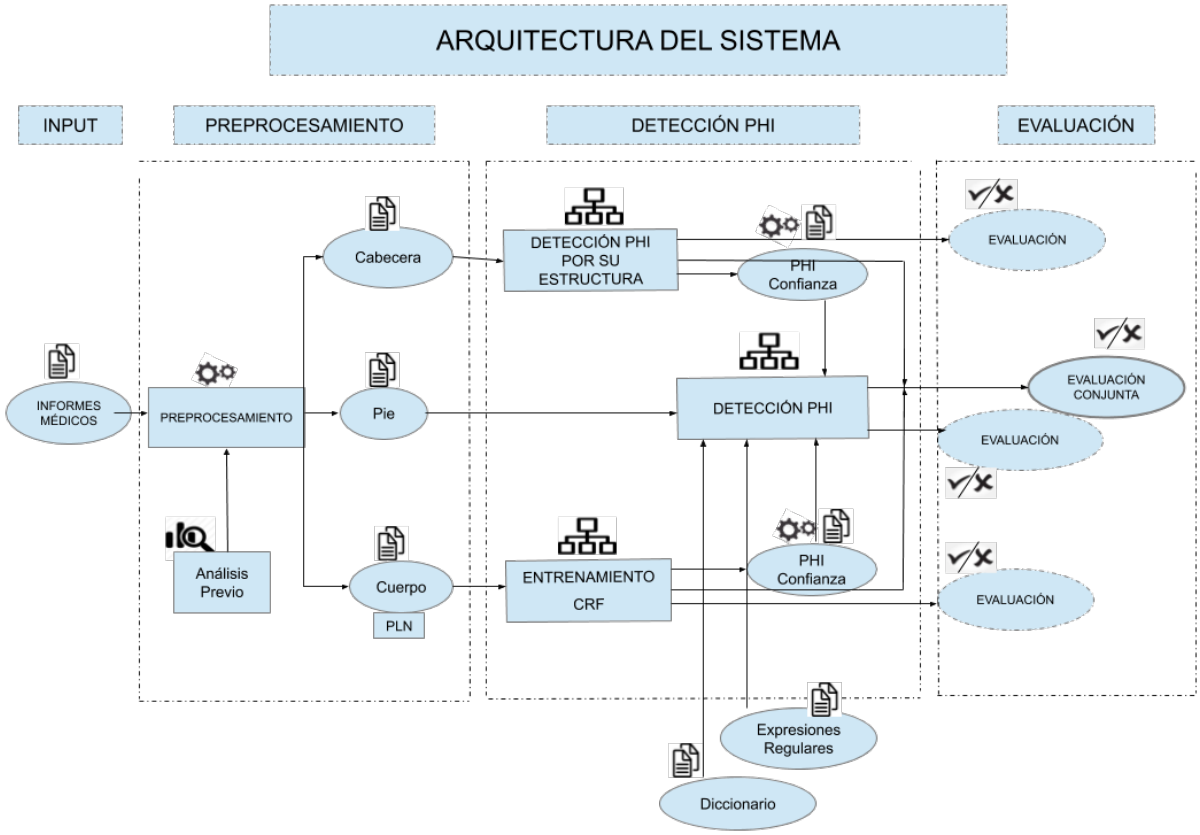

Figura 4.1: Gráfico-Resumen del sistema propuesto

ni todos los PHI que se podían encontrar eran de fácil extracción con uso de expresiones regulares y diccionarios. Por tanto, se ha optado por combinar las expresiones regulares y diccionarios con el uso de PHI de confianza.

La última sección viene dedicada a la evaluación del sistema. Por un lado, se han evaluado los resultados obtenidos de la extracción de PHI de cada una de las partes del módulo anterior. Por otro lado, se ha realizado una evaluación conjunta de todo el sistema. Cabe destacar que el mejor resultado se ha obtenido en el encabezado debido a su estructura bien definida. En cuanto al entrenamiento con el CRF para el cuerpo se ha observado que para ciertas categorías ha funcionado bastante bien mientras que para otras no tanto. El resultado obtenido para el pie también varía para algunos tipos de PHI pero no se han obtenido malos resultados.

4.2. Corpus MEDDOCAN

Como se ha comentado, el corpus que se ha usado para la realización del trabajo ha sido el que se ofrecía en la tarea de MEDDOCAN (`http://temu.bsc.es/meddocan/`). En esta tarea han preparado un corpus sintético de casos clínicos en español enriquecidos con expresiones PHI, llamado corpus MEDDOCAN. Este corpus se compone de 1,000 estudios de casos clínicos y fue seleccionado manualmente por un médico en ejercicio y documentalistas de salud aumentaron con frases de PHI, agregando información de PHI de los resúmenes de alta y los registros clínicos de médica genética.

La colección final de 1,000 casos clínicos que componen el corpus tenía alrededor de 33 mil oraciones, con un promedio de alrededor de 33 oraciones por caso clínico. El corpus de MEDDOCAN contiene alrededor de 495 mil palabras, con un promedio de 494 palabras por caso clínico. El corpus se ha distribuido en texto plano en la codificación UTF8, donde cada caso clínico se almacena como un solo archivo, mientras que las anotaciones de PHI se han publicado en el formato BRAT (ver Figura 4.2).

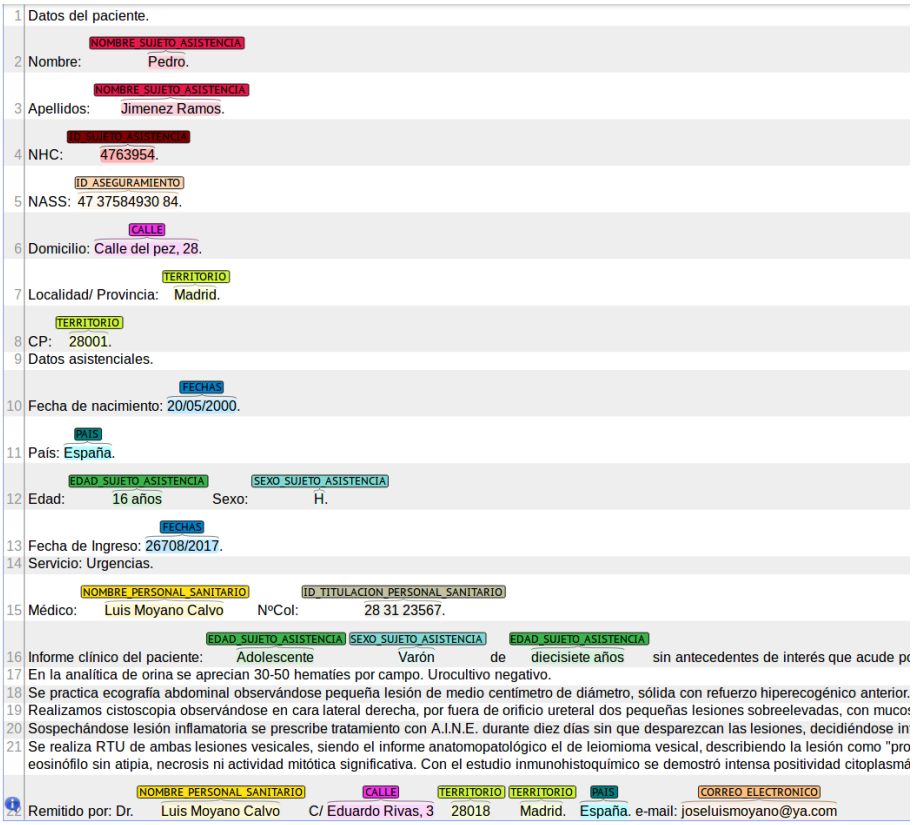

Figura 4.2: Ejemplo de anotación de MEDDOCAN

Se puede observar que los documentos vienen estructurados en tres partes bien definidas: cabecera, cuerpo y pie. Es por ello que ha sido conveniente estudiarlas y analizarlas cada una por separado. Por otro lado, se ha observado que la estructura del conjunto de informes no cambia, lo que evita la necesidad de realizar un análisis extra.

Las categorías de PHI que tienen en cuenta en la tarea de MEDDOCAN son 28, entre las que se encuentran nombres de pacientes, personal médico, números identificativos, hospitales, municipios, etc. (En el Anexo I de esta memoria se indican cada una de ellas junto con el valor que se ha asociado como equivalencia para simplificar cada una de las etiquetas a la hora de realizar el trabajo).

4.3. Análisis previo

Con el fin de plantear un sistema que pudiera funcionar adecuadamente para el corpus utilizado, se ha llevado a cabo un análisis previo de la estructura de los informes y las apariciones de PHI en cada una de sus partes.

