Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Detección de Negaciones en Informes Radiológicos escritos en Español

Tesis presentada para optar al título de
Licenciada en Ciencias de la Computación

Vanesa Stricker

Director: Viviana Cotik

Buenos Aires, 2016

DETECCIÓN DE NEGACIONES EN INFORMES RADIOLÓGICOS ESCRITOS EN ESPAÑOL

En el procesamiento de textos biomédicos se ha reconocido la importancia de identificar negaciones y su alcance, pues estas determinan si se informa de la presencia o ausencia de una condición clínica o *finding*.

Chapman et al. [5] desarrollaron el algoritmo NegEx basado en expresiones regulares para determinar si una condición clínica o *finding* está negada en textos médicos escritos en inglés. NegEx fue adaptado a otros idiomas. También se han desarrollado métodos basados en aprendizaje automático y técnicas que usan información sintáctica y semántica.

En este trabajo se presentan tres enfoques para detectar en informes radiológicos escritos en español si un *finding* está negado: se realiza una adaptación de NegEx para español con dos versiones (una adecuada para el ámbito radiológico, la otra adecuada para otros dominios) y se desarrollan dos métodos sintácticos. Uno utiliza el Part-of-Speech tag de las palabras para detectar las negaciones y se construyen reglas que determinen si el *finding* está alcanzado por la negación o no en base a sus posiciones dentro de una oración. El otro consiste en aplicar *shallow parsing* a las oraciones de los informes y utilizar la información obtenida para decidir si las oraciones mencionan *findings* negados o no. La adaptación de NegEx, el enfoque más simple, obtiene los mejores resultados.

Palabras claves: detección de negaciones, informes radiológicos, NegEx, español, NLP, PoS-tagging, shallow parsing

NEGATION DETECTION IN RADIOLOGY REPORTS WRITTEN IN SPANISH

The identification of negations and their scope has been recognized as important in the processing of biomedical texts, since these determine whether the presence or absence of a clinical condition or finding is reported.

Chapman et al. [5] have developed NegEx, an algorithm based on regular expressions in order to determine when a clinical condition or finding mentioned in clinical texts written in English is negated. NegEx has been adapted to other languages. Also machine learning techniques and methods based on syntactic and semantic information have been developed.

In this work three approaches are presented in order to detect negations of finding in radiological reports written in Spanish: an adaptation of NegEx to Spanish is performed with two versions (one adequated for the radiological field, the other is suitable for other domains) and two syntactic methods are developed. One of them uses the Part-of-Speech tag of the words to detect negations, and rules are developed in order to determine whether the finding is under the scope of the negation, based on their positions within a sentence. The other consists in applying shallow parsing to the report sentences and use the information obtained to decide whether the sentences mention negated findings. NegEx adaptation, the simplest approach, obtains the best results.

Keywords: negation detection, radiology reports, NegEx, spanish, NLP, PoS-tagging, shallow parsing

A mi familia, que me sostiene siempre.

A mi novio, quien me alienta, anima y acompaña.

A Dios, quien se merece todo y a quien le debo mi vida.

Índice general

Índice de cuadros

Índice de figuras

1. INTRODUCCIÓN

En la actualidad existe gran cantidad de información digitalizada de diversos dominios y características. Por ejemplo, existen documentos del dominio legal, económico, médico, así como manuales del dominio educativo enteramente digitalizados. Algunos de ellos tienen la característica de tener lenguaje formal como las leyes y artículos científicos, y otros tienen lenguaje informal como los blogs, los tweets y las publicaciones en la biografía de facebook. El lenguaje utilizado en unos y otros varía, y son de distintas longitudes: un tweet puede contener hasta 140 caracteres mientras que un texto literario, una novela o un manual educativo pueden llegar a abarcar cientos de páginas cada uno. Otra forma de caracterizar a los textos digitales es según la existencia o carencia de estructura, pudiendo ser estructurados, semi-estructurados o no estructurados.

En las últimas décadas ha crecido la cantidad de información en gran manera. De aquí se desprende la importancia de poder extraer la información de los textos digitales y analizarla. Es por esto que las disciplinas de extracción de información y *data mining* son áreas de estudio de mucho interés en la actualidad.

En el ámbito médico y del cuidado de la salud existen numerosos documentos médicos de diversos tipos (informes radiológicos, informes de ecocardiogramas, historias clínicas, entre otros). Estos documentos proveen información valiosa para la detección y caracterización de enfermedades. Ésta podría emplearse para estudios epidemiológicos, descriptivos, el descubrimiento de nuevas relaciones entre signos, síntomas y patologías, y podría impactar en los diagnósticos y tratamientos. Gran parte de estos documentos que permitirían mejorar el cuidado del paciente o la investigación médica, se encuentran en forma narrativa (en formato no estructurado) por lo que la información es de difícil acceso tanto para las personas como para los sistemas automatizados. Por lo tanto, debe estructurarse para ser de utilidad a los fines clínicos, de enseñanza, o de investigación.

Por este motivo el área de Procesamiento del Lenguaje Natural en Biomedicina (BioNLP, *Biomedical Natural Language Processing*) ha estado muy activa en las últimas décadas. Los sistemas de extracción y recuperación de información, que permiten indexar[1] y extraer automáticamente las condiciones clínicas de los documentos médicos almacenados en registros médicos electrónicos con el fin de facilitar la búsqueda de términos relevantes, han estado presentes en las últimas décadas [3, 17, 20, 37]. Además existen diversos inventarios y ontologías[2] digitales que sirven para identificar los términos de interés de los informes médicos. Por ejemplo, ICD (*International Classification of Diseases*, Clasificación Internacional de Enfermedades)[3], UMLS (*Unified Medical Language System*, Sistema Unificado de Lenguaje Médico)[4] y RadLex[5], entre otros.

En los informes clínicos, la mención de un término no necesariamente indica la presencia de la condición clínica representada por el mismo. Ésta podría referirse a la historia del

[1] Registrar ordenadamente datos e informaciones para elaborar su índice.

[2] Especificaciones formales explícitas de los términos en el dominio y las relaciones entre ellos [18].

[3] ICD es una herramienta utilizada para clasificar las enfermedades y otros problemas de salud en muchos tipos de registros. La versión más extendida se conoce como ICD-10, la 10° revisión.

[4] UMLS es un sistema que reúnen muchos vocabularios de salud, biomédicos y normas para permitir la interoperabilidad entre sistemas informáticos.

[5] RadLex es un léxico centrado en términos de radiología. Sólo está disponible en inglés y en una versión reducida en alemán.

paciente, pero ya no está presente. El informe también podría referirse al término como negado, es decir que explica que el paciente no padece esa condición clínica. Para conocer si un informe menciona a un término como presente en un paciente o no, es fundamental obtener la información que acompaña a dicho término (contexto del término) debido a que esta permite comprender el estado de dicho paciente. La negación es probablemente la característica más importante del contexto de un término en los informes clínicos. Sin embargo, los sistemas de recuperación y extracción de información por lo general para realizar la indexación no consideran el contexto de la condición, ni distinguen entre los términos que se mencionan como presentes y aquellos que están negados.

No obstante, Chapman et al. [6] exponen que gran parte de las condiciones indexadas en informes médicos están negadas, es decir que los médicos a menudo describen que una enfermedad en particular puede ser descartada o que una condición clínica consistente con una sospecha de enfermedad está ausente [6], indicando la importancia de diferenciar en un informe clínico las condiciones clínicas que están presentes, de aquellas que están ausentes, para la indexación precisa del informe.

Por esta razón, en el área de procesamiento del lenguaje natural (NLP, *Natural Language Processing*) se ha prestado mucha atención y se ha dado reconocimiento e importancia a las tareas de determinar el alcance de las negaciones y modalidad epistémica[6], ambas relacionadas con extracción y recuperación de información, y también con otras áreas como ser el análisis de sentimientos [11, 12, 24, 31, 47].

Se han desarrollado varios métodos para determinar si una condición clínica o *finding* está negada [5, 14, 21, 33] (llamamos *finding* a una observación clínicamente significativa, por lo general utilizada en relación con lo que se encuentra en el examen físico o análisis de laboratorio[7]). Las técnicas desarrolladas para abordar estas tareas pueden estar basadas en reglas diseñadas manualmente, o en algoritmos de aprendizaje automático (*machine learning*). Algunas utilizan información morfológica (los indicios que se pueden obtener de la estructura interna de una palabra en cuanto a la categoría que puede tener la misma), sintáctica (los indicios que se pueden obtener a partir de los contextos típicos en los que puede ocurrir una palabra) y semántica (información que se puede obtener de los significados de la palabra respecto a su categoría léxica). Otras técnicas usan expresiones regulares, reconocimiento de patrones, clasificadores o combinaciones de éstos.

Cabe destacar que las expresiones que denotan negaciones o incertidumbres varían según el idioma, y dado que las técnicas de procesamiento del lenguaje natural analizan el lenguaje para abordar estas tareas, el idioma en que se encuentra el texto es un factor muy importante. Por esta misma razón, las herramientas que existen para el procesamiento del lenguaje varían según el mismo. El idioma inglés cuenta con numerosas herramientas. Entre ellas se encuentra la colección de documentos de Bioscope[8] (Vincze et al. [45]). Ésta es descrita por los autores Cruz Díaz et al. [12] como un corpus público de acceso gratuito, que consiste de textos médicos y biológicos. En el mismo, cada frase se anota con información acerca de la negación y la especulación. La anotación indica los límites del alcance y las palabras clave. La colección está formada por tres tipos de documentos: do-

[6] La modalidad epistémica se define como la expresión del grado de certeza o duda que el emisor muestra con respecto a la verdad de la proposición contenida en su enunciado. `http://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/diccionario/modalidad.htm`

[7] Farlex Partner Medical Dictionary. Retrieved April 18 2016, 2012. URL `http://medical-dictionary.thefreedictionary.com/finding`.

[8] Presentado en junio de 2008 en *BioNLP 08 ACL Workshop: Themes in biomedical language processing* (Taller de ACL en BioNLP 08: Temas en el procesamiento del lenguaje biomédico) (Columbus, Ohio).

cumentos clínicos (informes de radiología en formato no estructurado), artículos científicos y resúmenes de artículos científicos.

Para este mismo idioma, se han desarrollado diversos algoritmos para emprender la tarea de detección de negaciones en textos médicos escritos en dicho idioma. Entre ellos, Chapman et al. [5] desarrollaron el algoritmo NegEx. Éste es un algoritmo sencillo basado en expresiones regulares y está disponible para su uso gratuitamente (hay varias implementaciones que se pueden descargar de http://code.google.com/p/negex).

Recientemente otros trabajos centraron sus esfuerzos en adaptar NegEx a diferentes idiomas, sin embargo, estas adaptaciones informan una performance degradada en comparación con la de la versión de NegEx en inglés.

A pesar de que el español es uno de los idiomas más hablado del mundo [1], los recursos de NLP que existen para este idioma, y en particular para análisis de textos médicos, son escasos, según la información obtenida de la investigación realizada al respecto para este trabajo. En particular, se desconoce la existencia de un corpus anotado de textos médicos del dominio radiológico escrito en español.

La propuesta de este trabajo es explorar, desarrollar y evaluar distintas técnicas para la detección de negaciones en informes médicos del ámbito radiológico escritos en idioma español. Además se propone analizar la aplicación de las técnicas sobre informes de otros dominios.

Para esto se utilizó un conjunto de datos que consiste de informes médicos radiológicos (de ecografías), realizados en el Hospital de Pediatría Garrahan. Este hospital es un centro de salud pública destinado a la atención de niños y adolescentes de entre 0 y 15 años. Es un centro de alta complejidad, que está destinado a la atención de patologías complejas.

Los informes están escritos en español en un formato no estructurado. Se caracterizan por tener ruido, es decir hay algunos informes que contienen errores de tipografía, abreviaciones no estándar, oraciones sintácticamente mal formadas, y en algunos casos falta de puntuaciones. Los informes describen qué se encontró en el paciente a partir del estudio radiológico realizado. Los términos de interés (que denominaremos *findings*) de las oraciones del conjunto de datos fueron previamente etiquetados automáticamente con una técnica de extracción de información que utiliza RadLex como léxico [10]. El conjunto de datos disponible se encuentra en un archivo de texto separado en párrafos. Cada párrafo es semi-estructurado y contiene un informe e información relacionada al mismo. El formato de un párrafo es el siguiente:

id de informe *edad del paciente* *fecha de realización del estudio* *informe*

A continuación se muestran dos informes de ejemplo (cuyos *findings* ya están etiquetados):

191523 10a 2m 20101015 A194532 HIGADO: tamano y ecoestructura normal. VIA BILIAR intra y extrahepatica: no <FINDING>dilatada</FINDING>. VESICULA BILIAR: <FINDING>alitiasica</FINDING>. Paredes y contenido normal. PANCREAS: tamano y ecoestructura normal. BAZO: tamano y ecoestructura normal. Diametro longitudinal: 9.3(cm) RETROPERITONEO VASCULAR: sin alteraciones. No se detectaron<FINDING> adenomegalias</FINDING>. No se observo <FINDING>liquido </FINDING> libre en cavidad. Ambos rinones de caracteristicas normales. RD Diam Long: 10.5 cm RI Diam long: 9.7 cm Vejiga poco replecionada. Se visualiza apéndice cecal, compresible, con diámetro anteroposterior de 0.4 cm. Ganglios en cadena iliaca derecha,

todos de estructura conservada. Útero no evaluable por vejiga vacia. No se logra identificar ovario izquierdo. Ovario derecho de ecoestructura conservada.(vol de 2.7 cm 3)

231755 12a 7m 20100702 A304514 Ambos rinones ortotopicos, de ecoestructura normal. Relacion corticomedular conservada. Se observa <FINDING>dilatacion </FINDING> pielica bilateral. der:0.8 cm, izq:1.9 cm. Vejiga de contorno neto, paredes finas y contenido normal. Se observan ureteres distales en forma intermitente: der de 0.4 cm, izq de 0.4 cm. DIMENSIONES: Rinon derecho: 7.3 x 3.3 x 2.6 cm Rinon izquierdo: 8.7 x 3.2 x 2.9 cm

El análisis de los informes se realiza por oración. Se busca identificar las oraciones que tengan *findings* relevantes, es decir, que se mencionen como presentes en el paciente, y descartar las oraciones que contengan *findings* que están negados.

En los ejemplos anteriores, en la oración *Se observa* ***dilatacion*** *pielica bilateral.* del segundo informe, el *finding dilatacion* se menciona como presente y por lo tanto es relevante, sin embargo en la oración *No se detectaron* ***adenomegalias*** del primer informe, el *finding adenomegalias* se menciona negado, es decir no hay adenomegalias en el paciente, por lo que la oración carece de interés y debe descartarse.

El objetivo es detectar cuando un término de interés está negado en una oración, para de esta forma saber que no tiene sentido considerarlo y descartarlo.

Para esto se propone construir un conjunto de datos anotados a partir de los informes radiológicos del hospital, y presentar tres enfoques distintos para afrontar la tarea de detección de negaciones en dicho conjunto de datos.

El primer enfoque consiste en realizar una adaptación de NegEx para español, adecuar dicha adaptación para el ámbito radiológico, evaluarla en este dominio y luego presentar una versión genérica de la adaptación para evaluarla en otros dominios.

Los otros dos enfoques se basan en la obtención y utilización de información morfológica y sintáctica de las oraciones de los informes para detectar negaciones en los mismos, y determinar el alcance de éstas, para luego poder decidir si los *findings* que se mencionan en los informes se encuentran negados o no.

1.1. Marco teórico

Modelo del lenguaje: El modelo del lenguaje es un mecanismo para definir la estructura del lenguaje. Éste solamente debería aceptar frases correctas y rechazar aquellas secuencias de palabras incorrectas. Un modelo del lenguaje estadístico asigna una probabilidad a una secuencia de m palabras. En general, son útiles en muchas aplicaciones de procesamiento de lenguaje natural. Se utiliza en reconocimiento de voz, en traducción automática, reconocimiento de escritura, recuperación de información y otras aplicaciones.

N-gramas: Un modelo del lenguaje estadístico clásico son los N-gramas, en donde se utilizan los N-1 tokens (palabras, cracteres, etc) anteriores para predecir el siguiente token. Para este modelo, si N=1 se denomina *unigramas*, si N=2 *bigramas*, N=3 *trigramas*. Una manera de interpretar los modelos de N-gramas es colocando una ventana sobre una frase o un texto, en el que sólo n tokens son visibles al mismo tiempo. Por ejemplo, el proceso de extracción de trigramas con la oración de entrada *El árbol de hojas verdes es frondoso.* se representa en la tabla 1.1

El	árbol	de	hojas	verdes	es	frondoso	.
El	árbol	de	hojas	verdes	es	frondoso	.
El	árbol	de	hojas	verdes	es	frondoso	.
El	árbol	de	hojas	verdes	es	frondoso	.
El	árbol	de	hojas	verdes	es	frondoso	.
El	árbol	de	hojas	verdes	es	frondoso	.

Tab. 1.1: Tabla que muestra el proceso de extracción de trigramas utilizando como tokens las palabras, con la oración de entrada *El árbol de hojas verdes es frondoso.*

Se coloca sobre la oración una ventana, y sólo tres palabras (tokens) están siempre disponibles a la vez. En este ejemplo, se extrajeron 6 trigramas de una oración de 8 palabras:

- (El, árbol, de)
- (árbol, de, hojas)
- (de, hojas, verdes)
- (hojas, verdes, es)
- (verdes, es, frondoso)
- (es, frondoso, .)

Información morfológica: Se denomina información morfológica a los indicios que se pueden obtener de la estructura interna de una palabra en cuanto a la categoría que puede tener la misma. Por ejemplo, *-mente* es un sufijo que se combina con un adjetivo para producir un adverbio de modo, por ejemplo, eficaz → eficazmente, fácil → fácilmente. Así que dada una palabra que termina en *-mente*, es muy probable que sea un adverbio [4].

Información sintáctica: La información sintáctica se refiere a las indicios que se pueden obtener a partir de los contextos típicos en los que puede ocurrir una palabra. Por ejemplo, dada la categoría de los sustantivos, un criterio sintáctico de un adjetivo en español es que puede ocurrir inmediatamente antes de un sustantivo. De acuerdo con esto, en la frase *la próxima calle*, la palabra *próxima*, debe ser categorizada como un adjetivo [4].

Información semántica: Los significados de las palabras brindan información respecto a su categoría léxica. En este caso, se denomina información semántica. Por ejemplo, la definición más conocida de un sustantivo es semántica: "El nombre de una persona, lugar o cosa" [4].

Parts-of-Speech (Clases de palabras): Se denomina *Parts-of-Speech* (PoS) a las clases de equivalencia en las que se pueden agrupar las palabras. Por ejemplo, sustantivo, verbo, adjetivo, preposición, adverbio, conjunción, etc. El *Part-of-Speech* de una palabra proporciona una cantidad significativa de información acerca de la palabra y sus vecinos, de cómo se pronuncia la palabra, del sentido de la misma, cuál es su género y su número. Esto puede ser útil en sistemas de reconocimiento de voz, en análisis sintáctico del lenguaje natural y en sistemas de recuperación y extracción de información. Por ejemplo, los corpus

que tienen marcado los *Parts-of-Speech* son muy útiles para ayudar a encontrar casos o frecuencias de construcciones particulares en grandes corpus [23].

PoS-tag: El PoS-tag es un código utilizado para codificar información morfológica. Este código está basado en la codificación propuesta por EAGLES[9] (*Expert Advisory Group on Language Engineering Standards*, Grupo Asesor de Expertos en Normas de Ingeniería del Lenguaje) [16].