4.3.1. Cabecera

Se ha podido ver que la cabecera tiene siempre una organización bien definida. Para asegurar que esta estructura es invariante en todos los documentos, se ha realizado un pequeño estudio confirmando finalmente que está muy estructurada y se pueden obtener un conjunto de PHI de manera sencilla y fiable. Para visualizar mejor la situación se presenta la Figura 4.3.

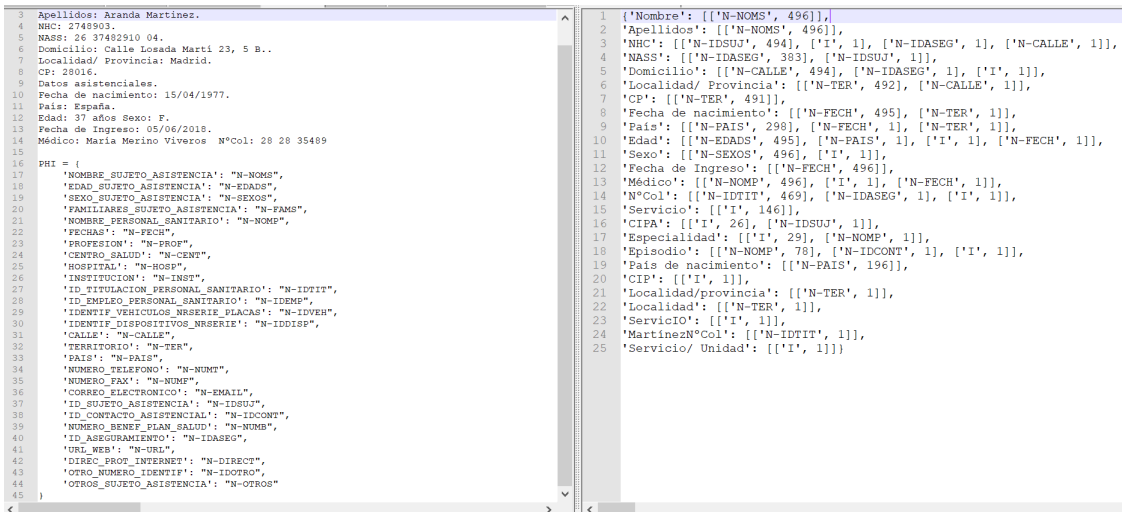

Figura 4.3: La parte izquierda representa un ejemplo del encabezado de un documento junto con todos los PHI catalogados. La parte derecha muestra el análisis realizado.

Se puede observar que en el ejemplo que se presenta en la parte izquierda superior de la imagen, posteriormente a la sentencia "*Apellidos*:", le sigue un apellido (que corresponde al apellido del paciente), posteriormente a "*NHC*:" le sigue un número (que corresponde al ID del sujeto de asistencia), y así una sentencia tras otra.

En la parte derecha, se examina el número de categorías que aparecen seguidas de una determinada palabra o palabras clave que se tomará como indicador de un PHI. Se

puede ver que en el total de documentos se pueden encontrar 25 indicadores distintos. Cada uno de ellos aparece con el tipo de PHI junto con el número de veces que se ha encontrado dicho PHI posteriormente a ese indicador. Por ejemplo, hay 496 PHI − NOMBRE_SUJETO_ASISTENCIA detrás de la palabra clave "*Nombre*:", de un total de 496 frases que lo contienen. Dicho de otra manera, *siempre* que se ha encontrado *Nombre*: venía seguido del PHI de la categoría NOMBRE_SUJETO_ASISTENCIA. Sin embargo, en algunos indicadores se han encontrado más de un tipo de categoría seguido de este. Por ejemplo, en 495 frases que aparecía "*Fecha de nacimiento*" le seguía un PHI de tipo fecha (FECH), salvo en un caso, en el cual se ha encontrado un PHI de tipo Territorio (TER). Sin embargo, estos casos son puntuales. De hecho, se han analizado algunos de estos casos y, en su mayoría, se deben a erratas en las anotaciones de los documentos o a un formato inesperado de algunos informes.

Por tanto, queda claro que la estructura del encabezado define los PHI que se van a encontrar salvo casos puntuales.

4.3.2. Cuerpo

Al igual que el encabezado, para el cuerpo se ha realizado un estudio previo para decidir cuál es la mejor manera para detectar los PHI que contiene.

Se han encontrado 18 PHI diferentes, contando solo el cuerpo. Debido a errores al leer los datos, se ha conseguido tratar correctamente 475 documentos, de los cuales en 8 no se han encontrado PHI anotados. El total de PHI encontrados en todos los documentos es de 2891. Es decir, obtenemos una media de cerca de 6 PHI en el cuerpo por documento. A priori, puede parecer un número aceptable pues solo estamos teniendo en cuenta una parte del documento. Sin embargo, si se considera el total de las frases leídas y la aparición de un PHI en una frase, la frecuencia de aparición es de 0.127 por frase. A esto hay que añadirle que puede pertenecer a una de las 18 categorías distintas.

Por tanto, puesto que hay muchas frases donde no se va a encontrar PHI y se ha observado que no hay un esquema definido ni ningún patrón de aparición en los PHI para la mayoría de las categorías, se pensó que el uso de expresiones regulares y diccionarios para su detección podría ser costoso tanto en implementación como en rendimiento. Se ha concluido entonces usar un modelo de aprendizaje supervisado que utiliza Campos Aleatorios Condicionales que, como se menciona en el capítulo 2, son modelos bastantes eficaces para la anonimización de textos.

Para entender mejor el funcionamiento de los CRFs se ha investigado sobre ellos y porqué son útiles en el tema que se trata.

Campos Aleatorios Condicionales

Los Modelos Gráficos proporcionan una metodología general de inferencia que explotan eficientemente la estructura del grafo para hallar probabilidades marginales y condicionales.

Tradicionalmente, estos modelos se han usado para representar la distribución conjunta $P(X, Y)$, pero esto requiere la modelización de la distribución marginal $P(X)$ la cual puede llevar a modelos complejos e intratables.

$$P(X, Y) = P(Y/X)P_X(X) \tag{4.1}$$

Una solución es modelizar directamente la distribución condicional $P(X/Y)$. Este es el enfoque que usan los Campos Aleatorios Condicionales (*Conditional Random Fields*, CRFs). Un CRF es simplemente una distribución condicional $P(Y/X)$ con una estructura de grafo asociada.

La probabilidad conjunta se factoriza como:

$$P(Y/X) = \frac{\prod_C \Psi_C(X_C, Y_C)}{Z(X)} \quad con \quad Z = Z(X) = \sum_{Y \in y} \prod_C \Psi_C(X_C, Y_C) \tag{4.2}$$

Donde:

- C se denomina *clique*. Un clique es un conjunto de vértices dentro de grafo tal que todo par de vértices distintos son adyacentes, es decir, existe una arista que los conecta.
- X_C es la variable observada en el clique C.
- Y_C variable multinomial a predecir en el clique C.
- $\Psi_C : V^n \longrightarrow \mathbb{R}^+$ son funciones potenciales definidas sobre un clique C. Estas funciones modelan relaciones entre padre e hijo dentro de un grafo.
- Z es una función de normalización.

El conjunto de factores en el grafo $\{\Psi_C\}$ se describen usualmente mediante combinaciones log-lineales de funciones características f que contienen la información relevante extraída de los datos:

$$\Psi_C(X_C, Y_C) = exp\{\sum_k \theta_{C_k} f_{C_k}(X_C, Y_C)\}$$

Estas funciones pueden comprender valores entre $-\infty$ y $+\infty$ aunque generalmente toman valores binarios $\{0, 1\}$. θ representa los pesos asociados a cada función característica y definen la importancia de una particular función f respecto al resto. Estos pesos constituyen los parámetros del modelo que son aprendidos durante el entrenamiento, minimizando la log-verosimilitud condicional $\log P(Y/X)$ de los datos etiquetados. La estimación de

dichos parámetros se realiza comúnmente mediante el método de máxima verosimilitud penalizado.

En el caso del trabajo que se está definiendo, Y representa la etiqueta asociada a cada palabra {PHI, No PHI}. La variable X es el vector con las características asociada a una palabra.

Para la estimación de los parámetros, se requiere el uso de Teoría de Probabilidad que exige que las variables sean independientes e idénticamente distribuidas. Es por esto que es necesario añadir información de las palabras vecinas como características. Sin embargo, existen otras variantes de CRF que usan estructuras gráficas generales especialmente útiles para el aprendizaje relacional que no requiere de este supuesto.

En la Figura 4.4 se puede ver representado un modelo gráfico de un CRF Lineal.

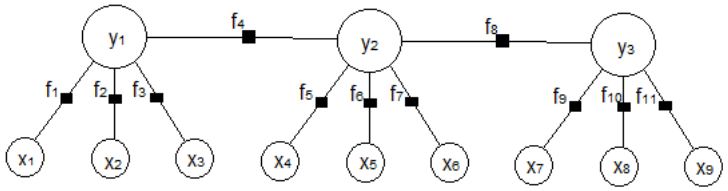

Figura 4.4: Representación de un Modelo Gráfico CRF Lineal

Una de las ventajas que tienen los CRFs es que la salida del modelo no es necesariamente binaria, es decir, puede entrenar diversas categorías. Esto evita el entrenamiento de un modelo por cada PHI.

Para entender mejor de dónde viene la fórmula de la probabilidad conjunta (4.2) y cómo se estiman los parámetros del modelo se aconseja leer el artículo *An Introduction to Conditional Random Fields* [18] y la tesis de *Reconocimiento de Acciones Humanas* [20]. Además, se recomienda el libro *Introducción a los Modelos Gráficos Probabilistas* [19] para conocer variaciones de este modelo y otros modelos gráficos que podrían ser de interés.

Se puede concluir que para que el modelo CRF funcione, se requiere un vector que contenga las características de las palabras. Además, solo serán relevantes aquellas oraciones que posean PHI pues es esta casuística la que va aportar más información al modelo.

4.3.3. Pie

En el pie de los documentos, se ha notado que se incluye, en general, información clínica así como el nombre del médico, hospital, dirección, municipio, país, fax, teléfono y e-mail.

El problema que se percibió fue que no seguían un orden de aparición específica ni tenían por qué incluir todas las categorías citadas. Además, tampoco seguían una estructura propia de una frase, es decir, cada información podía estar separada por un punto, una coma, un espacio o por un artículo.

Se llegó a la conclusión de que el uso de otro CRF para la detección de PHI en el pie podría no ser adecuada puesto que, al no tener la frase una organización estándar, la información de algunas características de las palabras podrían ser incorrectas, sobre todo en el cálculo de POS, NER o inicio y fin de la frase.

Por tanto, la mejor solución fue usar expresiones regulares para la detección de la dirección, teléfono y correo electrónico, y diccionarios para el municipio y país. Sin embargo, quedaban pendientes otras categorías difíciles de incluir en estas reglas. El uso de expresiones regulares para el fax era sencilla pero era complicado diferenciar cuándo era un número de fax o un teléfono. Se observó entonces que siempre que aparecía un fax venía precedido de la sentencia "*Fax:*" por lo que se procedió a detectarlo de manera similar al encabezado.

Seguía estando el problema para el resto de categorías, e incluso otras que podrían no haberse tenido en cuenta pero que podrían aparecer. Se vio que muchas de ellas ya aparecían en el mismo documento en la cabecera o en el cuerpo por lo que, finalmente, se ha tomado la idea del modelo de Yang y Garibaldi que se ha definido al final del capítulo 2 y se han usado PHIs de confianza.

Un PHI de confianza es aquella información dentro de un mismo documento que se acepta como PHI sin tener certeza de que lo sea. Yang y Garibaldi explican distintos métodos de obtención de PHI de confianza [6]. En este caso de estudio, para el pie del documento, se han considerado como PHI de confianza todos aquellos que el sistema ha detectado en la cabecera salvo el sexo del paciente. Es decir, si el sistema ha detectado un PHI en la cabecera, y la información aparece tal cual en el pie, lo tomará como PHI y lo clasificará con la misma categoría que el de la cabecera.

La razón por la que no se ha tenido en cuenta el sexo del paciente es porque se indica únicamente con H y M. Por tanto, podría llevar a errores en la búsqueda de estas letras en el pie del documento.

Además, se han tomado también como PHI de confianza aquellos que se ha sabido que funcionan bien en el modelo de CRF como los nombres propios o los territorios.

Aunque puede ser un método eficaz, hay que tener en cuenta que tiene riesgos. Por ejemplo, supóngase que se ha detectado *Rey Juan Carlos* en la cabecera y se ha clasificado como nombre propio de persona. Si en el pie apareciese Hospital Rey Juan Carlos o Univer-

sidad Rey Juan Carlos detectaría un PHI pero lo clasificaría erróneamente como nombre de persona en lugar de su correspondiente (en este caso Hospital e Institución respectivamente).

4.4. Módulo de Preprocesamiento

En este módulo se procesa el texto de los informes y se prepara de forma que el siguiente módulo pueda tratarlo para la detección de PHI. Se ha tenido en cuenta el estudio previo para preparar la salida según se quieran detectar los PHI para cada parte del documento.

Una vez leídos los informes, se ha creado una instancia de la clase *Document*. Se ha implementado esta clase para que contenga el nombre del documento, el texto instanciado a su vez de la clase *nlp* que usa la librería de *spaCy*, los PHI que contiene el documento junto con las posiciones donde se encuentran y cada una de las partes del documento por separado. Además, contiene algunos métodos adicionales para extraer características de las palabras que se encuentran en el cuerpo del documento:

- Para el encabezado, cada frase se ha dividido en dos subpartes, una que recoge la sentencia previa a los dos puntos y otra posterior. La primera se ha mantenido tal cual para luego escoger el tipo de PHI que le sigue. La segunda se ha divido en palabras ya que puede representar a un PHI con varias palabras. Finalmente, se obtiene un vector que contiene todas las frases de un documento divididas como se ha explicado.
- Para el cuerpo, como se ha indicado, se necesita la extracción de características de cada palabra para entrenar el modelo CRF para lo que se han empleado diferentes técnicas de procesamiento de texto.

  Las técnicas más comunes que se utilizan en PLN son:

  - *Tokenización*: Separación del texto en frases, palabras o carácteres denominados tokens.
  - *Lematización*: Dada una forma flexionada obtener el lema correspondiente que representa por convenio a todas esas formas flexionadas. (Por ejemplo, de las formas flexionadas *dije* o *dijo* se obtiene el lema *decir*)
  - Detección de signos de puntuación y stopwords o palabras vacías (como determinantes y preposiciones).
  - Detección de mayúsculas y caracteres no alfabéticos dentro de una palabra.
  - *Part-of-speech tagging* o *POS-tagging*: Etiquetar una palabra en función de su tipo (verbo, nombre, adjetivo...)

- *Named Entity Recognition (NER).* Consiste en el reconocimiento de Entidades Nombradas (NE). Las Entidades Nombradas son palabras o conjunto de palabras que corresponden a un "objeto del mundo real" a las que normalmente se les asocia a una categoría definida. Por ejemplo, Rafa Nadal y España serían dos NE. A la primera le correspondería la categoría de PERSONA y a la segunda de LUGAR.

Como se indicó en el Capítulo 2, la gran mayoría de los sistemas de anonimización de textos están basados en NER por su relación con la información sensible. Teniendo esto en cuenta se ha incluido un método en la clase *Document* centrado en la extracción de características de palabras dentro del cuerpo. El vector creado contiene las siguientes características: la palabra en sí, su lema, POS, booleano que indica si es alfanumérico, si contiene mayúsculas, si es un título (todas las letras son mayúsculas), si es un NE, si es inicio o fin de la frase y la dependencia dentro de la frase. La dependencia indica cómo actúa una palabra en conjunto con la frase. Además, se han añadido características de la palabra anterior y posterior: la palabra en sí, la dependencia y la etiqueta POS. A este vector se le ha asociado la anotación correspondiente, es decir, si es PHI y en tal caso su categoría asociada, que es lo que usará el modelo para entrenar. A continuación se muestra un ejemplo para la palabra "ingresada":

```
['label=I',
'word.lower=ingresada',
'lemma=ingresar',
'postag=ADJ',
'dependence=amod',
'word.isalpha=True',
'word.isupper=False',
'word.istitle=False',
'word.isentity=False',
'-1:word.lower=años',
'-1:postag=NOUN',
'-1:dependence=nmod',
'+1:word.lower=por',
'+1:postag=ADP',
'+1:dependence=case']
```

- Para el pie, puesto que va a usar expresiones regulares, un diccionario y PHIs de confianza, en la detección de PHI se va a necesitar buscar en el texto completo del pie para llevar a cabo estos métodos. El diccionario se ha formado con municipios de España que se ha obtenido de la Agencia Tributaria [22] añadiendo las Comunidades Autónomas. En total contiene 8273 datos de lugares. El preprocesamiento en este caso devuelve el pie instanciado en la clase *Document* para que el siguiente módulo

pueda usar tanto el texto plano como la división del mismo en tokens.