Lema: Un *lema* corresponde a un conjunto de formas léxicas que tienen la misma raíz y el mismo sentido. Esto permite tratar con las inflexiones de una palabra como *gato* y *gatos* tratándolas como instancias de una sola palabra abstracta, o lema [23].

Lematización: La lematización es un proceso que agrupa las distintas formas de una palabra (como por ejemplo *apareció, aparece, aparecerá* al lema de la palabra, también conocido como *forma canónica* de la palabra (para el ejemplo, *aparecer*) [4].

PoS-tagging: Se denomina *PoS-tagging* o *Parts-of-Speech tagging* (etiquetado gramatical) al proceso de asignación automática de *Parts-of-Speech* u otro marcador de clase léxica a cada palabra en un corpus [23].

Sintagma: Según el diccionario[10], un sintagma se define de la siguiente manera:

> Un sintagma es un grupo de palabras que forman un constituyente sintáctico. Dentro de un sintagma hay una palabra fundamental que recibe el nombre de núcleo sintáctico y es la que aporta las características básicas para la formación de ese grupo. Este núcleo será también el responsable de darle nombre al sintagma. Por ejemplo, si el núcleo de un sintagma es un verbo, estaremos frente a un grupo verbal. Todas las oraciones pueden descomponerse en diversos sintagmas que se encuentran vinculados mediante relaciones sintácticas y semánticas.
>
> Los sintagmas se clasifican en sintagmas léxicos y sintagmas funcionales. Dentro de los sintagmas léxicos se encuentran los siguientes subgrupos:
>
> - Sintagma Nominal (SN): es un conjunto de palabras que se encuentran vinculadas alrededor de un sustantivo, el cual es el núcleo del sintagma.
> - Sintagma Preposicional (SP): es el único que no recibe el nombre del núcleo del sintagma sino de una palabra cuya función es enlazar dos partes del sintagma, la cual siempre es una preposición.
> - Sintagma Adjetival (SAdj): este grupo de palabras se encuentran vinculadas a partir del núcleo que siempre es un adjetivo.
> - Sintagma Adverbial (SAdv): en este caso el núcleo es un adverbio, cuya fundamental función es brindarle coherencia a la oración.

[9] EAGLES es una iniciativa de la Comisión Europea, financiada dentro del programa *Investigación e Ingeniería lingüística*, que tiene como objetivo acelerar la provisión de estándares para recursos lingüísticos a gran escala (como corpus de textos, léxicos computacionales y corpus orales), medios de manipulación de esos conocimientos (a través de formalismos de lingüística computacional y diversas herramientas de software) y medios de apreciación y evaluación de los recursos, las herramientas y los productos. `http://www.ilc.cnr.it/EAGLES96/home.html`

[10] `http://definicion.de/sintagma/`

- Sintagma Verbal (SV): el núcleo es un verbo en torno al cual giran el resto de componentes de la oración.

Los sintagmas funcionales son aquellos cuyo núcleo no presenta significado léxico. Pueden dividirse en sintagmas de tiempo (un verbo auxiliar), sintagmas complementantes (un nexo subordinante) y sintagmas determinantes (un indicador de definición, cantidad o tipo de referencia).

Chunk: Un *chunk* es una unidad textual de tokens de palabras adyacentes[11]. Los tokens de palabras internos a un chunk comparten la propiedad de estar mutuamente vinculados a través de cadenas de dependencia que se pueden identificar sin ambigüedades, recurriendo únicamente a la información léxica obtenida del *part-of-speech* y el lema. Un chunk siempre es una unidad máxima[12].

Los chunks son casos particulares de los sintagmas. Por ejemplo un NP-chunk (*noun-phrase chunk*, chunk de sintagma nominal), es un sintagma nominal con la condición extra de que no puede contener otros sintagmas nominales.

Chunking: Aplicar *chunking* a un texto significa segmentarlo en una secuencia no estructurada de *chunks*, que muestran las relaciones que sostienen entre sus palabras internas [15].

Por ejemplo, para la frase *El árbol de hojas verdes es frondoso*, se tienen los siguientes chunks:

[**SN** El árbol] [**SP** de hojas verdes] [**SV** es] [**SAdj** frondoso]

Shallow parsing: Se denomina *shallow parsing* (análisis sintáctico superficial) a la tarea de asignar una estructura sintáctica parcial a oraciones. Este análisis identifica los componentes, pero no especifica su estructura interna, ni su papel en la oración principal. Para llevar a cabo esta tarea se requiere identificar frases sintácticas o palabras que participan en una relación sintáctica, es decir se requiere aplicar chunking [25, 32].

Los *shallow parsers* toman como entrada la salida de un analizador morfológico y computan la salida sobre una base de conocimiento lingüístico mínimo, que se suma a la información morfosintáctica, los lemas y el orden de las palabras, obtenida de la representación de los datos de entrada [15].

Por ejemplo, la salida de un shallow parser[13], para la oración de ejemplo anterior se ilustra en la figura 1.1
En el árbol del esquema se distinguen los chunks del ejemplo anterior en los nodos del primer nivel (sn: sintagma nominal, sp-de: sintagma preposicional, grup-verb: sintagma verbal, s-adj: sintagma adjetival, y además hay una categoría especial para el signo de puntuación). En particular, el significado de las etiquetas de cada nodo del árbol se muestra en la tabla 1.2. Los significados fueron extraídos de la página web de FreeLing[14].

[11] Los chunks discontinuos no están permitidos.

[12] Un chunk no se puede insertar dentro de otro chunk.

[13] Se utilizó el shallow parser de la herramienta FreeLing (un conjunto de herramientas que provee funciones de análisis del lenguaje, como *Morphological Analysis,* Part of Speech tagging *(PoS-tagging), Named Entity Recognition, Shallow Parsing*, entre otros, para varios idiomas, entre ellos, español) [36].

[14] La lista de significados de las etiquetas posibles de cada nodo usando la herramienta FreeLing se puede obtener de la página web. `https://github.com/iknow/FreeLing/blob/master/doc/grammars/esCHUNKtags`

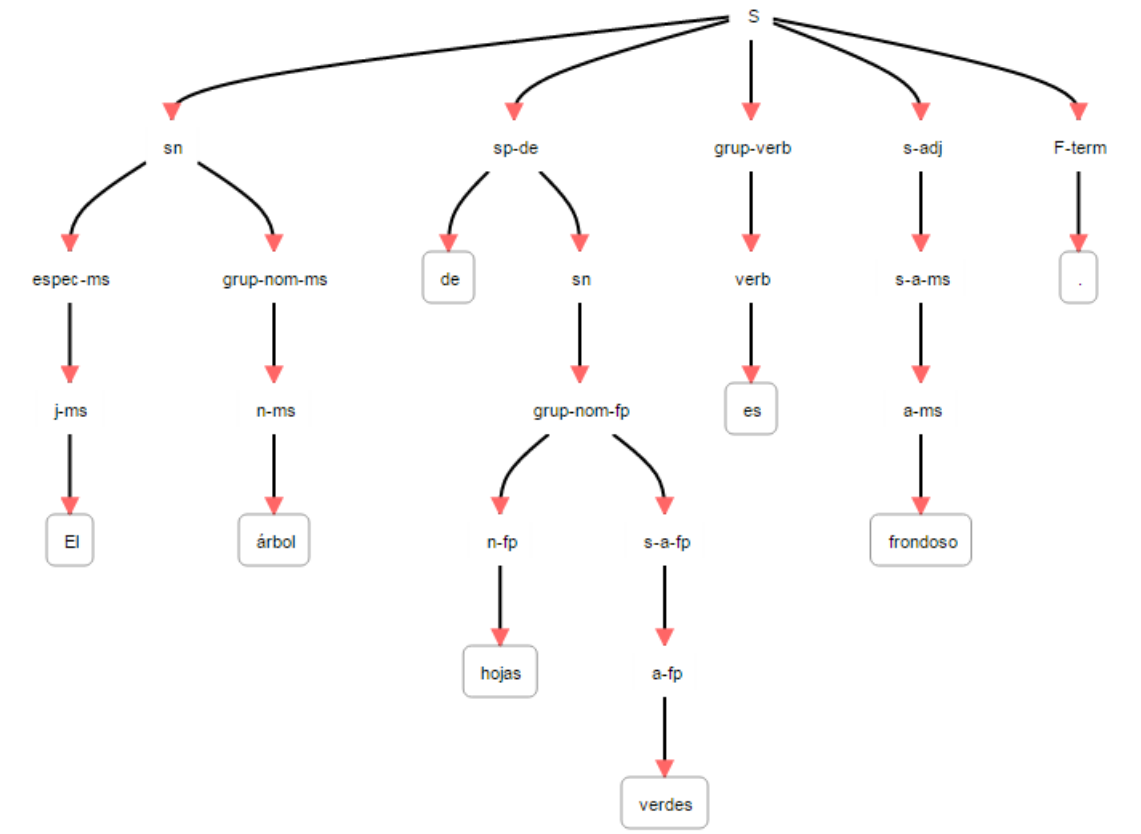

Fig. 1.1: Esquema que muestra la salida de un shallow parser para la oración *El árbol de hojas verdes es frondoso.* utilizando FreeLing.

Parsing gramático: Análisis formal de una cadena de palabras que resulta en un árbol de parsing, el cual muestra las relaciones sintácticas entre sí, y también puede contener información semántica u otro tipo de información.

1.2. Contribuciones

Las contribuciones de este trabajo se resumen de la siguiente manera:

- Se construyó un conjunto de datos anotados a partir de un corpus de datos conformado por informes de radiología escritos en español, provenientes de un importante hospital de pediatría público, cuyos *findings* están etiquetados.
- Se desarrollaron tres enfoques distintos para emprender la tarea de detección de negaciones en textos médicos escritos en idioma español. Éstos utilizan herramientas de NLP. Uno usa expresiones regulares, los otros dos utilizan métodos que obtienen información sintáctica de un texto para identificar cuáles de los *findings* se mencionan como negados en los informes radiológicos:

    1. Se presentó una adaptación de NegEx para español con dos variantes: una adecuada para el dominio de radiología y una genérica para ser aplicada en otros

Etiqueta	Significado
a-fp	*adjective, adjective feminine plural* (adjetivo, adjetivo femenino plural)
a-ms	*adjective, adjective masculine singular* (adjetivo, adjetivo masculino singular)
s-a-fp	*adjective, adjective phrase feminine plural* (adjetivo, frase adjetivo femenino plural)
s-a-ms	*adjective, adjective phrase masculine singular* (adjetivo, frase adjetivo masculino singular)
s-adj	*adjective, adjective phrase* (adjetivo, frase adjetivo)
j-ms	*determiner, definite determiner masculine singular* (determinador, determinador definido masculino singular)
espec-ms	*determiner, determiner masculine singular* (determinador, determinador masculino singular)
grup-nom-fp	*noun, nominal chunk feminine plural* (sustantivo, chunk nominal femenino plural)
grup-nom-ms	*noun, nominal chunk masculine singular* (sustantivo, chunk nominal masculino singular)
n-fp	*noun, noun feminine plural* (sustantivo, sustantivo femenino plural)
n-ms	*noun, noun masculine singular* (sustantivo, sustantivo masculino singular)
sn	*noun, noun phrase* (sustantivo, frase sustantivo)
sp-de	*preposition, prepositional phrase "de"* (preposición, frase preposicional "de")
F-term	*punctuation, sentence terminators* (puntuación, terminadores de oración)
verb	*verb* (verbo)
grup-verb	*verb, verbal chunk* (verbo, chunk verbal)

Tab. 1.2: Tabla que muestra el significado de las etiquetas de cada nodo del árbol obtenido al aplicar shallow parsing (utilizando FreeLing) a la oración de ejemplo *El árbol de hojas verdes es frondoso.*

dominios. Difieren de una adaptación existente para español (la adaptación de Costumero et al. [9], que se presenta en la sección 2) en que su propuesta es específica para el conjunto de datos que utilizan (éste no es del dominio radiológico, sino que consiste en textos médicos generales cuyos informes son más extensos), mientras que en la adaptación que se propone en este trabajo, una variante es específica para textos del dominio radiológico, la otra es genérica y puede aplicarse a textos de otros dominios. Además, las variantes propuestas presentan resultados distintos a los de aquellos.

2. Se desarrolló un método basado en PoS-tagging: la técnica consiste en utilizar el PoS-tag de las palabras para detectar las negaciones y en base a sus posiciones y el *finding* se construyen reglas que determinen si éste está alcanzado por la negación o no.
3. Se expuso un método basado en shallow parsing: la propuesta consiste en aplicar

shallow parsing a las oraciones de los informes y utilizar la información obtenida para decidir si las oraciones mencionan *findings* negados o no.

1.3. Difusión de los resultados

Una versión abreviada de los resultados presentados en esta tesis ha sido publicada originalmente en el *Workshop on Replicability and Reproducibility in Natural Language Processing: adaptive methods, resources and software at IJCAI 2015*[15] (Taller sobre la replicabilidad y reproducibilidad en el Procesamiento del Lenguaje Natural: métodos de adaptación, recursos y software en IJCAI 2015) [42].

Otra parte de los resultados expuestos en el presente trabajo serán publicados en el *15th Workshop on Biomedical Natural Language Processing at ACL 2016*[16] (15° Taller sobre Procesamiento del Lenguaje Natural en Biomedicina en ACL 2016).

1.4. Organización

El resto del trabajo se organiza de la siguiente manera: la sección Trabajos previos presenta los trabajos previos en la detección de negaciones en el ámbito médico, incluyendo el enfoque original de NegEx y las adaptaciones a otros idiomas distintos del inglés. También presenta otros enfoques para abordar la tarea. En la sección Metodología y Experimentación se discute el comportamiento de una implementación de NegEx utilizada como algoritmo base y se expone la metodología usada para lograr el objetivo propuesto, los experimentos realizados y las técnicas de evaluación. En la sección Resultados se muestran los resultados de los experimentos realizados y se presenta un análisis de los mismos. Finalmente, se presentan las conclusiones de este trabajo en la sección Conclusiones y el trabajo futuro. La sección Apéndice contiene un glosario de la terminología utilizada en este trabajo, un listado de abreviaturas y siglas comunes utilizadas, algunos cuadros con datos y resultados usados para el desarrollo del presente trabajo y un resumen de resultados de otros trabajos realizados sobre este tema.

Al comienzo de cada sección y de cada tema se presenta un resumen del contenido de la misma.

[15] IJCAI: *International Joint Conference on Artificial Intelligence* (Conferencia Internacional Conjunta sobre Inteligencia Artificial). `https://sites.google.com/site/adaptivenlp2015/`

[16] ACL: *Association for Computational Linguistics* (Asociación para Lingüística Computacional). `http://www.aclweb.org/aclwiki/index.php?title=BioNLP_Workshop`

2. TRABAJOS PREVIOS

En el ámbito médico y del cuidado de la salud existen numerosos documentos, que proveen información valiosa para la detección y caracterización de enfermedades, ya que reportan signos, síntomas, tratamientos y resultados. Esta información podría utilizarse por los médicos para la realización de diversos estudios e investigaciones, como por ejemplo estudios epidemiológicos, que podría llevar al descubrimiento de nuevas relaciones entre síntomas y patologías, entre otras cosas, y esto puede impactar en diagnósticos y tratamiento de enfermedades. Estos documentos suelen encontrarse escritos en formato de texto libre no estructurado, por lo cuál, con el fin de extraer la información, es necesario aplicar técnicas de procesamiento del lenguaje natural.

Pero, de acuerdo con Chapman et al. [6], aproximadamente la mitad de todas las condiciones clínicas descritas en los informes médicos están negadas. Es por esto que la identificación de negaciones en documentos del dominio médico es una tarea que ha recibido mucha atención en el área de NLP, en particular de BioNLP. El tema ha sido también el foco de diversas competencias, incluyendo SEM 2012 *Shared Task: Resolving the Scope and Focus of Negation* [29] (Tarea compartida: Resolviendo el alcance y enfoque de la negación) y 2010 i2b2/VA *challenge on concepts, assertions, and relations in clinical text* [44] (desafío en conceptos, afirmaciones, y las relaciones en textos clínicos).

La necesidad de determinar no sólo si un *finding* se menciona en los informes médicos narrativos sino también si tal *finding* está presente o ausente inspiró el trabajo de Chapman et al. [5], quienes desarrollaron un algoritmo basado en expresiones regulares llamado NegEx. NegEx define un conjunto de frases que indican negación (usualmente llamados *triggers, cues* o *lexicons* [7]) y clasifican a las oraciones en *Affirmed* o *Negated* de acuerdo a la aparición de las frases que indican negación en la oración, y la distancia que hay entre estas frases y los *findings*. Los resultados prometedores de este enfoque motivaron el desarrollo de otros trabajos basados en él (mejoras, adecuaciones a otros idiomas, entre otras cosas). NegEx también se ha integrado en varios sistemas de extracción de información de código abierto. Entre estos se encuentran MetaMap [2] un programa que provee acceso a los conceptos UMLS de textos biomédicos, cTAKES [39] un sistema de análisis de texto clínico y extracción de información, específico para el dominio clínico que crea anotaciones lingüísticas y semánticas, usadas para el procesamiento semántico de texto clínico no estructurado, y HITEX [50], una herramienta de extracción de texto en información sobre la salud y utilizado para extraer *findings* claves para un estudio de investigación sobre enfermedades de las vías respiratorias.

Además existen diversas alternativas para abordar esta tarea, como *machine learning* (aprendizaje automático) que utiliza clasificadores, técnicas que usan información léxica y sintáctica como análisis de dependencias, y también métodos híbridos que combinan distintos enfoques.

A continuación se describe NegEx y algunas de las adaptaciones del algoritmo. Además se presentan enfoques que aplican algunas de las técnicas alternativas mencionadas anteriormente. En el apéndice se muestra un cuadro con los resultados obtenidos por los diferentes trabajos reportados en esta sección.

2.1. NegEx y adaptaciones

Chapman et al. [5] diseñaron *NegEx*, un algoritmo simple para determinar si un *finding* mencionado dentro de un informe médico escrito en inglés está presente o ausente en el paciente, según lo dictaminado por el médico.

NegEx utiliza expresiones regulares que denotan frases que podrían indicar negación (triggers), para determinar si los *findings* de los informes se encuentran negados.

Las frases detectadas como posibles indicadores de negación se dividen en dos grupos:

- *pseudo-negation*: frases que aparentan indicar negación pero indican doble negaciones, modificadores y frases ambiguas. Ejemplos: *not necessarily* (no necesariamente), *without difficulty* (sin dificultad), *not rule out* (no se descarta).

- *negation*: frases que se utilizan para negar *findings* y enfermedades, que pueden aparecer antes (*pre-negation*) o después (*post-negation*) del *finding* con cinco tokens de por medio. Ejemplos de frases *pre-negation* son *absence of* (ausencia de), *no evidence of* (sin evidencia de), *without* (sin). Ejemplos de frases de *post-negation* son *declined* (disminuido), *unlikely* (improbable).

El algoritmo se aplicó sobre un conjunto de informes médicos provenientes de *discharge summaries*[1]. Para identificar los términos relevantes de los informes, se utilizó como ontología la intersección de UMLS [22] con ICD10[2], usando frases del tipo semántico *Finding, Disease or Syndrome* (Hallazgo, Enfermedad o Síndrome) y *Mental or Behavioral Dysfunction* (Disfunción Mental o de Comportamiento).

Este conjunto de datos tiene 1000 oraciones únicas, que contienen un total de 1235 *findings*, conformando un conjunto de 1235 oraciones ya que cuando una oración contiene más de un *finding* repiten la oración tantas veces como sea necesaria para analizar cada *finding* por separado.

Para poder evaluar las oraciones del conjunto de datos, tres médicos anotaron los informes estableciendo el Gold Standard[3].

Cada término UMLS en la oración (*finding*) es anotado como *Affirmed, Negated*, según si ese término se describe como presente o ausente en el paciente según la oración, o *Ambiguous* si no es posible determinar a partir de la misma el estado del paciente. Las 1000 oraciones fueron anotadas de forma superpuesta, de manera tal que cada médico anotó 400 oraciones, obteniendo 200 oraciones anotadas por dos pares de médicos. Si dos médicos no coincidían en la anotación de una misma oración, era anotada como ambigua.

De las 1235 oraciones, 343 fueron anotadas por los anotadores como *Negated*, 728 como *Affirmed* y 164 como *Ambiguous*. Las anotadas como *Ambiguous* luego se pasaron

[1] Informes clínicos elaborados por un médico u otro profesional de la salud como conclusión de una hospitalización o una serie de tratamientos (un resumen de alta). En ellos se escriben los principales síntomas por los que el paciente realizó la visita médica, los resultados de diagnóstico, la terapia administrada y la respuesta del paciente a la misma. Mosby's Medical Dictionary, 8th edition, 2009. Retrieved April 18 2016, `http://medical-dictionary.thefreedictionary.com/discharge+summary`

[2] International statistical classification of disease and related health problems, 10th revision. Volume 1. World Health Organization, `http://who.int/classifications/icd/en/`.

[3] Gold Standard Annotation es un corpus de confianza que se utiliza para el entrenamiento y evaluación significativa de algoritmos que usan anotaciones. Estas colecciones se llaman *Gold Standard Corpora* (GSC o GS, corpus estándar de oro). Sin embargo, la construcción de un GS es una tarea laboriosa y requiere mucho tiempo de proceso. Además el tamaño, la calidad y sobre todo la disponibilidad de GS de tareas específicas, influyen directamente en el desarrollo de algoritmos de procesamiento de lenguaje natural basados en aprendizaje automático [48].

a *Affirmed* debido a que NegEx no analiza esa categoría. De esta forma quedan 892 *Affirmed*. Los resultados obtenidos fueron los siguientes: *recall*: 77,84 %, *precision*: 84,49 %, *specificity*: 94,51 %, NPV: 91,73 %.

Harkema et al. [19] introdujeron y evaluaron un algoritmo llamado *ConText* que es una extensión de NegEx. A diferencia de éste, ConText identifica otros valores contextuales además de la negación: *temporalidad* y *experimentador*. Así, la negación especifica el estado de existencia de la condición clínica, la temporalidad ubica a la condición a lo largo de una línea de tiempo simple y el experimentador describe si el que experimenta la condición clínica es el paciente o algún otro. Por otra parte, se evalúa ConText en seis tipos distintos de documentos médicos: informes radiológicos, documentos del departamento de emergencia, documentos de patología quirúrgica, ecocardiogramas, documentos de procedimientos operativos y *discharge summaries*. El objetivo de los autores es evaluar la versatilidad del enfoque en dos dimensiones: que tan bien funciona para los nuevos valores contextuales y que tan bien funciona en distintos tipos de documentos.

ConText asigna a cada propiedad contextual un valor inicial (*Affirmed* para la negación, *Recent* para la temporalidad, y *Patient* para el experimentador). Luego busca triggers que preceden o siguen a la condición clínica. Si la condición clínica se ve afectada por un trigger, cambia el valor asumido inicialmente de la propiedad contextual correspondiente al valor indicado por ese trigger. Para esto los autores crearon una lista separada de triggers para cada propiedad contextual. Además una lista de pseudo-triggers y triggers de terminación por separado.

Para evaluar el algoritmo utilizaron anotaciones de referencia provistas por un médico. El médico anotó todas las condiciones clínicas en el conjunto de datos, y luego, a cada condición clínica le asignó un valor a cada propiedad contextual. Luego un segundo médico realizó el mismo proceso. Para cada desacuerdo entre los médicos, el primer médico revisó ambas anotaciones y corrigió los valores asignados por él, en los casos en los que creía que el segundo médico había anotado correctamente. Sobre este conjunto de datos se aplicó el algoritmo para evaluarlo.

Wu et al. [49] desarrollaron *RadReportMiner*, un método para realizar búsquedas en informes radiológicos, cuyo objetivo es mejorar la precisión en la búsqueda respecto al motor de búsqueda genérico Google Desktop. Para esto, cuando se realiza una búsqueda en informes radiológicos, se excluyen de los resultados de búsqueda aquellos informes que contienen términos de interés que se ven afectados por frases de negación o incertidumbre. La implementación se basa en una modificación de NegEx. Los autores agregaron triggers de negación específicos del dominio radiológico y triggers que denotan incertidumbre (*hedges*). De esta manera obtuvieron los siguientes grupos de triggers:

- *pseudo-negation* (los mismos triggers de NegEx).
- *negation*: *pre-negation* (por ejemplo *clear of*, libre de) y *post-negation* (por ejemplo *no longer present*, ya no está presente, *no longer visualized* ya no visualizado).
- *uncertainty*: *pre-possible negation* (por ejemplo *may represent* puede representar, *can not absolutely rule out* no se puede descartar absolutamente) y *post-possible negation* (por ejemplo *not excluded*, no excluidos, *not ruled out* no descartados, *should also be considered* también debe ser considerado).

Además computaron un puntaje de relevancia, según el trigger y el grupo al que pertenece el trigger, si es de negación o incertidumbre, etc., y lo utilizaron para clasificar cada informe en *Negated* o *Affirmed*.

Con el objetivo de evaluar el algoritmo, crearon un Gold Standard. Para hacerlo realizaron cinco búsquedas de cinco términos (*appendicitis, optic neuritis, pneumonia, hydronephrosis* y *fracture*) en una base de datos de informes radiológicos y utilizaron los primeros 100 resultados de búsqueda. Luego, un radiólogo clasificó manualmente el *finding* o diagnóstico en cuestión en *Affirmed* si podía afirmar con certidumbre que el *finding* en cuestión está presente en el paciente o *Negated* si el término se expresaba como ausente o con incertidumbre. Sobre este conjunto de informes aplicaron RadReportMiner y compararon los resultados con el Gold Standard. Los resultados de evaluar RadReportMiner son, según la cantidad de informes, *precision*: 81 % y *recall*: 72 %, y según la cantidad de términos, *precision*: 81 %, y *recall*:76 %.

Skeppstedt [40] realizó la adaptación de NegEx a sueco, traduciendo los triggers del inglés[4] a sueco. A diferencia del inglés, el sueco tiene dos géneros gramaticales, los tiempos verbales en inglés no se corresponden directamente con una forma verbal en sueco y además el sueco tiene el orden de las palabras invertidas para las cláusulas subordinadas. Por esto agregó cuantificadores negativos, generó inflexiones de todos los adjetivos, utilizó el *lemma* del verbo y luego generó todas las inflexiones del verbo, e invirtió el orden de las palabras en las frases de negación para contrarrestar las diferencias gramaticales.

Para evaluar el algoritmo utilizó un conjunto de oraciones elegidas aleatoriamente de la sección *assessment field* (evaluación sobre el terreno) de diferentes partes de *Swedish health records*[5] en el corpus Stockholm EPR [13] que contiene historias clínicas de los pacientes. En este corpus etiquetó los términos de interés con un algoritmo simple de *string matching*. Los términos de interés provienen de la traducción de términos de la ontología ICD10[6] y de términos que se incluyen en KSH97-P[7].

El proceso de anotación fue realizado por dos médicos y un anotador sin experiencia médica. Las oraciones fueron divididas en dos conjuntos, las que contienen algún trigger, y las que no contienen triggers. Para el primer grupo, uno de los médicos anotó 70 oraciones, el otro médico anotó las 488 oraciones restantes. El anotador sin experiencia médica anotó todas las oraciones. Se calculó el *inter rater agreement* (grado de acuerdo entre anotadores) para este clasificador y los médicos. La clasificación hecha por los médicos fue la que se utilizó para el Gold Standard. Para el segundo grupo, un médico anotó 95 oraciones. El anotador sin experiencia médica, anotó todas las oraciones. Para el Gold Standard, de las oraciones que fueron anotadas por dos anotadores, se utilizó la anotación del médico. En otro caso, se utilizó la anotación del anotador sin experiencia médica. Para el conjunto de 95 oraciones anotadas por dos personas, se calculó el *inter rater agreement*.

[4] Pueden descargarse de la página web de NegEx, `http://code.google.com/p/negex`.

[5] *Health record* es una recopilación de información que suele estar en el expediente médico, que abarca aspectos de la salud del paciente física, mental y social, que no necesariamente se relacionan directamente con la condición bajo tratamiento. Medical Dictionary for the Health Professions and Nursing, 2012. Retrieved April 18 2016, `http://medical-dictionary.thefreedictionary.com/health+record`

[6] International statistical classification of disease and related health problems, 10th revision. Volume 1. World Health Organization, `http://who.int/classifications/icd/en/`

[7] KSH97-P es una adaptación de los códigos de ICD10 para *primary care* (atención primaria), *mental disorders* (trastornos mentales) y *diseases* (enfermedades). Klassifikation av sjukdomar och hälsoproblem 1997 (KSH97), `http://www.socialstyrelsen.se/publikationer1997/1997-4-1`.

El conjunto de test utilizado está compuesto de dos grupos de oraciones, el primero de 558 oraciones que contiene triggers, el segundo de 342 oraciones sin triggers, todas las oraciones fueron clasificadas en *Negated, Affirmed* o *Ambiguous*, luego las *Affirmed* y *Ambiguous* se colapsaron en la categoría de *Not Negated*.

Se evaluó la adapatación de NegEx a sueco en los dos grupos de oraciones del conjunto de test por separado, obteniéndose para el primer grupo *precision*: 75,20 % y *recall*: 81,90 % y para el segundo grupo, NPV: 96,50 %.

Chapman et al. [7] tradujeron el léxico de NegEx a tres idiomas distintos: Francés, Alemán y Sueco, y modelaron una representación del léxico para facilitar la traducción a otros idiomas. Además realizaron un análisis léxico contando la frecuencia de los distintos tipos de triggers en los cuatro idiomas, comparándolos. A partir del análisis, los autores llegaron a la conclusión de que unos pocos triggers son muy frecuentemente usados y muchos triggers se usan muy pocas veces. Los grupos de triggers que utilizaron son:

- *negation triggers*: indica una negación. Por ejemplo *denies* (niega).
- *pseudo-negation triggers*: triggers de negación pero que no niegan la condición clínica. Por ejemplo *no increase* (no hay aumento).
- *termination terms*: determinan el alcance de un trigger de negación. Por ejemplo *but* (pero).
- *probable negation triggers*: triggers que indican posibilidad de negación. Por ejemplo *can be ruled out* (se puede descartar).

En dicho trabajo, cualquier condición clínica que se ve afectada por un trigger de negación está negada. Los *pseudo negation triggers* pueden modificar la información hacia la derecha o hacia la izquierda del término (siguiendo o precediendo en la oración). Los *termination terms* determinan el alcance de una negación. De otra manera la negación termina al final de la oración. Además, los autores sugirieron una representación más compleja del léxico, representando cada trigger como un concepto que puede tener una variedad de etiquetas distintas. Es decir, un concepto agrupa un término, sus sinónimos, frases alternativas y errores de ortografía. La traducción de los triggers fue realizada por un grupo de investigadores compuesto por médicos, informáticos y lingüistas.

En el trabajo, se mencionan los problemas y cuestiones encontrados al realizar las traducciones a los distintos idiomas: aglutinación, ambigüedad, terminología específica del lenguaje, el orden de las palabras y abreviaturas. Además, debido a que los informes para cada idioma difieren en tamaño y tipo (informes radiológicos y documentos del departamento de emergencias) las frecuencias de los distintos tipos de triggers son diferentes para cada idioma.

Costumero et al. [9] presentaron una adaptación de NegEx para la detección de negaciones de enfermedades mencionadas en documentos médicos escritos en español.

Para realizar la adaptación, los autores tradujeron los triggers de NegEx (de Harkema et al. [19] utilizaron los triggers de *hyphotesis* y *negation*), y agregaron sinónimos y frases comunes en español del ámbito clínico. A cada trigger le asignaron una de las siguientes categorías: *negation, uncertainty, pseudo-negation.*

Luego, obtuvieron 500 informes en español extraídos de SciElo[8] (*Scientific Electronic Library Online*), utilizando las secciones tituladas como *Reporte de caso, A propósito de un caso* y *Caso clínico*, y marcan los términos de interés utilizando la ontología ICD10. Para obtener un Gold Standard, anotaron manualmente los informes con *Negated* o *Affirmed* según si los términos de interés se mencionan negados o no.

Además, calcularon la frecuencia de los triggers en el corpus, y luego compararon los términos más frecuentes en español con los más frecuentes en inglés.

La versión de NegEx adaptada al español se evaluó con los 500 informes extraídos de SciElo, de los que se consiguieron 422 oraciones diferentes y 267 condiciones clínicas únicas, obteniéndose *precision*:49,47 %, *recall* : 55,70 %, *F1*: 52,38 %, *accuracy*: 83,37 % para el caso de *Negated* como conjunto positivo y *precision*: 86,86 %, *recall*: 95,20 %, *F1*: 90,84 %, *accuracy*: 84,78 % para el caso de *Affirmed*.

En la tabla 2.1 se presenta un resumen de las ontologías usadas por los trabajos descritos previamente.

Trabajo	Ontología
Chapman et al. [5]	términos que están en UMLS y en ICD10
Harkema et al. [19]	no utiliza una ontología particular, las condiciones clínicas fueron anotadas por un médico
Wu et al. [49]	no utiliza una ontología particular, se buscan cinco términos: *appendicitis, optic neuritis, pneumonia, hydronephrosis fracture*
Skeppstedt [40]	traducción a sueco de términos de ICD10 que se incluyen en KSH97-P
Chapman et al. [7]	-
Costumero et al. [9]	ICD10

Tab. 2.1: Resumen de las ontologías usadas para la detección de términos de interés en las distintas adaptaciones de NegEx.

2.2. Otros enfoques

Además de la propuesta de NegEx que utiliza expresiones regulares para abordar la problemática de detección de negaciones, existen enfoques que usan métodos alternativos o métodos híbridos en los que se combinan diferentes técnicas.

Algunos de ellos aplican técnicas de *machine learning*. Rokach et al. [38] extrajeron de forma automática expresiones regulares de los datos anotados y las utilizaron para crear un método de aprendizaje para la identificación automática del alcance de las negaciones en informes clínicos. Uzuner et al. [43] estudiaron la clasificación de afirmaciones en dos

[8] SciElo es una biblioteca virtual formada por una colección de revistas científicas. Esta plataforma virtual proporciona acceso completo a una colección de revistas, así como al texto completo de los artículos [35].

inventarios de *discharge summaries* y informes radiológicos obtenidos de dos competencias en BioNLP, con una técnica de aprendizaje automático que utiliza la información léxica y sintáctica. Morante and Daelemans [30] propusieron un sistema que combina varios clasificadores y trabaja en dos etapas: identificación de negaciones y determinación del alcance de las mismas. El sistema lo probaron en distintos tipos de corpus de BioScope (documentos clínicos, artículos científicos y resúmenes de artículos científicos). Cruz Díaz et al. [12] también aplicaron técnicas de aprendizaje automático sobre textos de BioScope para identificar las expresiones de negación y agregan detección de incertidumbre. Para la detección de negaciones en informes clínicos los primeros obtuvieron *precision*:100 %, *recall*: 98,0 % y F1:99,0 %, a diferencia de los segundos *precision*: 96,5 %, *recall*: 98,0 % y F1: 97,3 %. Para la definición del alcance de las negaciones en los informes clínicos los resultados obtenidos por los primeros fueron: *precision*:91,7 %, *recall*: 92,5 % y F1:92,1 %. En la detección de incertidumbre, los segundos lograron *precision*: 92,8 %, *recall*: 93,4 % y F1: 93,1 %. Cruz et al. [11] también presentaron un enfoque de aprendizaje automático para detectar negaciones, incertidumbre y su alcance en textos clínicos y agregaron análisis de sentimientos.

Otras propuestas son con técnicas que usan información léxica y sintáctica. Algunas de éstas realizan parsing gramático[9]. Mutalik et al. [33] desarrollaron un programa para identificar patrones de negaciones presentes en documentos médicos. Utilizaron UMLS para identificar los términos de interés en los documentos. Luego, identificaron las negaciones y usaron reglas gramaticales para determinar el alcance de las mismas y así definir si los términos de interés se ven afectados por las negaciones. Aplicaron el sistema a historias clínicas obteniendo *recall*: 91,0 % y *specificity*: 96,0 %. Huang and Lowe [21] describieron un enfoque híbrido para abordar la detección automática de negaciones en informes radiológicos, combinando expresiones regulares y parsing gramático. Construyeron una gramática para negaciones a partir de expresiones regulares y patrones de negaciones. Luego, clasificaron las negaciones basándose en la categorías sintácticas de las negaciones. A partir de esto, buscaron en los árboles de parsing las distintas clases de negaciones. En el conjunto de prueba de 120 informes, se identificaron correctamente 287 términos de interés como negados, 23 negaciones no fueron detectadas y 4 negaciones fueron detectadas erróneamente. El enfoque híbrido identificó negaciones con *accuracy*: 99,87 %, *precision*: 98,6 % y *recall*: 92,6 %.

Otras técnicas se basan en análisis de dependencias. Mehrabi et al. [28] probaron métodos de análisis de dependencias, para reducir los falsos positivos de NegEx teniendo en cuenta la relación de dependencia entre las palabras y la negaciones dentro de una oración. Los resultados que obtuvieron son: *precision*: 91,3 %, *recall*: 77,2 % y F1: 83,7 %. Sohn et al. [41] también aplicaron técnicas de análisis de dependencias. Para esto construyeron manualmente reglas de negación, basándose en los caminos de dependencia. Estas reglas no utilizan cantidad de palabras para limitar el alcance de la negación, sino que se basan en el contexto sintáctico. Los resultados que obtuvieron son: *precision*: 96,6 %, *recall*: 73,9 % y F1: 83,8 %.

[9] Parsing: Dentro de la lingüística computacional el término se utiliza para referirse al análisis formal de una cadena de palabras que resulta en un árbol de parsing, el cual muestra las relaciones sintácticas entre sí, y también puede contener información semántica u otro tipo de información.

3. METODOLOGÍA Y EXPERIMENTACIÓN

La detección de negaciones en textos médicos es una tarea de interés en el ámbito del procesamiento del lenguaje natural. A pesar de que el español es uno de los idiomas más hablado del mundo [1], los recursos y herramientas de NLP que existen para dicho idioma, y en particular para análisis de textos médicos, son escasos.

Por estos motivos se proponen y desarrollan tres técnicas distintas para la detección de negaciones en informes médicos del ámbito radiológico escritos en idioma español. Éstas utilizan herramientas de NLP como expresiones regulares y métodos que obtienen información sintáctica de un texto para identificar cuáles de los *findings* se mencionan como negados en los informes radiológicos.

La primer técnica propuesta consiste en realizar una adaptación de NegEx para español con dos variantes: una adecuada para el dominio de radiología y una genérica para ser aplicada en otros dominios.