Cada una de las salidas se resume en un vector con el preprocesamiendo descrito. Además, tienen asociadas las anotaciones del documento que se trata para posteriormente testear los resultados y un identificador único (el nombre del documento) para poder acceder a ellos a la hora de realizar el testeo de todas las partes conjuntamente.

Se ha decidido asociar las anotaciones a esta salida porque la finalidad del trabajo no tiene en cuenta el procesamiento de informes sin anotaciones. Es decir, solo se llega a un análisis del sistema usando los informes que tienen información útil para el testeo.

4.4.1. Módulo de Detección

Este módulo se divide en tres partes, cada una orientada a la detección de PHI para cada una de las secciones de los informes. Para entender el etiquetado que se ha seguido, primero se debe explicar el procesamiento que se ha realizado en el cuerpo:

- La parte del procesamiento del cuerpo se centra en el entrenamiento de un modelo CRF. Como se ha visto en el estudio de los CRFs, la ventaja de este es que el mismo modelo es capaz de aprender diferenciando distintas categorías de PHI introducidas. Esto es importante pues es de esperar que si solo se categoriza como PHI o no PHI la eficacia del modelo será menor, pues las características de un nombre propio difieren, por ejemplo, de un número de identificación. Por otro lado, entrenar un CRF distinto para cada clase de etiqueta sería pesado.

  Otra cosa a tener en cuenta es el tipo de etiquetado. Según lo aprendido en los artículos mencionados en el Capítulo 2, hay algoritmos que se han entrenado diferenciando la organización de las palabras dentro de un PHI. Esto es, si un PHI incluye varias palabras, la primera se etiquetará como “Comienzo PHI” mientras que el resto de palabras que le siguen (en caso de haberlas) se etiquetarán como “Dentro PHI”. Hay incluso autores que también tienen en cuenta si la palabra es final de PHI.

  En este trabajo se han entrenado diversos modelos para ver la eficacia de cada uno según la forma de etiquetado escogida:

  1. Etiquetado *N-I* (Nominal - Irrelevante), sin diferenciar entre las distintas categorías.
  2. Etiquetado *CDI* (Comienzo, Dentro, Irrelevante) sin diferencias entre categorías.
  3. Etiquetado *N-I* diferenciando por categorías.
  4. Etiquetado *CDI* diferenciando por categorías.

Finalmente se decidió el último planteado porque es donde se obtuvieron mejores resultados. Esto es importante tenerlo en cuenta pues, según esta elección, así se ha ido etiquetando la cabecera y el pie para que al realizar la evaluación conjunta no diera problemas.

Otra cuestión a considerar es la elección de los parámetros de entrada para el modelo. El algoritmo de entrenamiento que utiliza la librería *CRFsuite* maximiza el logaritmo de la probabilidad de los datos de entrenamiento con los términos de regularización L1 y L2 utilizando el método de *Limited-memory Broyden-Fletcher-Goldfarb-Shanno* (L-BFGS). Cuando se especifica un coeficiente distinto de cero para el término de regularización L1, el algoritmo cambia al método de *Quasi-Newton Orthant-Wise Limited-memory (OWL-QN)*. En la práctica, este algoritmo mejora los pesos de las características muy lentamente al comienzo de un proceso de entrenamiento, pero al final converge rápidamente en los pesos óptimos de las características.

Se concluyó que una elección de L1 = 0.01, L2 = 0.1 y realizando a lo sumo 5000 iteraciones, se obtenía un mejor resultado en el entrenamiento.

Por tanto, esta parte del módulo transforma el etiquetado según lo indicado y entrena el CRF con los parámetros descritos.

- Posteriormente, se ha realizado la detección de PHI en la cabecera. Para ello, se ha creado un diccionario de Python indicando el tipo de categoría que le corresponde a una oración seguida de un determinado indicador o palabra clave. Se debe recordar que la entrada a esta parte del módulo es [indicador, tokens de palabras que siguen al indicador]. La división en tokens es precisamente para usar el etiquetado *CDI*. Por ejemplo, si se recibe el vector [“Domicilio:”, [“Calle”,“Losada”]], como el diccionario indica que después de *Domicilio* se va a encontrar una calle, el módulo se encargará de asignar las etiquetas *Calle*: C-CALLE, *Losada*: D-CALLE.

- La última parte del módulo se centra en la detección de PHI para el pie. Aquí se han definido las expresiones regulares que se van a usar para localizar las categorías de dirección, teléfono y correo electrónico. Además, se ha añadido el diccionario para la detección de las categorías de territorio. También, realiza la búsqueda de números de faxs en función de si vienen seguidos del indicador *Fax:*. Por último, incluye una sección para obtención de PHI de confianza y posterior utilización.

  Como se ha indicado, los PHIs de confianza se obtienen después del procesado de la cabecera y el pie. Por tanto, una vez ejecutadas las dos primeras partes del módulo,

esta última sección recoge los PHI obtenidos. De la cabeza se toman todos los PHI encontrados salvo el sexo mientras que del cuerpo solo los PHI que ha clasificado como Hospital o Institución. Esto se debe a que el resultado del CRF no es tan bueno como el de la cabecera, por tanto se han escogido estas categorías porque han dado buenos resultados en la evaluación del modelo y son las más difíciles de localizar en el pie. Para evitar PHI de confianza incoherentes y que puedan afectar a una mala detección, solo se han tenido en cuenta aquellos que superen una longitud de 3 caracteres.

Finalmente, se aúna todo lo anterior siguiendo un orden:

1. PHI de Confianza.
2. Expresión regular para e-mail.
3. Búsqueda de Fax por sentencia.
4. Expresión regular para teléfonos.
5. Diccionario de municipios.
6. Expresión regular para categoría “CALLE”.

Este orden es importante pues se ha implementado de forma que, si una palabra ya ha sido clasificada y se realiza otra detección de la misma palabra al usar otro método que se ejecuta posteriormente, ignorará la clasificación obtenida de este último. Esto es necesario ya que, por ejemplo, en el caso del fax y el teléfono, si se utilizara primeramente la expresión regular implementada para el teléfono también podría coincidir con el fax. Por ello, primero deberá intentar buscar el fax y posteriormente ignorar la clasificación de la expresión regular. Por otro lado, la decisión de poner el uso del diccionario al final resultó tras comprobar que existen municipios con nombres de persona (como el municipio Javier en Navarra).

Finalmente, se ha realizado un módulo para la evaluación de cada una de las partes diferenciadas anteriormente y para el conjunto de todas ellas. Este módulo se detalla en el siguiente capítulo donde se incluyen los resultados de cada uno de ellos.

Capítulo 5

Evaluación del Sistema

En este capítulo se presentan los resultados obtenidos en cada una de las partes que se han definido dentro del sistema y el resultado del sistema completo. Para ello, se ha creado un módulo que se apoya en la función de medida que ofrece la misma librería que se usa para entrenar el modelo CRF, *CRFSuite*. Esta función devuelve un resumen de la *precisión*, la *sensibilidad* (recall), el *valor-F* (F1 score) y el *soporte* (support) para cada clase.

5.1. Medidas de evaluación

Se ha considerado que las medidas de evaluación que devuelve la función de la librería *CRFSuite* son suficientes para cuantificar la eficacia del sistema propuesto.

Considérese TP (True Positives) el número de verdaderos positivos obtenidos, FN (False Negatives) el número de falsos positivos, TN (True Negatives) el número de verdaderos negativos y FP (False Positives) el número de falsos positivos. Entonces:

- La *precisión* es la capacidad del clasificador de no etiquetar como positiva una muestra que es negativa. Se calcula como $\frac{TP}{TP+FP}$.
- La *sensibilidad* o *recall* es la capacidad del clasificador para encontrar todas las muestras que son positivas. Se calcula como $\frac{TP}{TP+FN}$.
- El *valor F* o *F1 score* se considera como una media armónica que combina los valores de la precisión y de la sensibilidad. Se calcula como $2\frac{Precision \cdot Recall}{Precision+Recall}$. El uso de esta medida se ha extendido sobretodo en estudios referentes a PLN como es el caso [21].
- El *soporte* o *support* es el número de ocurrencias de cada clase dentro del vector de anotaciones reales.

Por otro lado, incluye medidas conjuntas o medidas multiclase para un modelo:

- *Macro average* que calcula el promedio de la media no ponderada por etiqueta. Esto no tiene en cuenta el desequilibrio de la etiqueta. Da más peso a las etiquetas más comunes.

- *Micro average* que calcula el promedio de los verdaderos positivos, falsos negativos y falsos positivos totales, es decir, calcula las condiciones para cada clase y realiza el promedio. Da el mismo peso a todas las clases y por tanto, tiene más en cuenta las clases menos frecuentes y en general más difíciles. Puesto que hay un desequilibrio en el número de clases, esta será de mayor utilidad que la anterior.
- *Weighted average* es el promedio de la media ponderada del *support*.
- *Sample average* es el promedio de la muestra (solo es significativo para la clasificación de múltiples etiquetas donde esto difiere del *accuracy*).

5.2. Resultados

Como se ha mencionado, las primeras pruebas se hicieron con el CRF para el cuerpo de los documentos pues había que comprobar qué etiquetado era más conveniente usar. Posteriormente se han obtenido los resultados para la detección de PHI por estructura del encabezado seguido de los obtenidos en la detección de PHI con expresiones regulares, diccionario y PHI de confianza para el pie, y finalmente una evaluación conjunta de las tres partes.

5.2.1. Resultados Cuerpo - CRF

Para el entrenamiento del modelo, del total de documentos, se han creado dos subconjuntos, el 75 % para entrenar el clasificador CRF (usando solo aquellas frases que contenían, al menos, un PHI) y el 25 % restante para predecir los resultados del modelo y testear su eficacia. Es conveniente indicar que en primeras pruebas, se dividió en un $80-20$. Sin embargo, a la hora de evaluar el test resultaba que para algunas clases se testeaban con pocos datos. Por tanto, al comprobar que la eficacia del modelo apenas sufría una mínima variación, se optó por la división $75-25$.

En las tablas Tabla 5.1, Tabla 5.2, Tabla 5.3 y Tabla 5.4 se observan los resultados obtenidos al escoger etiquetado N-I sin diferenciación de clases, CDI sin diferenciación de clases, N-I con diferenciación de clases y CDI con diferenciación de clases respectivamente.

De aquí se pueden extraer diversas conclusiones:

La primera es que la opción del etiquetado CDI diferenciando por clases es ligeramente mejor que el resto. Las medidas generales rondan el 0.98 mientras que en el resto se mantienen sobre el 0.97, con la salvedad de la *macro average*. Recuérdese que esta hace la media de la precisión de todas las etiquetas. Es de esperar por tanto, que las clases donde no se han obtenido aciertos hagan bajar esta media.

Tabla 5.1: N-I sin diferenciación de clases

	precision	recall	f1-score	support
I	0.98	0.99	0.99	5292
N	0.90	0.82	0.86	613
accuracy			0.97	5905
macro avg	0.94	0.90	0.92	5905
weighted avg	0.97	0.97	0.97	5905

Tabla 5.2: CDI sin diferenciación de clases

	precision	recall	f1-score	support
I	0.98	0.99	0.99	4990
C-N	0.91	0.84	0.87	340
D-N	0.87	0.80	0.83	265
micro avg	0.97	0.97	0.97	5595
macro avg	0.92	0.88	0.90	5595
weighted avg	0.97	0.97	0.97	5595
samples avg	0.97	0.97	0.97	5595

Tabla 5.3: N-I con diferenciación de clases

	precision	recall	f1-score	support
N-CALLE	0.00	0.00	0.00	27
N-EDADS	0.90	0.90	0.90	208
N-EMAIL	0.50	1.00	0.67	1
N-FAMS	0.78	0.66	0.72	65
N-FECH	1.00	0.99	0.99	70
N-HOSP	0.79	0.94	0.86	36
I	0.98	0.99	0.99	5102
N-IDSUJ	0.00	0.00	0.00	9
N-INST	0.69	0.69	0.69	26
N-NOMP	0.43	1.00	0.60	3
N-NOMS	0.00	0.00	0.00	4
N-PAIS	1.00	0.67	0.80	15
N-PROF	1.00	0.43	0.60	14
N-SEXOS	0.96	0.96	0.96	84
N-TER	0.73	0.40	0.52	20
micro avg	0.97	0.97	0.97	5684
macro avg	0.65	0.64	0.62	5684
weighted avg	0.97	0.97	0.97	5684
samples avg	0.97	0.97	0.97	5684

Tabla 5.4: CDI con diferenciación de clases

	precision	recall	f1-score	support
C-EDADS	0.93	0.97	0.95	100
D-EDADS	0.91	0.95	0.93	111
C-EMAIL	0.00	0.00	0.00	1
C-FAMS	0.80	0.77	0.79	43
D-FAMS	0.72	0.36	0.48	36
C-FECH	0.97	0.95	0.96	39
D-FECH	1.00	0.96	0.98	46
C-HOSP	1.00	0.75	0.86	4
D-HOSP	1.00	0.75	0.86	12
I	0.98	0.99	0.99	4852
C-IDSUJ	0.50	0.17	0.25	6
D-IDSUJ	0.00	0.00	0.00	12
C-INST	0.71	0.62	0.67	8
D-INST	0.59	0.71	0.65	14
C-NOMP	1.00	1.00	1.00	1
D-NOMP	0.50	1.00	0.67	3
C-NOMS	0.00	0.00	0.00	2
C-NUMT	0.00	0.00	0.00	2
C-PAIS	0.80	0.67	0.73	6
D-PAIS	0.00	0.00	0.00	1
C-PROF	0.00	0.00	0.00	6
D-PROF	0.00	0.00	0.00	9
C-SEXOS	1.00	0.98	0.99	88
C-TER	0.50	0.60	0.55	5
D-TER	1.00	0.33	0.50	3
micro avg	0.98	0.98	0.98	5410
macro avg	0.60	0.54	0.55	5410
weighted avg	0.97	0.98	0.97	5410
samples avg	0.98	0.98	0.98	5410

La segunda y la más importante es que el uso del CRF da, a priori, buenos resultados. Sin embargo, estos resultados vienen influidos por la categoría mayoritaria y la que menos interesa: las etiquetas I ("Irrelevante"). Por tanto, aumenta el support e infla las medidas conjuntas.

Entonces, se han recalculado las medidas de precisión sin tener en cuenta estas etiquetas y, aunque el modelo no funciona mal, se puede ver en la Tabla 5.5 que la eficacia baja.

De manera individual se puede observar que las categorías que peor han funcionado han sido aquellas con pocos datos anotados como "C-NOMP" (Nombre Personal Sanitario) donde se ha obtenido un *F1-score* de 0.33 si es inicio de PHI y 0.40 si no está al comienzo. Sin embargo, si nos fijamos en el número de clases "NOMP" que hay anotados dentro del cuerpo en el total de documentos (ver Anexo A.2) son 45. Si el support nos indican que solo ha testeado con 10 quiere decir que el modelo ha entrenado con 35 variables. Teniendo esto en cuenta, podría decirse incluso que el aprendizaje iba bien encaminado.

Por otro lado, es interesante resaltar que ha funcionado bastante bien para "C-FECH" y "D-FECH" (*F1-score* 0.91 y 0.90 respectivamente) que corresponde a fechas, se intuye que debido a las características morfológicas que particularizan este tipo de datos.

5.2.2. Resultados Cabecera - Detección PHI según estructura

Cuando se realizó el estudio previo para plantear un método de detección para los PHI de la cabecera, una vez escogido dicho método, prácticamente se indicaban los resultados que se iban a obtener al realizar finalmente el método. Sin embargo, para seguir las mismas medidas que para el cuerpo, en la Tabla 5.6 se muestran los resultados obtenidos. En este caso se ha testeado con todos los datos.

En general, la metodología escogida funciona bien en el 99 % de los casos. Si bien, el porcentaje restante como se ha mencionado ya, se debe en general a erratas en las anotaciones (como indicaciones de una categoría que corresponde a otra) o un formato inesperado (saltos de línea o tabulación que no se esperaba).

Además, esta tabla deja ver que la aparición de datos PHI es mayor que los datos Irrelevantes, al contrario que pasaba en el cuerpo.

5.2.3. Resultados Pie - Detección PHI con varios recursos

Para obtener los resultados del pie, como se requieren PHIs de confianza obtenidos del cuerpo y del encabezado, se ha tenido que usar el modelo de CRF entrenado. Por tanto, los resultados se han realizado con los datos test preparados para el modelo CRF.