La segunda propuesta consiste en utilizar el PoS-tag[1] de las palabras para detectar negaciones en el texto y en base a sus posiciones y la posición del *finding* en una oración, armar reglas que determinen si el mismo está alcanzado por la negación o no.

La tercer propuesta consiste en aplicar *shallow parsing*[2] a las oraciones y utilizar la información obtenida del parser para decidir si las oraciones mencionan *findings* negados o no.

En esta sección se presentan los conjuntos de datos utilizados para elaborar y evaluar los algoritmos, se explican los métodos propuestos para abordar la problemática de la detección de negaciones, los experimentos realizados y las técnicas usadas para medir los resultados.

3.1. Datos

El conjunto de datos del que se dispone consiste de aproximadamente 85600 informes médicos radiológicos (de ecografías), realizados en el Hospital de Pediatría Garrahan. Este hospital es un centro de salud pública destinado a la atención de niños y adolescentes de entre 0 y 15 años. Es un centro de alta complejidad, que está destinado a la atención de patologías complejas.

Los informes están escritos en español en un formato no estructurado. Este conjunto se caracteriza por tener ruido, es decir hay algunos informes que contienen errores de tipografía, abreviaciones no estándar, oraciones sintácticamente mal formadas, y en algunos casos falta de puntuaciones. Los informes describen qué se encontró en el paciente a partir del estudio radiológico realizado. El conjunto de datos disponible se encuentra en un archivo de texto separado en párrafos. Cada párrafo contiene un informe e información relacionada al mismo. El formato de un párrafo es el siguiente:

[1] Como se explicó previamente, el PoS-tag es un código utilizado para codificar información morfológica.

[2] Como se describió anteriormente, se denomina *shallow parsing* (análisis sintáctico superficial) a la tarea de asignar una estructura sintáctica parcial a oraciones. Este análisis identifica los componentes, pero no especifica su estructura interna, ni su papel en la oración principal. Para llevar a cabo esta tarea se requiere identificar frases sintácticas o palabras que participan en una relación sintáctica, es decir se requiere aplicar chunking [25, 32].

id de informe *edad del paciente* *fecha de realización del estudio* *informe*

A continuación se muestran dos informes de ejemplo (cuyos *findings* ya están etiquetados):

191523 10a 2m 20101015 A194532 HIGADO: tamano y ecoestructura normal. VIA BILIAR intra y extrahepatica: no <FINDING>dilatada</FINDING>. VESICULA BILIAR: <FINDING>alitiasica</FINDING>. Paredes y contenido normal. PANCREAS: tamano y ecoestructura normal. BAZO: tamano y ecoestructura normal. Diametro longitudinal: 9.3(cm) RETROPERITONEO VASCULAR: sin alteraciones. No se detectaron<FINDING> adenomegalias</FINDING>. No se observo <FINDING>liquido </FINDING> libre en cavidad. Ambos rinones de caracteristicas normales. RD Diam Long: 10.5 cm RI Diam long: 9.7 cm Vejiga poco replecionada. Se visualiza apéndice cecal, compresible, con diámetro anteroposterior de 0.4 cm. Ganglios en cadena iliaca derecha, todos de estructura conservada. Útero no evaluable por vejiga vacia. No se logra identificar ovario izquierdo. Ovario derecho de ecoestructura conservada.(vol de 2.7 cm 3)

231755 12a 7m 20100702 A304514 Ambos rinones ortotopicos, de ecoestructura normal. Relacion corticomedular conservada. Se observa <FINDING>dilatacion </FINDING> pielica bilateral. der:0.8 cm, izq:1.9 cm. Vejiga de contorno neto, paredes finas y contenido normal. Se observan ureteres distales en forma intermitente: der de 0.4 cm, izq de 0.4 cm. DIMENSIONES: Rinon derecho: 7.3 x 3.3 x 2.6 cm Rinon izquierdo: 8.7 x 3.2 x 2.9 cm

Existen diversos inventarios y ontologías[3] que sirven para identificar los términos de interés de los informes. Por ejemplo, la Clasificación Internacional de Enfermedades (ICD, *International Classification of Diseases*) es una herramienta utilizada por médicos, enfermeros, investigadores, trabajadores en tecnología de información de salud, legisladores, aseguradoras, entre otros para clasificar las enfermedades y otros problemas de salud en muchos tipos de registros, incluyendo certificados de defunción y registros de salud. Además de permitir el almacenamiento y recuperación de información de diagnósticos para ser usados en procedimientos clínicos, con fines epidemiológicos y de calidad, estos registros sirven para la elaboración de estadísticas nacionales de mortalidad y morbilidad[4] en los estados miembros de la OMS (Organización Mundial de la Salud). La versión más extendida se conoce como ICD-10, la 10° revisión. Otro ejemplo es SNOMED CT (*Systematized Nomenclature of Medicine - Clinical Terms*, Nomenclatura Sistematizada de Medicina - Condiciones clínicas). Éste es un vocabulario estándar de terminología de salud con términos del dominio clínico. UMLS (Sistema Unificado de Lenguaje Médico) es un sistema compuesto de archivos y software que reúnen diversos vocabularios de salud, biomédicos y normas para permitir la interoperabilidad entre sistemas informáticos. RadLex es un léxico centrado sólo en términos de radiología, recopilados por la RSNA (*Radiological Society of North America*, Sociedad Radiológica de América del Norte). SNOMED CT, UMLS y la ICD-10 están disponibles en español, RadLex sólo está disponible en inglés y en una versión reducida en alemán.

En la implementación original de NegEx [5] se utiliza UMLS para detectar términos de

[3] Especificaciones formales explícitas de los términos en el dominio y las relaciones entre ellos [18].

[4] Cantidad de personas que enferman en un lugar y un período de tiempo determinados en relación con el total de la población.

interés. La adaptación a sueco de NegEx [40] utiliza UMLS, KSH97-P[5] y MeSH[6] (*Medical Subject Headings*, Encabezados de Temas Médicos). La adpatación de NegEx a español hecha por Costumero et al. [9] usa la terminología ICD-10.

Como se ve en los informes de ejemplo, los términos de interés (que denominaremos *findings*) de las oraciones del conjunto de datos fueron previamente etiquetados automáticamente con un algoritmo de extracción de información que utiliza RadLex como léxico[10]. El proceso de etiquetado (*tagging*) automático previo a la realización de este trabajo puede introducir errores (términos etiquetados que no son *findings* radiológicos o que son solo parte de uno). Sin embargo, el enfoque que se considerará en este trabajo es que la salida del experimento es determinar cuando un término dado está negado, independientemente de si el término etiquetado representa un *finding* o no.

Este conjunto de datos tiene la característica de poseer símbolos propios del lenguaje (como signos de interrogación, acentuaciones, entre otros) y errores de tipografía. Estos presentan inconvenientes para las herramientas de procesamiento del lenguaje a utilizar. Por este motivo se requiere de un pre-procesamiento para poder utilizarlas. A continuación se describen los símbolos que generan inconvenientes y el tratamiento aplicado para cada uno:

1. Símbolos reemplazados: el reemplazo de estos símbolos no cambia la interpretación de los informes que se tiene en los algoritmos.
   - ü por u
   - á, à, 'a, é, è, 'e, í, ì, 'i, ó, ò, 'o, ú, ù, 'u, por la vocal correspondiente sin acentuación
   - ç por c
   - ê por e
2. Símbolos reemplazados con pérdida de información: el reemplazo de estos símbolos cambia la interpretación de los informes que se tiene con algunos de los algoritmos. Sin embargo, algunas de las herramientas utilizadas por los algoritmos no pueden procesar este símbolo.
   - ñ por n
3. Símbolos eliminados sin pérdida de información: estos símbolos no aportan información lingüística para los algoritmos
   - ^{o}y $^{\circ}$
   - +
4. Símbolos eliminados con pérdida de información: estos símbolos en algunos casos brindan información relevante respecto al comienzo o fin de la oración, al alcance de las negaciones, o de la certidumbre que se tiene de las expresiones, y en otros

[5] Una adaptación sueca de ICD-10. Klassifikation av sjukdomar och hälsoproblem 1997 (KSH97), `http://www.socialstyrelsen.se/publikationer1997/1997-4-1`.

[6] Es un vocabulario terminológico controlado de la Biblioteca Nacional de Medicina (NLM) utilizado para publicaciones de artículos y libros de ciencia. Proporciona una terminología organizada jerárquicamente para la indexación y catalogación de la información biomédica, de distintas bases de datos de la NLM. `http://www.ncbi.nlm.nih.gov/mesh`

son errores de tipografía. Eliminar estos símbolos podría afectar los resultados. Estos casos fueron revisados en posteriores procesos manuales realizados (proceso de anotación), agregándose signos de puntuación de ser necesario.

- ¿ , ?

Además, se convirtió todo el conjunto de datos a minúscula, para considerar casos que son iguales pero están escritos distinto, como un mismo caso.

A partir de este conjunto procesado de datos se armaron dos subconjuntos, el de análisis para realizar experimentos, buscar alternativas, analizar errores y resultados para mejorar los algoritmos, y el de test para evaluar dichos algoritmos y comparar resultados con otros trabajos.

Para obtener los conjuntos de datos mencionados y que dichos conjuntos tengan características similares al corpus de NegEx (conjuntos de 1000 oraciones, de las cuales 500 oraciones contengan triggers y 500 oraciones no), se definieron dos criterios de selección de informes.

3.1.1. Conjunto de Análisis

A continuación se mencionan los pasos realizados para obtener el conjunto de análisis. Más adelante, se detalla cada uno de ellos. 1) Se ordenó de manera aleatoria el conjunto de datos. 2) Se seleccionó un subconjunto de informes, de los cuáles se extrajeron las oraciones mediante un proceso de tokenización. 3) De las oraciones obtenidas, se tomó un subconjunto de ellas, de manera tal que contuvieran *findings*. 4) Estas oraciones fueron luego anotadas por no expertos del dominio, generándose así el conjunto de análisis.

Para ordenar los informes de forma aleatoria se asignó un número aleatorio entre 0 y 1 a cada informe, y éste número se uso para ordenarlos. El número aleatorio fue generado usando la función *ALEATORIO()* que brinda el software Microsoft Excel starter 2010.

Para elegir el subconjunto de informes a los cuales extraerles oraciones se realizó un análisis manual de 10 informes, elegidos de forma aleatoria para analizar cuántas oraciones hay por informe, la cantidad de términos patológicos que hay en cada informe y la cantidad de negaciones. Usando esta información, se realizó una estimación de cuántos informes es necesario extraer del conjunto total de datos, para obtener un conjunto de 1000 oraciones luego de procesarlo, de las cuales aproximadamente 500 tengan una negación y aproximadamente 500 no contengan una negación. La tabla 3.1 muestra la composición de los 10 informes evaluados.

Del análisis de 10 informes se encontró que la cantidad total de oraciones es 66, la cantidad total de oraciones que contienen *findings* es 31 y que la cantidad total de *findings* en los informes es 35. Además, se encontró que la cantidad de negaciones que aparecen en los informes son 20 en 20 oraciones de las que contienen términos patológicos.

Debido a que solo interesa trabajar con oraciones que tienen *findings*, de las 66 oraciones encontradas en los 10 informes, solo se quieren las 31 con *findings*. De las 31, 11 no tienen negaciones, y 20 si contienen. Si las proporciones se mantienen, tomando 500 informes, aplicando los procesos correspondientes podrían obtenerse aproximadamente 500 oraciones sin negaciones, y 1000 oraciones con negaciones.

Por lo tanto, se extrajo del conjunto total un subconjunto de 500 informes.

Informe	#oraciones	#findings	#oraciones con findings
1	13	7	6
2	4	1	1
3	3	1	1
4	4	7	4
5	10	7	7
6	9	4	4
7	12	5	5
8	4	2	2
9	6	1	1
10	1	0	0
TOTAL	66	35	31

Tab. 3.1: Composición de 10 informes tomados de forma aleatoria del conjunto total de datos. Los *findings* corresponden a los términos detectados automáticamente por el algoritmo de detección de *findings*. El símbolo # denota cantidad.

A partir de estos informes se aplicó el proceso de tokenización o segmentación de oraciones y luego se extrajeron aquellas oraciones que contenían términos patológicos y fueron anotadas.

Para este proceso se utilizó la herramienta NLTK (Natural Language ToolKit) Loper and Bird [26] que provee la función *sentence tokenizer (sent_tokenice() function)*. Pero el resultado de aplicar esta función presenta fallas (es decir, que no todas las oraciones quedan bien segmentadas; por ejemplo, al aplicar la funión de tokenización de NLTK se podría obtener una oración que en verdad corresponde a dos oraciones) debido a dos tipos de problemas, uno de ellos se debe a errores en los datos de entrada (falta de puntuaciones, oraciones mal formadas, sintaxis incorrecta, por ejemplo), el otro se debe a la dificultad propia de la tarea de establecer cuándo un signo de puntuación representa el final de una oración, y cuándo no (un signo de puntuación podría representar el separador decimal de un número, por ejemplo). Por esta razón al finalizar el proceso, las frases obtenidas no siempre corresponden a una única oración.

Para solventar este problema, se analizó cada oración al momento de ser anotada y en caso de encontrar una frase que contenían más de una oración se separó en oraciones manualmente. Se descartaron las oraciones que no contenían términos de interés.

De las oraciones obtenidas, se extrajeron las primeras 1160 oraciones para proceder al proceso de anotación. Para este proceso, dos anotadores, a los que llamaremos no expertos en el dominio médico, anotaron las oraciones como *Affirmed* cuando es posible inferir de la misma que el *finding* se encuentra presente en el paciente, o *Negated* si el *finding* está ausente.

La anotación se realizó en dos partes. En primer lugar, de las 1160 oraciones se tomaron 160 para definir y revisar el criterio de anotación. Se analizó y refinó el criterio. Luego se procedió a anotar las 1000 oraciones restantes, que conformarían el corpus de análisis (Las 160 oraciones usadas para refinar el criterio no se incluyeron en el corpus de análisis).

Para la revisión del criterio, cada anotador anotó 100 oraciones, 60 oraciones fueron anotadas por el primer anotador, 60 por el segundo y 40 oraciones fueron anotadas por

ambos, para poder calcular el *Inter Rater Agreement* (IRA, Grado de Acuerdo entre Anotadores) [27] entre los anotadores de manera tal de poder medir el nivel de acuerdo entre ellos (Más adelante, en la sección 3.6.1 se explica en detalle este cálculo). Para la segunda etapa, cada anotador individualmente anotó 375 oraciones, y 250 fueron anotadas por ambos.

Debido al ruido del corpus, luego del proceso de extracción y anotación de oraciones se descartaron algunas de ellas (21), por resultar mal formadas o mal anotadas, resultando así un conjunto de 979 oraciones. Este contiene 571 oraciones que fueron reportadas como *Affirmed* según los anotadores, y 408 como *Negated* y conforma el conjunto de análisis.

Este corpus resulta de no muy buena calidad ya que:

- Las términos etiquetados como *findings* no siempre son verdaderos *findings*. Estas fallas provienen del algoritmo usado para la detección de *findings*.
- Hay muchas oraciones repetidas textualmente o muy parecidas. Con lo cual se reducen mucho las oraciones distintas.

A pesar de lo mencionado, este corpus sirve para analizar los algoritmos propuestos, realizar pruebas, detectar errores en las implementaciones, evaluar oportunidades de mejora, y fue utilizado para este fin.

3.1.2. Conjunto de Test

El objetivo final es contar con 1000 oraciones que mencionen algún *finding*, 500 con una negación, 500 sin negaciones. Para identificar las oraciones con negaciones se utilizó como base el trabajo de Chapman et al. [5] (se extrajeron manualmente del conjunto de análisis frases que denotan negaciones y se buscaron estos términos en el nuevo conjunto de datos). Este paso requiere que se examine el corpus de análisis, para detectar los triggers (términos que denotan negación) que aparecen en ese conjunto, y extraer del nuevo corpus 500 oraciones que tengan alguno de esos triggers, y 500 oraciones que no tengan ninguno de ellos.

Los pasos realizados para el armado del conjunto de test fueron:

1. segmentación en oraciones (tokenización): los informes fueron separados en oraciones usando la función de tokenización de NLTK. A partir de los *findings* etiqeutados en los informes, obtenidos del algoritmo de detección de *findings*, solo se consideró como *finding* el primero que aparece en la oración. Así, cada oración que tiene un *finding* queda asociada a un único *finding*.

2. eliminación de oraciones sin términos de interés: de todas las oraciones obtenidas en el paso anterior, algunas no contienen ningún *finding* etiquetado, con lo cual no resultan de interés para este trabajo. Por este motivo fueron descartadas. El resultado son solo oraciones que tienen etiquetado un único *finding*[7].

3. clasificación de oraciones en oraciones con negaciones y oraciones sin negaciones. El conjunto de negaciones usado para realizar la clasificación fue obtenido manualmente del corpus de análisis. Estas son:

[7] Por este motivo en el resto del trabajo se hará referencia al *finding* de la oración, y deberá interpretarse como el primer *finding* que aparece en la oración.

- no,
- no se detectaron,
- no se observo,
- no se detecto,
- sin,
- no se visualizan,
- no se observa,
- ni,
- no se observan,
- no se identifica,
- no se visualiza,
- no observandose,
- no presenta,
- no se lograron individualizar y
- no se pudo evaluar

4. eliminación de oraciones repetidas: Muchas oraciones aparecen repetidas textualmente o son muy similares. Manualmente se listaron aquellas que se encontraron repetidas y se filtraron las repetidas usando expresiones regulares.

5. eliminación de oraciones con términos incorrectamente etiquetados como *findings*. Estos términos mal etiquetados fueron identificados manualmente. Con un algoritmo se descartaron del conjunto de datos todas las oraciones que tenían asociado alguno de estos *findings*.

6. limpieza de *findings*: el uso del algoritmo de detección de *findings* para extraer los *findings* de las oraciones nos daba como resultado un lema del *finding* que aparece en la oración. Se transformó el *finding* a la expresión textual usada en la oración. Por ejemplo, en la oración *No se detectaron* ***adenomegalias***, en donde el *finding* es *adenomegalias*, pero el término etiquetado por el algoritmo de detección de *findings* es *adenomegalia*. Para este caso, el *finding* obtenido del algoritmo se transforma a *adenomegalias*.

7. mejora de oraciones mal segmentadas. Se analizaron manualmente algunos casos en los que las oraciones estaban mal segmentadas. Se detectaron los siguientes patrones de falla:

   - El signo de puntuación correspondiente a una oración está separado de la última palabra de la oración y sin espacio de la primer palabra de la siguiente oración: *Primer oración .Segunda oración.*
   - No hay espacio entre la última palabra de la primer oración, el signo de puntuación y la primer palabra de la segunda oración: *Primer oración.Segunda oración.*
   - No hay signo de puntuación: *Primer oración Segunda oración*

los primero dos casos se corrigieron usando expresiones regulares. Para el último caso se realizó análisis y corrección manual en una etapa posterior.

8. ordenamiento aleatorio de oraciones. Las oraciones fueron ordenadas de forma aleatoria, para luego poder obtener las 1000 oraciones que conformen el conjunto de test.

3.1.3. Gold Standard

Para el armado del Gold Standard, tres anotadores: un experto en el dominio médico (anotador 3), y dos no expertos en el ámbito médico (anotador 1 y anotador 2), realizaron un proceso de anotación sobre el conjunto de datos de test. Luego se calculó el *Inter Rater Agreement* (IRA) usando el coeficiente de Kappa [27], para medir el grado de acuerdo entre los anotadores.

Para el proceso de anotación se estableció un criterio de anotación, y se realizó una prueba de anotación para revisarlo. Luego de refinar el mismo se procedió a la anotación.

El criterio consiste en anotar cada oración en alguna de las siguientes categorías:

- ***Affirmed:*** La oración afirma que el *finding* asociado se encuentra presente en el paciente.
- ***Negated:*** De la oración se entiende que el *finding* asociado no está presente en el paciente.
- ***Probable:*** En la oración el *finding* asociado no está ni afirmado ni negado, sino que denota una probabilidad.
- ***Doubt:*** Corresponde a los siguientes casos: 1) para el anotador no está claro si el paciente tuvo el síntoma o enfermedad (*finding*), 2) no se puede determinar a partir de la oración si lo tuvo o no, 3) el *finding* no hace referencia al paciente o, 4) el *finding* no hace referencia a una condición presente del paciente (por ejemplo si se habla en pasado).
- ***-:*** El término asociado a la oración no es verdaderamente un *finding*.

El armado del Gold Standard se hizo sobre el conjunto de test, compuesto de 1000 oraciones. Cada anotador anotó 400 oraciones, de la siguiente forma: 100 oraciones fueron anotadas por el anotador 1 (no experto) y el anotador 3 (experto en dominio médico), otras 100 oraciones fueron anotadas por el anotador 2 (no experto) y el anotador 3. Luego, 300 oraciones fueron anotadas solo por el anotador 1, otras 300 oraciones solo fueron anotadas por el anotador 2 y las 200 oraciones restantes fueron anotadas solo por el anotador 3. La figura 3.1 ilustra los conjuntos de oraciones anotadas por cada anotador.
Las oraciones anotadas con el símbolo "-" se descartan debido a que se considera que no tiene sentido evaluarlas. Se incorporaron nuevas oraciones, tantas como fueron descartadas, para completar el conjunto de 1000 oraciones, y estas fueron anotadas por los anotadores 1 y 2.

Las oraciones anotadas como *Probable*, se consideraron como afirmadas, ya que se cree que a los médicos les interesará analizarlas (se cambió la anotación de *Probable* a *Affirmed*).

Para las oraciones que fueron anotadas por dos anotadores (el experto en el dominio médico o anotador 3 y otro de los anotadores), se evaluaron los casos de desacuerdo en la

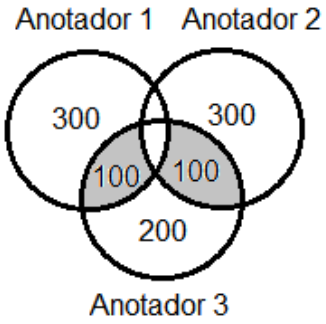

Fig. 3.1: Diagrama de Venn que muestra el conjunto de oraciones anotada por cada anotador. Anotador 1 y 2 son los no expertos, el anotador 3 es el experto en dominio médico. Cada anotador anotó en total 400 oraciones.

anotación. El anotador 3 revisó la oración y volvió a anotarla. Luego de este proceso, se optó por utilizar para el Gold Standard la anotación del anotador 3.

El conjunto resultante está compuesto por 487 oraciones que contienen alguno de los términos que denotan negación extraídos del conjunto de análisis y 513 oraciones sin ninguno de estos términos. De las 1000 oraciones 775 fueron anotadas como *Affirmed* por los anotadores y 225 fueron anotadas como *Negated*.

3.1.4. Otro conjunto de datos

Costumero et al. [9] hicieron una adaptación de NegEx para la detección de negaciones en los registros clínicos escritos en español. El conjunto de datos utilizado por ellos fue extraído de SciElo[8] [35] utilizando secciones tituladas "Reporte de caso", "A propósito de un caso" y "Caso clínico", entre otros. Los artículos obtenidos de SciElo son textos científicos escritos con lenguaje formal, extensos, contienen imágenes, se dividen en secciones (por ejemplo: introducción, caso clínico, discusión, y conclusiones) y el formato del texto no es estructurado. De estos artículos, los autores de la adaptación de NegEx extrajeron 500 informes que tienen 422 oraciones diferentes. Los *findings* se detectaron usando la terminología ICD-10. Su adaptación de NegEx consiste en adaptar los triggers de NegEx al español. Para esto utilizaron los triggers de negación de NegEx [5] y ConText [19] (estos trigger están en inglés) y un grupo multidisciplinario (de informáticos, médicos y lingüistas computacionales) los tradujeron al español. Los términos resultantes fueron enriquecidos con una recopilación de términos que denotan negaciones, extraídas de un subconjunto de los informes clínicos, los cuales fueron ampliados con sinónimos conocidos. Los términos resultantes fueron clasificados con las mismas etiquetas de NegEx (*pre-negation, post-negation*, etc). No modificaron el algoritmo NegEx.

Con el propósito de comparar el enfoque propuesto en este trabajo y la implementación de NegEx para español de Costumero et al. [9] se obtuvo el corpus de datos utilizado por éstos, a través de comunicación personal con los autores. Como el conjunto de datos utilizado para este trabajo aún no está disponible, es decir, aún no es público, y el conjunto de SciElo es de libre acceso, tener este conjunto de datos permite comparar los algoritmos propuestos en el presente trabajo con otros algoritmos.

[8] La biblioteca virtual formada por una colección de revistas científicas, descrita en la sección 2.1.

3.2. Algoritmo original: NegEx

NegEx es un algoritmo pensado originalmente para detectar negaciones en informes clínicos con *findings* etiquetados, que utiliza expresiones regulares para decidir si un *finding* está presente o ausente en un paciente de acuerdo a la descripción del informe médico, como se mencionó en la sección 2. El algoritmo toma como entrada una oración con un término de interés (un *finding*) etiquetado.

Este busca en la oración triggers de negación (términos que usualmente se utilizan para denotar negación), por ejemplo *no hay señales de*. El algoritmo verifica si el *finding* se encuentra dentro del alcance del trigger de negación usando distancia de palabras entre el *finding* y el trigger e información de la ubicación de los mismos en la oración.

En la versión original NegEx [5], se identificaron 35 triggers de negación y se dividieron en dos grupos (*pseudo-negation* y *negation*). Una extensión NegEx Chapman et al. [7] añade nuevos grupos (*termination terms* y *probable-negation*). Se utiliza una etiqueta para clasificar cada trigger en la clase correspondiente. Los triggers de NegEx y sus etiquetas son los siguientes:

- *negation triggers* (triggers de negación): indica una negación, etiquetados como PREN y POST según si aparecen antes o después del termino que están negando respectivamente. Por ejemplo *denies* (niega)
- *pseudo-negation triggers* (triggers de psuedo-negación): triggers de negación pero que no niegan la condición clínica, etiquetados como PSEU. Por ejemplo *no increase* (no hay aumento)
- *termination terms* (términos de terminación): determinan el alcance de un trigger de negación, etiquetados como CONJ. Por ejemplo *but* (pero)
- *probable negation triggers* (triggers de probables negaciones): triggers que indican posibilidad de negación, etiquetados como PREP y POSP según si aparecen antes o después del término que podrían estar negando. Por ejemplo *can be ruled out* (se puede descartar)

El algoritmo 1 describe la implementación original NegEx en pseudo código[9].
A partir del algoritmo 1 y del código en python obtenido de la página web de NegEx, se deducen los siguientes comportamientos:

- Si un *finding* aparece más de una vez en la oración, y alguna de las apariciones está negada, el algoritmo asume que todas las apariciones del *finding* están negadas.
- Si un trigger aparece en el listado de triggers más de una vez, pero con distinta etiqueta, el algoritmo siempre utilizará la etiqueta por la que se evalúa primero. Debido a que primero se chequea por la etiqueta PREN, esta tiene precedencia sobre la etiqueta POST. Similarmente, POST tiene precedencia sobre PREP, y PREP sobre POSP.
- Si un trigger de etiqueta PSEU aparece en una oración, y el *finding* queda dentro del alcance de este trigger, el algoritmo determina que el *finding* está afirmado.

[9] Los detalles del algoritmo fueron tomados de la página web de NegEx `http://code.google.com/p/negex`.

Algoritmo 1 Algoritmo original NegEx

```
for cada oración do
    for cada negation trigger (Neg1) do
        if Neg1 es un pseudo-negation trigger then
            Ir al siguiente negation trigger en la oración
        else if Neg1 es un pre-negation trigger then
            // Definir el alcance de Neg1 hacia adelante
            if Si se encuentra (un termination term or
                                otro negation o pseudo-negation trigger or
                                el final de la oración) then
                Terminar el alcance de Neg1
            end if
        else if Neg1 es un post-negation trigger then
            Definir el alcance de Neg1 hacia atrás basado en la distancia de palabras
        end if
    end for
end for
```

- Si un trigger con etiqueta PREP o POSP aparece asociado al *finding*, el algoritmo determina que el *finding* está afirmado.

3.3. Adaptación NegEx

Uno de los objetivos del presente trabajo es adaptar NegEx al español, para poder aplicarlo al conjunto de datos del dominio radiológico. En el trabajo de Costumero et al. [9] se expone una adaptación al español, sin embargo, como describen los autores, si se mejoran las reglas para determinar negaciones en dicha adaptación, deberían mejorar los resultados obtenidos hasta ahora. En base al desarrollo realizado por Costumero et al. [9] y otros trabajos previos, se elaboró una nueva adaptación de NegEx para español, para intentar lograr mejorar los resultados. Esta adaptación incluye la traducción de los triggers usados en la versión de NegEx para inglés, la revisión de estos y la adaptación de sus etiquetas para español, nuevas propuestas de conjuntos de triggers, específicos para el dominio radiológico y otros genéricos para ser usados en otros dominios, y adecuación del algoritmo que evalúa las oraciones. A continuación se describe la nueva adaptación.

3.3.1. Triggers

Siguiendo un enfoque similar al de Costumero et al. [9] y Skeppstedt [40] para realizar la adaptación de NegEx a otro idioma, se tradujeron los triggers de NegEx[10] de inglés a español usando Google Translate[11]. Debido a que esta traducción es automática, algunas de las traducciones obtenidas son incorrectas. En algunos otros casos, hay traducciones que son correctas pero que en español no representan triggers de negación. Además la gramática de inglés difiere con la de español en cuanto a género (en inglés hay uno solo, en español femenino y masculino) y número. Por lo tanto, en algunos casos para un mismo trigger en inglés, existe más de un trigger en español de acuerdo al género y número.

[10] `https://code.google.com/p/negex/`
[11] `https://www.translate.google.com/`

Otro aspecto a tener en cuenta es que una misma frase de negación en inglés puede tener más de una traducción y en algunos casos distintas frases en inglés tienen la misma traducción en español. Las traducciones automáticas fueron revisadas por no expertos en el dominio lingüístico, quienes corrigieron las traducciones incorrectas, descartaron las que no representan triggers de negación en español y las repeticiones, incorporaron nuevos triggers según género y número, y para los triggers que tienen más de una traducción a español, agregaron las traducciones que consideraron adecuadas.

El conjunto resultante fue de 210 triggers traducidos. A lo largo de este trabajo llamaremos a este conjunto de triggers *conjunto de traducciones* (en las tablas, se mencionará como *traducciones*).

Sin embargo, al analizar estos triggers, es evidente que faltan triggers relevantes para el dominio (por ejemplo *no se visualizan*, que está relacionado con el tipo de informe). Esto es debido a que las frases más utilizadas por médicos para denotar negaciones en inglés son distintas a las utilizadas por los médicos cuando escriben informes en español, sumado a que los triggers traducidos provienen de informes de otro tipo (no radiológicos). Por lo que se decidió incorporar nuevos triggers para español, provenientes de dos fuentes distintas (del corpus del dominio radiológico y de un médico radiológico).

Del conjunto de análisis, se obtuvieron los bigramas y trigramas. A partir de ellos, se buscaron posibles triggers obteniendo los bigramas y trigramas cuya primer palabra fuera *no*. El conjunto de frases resultantes se analizó manualmente y se descartaron aquellas frases que no corresponden a triggers, por ejemplo trigramas que contienen un *finding*, como los trigramas (*no, dilatada, vesicula*) y (*ni, liquido, libre*). De este proceso se obtuvieron 94 triggers. Este conjunto de triggers lo denominaremos en lo que resta del trabajo como *conjunto de bi-trigramas* (en las tablas se referenciará como *bi-trigramas* solamente).

Por otro lado, un médico radiólogo proveyó una lista de triggers frecuentemente utilizados en el dominio. Esta lista consiste de 57 triggers, y lo mencionaremos en lo que sigue como *conjunto radiológico* (en las tablas se llamará solamente *radiológico*).

Cabe destacar que entre los tres conjuntos (*conjunto de traducciones, conjunto de bi-trigramas* y *conjunto radiológico*) hay triggers repetidos.

Con estos tres conjuntos, se construyó un nuevo conjunto que contiene a todos los triggers (las repeticiones fueron eliminadas).

Es importante notar que en trabajos previos [5, 40] se menciona que la *precision* del trigger *no* es bastante baja. Debido a que este trigger se considera una negación muy frecuentemente usada en español, se hicieron pruebas para evaluar la posibilidad de incorporarlo en este nuevo conjunto de triggers que contiene a los tres conjuntos anteriores y estudiar la etiqueta más adecuada para el mismo (las pruebas con este trigger se describirán más adelante). Así conformado, a este conjunto lo llamaremos *conjunto compilado* (en las tablas *compilado*) y contando al trigger *no*, consiste de 350 triggers únicos.

Chapman et al. [7] observan que para los triggers de negación ocurre que algunos de ellos aparecen un gran número de veces en los textos, y un gran número de frases de negación se utilizan muy pocas veces. Por este motivo, se analizó la cantidad de veces que se utiliza cada trigger en los informes, lo que permite evaluar junto con los resultados, la utilidad de cada conjunto de triggers. A partir de la frecuencia de aparición de los triggers, se evaluaron posibles subconjuntos de triggers que podrían mejorar la perfomance de la adaptación de NegEx para español. Este análisis se explica en la sección 4.

3.3.2. Etiquetas

Cada trigger tiene una etiqueta asociada que se utiliza para describir la función del mismo en la oración. Las etiquetas posibles que se utilizan en la versión de NegEx para inglés son:

- PREN: Trigger que indica negación y precede al *finding* en la oración (por ejemplo: *no se evidencia, no se observa*)
- POST: Trigger que indica negación y aparece después del *finding* en la oración (por ejemplo: *negado, tiene que ser descartado*)
- PREP: Trigger que indica posible negación y precede al *finding* en la oración (por ejemplo: *habría que descartar, no se correspondería*)
- POSP: Trigger que indica posible negación y aparece después del *finding* en la oración (por ejemplo: *podría ser descartado*)
- PSEU: Trigger que indica pseudo-negación y puede preceder al *finding* o aparecer después del él en la oración (por ejemplo: *disminuye, no se incrementa*)
- CONJ: Trigger que indica conjunción o terminación (por ejemplo: *pero, aunque*)

Al traducir los triggers de inglés a español, estos no necesariamente cumplen la misma función dentro de la oración. En algunos casos, la etiqueta para un trigger en inglés no es la adecuada para su traducción, por lo que las etiquetas de las traducciones fueron revisadas y corregidas. Para esto se consideró el comportamiento deducido del algoritmo y las etiquetas explicados en la sección 3.2.

Además, se observó que el algoritmo NegEx para inglés no funciona bien con el trigger *ni* con las etiquetas antes mencionadas. Por lo tanto a la versión en español se le incorporó una nueva etiqueta para el conjunto de triggers, a la que llamamos CONJN (de Conjunción Negada) para denotar triggers que indican una conjunción negada.

También se consideraron algunos casos particulares para evaluar las etiquetas a usar en esos casos.

Cuando hay triggers que están expresados en pasado, hacen referencia a la historia del paciente y no al estado actual, con lo cual no podríamos afirmar o negar si el paciente tiene o no tiene el *finding*. Suponemos que al médico no le interesa analizar ese caso, por lo cual quisiéramos descartar la oración, por lo tanto queremos que el trigger tenga asignado una etiqueta de negación. Por ejemplo, *El paciente* ***no tenía*** *gripe.*, *no tenía* está en pasado. En este caso la etiqueta correspondiente para el trigger será PREN (si se le asignara PSEU, el algoritmo determinaría que el *finding* está afirmado).

En el caso que un trigger denote una expresión en futuro, implica que no hay certeza de que el *finding* esté ausente o presente en el paciente, aunque hay algún signo que podría indicar que lo tiene. Como no hay certeza, es un caso que podría resultar de interés para ser analizado por un médico, por eso se busca que el algoritmo determine que el *finding* está afirmado. Por ejemplo, *Se le receta análisis* ***para excluir*** *alergia.*, *para excluir* indica que no hay certeza de que el *finding* esté ausente o presente en el paciente.

3.3.3. Algoritmo

Para la adaptación al español del algoritmo, se tomó la implementación en python que se obtiene de la página web de NegEx[12]. Se modificó el código en dos líneas para contemplar el caso de un trigger que contenga la etiqueta CONJN. A este algoritmo lo llamaremos en el resto del trabajo *NegExMod*. Debido a que CONJN se utiliza para conjunciones negadas, si un trigger con esta etiqueta aparece en una oración, es porque antes en la oración hay un trigger de etiqueta PREN. Por lo tanto, la etiqueta CONJN se verifica en el caso de haber encontrado un trigger PREN antes. Esta fue la implementación utilizada para detectar negaciones en textos médicos escritos en español.

3.4. Método basado en PosTagging

Una alternativa propuesta para la detección de negaciones en informes radiológicos escritos en español, distinta al enfoque de NegEx, es buscar, detectar y analizar patrones en la forma de las oraciones que indiquen posibles negaciones de un *finding* y desarrollar un algoritmo que detecte estos patrones en las oraciones.

Para definir los patrones se propusieron algunos como base, de acuerdo al conocimiento que se tiene del idioma español. A partir de éstos se analizó la estructura de un subconjunto de oraciones usando información morfológica de las palabras de cada oración en busca de los patrones propuestos (o similares) con el objetivo de encontrar los componentes que forman la negación de la oración. Usando estos componentes se formaron reglas para detectar los patrones, y así establecer si el *finding* de una oración está negado o no.

Los patrones propuestos fueron:

- no ... verbo ... *finding*
- sin ... *finding*
- ni ... *finding*

Se cree que en una oración en la que aparecen esos componentes, el *finding* de la misma está negado, donde "..." indica que pueden haber 0 o más palabras entre medio.

El análisis de la estructura de las oraciones se realizó sobre 50 oraciones del conjunto de análisis anotadas como *Negated*. Para determinar los componentes que forman una negación en una oración, se utiliza FreeLing [36], un conjunto de herramientas de análisis del lenguaje, como *Morphological Analysis, Part-of-Speech tagging (PoS-tagging), Named Entity Recognition, Shallow Parsing*, entre otros, para varios idiomas, entre ellos, español (su diccionario está basado en EuroWordNet [46], una ontología léxica multilingüe).