Recuérse que cada vector con los datos de cada parte tiene asociado el documento del

Tabla 5.5: CDI sin incluir clase I

	precision	recall	f1-score	support
C-CALLE	0.00	0.00	0.00	2
D-CALLE	0.00	0.00	0.00	8
C-EDADS	0.89	0.90	0.89	107
D-EDADS	0.85	0.90	0.88	110
C-EMAIL	0.33	1.00	0.50	1
C-FAMS	0.81	0.70	0.75	56
D-FAMS	0.55	0.24	0.33	25
C-FECH	0.97	0.86	0.91	36
D-FECH	0.95	0.86	0.90	44
C-HOSP	1.00	1.00	1.00	6
D-HOSP	0.84	0.89	0.86	18
C-IDSUJ	0.75	0.43	0.55	7
D-IDSUJ	0.00	0.00	0.00	4
C-IDTIT	0.00	0.00	0.00	1
D-IDTIT	0.00	0.00	0.00	2
C-INST	0.75	0.55	0.63	11
D-INST	0.83	0.59	0.69	17
C-NOMP	0.50	0.25	0.33	4
D-NOMP	0.33	0.50	0.40	6
C-NOMS	0.00	0.00	0.00	1
C-OTROS	0.00	0.00	0.00	3
D-OTROS	0.00	0.00	0.00	1
C-PAIS	1.00	0.56	0.71	9
C-PROF	0.00	0.00	0.00	4
D-PROF	0.00	0.00	0.00	7
C-SEXOS	0.96	0.96	0.96	82
D-SEXOS	0.00	0.00	0.00	1
C-TER	0.40	0.15	0.22	13
D-TER	0.00	0.00	0.00	2
micro avg	0.85	0.75	0.80	588
macro avg	0.44	0.39	0.40	588
weighted avg	0.80	0.75	0.77	588
samples avg	0.08	0.08	0.08	588

Tabla 5.6: Resultados Cabecera

	precision	recall	f1-score	support
C-CALLE	0.99	0.99	0.99	490
D-CALLE	1.00	1.00	1.00	2062
C-EDADS	0.98	1.00	0.99	480
D-EDADS	0.98	1.00	0.99	454
C-FECH	1.00	1.00	1.00	982
D-FECH	1.00	0.50	0.67	6
I	0.93	0.92	0.93	217
C-IDASEG	0.99	0.99	0.99	382
D-IDASEG	1.00	0.99	1.00	741
C-IDCONT	0.96	1.00	0.98	77
C-IDSUJ	0.99	0.99	0.99	487
D-IDSUJ	0.29	1.00	0.44	4
C-IDTIT	0.99	1.00	1.00	461
D-IDTIT	1.00	1.00	1.00	894
C-NOMP	0.99	1.00	0.99	486
D-NOMP	1.00	1.00	1.00	1077
C-NOMS	1.00	0.99	0.99	980
D-NOMS	1.00	1.00	1.00	539
C-PAIS	1.00	0.99	0.99	494
D-PAIS	1.00	0.14	0.25	7
C-SEXOS	0.99	1.00	0.99	484
C-TER	1.00	1.00	1.00	975
D-TER	1.00	1.00	1.00	91
micro avg	0.99	0.99	0.99	12870
macro avg	0.96	0.93	0.92	12870
weighted avg	0.99	0.99	0.99	12870
samples avg	0.99	0.99	0.99	12870

Tabla 5.7: Resultados Pie

	precision	recall	f1-score	support
C-CALLE	1.00	0.76	0.86	41
D-CALLE	0.85	0.67	0.75	200
C-EMAIL	1.00	1.00	1.00	42
C-HOSP	1.00	0.50	0.67	22
D-HOSP	0.91	0.42	0.57	69
C-INST	0.00	0.00	0.00	7
D-INST	0.00	0.00	0.00	31
C-NOMP	1.00	0.98	0.99	47
D-NOMP	0.99	0.97	0.98	106
C-NUMF	0.33	1.00	0.50	1
D-NUMF	0.00	0.00	0.00	2
C-NUMT	0.00	0.00	0.00	2
D-NUMT	0.50	1.00	0.67	4
C-PAIS	0.87	0.95	0.91	21
D-PAIS	1.00	1.00	1.00	1
C-TER	0.77	0.87	0.82	95
micro avg	0.89	0.70	0.78	720
macro avg	0.60	0.60	0.57	720
weighted avg	0.81	0.70	0.74	720
samples avg	0.39	0.39	0.39	720

que se trata. Así, se puede recuperar del vector de test para el CRF los PHI para cada documento. Recuérdese además que esto es necesario pues los PHI de confianza se usan únicamente en los mismos documentos de donde se han obtenido. Dicho de otro modo, si del *documento 1* se ha obtenido un PHI de confianza, este no será válido para la detección de PHI en el *documento 2*.

Los resultados obtenidos pueden verse en la Tabla 5.7.

Se pueden obtener las siguientes conclusiones:

1. El uso de PHI de confianza centrado en detectar esencialmente las categorías de "NOMS" (Nombre del sujeto) ha funcionado bastante bien, obteniendo un *F1-score* alrededor de 0.985.

2. Los resultados obtenidos de las clases que se buscaba mediante expresiones regulares no han resultado del todo efectivas. Para la calle, por ejemplo, obtiene mejor resultado si es inicio pues es fácil detectarlo (si inicia con c/, calle, Av,...) pero más difícil saber donde termina. Normalmente debiera ser con un código postal (que corresponde con la categoría "TER") pero no se puede generalizar esta regla. Por otro lado, los números de teléfono y de fax tampoco han funcionado bien, puede ser por un tipo de

formato no tenido en cuenta o caracteres inesperados. Sin embargo, son pocos los datos de test para llegar a conclusiones claras y se desconoce el motivo exacto.

3. Por último el uso del diccionario para la detección de la clase "TER" (territorio, incluyendo entre otros municipios, ciudades y código postal) no ha sido del todo óptimo, obteniendo un *F1-score* de 0.82. Se recuerda que el diccionario incluía los municipios de España. Esto puede ser debido a que el municipio en el texto no se encuentre igual forma escrito en el diccionario, que se indique la localidad en lugar del municipio o que directamente pertenezca a otro país.

A pesar de todo ello, se debe tener en cuenta que la estructura del pie era la menos definida en comparación con el encabezado y el cuerpo.

5.2.4. Resultado total del sistema

Finalmente, se presenta el resultado que se ha obtenido en la evaluación del sistema uniendo los métodos realizados para la detección de la cabecera, el cuerpo y el pie.

Para ello, como se ha mencionado, se han usado el 75 % de los documentos para entrenar el modelo y el 25 % para testear. De ese 25 % (alrededor de unos 117 documentos) se ha obtenido el cuerpo para predecir las etiquetas introduciéndolos en el modelo de CRF entrenado, se ha cogido la cabecera para detectar los PHI según su estructura y, por último, se han obtenido los PHI de confianza e introducido el pie en el sistema de detección restante. Finalmente, se han unido los resultados de cada uno de los métodos y se ha usado la función de *CRFsuite* para calcular las medidas de evaluación. El resultado se puede observar en la Tabla 5.8.

Las conclusiones que se han obtenido para cada una de las partes evaluadas se pueden extender al sistema completo. Por un lado, se puede ver que las categorías con pocas anotaciones tienden a no ser halladas por el sistema. Por otro lado, el buen funcionamiento de la cabecera eleva las puntuaciones finales. En la tabla 5.6 se veían los resultados de la cabecera introduciendo todos los documentos. En cambio, si se hubieran dividido los datos en train y test habríamos obtenido un soporte igual a 1382 que es más de la mitad del soporte total que se ha introducido para testear (2407). Es decir, más del 50 % de los PHI se encuentran en el encabezado de los informes y, como el método propuesto para la detección de esta parte funciona muy bien, hace que tenga bastante peso en los resultados globales del sistema.