En particular, se utiliza el analizador de FreeLing para obtener el *Part-of-Speech tag* (PoS-tag) de las palabras de cada oración. Anteriormente se definió PoS-tag. El manual de usuario de FreeLing describe al PoS-tag de la siguiente manera:

> El PoS-tag es un código utilizado para codificar información morfológica. Este código está basado en la codificación propuesta por EAGLES[13]. Los PoS-tags propuestos por EAGLES consisten en etiquetas de longitud variable donde

[12] `http://code.google.com/p/negex`

[13] EAGLES busca codificar todas las características morfológicas existentes para la mayoría de los lenguajes europeos. `http://www.ilc.cnr.it/EAGLES96/home.html`

cada caracter corresponde a una característica morfológica. El primer caracter en la etiqueta siempre es la categoría (PoS). La categoría determina la longitud de la etiqueta y la interpretación de cada caracter en la etiqueta. Por ejemplo, para la categoría sustantivo se podría tener la definición:

Posición	Atributo	Valor
0	categoría	N: *noun*
1	tipo	C: *common*; P: *proper*
2	género	F: *female*; M: *male*; C: *common*
3	número	S: *singular*; P: *plural*; N: *invariable*

Tab. 3.2: Definición de PoS-tag de la categoría sustantivo según EAGLES, FreeLing User Manual, Tagset for Spanish [16], donde los significados de las palabras en inglés son los siguientes: *noun* es sutantivo, *common* es común, *proper* es propio, *female* es femenino, *male* es masculino, *singular* es singular, *plural* es plural e *invariable* es invariable.

> Esto permitiría PoS-tags como NCMS (para sustantivo - común - masculino - singular). Para las características que no son aplicables o no están especificadas para una palabra en particular se utiliza el 0 (cero). Por ejemplo, la etiqueta NC00 corresponde a un sustantivo - común - género no especificado - número no especificado. (FreeLing User Manual, Tagset for Spanish [16])

De la misma manera, se definen etiquetas para las distintas categorías (PoS). En el manual de FreeLing[14] se definen los PoS-tags utilizados por dicha herramienta.

El analizador de FreeLing toma como entrada el texto a analizar. Y como salida devuelve el texto analizado. Por ejemplo, dada la oración *No se detectaron adenomegalias.* la salida del analizador de FreeLing es:

```
No no RN
se se P00CN000
detectaron detectar VMIS3P0
adenomegalias adenomegalia NCFP000
. . Fp
```

Donde cada linea tiene el análisis de cada palabra según el orden en el que aparece en la oración, así la primer línea corresponde a la primer palabra, etc. Y cada línea contiene tres columnas. La primera corresponde a la palabra tal como aparece en la oración, la segunda corresponde al lema de la palabra y la tercera corresponde al PoS-tag de la palabra.

Para lograr el objetivo propuesto de este método se requiere conocer los PoS-tags de las palabras *no, sin, ni*, de los *verbos* y saber cuáles son los *findings* que fueron obtenidos del manual del usuario de FreeLing [16]:

- no: RN (la etiqueta de adverbio negativo (RN) está reservada exclusivamente para el adverbio no)

[14] `https://talp-upc.gitbooks.io/freeling-user-manual/content/tagsets/tagset-es.html`

- sin: SPS00 (corresponde a la categoría de Adposición del tipo Preposición, forma simple)
- ni: CC (pertenece a la categoría Conjunciones coordinadas)
- verbos: existe una categoría para los verbos, y tienen asignado el caracter V. Por lo tanto todos los verbos tienen PoS-tag cuyo primer caracter es V.
- *findings*: es útil notar que los *findings* siempre corresponden a sustantivos, y la categoría sustantivos se denota con el carcater N.

Para este análisis es conveniente tener en lugar de la oración compuesta de palabras, una oración compuesta por los PoS-tags de las palabras correspondientes a la oración. Por ejemplo, para *No se detectaron adenomegalias.* tendríamos RN P00CN000 VMIS3P0 NCFP000 Fp.

De esta forma, tomando como *finding* las *adenomegalias*, podríamos analizar si se cumple el patrón propuesto "no ... verbo ... *finding*" buscando que la oración de PoS-tags contenga RN (por no), algún PoS-tag que comience con V (de verbo) y un PoS-tag que empiece con N (de sustantivo, por el *finding*). La oración contiene **RN**, **V**MIS3P0 y **N**CFP000. Esta oración cumple el patrón propuesto y se puede dictaminar que el *finding* de la oración está negado debido a que según la oración el paciente no presenta adenomegalias. Por lo cual, se tiene una oración que cumple el patrón y su *finding* está negado. El formato de la información resultó útil.

Sin embargo este enfoque presenta inconvenientes, por que si bien el PoS-tag RN está reservado exclusivamente para *no*, no ocurre lo mismo con SPS00 para *sin*(la preposición *con*, que denota lo contrario a *sin*, lleva el mismo PoS-tag), como tampoco CC para *ni* (a la conjunción *y* que expresa la idea opuesta a *ni*, le corresponde el mismo PoS-tag). Además, con los *findings* como sustantivos tenemos inconvenientes ya que pueden haber muchos sustantivos en una oración y no sabríamos cual corresponde al *finding*.

Por ejemplo para la oración *Via biliar intrahepatica visible con dilatación de coledoco.*, cuyo *finding* es dilatación y cuya oración de PoS-tags es NCFS000 AQ0CS0 NCFS000 AQ0CS0 SPS00 NP00000 SPS00 RG Fp, se podría decir que se encuentra el patrón "sin ... *finding*" por contener el PoS-tag de *sin* (SPS00) (que en este caso corresponde a *con*)y un sustantivo correspondiente al *finding*: (NP00000) NCFS000 AQ0CS0 NCFS000 AQ0CS0 **SPS00 NP00000** SPS00 RG Fp, y de esta forma dictaminar que el *finding* de la oración está negado, es decir que no hay dilatación, sin embargo ocurre lo contrario, hay dilatación.

Pero se pueden evitar estos problemas haciendo modificaciones simples a la propuesta anterior. Para la oración a analizar, se reemplaza cada palabra por su correspondiente PoS-tag exceptuando las palabras sin y ni. Además, el *finding* lo reemplazaremos por la etiqueta FINDING. De esta forma se puede continuar utilizando el análisis anterior pero sin inconvenientes. Y a las oraciones así formadas por PoS-tags, sin, ni y FINDINGS se mencionarán como *oraciones de PoS-tags*.

Utilizando el analizador de FreeLing y el formato propuesto para analizar las oraciones, se obtuvieron las oraciones de PoS-tags de las 50 oraciones, y se estudiaron manualmente en busca de los patrones propuestos inicialmente y nuevos patrones.

Los patrones encontrados fueron:

1. RN ... V ... FINDING (no ... verbo ... *finding*)
2. sin FINDING (sin *finding*)

3. RN FINDING (no *finding*)

4. RN ... V ... N ... ni ... FINDING (no ... verbo ... sustantivo ... ni ... *finding*)

Es decir que en el conjunto de análisis, muchas de las oraciones que tienen *findings* que fueron anotados como negados tienen esa estructura, donde "..." indica que puede haber 0 o más palabras en el medio (de cualquier categoría de palabra).

Utilizando esto, se construyó un algoritmo que verifica si, dada una oración de PoS-tags, cumple alguno de los cuatro patrones mencionados. Si una oración cumple alguno, el algoritmo determina que el *finding* de la oración está negado. De lo contrario, determina que está afirmado.

Se utilizó este algoritmo sobre el conjunto de análisis para evaluar que la implementación fuera correcta, así como para detectar fallas y posibles mejoras del algoritmo.

De los resultados obtenidos al aplicar el algoritmo se detectó que hay un patrón en la estructura de las oraciones que genera algunos falsos positivos: RN ... V ... por ... FINDING. Por ejemplo:

- En la oración *No se pueden evaluar ovarios por mala de ventana.* el *finding* es *mala*, y la oración está afirmada, pero debido al patrón no ... verbo ... *finding*, el algoritmo determina que está negada.

- En la oración *Pancreas: no evaluable por interposicion gaseosa.* el *finding* es *gaseosa* y ocurre lo mismo que en el ejemplo anterior.

Podría considerarse agregar al algoritmo la posibilidad de detectar este caso, y de encontrarse con una oración que lo cumpla, dictaminar que el *finding* de la oración está afirmado. Sin embargo, no es un patrón de negación.

Además, se observó que el patrón "no ... verbo ... *finding* ... ni ... *finding*" puede extenderse a "sin ... *finding* ... ni ... *finding*" y "no ... *finding* ... ni ... *finding*".

Otra cuestión a analizar es el uso de los paréntesis en el texto. Este factor es importante debido a que podría aportar información respecto al alcance de una negación (para este algoritmo las negaciones que consideramos son las que usan los patrones: *no, sin, ni.* Si la negación se encuentra dentro del paréntesis, pero el *finding* se encuentra fuera del mismo, dentro de la oración, la negación en ese caso no afecta al *finding*, y el paréntesis funciona como delimitador del alcance de la negación. Por ejemplo, en la oración *del análisis realizado (el paciente no presenta síntomas) se detecta enfermedad*, tenemos la negación *no*, el verbo *presenta*, y el *finding enfermedad*, cumpliendo el patrón "no ... verbo ... *finding*", sin embargo por los paréntesis sabemos que el alcance de *no presenta* terminó con el paréntesis y no afecta al *finding*. Con lo cual, el *finding* de la oración debería estar afirmado, pero con los patrones descritos hasta el momento, el algoritmo determinará que está negado.

En una oración que contiene paréntesis podría haber diversos casos evaluando en qué parte de la oración se encuentra (si es que hay) una negación y dónde se encuentra el *finding*, y dependiendo de la oración misma eso podría significar que está afirmada o negada según el contexto. Dada una oración que presenta una negación, las decisiones tomadas según los casos son:

- Si la negación y el *finding* están ambos dentro del paréntesis, entonces con ventana de 6 tokens evaluar si la negación afecta al *finding* según las reglas de negación establecidas.[15]
- Si la negación esta dentro del paréntesis, y el *finding* fuera, determinamos que la negación no afecta al *finding*.
- Si la negación está fuera del paréntesis y el *finding* dentro, determinamos que la negación no afecta al *finding*.

Los PoS-tag utilizados por FreeLing para los paréntesis son Fpa para el paréntesis de apertura y Fpt para el paréntesis de cierre. Por lo tanto al algoritmo de detección de patrones de negaciones se le incorporó la posibilidad de detectar si la oración contiene paréntesis, y que la salida del algoritmo en esos casos cumpla las decisiones anteriores.

Una diferencia entre este enfoque y NegEx, es que en los informes existen muchas oraciones que tienen palabras mal escritas. Esto en NegEx genera un problema importante ya que si la negación de la oración (trigger) está mal escrita no se detecta. Sin embargo, para esta propuesta, se corroboró que FreeLing obtiene el PoS-tag correcto para los verbos (aún estando mal escritos), debido a que lo que se considera del PoS-tag es que pertenezca a la categoría verbo independientemente de las otras características. Como ejemplos, se usaron las siguientes palabras: detecron, obsesrvo, bservan. Aunque en verdad hay otro problema: si las palabras "sin" o "ni" están mal escritas, pero es un problema mucho menor que el que presenta NegEx. Ya que se reduce a que se escriban mal dos palabras cortas.

3.5. Método basado en Shallow Parsing

Este enfoque se basa en la información sintáctica obtenida al aplicar la técnica de *shallow parsing* a las oraciones.

Como se estudió anteriormente, se denomina *shallow parsing* (análisis sintáctico superficial) a la tarea de asignar una estructura sintáctica parcial a oraciones. Este análisis identifica los componentes, pero no especifica su estructura interna, ni su papel en la oración principal. Para llevar a cabo esta tarea se requiere identificar frases sintácticas o palabras que participan en una relación sintáctica, es decir se requiere aplicar chunking [25, 32].

También se mencionó anteriormente la definición de *chunking*. Bird et al. [4] describen esta técnica de la siguiente manera:

> Chunking es una técnica que se usa frecuentemente para la tarea de reconocimiento de entidades (*entity recognition*). Los *chunkers* segmentan secuencias de palabras (tokens) y los etiquetan con el tipo de entidad apropiada. En la figura 3.2, las cajas mas pequeñas muestran la tokenización a nivel palabra y el resultado reducido (clase de palabra) de la aplicación de PoS-tagging, mientras que las cajas más grandes muestran los resultados de chunking, que son de más alto nivel. Cada una de estas cajas grandes se denomina *chunk*. En un texto, los *chunks* correspondientes no se superponen entre si. (Bird et al. [4])

[15] La elección de utilizar una ventana de 6 tokens se desprende de la implementación original de NegEx, en la que utilizan una ventana de 6 tokens para definir el alcance de una negación.

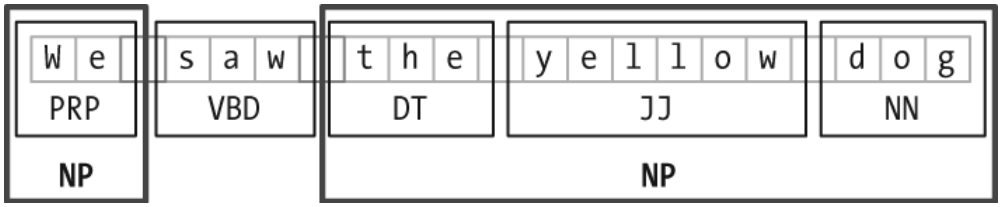

Fig. 3.2: Segmentación y etiquetado en dos niveles: Token, PoS-tag y Chunk. Este ejemplo fue tomado de Bird et al. [4]

El objetivo de utilizar esta técnica es determinar si dada una oración, ésta enuncia que el *finding* asociado a la misma se encuentra negado o no, según la información sintáctica (o según la estructura de la oración). A continuación se describen los pasos realizados para lograr esto, y más adelante se explican detalladamente: 1) A partir de los patrones definidos para la estrategia de PoS-tagging (y un patrón nuevo), se analizó la estructura de los árboles obtenidos al aplicar shallow parsing a oraciones de ejemplo que cumplen esos patrones, para definir la estructura de los patrones con este método. 2) Se aplicó la función de shallow parsing del analizador de FreeLing a cada oración del corpus de datos. 3) Se utilizó el árbol obtenido al aplicar shallow parsing para establecer si la oración determina que el *finding* está negado o no en base a los árboles obtenidos en el paso 1).

Cuando se aplica el analizador de FreeLing a un texto, se obtiene como salida un archivo de texto en donde se describe una estructura de árbol. En este árbol, los nodos más altos corresponden a los chunks, y las hojas se corresponden a los PoS-tags de las palabras de la oración. Por ejemplo, en la figura 3.3 se visualiza el árbol obtenido para la oración *No se detectó dilatación ureteral.*

```
S
  neg
    (No no RN -)
  grup-verb
    morfema-verbal
      (se se P00CN000 -)
    grup-verb
      verb
        (detecto detectar VMIP1S0 -)
  sn
    grup-nom-ms
      w-ms
        (FINDING finding NP00000 -)
      s-a-ms
        a-ms
          (ureteral ureteral AQ0CS0 -)
  F-term
    (. . Fp -)
```

Fig. 3.3: Esquema que muestra el árbol obtenido en la salida de texto al aplicar la función de shallow parsing del analizador de FreeLing en la oración *No se detecto* ***dilatacion*** *ureteral.*, cuyo *finding* fue reemplazado por la etiqueta FINDING.

En el árbol, cada nivel del árbol se representa con tabulaciones. El nodo raíz es S. Los nodos intermedios están etiquetados según el análisis realizado por el parser[16]. Las hojas

[16] La lista de significados de las etiquetas posibles de cada nodo usando la herramienta FreeLing se puede obtener de la página web `https://github.com/iknow/FreeLing/blob/master/doc/grammars/esCHUNKtags`.

del árbol contienen el PoS-tag de cada palabra de la oración. Notar que en el árbol, se mantiene el orden de las palabras según el orden en el que aparecen en la oración.

Para poder diseñar un algoritmo que utilice la información obtenida de aplicar shallow parsing, se necesita conocer el árbol obtenido con el analizador de FreeLing, para luego evaluar posibles implementaciones del algoritmo de detección de negaciones y establecer cuál es el formato que cumplen los patrones de negaciones.

Entonces se realizó una prueba de FreeLing utilizando la opción de shallow parsing del analizador, con oraciones que tengan negaciones que cumplen las reglas establecidas para hacer PoS-tagging. Las oraciones utilizadas para esta prueba fueron:

- No se detecto **dilatacion** ureteral. (Patrón 1)
- Higado: ligeramente heterogeneo en forma difusa, sin **lesiones** focales. (Patrón 2)
- VIA BILIAR intra y extrahepatica: no **dilatada**. (Patrón 3)
- No se detectaron **adenomegalias** ni **liquido** libre. (Patrón 4)

Para realizar esta prueba, los *findings* de cada oración (los resaltados) fueron reemplazados por la etiqueta *FINDING* (al igual que se hizo para la técnica de PoS-tagging). La etiqueta *FINDING* es considerada por FreeLing como un sustantivo. Se cree que esto no afecta a los resultados, debido a que se espera que los *findings* de las oraciones siempre sean sustantivos. En la sección 6.4, se presentan los árboles obtenidos con FreeLing para cada oración de esta prueba.

Utilizando los árboles generados con FreeLing se observó que éstos cumplen los patrones descritos a continuación[17]:

Patrón 1

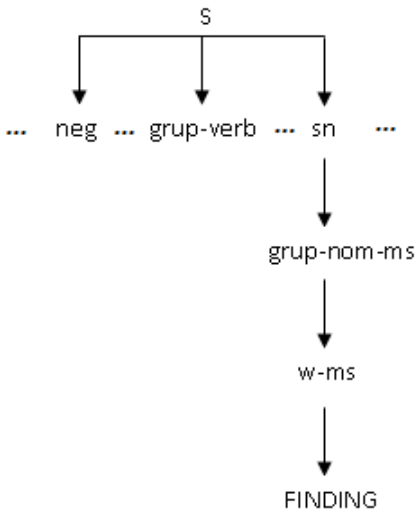

Fig. 3.4: Esquema que muestra el patrón *No...verbo...FINDING*

[17] Los puntos suspensivos indican que puede haber 0 o más ramas del árbol en el medio, correspondientes a otros chunks o tokens, que no están relacionados con el patrón.