Se debe mencionar que la media total del sistema dada por *sample average*, se está calculando como número de aciertos totales entre total de etiquetas introducidas (*support* total). Sin embargo, de alguna forma, está teniendo en cuenta el total de etiquetas incluyendo la etiqueta "I" que correspondía a todo lo que no era PHI pero no cuenta los aciertos obtenidos en esta etiqueta. Esto mismo se podía ver en la Tabla 5.5 donde se eliminó esta etiqueta para que no influyera en las demás medidas. Por ello, el *sample avg* tiene un

Tabla 5.8: Resultados Globales del Sistema

	precision	recall	f1-score	support
C-CALLE	0.99	0.88	0.93	92
D-CALLE	0.93	0.84	0.88	445
C-EDADS	0.98	1.00	0.99	105
D-EDADS	0.98	1.00	0.99	99
C-EMAIL	0.98	0.98	0.98	43
C-FAMS	0.98	0.94	0.96	67
D-FAMS	1.00	0.81	0.90	32
C-FECH	1.00	0.99	1.00	116
D-FECH	1.00	1.00	1.00	16
C-HOSP	1.00	0.72	0.84	25
D-HOSP	0.93	0.57	0.70	88
C-IDASEG	0.98	0.95	0.97	44
D-IDASEG	1.00	0.98	0.99	84
C-IDCONT	1.00	1.00	1.00	7
C-IDSUJ	0.96	0.98	0.97	55
D-IDSUJ	0.42	1.00	0.59	5
C-IDTIT	1.00	1.00	1.00	48
D-IDTIT	1.00	1.00	1.00	90
C-INST	1.00	0.71	0.83	7
D-INST	1.00	0.90	0.95	10
C-NOMP	0.99	0.98	0.99	101
D-NOMP	0.99	1.00	0.99	230
C-NOMS	1.00	0.95	0.97	100
D-NOMS	1.00	1.00	1.00	56
C-NUMT	0.00	0.00	0.00	3
D-NUMT	0.56	1.00	0.71	5
C-OTROS	0.00	0.00	0.00	1
C-PAIS	0.98	0.98	0.98	81
C-PROF	1.00	1.00	1.00	4
D-PROF	1.00	1.00	1.00	4
C-SEXOS	1.00	0.99	1.00	101
C-TER	0.91	0.89	0.90	203
D-TER	1.00	0.33	0.49	40
micro avg	0.96	0.91	0.94	2407
macro avg	0.90	0.86	0.86	2407
weighted avg	0.97	0.91	0.93	2407
samples avg	0.35	0.35	0.35	2407

Tabla 5.9: Resultados de sistemas que participaron en la tarea MEDDOCAN

Organization	precision	recall	F1-score
Bosch Center for Artificial Intelligence (Germany)	0.97508	0.97474	0.97491
Universitat Rovira i Virgili, CRISES group (Spain)	0.97529	0.96202	0.96861
Vicomtech (Spain)	0.97187	0.96414	0.96799
FSL - Unaffiliated (Spain)	0.96315	0.96502	0.96409

valor tan bajo (0.35). Sin embargo, si hubiésemos tenido en cuenta los aciertos de la clase "I", se habría obtenido un *F1-score* de 0.96 para la clase "I" y un *sample avg* de 0.95 total.

Por último, en la Tabla 5.9 se muestran los resultados de los mejores sistemas que se presentaron en la tarea MEDDOCAN [23]. Las evaluaciones se realizaron con un sistema de evaluación propio de la tarea. El que obtuvo mejor puntuación fue el presentado por la compañía *Bosch Center for Artificial Intelligence* [25] de Alemania obteniendo un *F1-score* total de 0.97491. El *F1-score* que utilizan es el que corresponde con el *micro-average F1-score*. Si se comprara con el obtenido en el sistema de este trabajo (0.94) podríamos considerar que la propuesta no ha estado mal encaminada. Sin embargo, hay que tener en cuenta que el conjunto de test usado en la evaluación no es el mismo. La tarea utiliza un conjunto para test con datos no etiquetados para los participantes mientras que en este trabajo se han usado una parte del conjunto de train.

Capítulo 6

Conclusiones

El objetivo de este trabajo era estudiar sistemas y métodos de anonimización usados para textos en inglés y así proponer un modelo que pudiera funcionar para historias clínicas en español.

Una de las cosas aprendidas durante la realización de este trabajo ha sido que existen numerosos métodos que se podrían usar en un trabajo de anonimización junto con numerosas casuísticas posibles en cuanto a tipos de texto, estructura, PHI, etc. Además, se ha advertido la complejidad que tiene, por un lado, el tratamiento de texto plano y por otro, la extracción de información útil del mismo.

Tras realizar el estudio previo y el preprocesamiento de los informes, se ha conseguido realizar un sistema híbrido para la detección de información de salud protegida, el cual ha usado un modelo de aprendizaje con Campos Aleatorios Condicionales, expresiones regulares, un diccionario, PHIs de confianza y la propia estructura de los informes.

6.1. Ventajas y desventajas

Las ventajas y desventajas que se han concluido en la realización del trabajo para cada método implementado varían para cada uno de ellos:

- Se ha comprobado que el uso de un CRF como modelo de aprendizaje es bastante eficaz en el tema de anonimización de datos. En algunos casos incluso se ha visto que no ha necesitado un gran número de anotaciones de una clase para aprender aceptablemente. Por contra, no siempre esto es así. Como cualquier modelo de aprendizaje es necesario un elevado número de anotaciones y aún así, no se asegura que pueda aprender bien.
- Como es de esperar, las expresiones regulares han sido muy útiles cuando se ha tratado de una clase bien definida, como números de teléfono, códigos postales, inicio de direcciones o e-amil. Sin embargo, cuanta más variabilidad morfológica tenga un

tipo de dato más costoso es formar una expresión regular adecuada para encontrar este tipo de dato.

- Al uso de diccionarios le sucede algo similar a las expresiones regulares, son muy útiles para algunas categorías pero inviables para otras. Son útiles sobretodo cuando se puede acceder a una base de datos existente con información de determinado tipo. Por ejemplo, base de datos de municipios, países o incluso nombres propios y hospitales. La problemática se encuentra en que estas bases de datos pueden ser de gran tamaño, lo que crearía un método poco eficiente.
- Los PHI de confianza, en este modelo, han sido muy útiles para el propósito para el que se han usado, este es, ayudar a detectar algunas categorías de PHI que habrían sido costosas de encontrar si se hubiera usado otro método. Sin embargo, como se explicó cuando se definieron estos PHI de confianza, se deben tratar con cuidado para evitar clasificar de manera incorrecta un PHI con otra categoría distinta a la que le corresponde. De ahí que sobretodo sea útil su uso dentro del documento que se está analizando.

Por tanto, la ventaja que tiene el sistema propuesto es que ha tenido en cuenta los puntos fuertes de cada método con la intención de usarlos siempre y cundo tuvieran sentido. Aunque no siempre se pueda saber, es importante intuir cuándo la implementación tiene sentido o no. Por ejemplo, para los informes que se han tratado en este trabajo, no se ha requerido ningún algoritmo complejo para la detección de PHI en la cabecera más allá que el uso de su propia estructura.

Por contra, como en el cuerpo solo se ha usado un modelo de aprendizaje, se han arrastrado las desventajas del mismo. La forma de proceder hubiera sido fijarse en las categorías que no haya entrenado bien y pensar alguna otra manera de tratarlas y detectarlas. Sin embargo, este estudio hubiera requerido un tiempo adicional que habría impedido desarrollar el resto del sistema.

Se concluye por tanto que el resultado del sistema propuesto, aunque mejorable, ha sido óptimo y ha conseguido presentar distintos métodos útiles en la tarea de anonimización de informes médicos.

6.2. Trabajo futuro

Aún habiendo conseguido un sistema de anonimización razonable, todavía quedarían posibles mejoras o implementaciones que podrían tratarse:

1. Quizás el punto más necesario a realizar como trabajo futuro sería la creación de un módulo que sea capaz de, al introducir un nuevo informe, usar el sistema para sustituir el PHI detectado por un valor que no relacione al mismo (en caso de que se

quiera anonimizar completamente) o por un tipo de identificador (en caso de que solo quiera disociarse la información). En este trabajo, se ha ido arrastrando el texto instanciado en la clase que ofrece *spaCy* para su procesamiento. La ventaja de esta clase es que siempre guarda la posición del token que se está tratando. Por tanto, haber realizado este módulo no habría sido complejo pero hubiera llevado bastante tiempo por lo que se ha considerado realizar otras implementaciones en su defecto.

2. Uno de los peores resultados que se podría mejorar es el procesamiento del pie. Posibles soluciones para ello son la mejora de expresiones regulares en la detección de direcciones y el uso de más tipos de diccionarios como por ejemplo diccionarios con hospitales e instituciones. Incluso, sería interesante comprobar cómo funcionaría si se usa otro modelo CRF.

3. En la parte del cuerpo quedaría estudiar las clases que no funcionan bien en el modelo de aprendizaje y tratarlas de forma especial. Complementariamente, podrían usarse algunos de los otros métodos que se han visto para mejorar o corregir el resultado que se obtiene del CRF.

4. Añadir o eliminar características a las palabras antes del entrenamiento del CRF para comprobar cómo funciona podría resultar un estudio interesante.

5. Últimamente el uso de redes neuronales se está extendiendo para clasificación o traducción de texto [24]. Una propuesta interesante sería probar su funcionamiento en la anonimización de textos en español. Si bien es cierto, requiere de una gran cantidad de datos para su aprendizaje y obtener un alto número de informes médicos para ello es difícil.

Bibliografía

[1] Cristina Gómez Piqueras. 2009. Artículo DisociaciónAnonimización de los Datos de Salud.

[2] Real Academia Española. Diccionario en línea de la real academia española. 2014. `https://dle.rae.es/?id=2jjMiRi`.

[3] Reino de España. Ley 14/2007, de 3 de julio, de investigación biomédica. Colección Textos legales. Ministerio de Sanidad y Consumo, 2007. ISBN 9788476706886. `http://books.google.com.uy/books?id=eP4xQwAACAAJ`.

[4] IberLEF 2019 Workshop SEPLN 2019 (Bilbao), 24th Sep 2019. MEDDOCAN: Medical Document Anonymization task. `http://temu.bsc.es/meddocan/`.

[5] Fadi Hassan, Josep Domingo-Ferrer, Jordi SoriaComas. 2017. Anonimización de datos no estructurados a través del reconocimiento de entidades nominadas.

[6] Hui Yang, Jonathan M. Garibaldi. 2015. Automatic Detection of Protected Health Information from Clinic Narratives. *Journal of Biomedical Informatics, 58, Supplement, S30S38.*

[7] Zengjian Liua, Yangxin Chenb, Buzhou Tanga, Xiaolong Wanga, Qingcai Chena, Haodi Lia, Jingfeng Wangb, Qiwen Dengc and Suisong Zhuc. 2015. Automatic De-identification of Electronic Medical Records using Tokenlevel and Characterlevel Conditional Random Fields. *Journal of Biomedical Informatics, 58, Supplement, S47–S52.*

[8] Steven Bird, Ewan Klein and Edward Loper. 2009. Natural Language Processing with Python. `http://www.nltk.org/book_1ed/`.

[9] spaCy 101: Everything you need to know. `https://spacy.io/usage/spacy-101`.

[10] CRFsuite - Documentation. `http://www.chokkan.org/software/crfsuite/manual.html#idp8849114176`.

[11] Sweeney, Latanya. 1996. Replacing Personally-Identifying Information in Medical Records, the Scrub System. *Journal of the American Medical Informatics Association, 333-337.*

[12] Bernan, Jules J. 2003. Concept-Match medical data scrubbing. How pathology text can be used in research. *Arch. Pathol. Lab. Med., 127(6), 680 - 686.*

[13] Wellner, Ben, Huyck, Matt, Mardis, Scott, Aberdeen, John, Morgan, Alex, Pershkin, Leonid, Yeh, Alex, Hitzeman, Janet, Hirschman, Lynette. 2007. Rapidly Retargetable Approaches to Deidentification in Medical Records. *Journal of the American Medical Informatics Association, 14(5), 564 - 573.*

[14] Szarvas, Goyorgy, Farkasb, Richárd, Busa-Feketeb and Róbert. 2007. State-of-the-Art Anonymization of Medical Records Using an Iterative Machine Learning Framework. *Journal of the American Medical Informatics Association, vol. 14 (574 - 580).*

[15] Lafferty, John D., McCallum, Andrew, Pereira and Fernando C. N. 2001 (June). Conditional random fields: probabilistic models for segmenting and sequence data. *DANYLUK, AP. (ed), Proceedings of International Conference on Machine Learning (282 - 289).*

[16] Hara, Kazuo. 2006. Applying a SVM based chunker and a text classifier to the Deid Challenge. *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data (10 - 11).*

[17] Garner, James, Xiong Li. 2008. HIDE: An Integrated System for Health Information DEidentification. *Computer-Based Medical Systems (254 - 259).* IEEE Computer Society.

[18] Charles Sutton and Andrew McCallum. 2012. An Introduction to Conditional Random Fields.

[19] Francisco Javier Díez. 2014. UNED. Introducción a los Modelos Gráficos Probabilistas.

[20] María Ángela Mendoza Pérez. 2010. Universidad de Granada. Tesis de Reconocimiento de Acciones Humanas Basado en Modelos Probabilísticos de Espacio de Estado.

[21] Leon Derczynski. 2016. Complementarity, F-score, and NLP Evaluation. `https://www.aclweb.org/anthology/L16-1040`.

[22] Agencia Tributaria. Gobierno de España. Tabla de municipios. `https://www.agenciatributaria.es/AEAT.internet/Inicio/Ayuda/Tablas_auxiliares_de_domicilios__provincias__municipios____/Tabla_de_Municipios/Tabla_de_Municipios.shtml`

[23] Montserrat Marimon, Aitor Gonzalez-Agirre1, Ander Intxaurrondo, Heidy Rodríguez, Jose Antonio Lopez Martin, Marta Villegas, and Martin Krallinger. 2019. Automatic De-Identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results. Centro Nacional de Investigaciones Oncol´ogicas (CNIO).

[24] Periódico *El Español*. *Google Translate mejora estrenando red neuronal*. `https://elandroidelibre.elespanol.com/2016/11/google-translate-red-neuronal.html`

[25] Compañía *Bosch Center for Artificial Intelligence*. `https://www.bosch-ai.com/`.

Anexos

Apéndice A

Anexo I: Información PHI

A.1. Tipo PHI diferenciados

```
PHI =
    'NOMBRE_SUJETO_ASISTENCIA': "NOMS", (Nombre del paciente)
    'EDAD_SUJETO_ASISTENCIA': "EDADS", (Edad del paciente)
    'SEXO_SUJETO_ASISTENCIA': "SEXOS", (Sexo del paciente H o M)
    'FAMILIARES_SUJETO_ASISTENCIA': "FAMS", (Parentescos)
    'NOMBRE_PERSONAL_SANITARIO': "NOMP", (Nombre personal sanitario
            -Médico, Enfermero... -)
    'FECHAS': "FECH", (Fecha)
    'PROFESION': "PROF", (Profesión)
    'CENTRO_SALUD': "CENT", (Centro de Salud sin incluir Hospitales)
    'HOSPITAL': "HOSP", (Hospital)
    'INSTITUCION': "INST", (Instituciones -Universidad, Clínicas,...)
    'ID_TITULACION_PERSONAL_SANITARIO': "IDTIT", (Identificador
            numérico del personal sanitario)
    'ID_EMPLEO_PERSONAL_SANITARIO': "IDEMP", (Identificador numérico
            del empleo del personal sanitario)
    'IDENTIF_VEHICULOS_NRSERIE_PLACAS': "IDVEH", (Identificador
            numérico de vehículos)
    'IDENTIF_DISPOSITIVOS_NRSERIE': "IDDISP", (Número de serie de
            dispositivos)
    'CALLE': "CALLE", (Nombre de direcciones postales)
    'TERRITORIO': "TER", (Municipios, Localidades y Código Postal)
    'PAIS': "PAIS", (Países)
    'NUMERO_TELEFONO': "NUMT", (Número de teléfono)
    'NUMERO_FAX': "NUMF", (Número de fax)
    'CORREO_ELECTRONICO': "EMAIL", (Dirección de correo electrónico)
    'ID_SUJETO_ASISTENCIA': "IDSUJ", (Identificador del paciente)
```

'ID_CONTACTO_ASISTENCIAL': "IDCONT", (Se desconoce)
'NUMERO_BENEF_PLAN_SALUD': "NUMB", (Se desconoce)
'ID_ASEGURAMIENTO': "IDASEG", (Se desconoce)
'URL_WEB': "URL", (Direción de página Web)
'DIREC_PROT_INTERNET': "DIRECT", (Se desconoce)
'OTRO_NUMERO_IDENTIF': "IDOTRO", (Identificador distinto a los ya considerados)
'OTROS_SUJETO_ASISTENCIA': "OTROS" (Se desconoce)

A.2. Total PHI en el cuerpo de los documentos

'C–SEXOS': 406,
'C–EDADS': 509,
'D–EDADS': 534,
'C–FAMS': 236,
'D–FAMS': 134,
'C–FECH': 193,
'D–FECH': 221,
'C–NOMP': 15,
'D–NOMP': 30,
'C–IDSUJ': 41,
'D–IDSUJ': 27,
'C–INST': 50,
'D–INST': 77,
'C–TER': 59,
'C–PAIS': 43,
'C–PROF': 24,
'C–NOMS': 11,
'D–SEXOS': 2,
'D–TER': 10,
'D–PROF': 28,
'C–OTROS': 9,
'D–OTROS': 23,
'D–PAIS': 4,
'C–HOSP': 26,
'D–HOSP': 81,
'C–IDTIT': 4,
'D–IDTIT': 7,
'C–CALLE': 9,
'D–CALLE': 45,
'C–NUMT': 5,

'D–NUMT': 8,
'C–EMAIL': 7,
'D–EMAIL': 1,
'D–NOMS': 3,
'C–IDASEG': 1,
'D–IDASEG': 2