Patrón 2

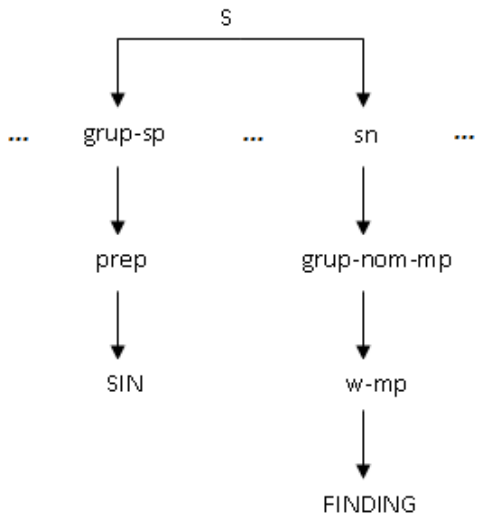

Fig. 3.5: Esquema que muestra el patrón *sin FINDING*

Patrón 3

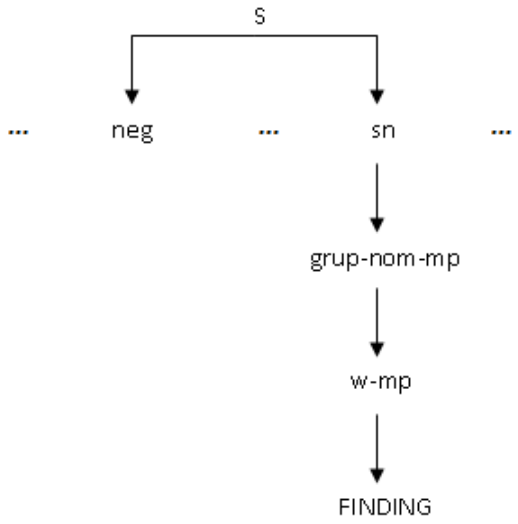

Fig. 3.6: Esquema que muestra el patrón *no FINDING*

Patrón 4

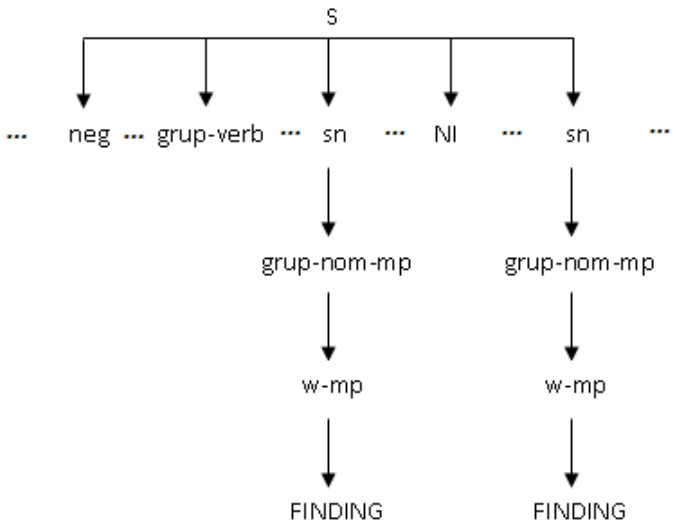

Fig. 3.7: Esquema que muestra el patrón *no...verbo...FINDING...ni...FINDING*

Patrón 5

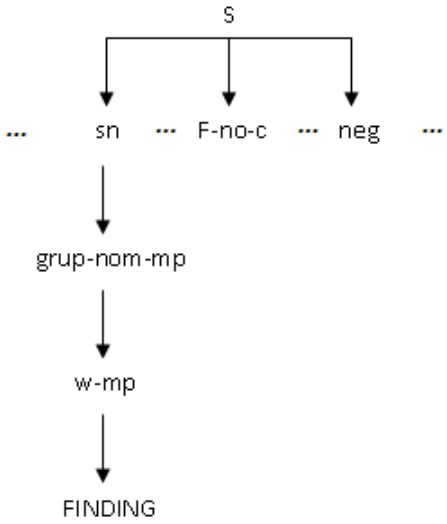

Fig. 3.8: Esquema que muestra el patrón *FINDING: no*

A partir de estos patrones se diseñó un algoritmo que, dada una oración, analiza el árbol que se obtiene de aplicarle la función de shallow parsing de FreeLing, y verifica si el árbol satisface alguno de los formatos que cumplen los patrones de negaciones, en cuyo caso determina que el *finding* de la oración está negado. Caso contrario establece que está afirmado. Luego se comparó el resultado con el Gold Standard.

La implementación del algoritmo consta de 3 pasos aplicados a cada oración del conjunto de datos. Los pasos son los siguientes:

1. ejecución de shallow parsing
2. armado de estructura de datos para representar el árbol
3. chequeo de patrones

3.5.1. Ejecución de shallow parsing

Dada una oración del conjunto de datos, se realizó el mismo proceso que con la técnica de PoS-tagging: se reemplazó el *finding* de la oración por la etiqueta *FINDING*, y luego utilizando el analizador de FreeLing se obtuvo el árbol de shallow parsing. La salida es texto que describe la estructura del árbol (como las oraciones de ejemplo), y se guarda en un archivo de texto.

3.5.2. Armado de estructura de datos para representar el árbol

Para poder analizar con mayor facilidad la salida obtenida de FreeLing, se armó una estructura de datos a partir del árbol representado con formato de texto. A continuación se describe la estructura construida. Cada nodo del árbol se representa con un array, que contiene información del mismo: nombre del chunk (o etiqueta) dado por FreeLing[18], la altura en la que se encuentra en el árbol, número de orden dado por la posición en la oración (se utilizó para poder armar la estructura correctamente, y para posteriores análisis), y los nodos de los hijos. Esquemáticamente, la estructura de datos es la siguiente:

[chunk, altura, orden, hijo 1, hijo 2, ...]

donde cada hijo tiene la misma estructura.

FreeLing tiene como salida un árbol en donde el primer nodo es etiquetado por *S*, y contiene a toda la oración (nodo raíz). Los nodos hoja son aquellos que no tienen hijos y por lo tanto tienen longitud 3. En la figura 3.3, se ve el nodo raíz S, que contiene a toda la oración, y los nodos hoja corresponden a los PoS-tag de las palabras de la oración.

3.5.3. Chequeo de patrones

El objetivo es determinar si las oraciones se encuentran negadas a partir de verificar si la oración cumple alguno de los patrones de negación detectados. Para esto se construye un algoritmo que etiqueta al *finding* de la oración como *negated* o *affirmed* según los patrones, y le asigna a la oración el número del patrón que cumple, si fue etiquetado como *negated*, o un tag que indica que no cumple ninguna regla si fue etiquetado como *affirmed*.

El algoritmo utiliza la estructura construida (árbol de arrays o simplemente árbol), y verifica si se cumplen los patrones antes descritos. Los patrones se verifican en orden, es decir, se chequea de a un patrón a la vez, primero el patrón 1, luego el patrón 2, y así siguiendo hasta verificar todos los patrones. Debido a esto, si el algoritmo determina que la oración está negada y la misma cumple más de un patrón, se reportará que el número

[18] `https://github.com/iknow/FreeLing/blob/master/doc/grammars/esCHUNKtags`

de patrón es el último patrón verificado por el cual la oración se encuentra negada. Esto es una decisión tomada, y puede afectar a los resultados que se obtengan, si es que falla.

Para verificar cada patrón, se recorre el árbol buscando los nodos de interés según el patrón que se esté analizando. Así, por ejemplo, para analizar si una oración cumple el patrón 1, se verificará si el nodo con etiqueta *S* tiene como hijos un nodo con etiqueta *neg*, uno con etiqueta *grup-verb*, y uno con etiqueta *sn*, en ese orden de aparición (para esto se utiliza el número de orden de cada nodo). Y, además, que el nodo etiquetado con *sn*, tenga como hijo un nodo etiquetado *grup-nom-ms*, y este a su vez tenga un nodo hijo *w-ms*, y este tenga como hijo el nodo *FINDING*. Si la oración cumple esto, se determinará que está negada por el patrón 1.

Para realizar este análisis hay que tomar la siguiente decisión: para verificar que un mismo nodo tenga determinados hijos en cierto orden, hay que decidir si solo importa que aparezcan esos hijos en ese orden con la posibilidad de que existan nodos intermedios entre esos hijos, es decir que el nodo tenga más hijos y que aparezcan entre medio de los nodos de interés, o si se quiere que esos hijos aparezcan en ese orden sin ningún nodo intermedio (es decir, si los "..." de los patrones representan 0 palabras o 0 ó más palabras).

Esta decisión debe ser tomada para cada patrón.
En primer instancia se decidió verificar que los patrones se cumplan en el orden mencionado antes, con la posibilidad de que existan nodos intermedios.

Definida la implementación, se realizaron pruebas utilizando el conjunto de análisis, para verificar su correctitud así como evaluar las decisiones tomadas, el patrón adicional incorporado para esta técnica (patrón 5) y posibles mejoras para el algoritmo. De estas pruebas se detectaron los siguientes inconvenientes[19]:

- El patrón 4 nunca es detectado ya que está precedido por el patrón 1.
- En las siguientes oraciones los *findings* fueron etiquetados incorrectamente por el algoritmo como negados, porque las oraciones cumplen los patrones pero teniendo nodos intermedios que alteran el estado de negación del *finding* (es decir que estos nodos intermedios terminan el alcance de la negación): *No se pueden evaluar ovarios por **mala** de ventana.*, *Varicocele bilateral, leve en TD y moderada en TI ,con vaso de mayor calibre de 3.3 cm de diametro AP,el cual no se incrementa con las maniobras de valsalva y **reflujo** venoso presente.*

  Sin embargo, se encontraron oraciones con nodos intermedios que no afectan el alcance de la negación. Algunos ejemplos son: *No se lograron individualizar por este metodo las **imagenes** nodulares descriptas por tomografia.*, *No presenta otras **lesiones** solidas ni quisticas en la region mamaria.*, *No se observan signos de **hiper**tension portal.* En estos ejemplos, los *finidngs* están negados.
- Se encontraron oraciones que aparentemente cumplen el patrón 1, sin embargo el algoritmo no las detecta con este patrón. Por ejemplo, la oración *Se exploro cara posterior y externa de pierna derecha sobre zona dolorosa en forma comparativa con la contralateral, no observandose **lesiones** a nivel del TCSC ni en el plano muscular en forma comparativa.* Analizando el árbol de shallow parsing puede verse que la palabra *observandose* es considerada como dos palabras distintas, que forman

[19] Los *findings* de los ejemplos de los inconvenientes encontrados en las pruebas se encuentran resaltados en las oraciones.

un gerundio, y no un grupo verbal, y por lo tanto para el algoritmo la oración no cumple el patrón 1, pero el *finding* se encuentra negado.

- La oración *No **liquido** libre en cav.* no se ajusta a ninguno de las patrones propuestos. Es una oración que no está del todo bien formada, más allá de la abreviación, debido a que falta el verbo de la oración, sin embargo se puede comprender que el *finding líquido* está negado.

A partir de este análisis se realizaron las siguientes pruebas:

- Verificar que los hijos aparezcan en el orden establecido por las reglas no pudiendo existir nodos intermedios para la verificación de todos los patrones (es decir que en los patrones, los "..." representan 0 palabras): Esta prueba produjo más errores que antes, por lo que se optó por la implementación original.
- Si cumplió alguno de los patrones 1, 2 ó 3, igual se verifica el patrón 4. Esto hizo que el patrón 4 fuera detectado correctamente, por lo cual se incorpora esta modificación al algoritmo.

De esta forma queda definido el algoritmo basado en shallow parsing.

3.6. Técnicas de evaluación y medidas

3.6.1. Inter Rater Agreement (IRA)

El proceso de anotación se realizó en dos etapas, con el objetivo de poder revisar el criterio de anotación. Algunas de las oraciones fueron anotadas por dos anotadores distintos, para poder calcular el grado de acuerdo entre éstos usando el *Inter Rater Agreement* (IRA) para dos anotadores.

Hay una serie de estadísticas que se han utilizado para medir el IRA entre anotadores. Una de las medidas más comunes es el coeficiente de *Kappa* de Cohen (κ) para dos anotadores. Existe una adaptación de esta medida para 3 ó más anotadores, pero en este trabajo se decidió utilizar el coeficiente de *Kappa* de Cohen para dos anotadores y medir el acuerdo entre pares de anotadores.

La fórmula de este coeficiente se presenta en la ecuación 3.1.

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \tag{3.1}$$

donde $\Pr(a)$ representa el acuerdo observado real, y $\Pr(e)$ representa la posibilidad de acuerdo por casualidad (se calcula utilizando los datos observados para calcular las probabilidades de cada observador eligiendo cada categoría al azar[20]).

La diferencia $\Pr(a) - \Pr(e)$ representa la proporción de casos en que hubo acuerdo entre los anotadores más allá de la casualidad.

El coeficiente κ es simplemente la proporción de casos de desacuerdos esperados que no se producen, dicho de otra manera, es la proporción de acuerdo después de eliminar la posibilidad de acuerdo por casualidad [8].

Cohen sugiere que el resultado del cálculo del coeficiente de *Kappa* puede interpretarse como se muestra en la tabla 3.3 [27].

[20] `https://en.wikipedia.org/wiki/Cohen%27s_kappa`

Valor de *Kappa*	Nivel de acuerdo	% de datos que son confiables
0-0,20	Ninguno	0-4 %
0,21-0,39	Mínimo	4-15 %
0,40-0,59	Débil	15-35 %
0,60-0,79	Moderado	35-63 %
0,80-0,90	Fuerte	64-81 %
Arriba de 0,90	Casi perfecto	82-100 %

Tab. 3.3: Interpretación del coeficiente de *Kappa* de Cohen, según McHugh [27]

3.6.2. Medición de los resultados

Se evaluó la performance de todos los algoritmos según el valor de negación (afirmado o negado) asignado al *finding* de cada oración.

Las medidas, *Accuracy, Precision, Recall* y F1 son las que se utilizan usualmente en este área para medir la performance de los algoritmos. En este trabajo, éstas se basan en la interpretación de *findings* verdaderamente negados. Estas medidas se calculan usando los verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN). La tabla 3.4 muestra el significado de TP, FP, TN y FN.

		Algoritmos	
		predicción Negated	predicción Affirmed
Gold Standard	Negated	TP	FN
	Affirmed	FP	TN

Tab. 3.4: Significado de TP, TN, FP y FN según las anotaciones del Gold Standard, y la predicción en la salida de los algoritmos.

Dado un algoritmo, para cada oración a cuyo *finding* el algoritmo le asignó el valor *Negated*, y el Gold Standard determina que el valor de negación del *finding* de la oración es *Negated*, se cuenta como *True Positive (TP, verdadero positivo)*. Si el algoritmo designa su valor como *Affirmed* y el Gold Standard determina lo mismo, se cuenta como *True Negative (TN, verdadero negativo)*. Si el algoritmo erróneamente clasifica como *Negated* el *finding* de una oración cuyo valor designado por el Gold Standard es *Affirmed*, se cuenta como *False Positive (FP, falso positivo)*. Si la clasificación otorgada por el algoritmo es *Affirmed*, cuando el Gold Standard determina *Negated*, se considera como *False Negative (FN, falso negativo)*.

A continuación se muestran las fórmulas de las medidas *accuracy, precision, recall* y F1.

- Accuracy: Proporción de oraciones correctamente clasificadas

$$accuracy = \frac{\#TP + \#TN}{total} \tag{3.2}$$

donde *total* es la cantidad total de oraciones evaluadas.

- Precision: De todas las oraciones que el algoritmo etiquetó como negadas, proporción de cuántas de esas oraciones están verdaderamente anotadas como negadas en el Gold Standard.

$$precision = \frac{\#TP}{\#TP + \#FP} \tag{3.3}$$

- Recall: De todas las oraciones que el Gold Standard determina que están negadas, proporción de las oraciones que el algoritmo detecta correctamente que están negadas.

$$recall = \frac{\#TP}{\#TP + \#FN} \tag{3.4}$$

- F-score o F1: Mide el balance entre *precision* y *recall*, e indica cuántas de las oraciones identificadas como negadas, realmente lo están.

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{3.5}$$

3.7. Experimentos realizados

A partir de los algoritmos descritos y sus aspectos particulares (conjuntos de triggers y traducciones para NegEx, orden de los PoS-tags para el método basado en PoS-tagging, los árboles y los nodos considerados en la técnica basada en shallow parsing, entre otros), se realizaron varios experimentos, con el objetivo de poder analizar cada enfoque; evaluar ventajas, desventajas de cada técnica y las decisiones tomadas para cada caso; y estudiar la posibilidad de mejorar las distintas propuestas. Se utilizará F1 como medida para comparar los resultados de los distintos algoritmos debido a que esta medida es un balance entre *precision* y *recall*.

Se desarrollarán experimentos con la adaptación de NegEx a español propuesta en este trabajo (NegExMod) utilizando distintos conjuntos de triggers, sobre el conjunto de test de 1000 oraciones. A partir de estos experimentos se estudiará la posibilidad de construir dos conjuntos de triggers distintos, uno más específico del dominio radiológico, y otro más genérico pensado para usarlo en otros dominios. También se evaluarán las técnicas que usan información sintáctica (PoS-tagging y Shallow Parsing).

Existe una adaptación de NegEx al español [9], como se mencionó anteriormente. Con el objetivo de realizar comparaciones, se obtuvo su corpus de datos[21] a través de comunicación personal con los autores y se realizaron experimentos y comparaciones usando este corpus.

Debido a que el conjunto de datos utilizado en este trabajo no está disponible (no es público), utilizar el corpus utilizado por Costumero et al. [9] permite realizar comparaciones con otros algoritmos y con otro tipo de informes médicos.

[21] El corpus utilizado en el trabajo de Costumero et al. [9] es de características diferentes al utilizado en este trabajo: no es específico del dominio radiológico, los informes son de mayor longitud, algunos utilizan un lenguaje más formal, y otros más informal.

4. RESULTADOS

En esta sección se presentan los resultados del cálculo del grado de acuerdo entre los anotadores del Gold Standard, así como también los de la aplicación de los métodos propuestos. Para calcularlos se contrastó a los mismos con el Gold Standard. Adicionalmente, se muestran las comparaciones entre los distintos enfoques, los conjuntos de datos y los resultados de los trabajos realizados para otros idiomas.

4.1. Inter Rater Agreement (IRA)

El proceso de anotación consistió de dos tareas:

1. Determinar si el término marcado por el algoritmo de detección de *findings* corresponde a un verdadero *finding*.

2. Anotar como *Affirmed* o *Negated* un *finding*, según el criterio establecido.

Para calcular el grado de acuerdo entre anotadores se consideró solo el resultado de la segunda tarea. Para este cálculo, las oraciones que fueron anotadas con "-" por alguno de los anotadores, fueron descartadas, ya que no se puede inferir de las mismas si hubo acuerdo o no en la anotación, debido a que no tiene sentido una anotación si el término marcado no corresponde a un *finding*. Las oraciones anotadas como *Doubt* no pueden contabilizarse en el cálculo, por lo cual descartamos las clasificadas con esta categoría (si uno de los dos anotadores anotó *Doubt* se descarta) y las anotadas como *Probable* se cuentan como *Affirmed* tanto para el cálculo así como para armar el corpus (debido a que para un médico podría llegar a ser de interés evaluar la oración).

El cálculo del coeficiente de *Kappa* (κ) se hizo sobre el Gold Standard, para el conjunto de 100 oraciones anotadas por el anotador 1 y el anotador 3, y para el conjunto de 100 oraciones anotadas por el anotador 2 y el anotador 3.

De las 100 oraciones en común entre los anotadores 1 y 3, 14 oraciones se descartaron porque alguno de los anotadores la anotó con "-" (no la anotó), y 4 se descartaron porque alguno de los anotadores la anotó como *Doubt*. En la tabla 4.1 se muestra la clasificación de las 78 oraciones restantes según estos anotadores.

		Anotador 3		Total
		Affirmed	Negated	
Anotador 1	Affirmed	67	0	67
	Negated	2	13	15
	Total	69	13	82

Tab. 4.1: Oraciones anotadas como *Negated* o *Affirmed* (*Affirmed* o *Probable*) por los anotadores 1 y 3

De las 100 oraciones en común entre los anotadores 2 y 3, se descartaron 20 oraciones, porque alguno de los dos anotadores la anotó como "-" (no la anotó), y se descartaron otras 2 porque alguno de los anotadores la anotó como *Doubt*. En la tabla 4.2 se muestra la clasificación realizada por los anotadores de las 82 oraciones restantes.

		Anotador 3		Total
		Affirmed	Negated	
Anotador 2	Affirmed	56	1	57
	Negated	1	20	21
	Total	57	21	78

Tab. 4.2: Oraciones anotadas como *Negated* o *Affirmed* (*Affirmed* o *Probable*) por los anotadores 2 y 3

El coeficiente de *Kappa* obtenido para los dos conjuntos de oraciones anotadas por dos anotadores cada conjunto se muestra en la tabla 4.3.

Anotadores	#oraciones	κ
A1-A3	82	0,97
A2-A3	78	0,96

Tab. 4.3: Coeficiente de *Kappa* entre los anotadores A1-A3 y A2-A3

4.2. Experimentos sobre el conjunto de test

Uno de los experimentos realizados fue evaluar la performance del trigger *no*, con el objetivo de decidir si este trigger debería incluirse en el conjunto de triggers *compilado* (el conjunto de triggers conformado por los triggers obtenidos de bigramas y trigramas, los triggers provistos por el médico radiólogo y los triggers traducidos y adaptados de la versión para inglés de NegEx), y analizar la etiqueta más adecuada para el mismo. Para esto, se aplicó el algoritmo NegExMod (la versión modificada para español de NegEx, presentada en este trabajo) con el conjunto de test de 1000 oraciones, utilizando el conjunto de triggers *compilado* excluyendo el trigger *no* para un caso, incluyéndolo en los otros, en un caso con etiqueta PREN y en el otro con etiqueta POST. Los resultados de este experimento se muestran en la tabla 4.4.

Puede verse que en el caso en el que se utiliza el conjunto de triggers *compilado* incluyendo al trigger *no* con etiqueta POST, los resultados en general son peores. Comparando los resultados de aplicar el algoritmo NegExMod usando el conjunto de triggers que incluye al trigger *no* con etiqueta PREN, y usando el conjunto de triggers que lo excluye, la *precision* y el *accuracy* obtenidos son muy similares. Sin embargo, la *recall* para el caso en que el trigger *no* está excluído del conjunto de triggers es peor, afectando también el F1. El mejor caso es el conjunto de triggers que incluye al trigger *no* con la etiqueta PREN. Por lo tanto, el trigger *no* será incluido en el conjunto de triggers *compilado* con etiqueta PREN. Así este conjunto queda conformado de esta manera por un total de 350 triggers.

En la tabla 4.5 se presentan los resultados de aplicar la versión modificada de NegEx para español (NegExMod) utilizando los distintos conjuntos de triggers construidos. Puede verse que utilizando el conjunto de triggers de bigramas y trigramas la cantidad de TP obtenidos es muy baja, es decir la cantidad de oraciones que el algoritmo anotó como negadas y verdaderamente estaban negadas (según el Gold Standard) fueron pocas (26), generando gran cantidad de FN (199). Además hay muy pocos casos de FP, es decir, que las veces que el algoritmo determinó que en una oración el *finding* estaba negado pero en verdad estaba

	sin trigger *no*	con trigger *no*	
	-	PREN	POST
TP	194	220	195
FP	23	31	54
FN	31	5	30
TN	752	744	721
Accuracy	0,95	0,96	0,92
Precision	0,89	0,88	0,78
Recall	0,86	0,98	0,87
F1	0,88	0,92	0,82

Tab. 4.4: Tabla comparativa de los resultados de NegExMod aplicado sobre el corpus de test de 1000 oraciones evaluando el trigger *no* en el conjunto de triggers compilado. La columna *sin trigger no* corresponde a la aplicación del algoritmo usando el conjunto de triggers compilado (triggers obtenidos de bigramas y trigramas, triggers provistos por el médico radiólogo y triggers traducidos y adaptados de la versión para inglés de NegEx), que no contiene el trigger *no*. La columna *con trigger no*, corresponde a los resultados de aplicar el algoritmo con el conjunto de triggers compilado incluyendo el trigger *no* con distintas etiquetas.

presente fueron muy pocas (2). Algo similar ocurre con el conjunto de triggers provisto por el médico radiólogo, aunque en este puede observarse un aumento importante en el número de casos que el algoritmo detecta negaciones correctamente (118). Con el conjunto de triggers obtenidos de traducir y mejorar los triggers de la versión original de NegEx se ve que crece considerablemente la cantidad de casos de FP respecto de los dos conjuntos anteriores. Es decir se detectan mayor cantidad de negaciones equivocadamente. Respecto del conjunto proveniente de la compilación de los tres conjuntos de triggers anteriores, la cantidad de TP aumenta en gran medida, y se reducen los FN. Esto significa que con en este conjunto se detectan mayor cantidad de negaciones. Pero además, si bien tiene bastantes FP comparado con los conjuntos de triggers de bigramas y trigramas y el provisto por el médico radiólogo, estos disminuyen respecto al conjunto de triggers traducidos. Con este conjunto es con el que se obtiene mayor F1, es decir hay un mejor balance en la cantidad de negaciones que detecta correctamente y la cantidad de negaciones erróneamente detectadas.

Una vez obtenidos los resultados de aplicar el algoritmo con los distintos conjuntos de triggers, se obtuvo la frecuencia de aparición de los mismos en las oraciones. A partir de los resultados se decidió estudiar si es posible construir subconjuntos de triggers, a partir de los conjuntos anteriores, que mejoren la performance del algoritmo y también analizar si es posible generalizar el algoritmo para aplicarlo a otros dominios distintos del radiológico.

En la tabla 6.2 de la sección Cantidad de apariciones de los triggers en español (6.3), se presenta la cantidad de apariciones de los triggers del conjunto de traducciones, en el corpus de test. De estos triggers, muchos se usan muy poco o no se usan (en el corpus solo aparecen 9 triggers distintos de los cuales 6 de ellos aparecen menos de 5 veces, los otros 201 triggers no aparecen). Se decidió analizar este conjunto debido a que no contiene triggers específicos de radiología, para poder evaluar la posibilidad de generalizar el algoritmo. También se analizó la frecuencia de aparición de los triggers del conjunto *compilado* (que se muestra en la tabla 6.1 de la sección 6.3).

Medidas	Bi-trigramas	Radiológico	Traducciones	Compilado
TP	26	118	81	220
FP	2	9	76	31
FN	199	107	144	5
TN	773	766	699	744
Accuracy	0,80	0,88	0,96	0,96
Precision	0,93	0,93	0,86	0,88
Recall	0,12	0,52	0,98	0,98
F1	0,21	0,67	0,91	0,92

Tab. 4.5: Experimentos con NegExMod. Se muestran los resultados de aplicar la versión para español de NegEx usando distintos conjuntos de triggers, sobre el conjunto de test de 1000 oraciones. La columna de *traducciones* corresponde a las traducciones y adaptaciones de los triggers para NegEx en inglés al español; la columna *radiológico* corresponde a triggers provistos por un médico radiólogo; La columna *bi-trigramas* corresponde a los triggers generados a partir de obtener bigramas y trigramas del conjunto de análisis y seleccionar aquellos que corresponden a posibles negaciones; la columna compilado corresponde a la combinación de los tres conjuntos anteriores.

A partir de la información obtenida se seleccionaron los triggers más usados. Estos son: "no" y "sin". Se incorpora el trigger "ni" (que fue incorporado en el conjunto de triggers compilado) ya que se encuentra dentro de los triggers más usados y no es específico de radiología. Además, analizando el trigger "sin" utilizado para denotar una negación, puede deducirse que la frase "con" (antónimo de "sin") debe ser muy utilizada para mencionar como presente un *finding*. Por esto, se incorpora este trigger a la lista utilizando la etiqueta PSEU[1]. De la misma manera se incorporan frases que cumplen una función similar al trigger "con", generándose así una lista total de 16 triggers que incluyen "no", "sin", "ni", "con" y triggers derivados de "con". Con este subconjunto de triggers, que se llamará *genTriggers* se evaluó el algoritmo sobre el conjunto de test. Los resultados de este experimento se presentan junto con los resultados de aplicar el mismo algoritmo utilizando el conjunto resultante de compilar los tres conjuntos de triggers (con el que se obtuvo mayor F1 de los tres conjuntos de triggers) en la tabla 4.6.
Puede verse que, aunque con el conjunto *genTriggers* se obtiene menor número de oraciones cuyos *findings* son correctamente negados por el algoritmo y menor número de oraciones con *findings* incorrectamente negados, los resultados son muy similares entre los dos conjuntos. Sin embargo, con el conjunto compilado se obtiene mayor F1. Por este motivo utilizaremos el conjunto de triggers compilado para evaluar y comparar con los otros enfoques de detección de negaciones propuestos en este trabajo.

La tabla 4.7 presenta los resultados de aplicar los tres enfoques propuestos para la detección de negaciones en informes radiológicos escritos en español. Para el caso de NegEx, se utilizó el conjunto de triggers compilado debido a que es aquel con el que se obtuvo mayor F1 respecto de todos los conjuntos de triggers evaluados.
Puede verse que los resultados de NegEx y PoS-tagging son casi idénticos, siendo NegEx el que da mayor TP, y mayor *recall*, sin embargo con ambos se obtiene el mismo F1. Con el

[1] Si un trigger tiene la etiqueta PSEU, y el *finding* se encuentra dentro del alcance de este trigger el algoritmo lo considerará como *Affirmed*.

Medidas	genTriggers	compilado
TP	207	220
FP	25	31
FN	18	5
TN	750	744
Accuracy	0,96	0,96
Precision	0,89	0,88
Recall	0,92	0,98
F1	0,91	0,92

Tab. 4.6: Tabla comparativa de los resultados de NegExMod aplicado sobre el corpus de test de 1000 oraciones utilizando dos conjuntos de triggers: la columna *genTriggers* corresponde al conjunto de triggers reducido *genTriggers*, la columna *compilado* corresponde a la ejecución del algoritmo con el conjunto de triggers compilado (triggers obtenidos de bigramas y trigramas, triggers provistos por el médico radiólogo y triggers traducidos y adaptados de la versión para inglés de NegEx).

Medidas	NegExMod	PoS-tagging	Shallow Parsing
TP	220	219	200
FP	31	31	19
FN	5	6	25
TN	744	744	756
Accuracy	0,96	0,96	0,96
Precision	0,88	0,88	0,91
Recall	0,98	0,97	0,89
F1	0,92	0,92	0,90

Tab. 4.7: Tabla comparativa de resultados de la detección de negaciones en el corpus de test de 1000 oraciones con distintos enfoques: NegExMod (segunda columna), utilizando el conjunto de triggers compilado, PoS-tagging (tercer columna) y Shallow Parsing (cuarta columna).

algoritmo de Shallow Parsing hay menor cantidad FP y más TN, pero baja la cantidad de TP respecto de los otros dos algoritmos. Por lo tanto los resultados obtenidos son peores en general, excepto en *precision*.

4.3. Experimentos con otro conjunto de datos

En esta sección se presentan los resultados de aplicar los distintos enfoques y propuestas al conjunto de datos utilizados por Costumero et al. [9], y se comparan con los resultados obtenidos con el conjunto de test desarrollado en este trabajo.

Caben destacar las similitudes y diferencias entre ambos conjuntos de datos. El conjunto de datos utilizado por Costumero et al. [9] (extraído de SciElo[2]) proviene de textos científicos escritos en lenguaje formal y de longitud extensa. Los datos del conjunto de test desarrollado para este trabajo provienen de informes de radiología, escritos por médicos, contienen abreviaturas no usuales y errores de tipografía, no tienen lenguaje formal y son

[2] `http://www.scielo.org/`

de longitud corta. Los dos conjuntos son de formato no estructurado.

En la tabla 4.8 se muestran los resultados obtenidos por Costumero et al. [9] utilizando su adaptación de NegEx a español sobre el conjunto de datos que extrajeron de SciElo[3], la adaptación de NegEx a español propuesta en este trabajo (NegExMod) utilizando el conjunto de triggers compilado, aplicada sobre el conjunto de datos de SciElo, y este mismo utilizando los datos del conjunto de test.

Algoritmo	NegEx de Costumero et al. [9]	NegExMod	
Datos	SciElo [9]	SciElo [9]	conjunto de test
Total	500	500	1000
TP	61	63	220
FP	25	47	31
FN	18	16	5
TN	350	374	744
Accuracy	0,82	0,87	0,96
Precision	0,71	0,57	0,88
Recall	0,77	0,80	0,98
F1	0,74	0,67	0,92

Tab. 4.8: Tabla comparativa de resultados, aplicando NegExMod al corpus de test con el conjunto de triggers compilado, este mismo aplicado al corpus de datos utilizado por Costumero et al. [9], y la adaptación de NegEx y los datos usado por Costumero et al. [9].

Se ve que los resultados de Costumero et al. [9] sobre su conjunto de datos es superior al obtenido con NegExMod sobre los mismos datos, en particular el F1, sin embargo, hay una diferencia considerable con respecto a los resultados obtenidos con NegExMod sobre el conjunto de test, siendo NegExMod con el conjunto de test el de mayor F1. Cabe destacar que el conjunto de triggers utilizado por Costumero et al. [9] contiene triggers propios de su conjunto de datos, y el compilado de triggers aquí propuesto tiene triggers específicos de radiología, de donde proviene el conjunto de test.

A partir de estos resultados, se decidió analizar como funciona el conjunto de triggers *genTriggers* que resulta más genérico por no tener triggers específicos del dominio radiológico, con el conjunto de datos de Costumero et al. [9].

La tabla 4.9 presenta los resultados de Costumero et al. [9] sobre los datos de SciElo, y NegExMod utilizando el conjunto *genTriggers* de triggers sobre los datos de SciElo.
Se ve que con NegExMod y el conjunto de triggers genéricos se obtiene menor F1, aunque resulta casi igual que el resultado obtenido por Costumero et al. [9]. Las medidas *accuracy* y *precision* son mejores con *genTriggers*, pero el *recall* es peor.

La tabla 4.10 provee los resultados de evaluar los tres enfoques propuestos sobre el conjunto de datos de SciElo de Costumero et al. [9]
Con los tres enfoques se obtiene el mismo *accuracy*, con shallow parsing se obtiene mayor *precision* pero el menor *recall*. Y con NegExMod es con el que se obtiene mayor *recall* y mayor F1. Además utilizando el conjunto de triggers *genTriggers* con NegExMod se obtienen mejor *accuracy, precision* y F1 que con las otras técnicas.

[3] El conjunto de datos utilizado por Costumero et al. [9] fue obtenido a través de comunicación personal con los autores.

Algoritmo	NegEx de Costumero et al. [9]	NegExMod
Triggers	Triggers de Costumero et al. [9]	genTriggers
TP	61	55
FP	25	17
FN	18	24
TN	350	404
Accuracy	0,82	0,92
Precision	0,71	0,76
Recall	0,77	0,70
F1	0,74	0,73

Tab. 4.9: Tabla comparativa de resultados utilizando los datos de SciElo [9], el algoritmo NegExMod con triggers *genTriggers* y el algoritmo de Costumero et al. [9] con los triggers de Costumero et al. [9]

.

Medidas	NegExMod	PoS-tagging	Shallow Parsing	NegExMod (genTriggers)
TP	63	53	37	55
FP	47	40	22	17
FN	16	26	42	24
TN	374	381	399	404
Accuracy	0,87	0,87	0,87	0,92
Precision	0,57	0,57	0,63	0,76
Recall	0,80	0,67	0,47	0,70
F1	0,67	0,62	0,54	0,73

Tab. 4.10: Tabla comparativa de resultados de aplicar los algoritmos propuestos para la detección de negaciones al corpus de datos utilizado por Costumero et al. [9] de 500 oraciones, provenientes de SciElo.

4.4. Comparaciones con otros trabajos

En la tabla 4.11 se presentan los resultados de dos versiones distintas de NegEx: la versión original desarrollada por Chapman et al. [5] y la adaptación de NegEx para español presentada en este trabajo (NegExMod).
Puede verse que NegExMod presenta mejores resultados, tanto en *precision*, *recall* y F1.

En la tabla 4.12 se presentan los resultados obtenidos por otros trabajos en la tarea de detección de negaciones en textos médicos, y los resultados obtenidos en este trabajo, con las distintas propuestas.
De la tabla puede verse que en los trabajos previos los enfoques que se basan en NegEx obtienen peores resultados que los que usan otros enfoques, a diferencia de los resultados obtenidos con los enfoques de este trabajo, en el que la técnica que se basa en NegEx (NegExMod) es con la que se obtienen mejores resultados. Los resultados de las técnicas propuestas en este trabajo superan a los resultados de todas las técnicas basadas en NegEx de los trabajos expuestos. Se puede destacar que el algoritmo NegExMod es aquel con el que se obtiene mayor *recall*, y que con la técnica basada en PoS-tagging se obtienen

Medidas	NegEx de Chapman et al. [5]	NegExMod
Precision	0,84	0,88
Recall	0,78	0,98
F1	0,81	0,92

Tab. 4.11: Tabla comparativa de resultados de distintas versiones de NegEx. La primer columna corresponde a los resultados de NegEx de Chapman et al. [5] utilizando un conjunto de datos de 1000 oraciones provenientes de *discharge summaries* escritos en inglés. La segunda columna corresponde a los resultados de aplicar NegExMod (la versión de NegEx adaptada a español, presentada en este trabajo) usando el conjunto de test de 1000 oraciones provenientes de informes radiológicos escritos en español.

resultados muy similares. La propuesta de Huang and Lowe [21] es aquella con la que se obtiene mayor *precision* y F1.

4.5. Discusión

En esta sección se discuten los resultados obtenidos en los experimentos realizados. La primer subsección hace referencia al cálculo del grado de acuerdo entre los anotadores del Gold Standard. En las siguientes dos subsecciones se estudian los resultados de los experimentos de detección de negaciones, sobre el conjunto de test (del dominio radiológico) y sobre otro conjunto de datos (de otro dominio). A continuación de ese estudio, se comparan los resultados obtenidos con las técnicas propuestas en este trabajo, con los resultados de trabajos previos en el tema. Luego se presenta un análisis de los errores en los resultados sobre el conjunto de test. En la última subsección se mencionarán las limitaciones de los enfoques propuestos.

4.5.1. Inter Rater Agreement

Como se ve en la tabla 4.3, los valores obtenidos en la anotación del Gold Standard son altos. Según el criterio de Cohen presentado en la tabla 3.3, son casi perfectos. En cada grupo de anotaciones compartidas, en solo dos oraciones hubo desacuerdo, y parecen ser errores de distracción (las oraciones son cortas y claras). Es posible que la razón por la que los valores de κ obtenidos sean altos, es porque al descartarse muchas oraciones por ser anotadas como *Doubt* o con -, las diferencias entre las anotaciones de esas oraciones no se contabilizaron. Entre los anotadores 1 y 3 se descartan 21 oraciones, y entre los anotadores 2 y 3 se descartan 18. Estos casos no pueden contabilizarse utilizando el coeficiente de Kappa, debido a que esta medida sirve para clasificación binaria. Se puede concluir que para todos los casos que si pueden contabilizarse, la tarea de anotar esta clase de informes (de radiología, de longitud corta, de lenguaje informal y formato no estructurado) como *Affirmed* y *Negated* es una tarea sencilla aún para no expertos en el dominio médico.

Por esta última razón se considera que podrían realizarse anotaciones automáticas, a partir de las anotaciones realizadas y obtener buenos resultados. En el trabajo presentado en *Workshop on Replicability and Reproducibility in Natural Language Processing: adaptive methods, resources and software IJCAI 2015*[4] (Taller sobre la replicabilidad y

[4] `https://sites.google.com/site/adaptivenlp2015/`

Trabajo	Recall	Precision	F1	Accuracy
NegExMod (con triggers *compilado*)	0,98	0,88	0,92	0,96
Método basado en PoS-tagging	0,97	0,88	0,92	0,96
Método basado en shallow parsing	0,89	0,91	0,90	0,96
Chapman et al. [5] (NegEx)	0,78	0,84	0,81	-
Wu et al. [49] (RadReportMiner)	0,72	0,81	0,76	-
Skeppstedt [40] (NegEx to Swedish)	0,82	0,84	0,83	-
Costumero et al. [9] (NegEx to Spanish)	0,77	0,71	0,74	0,82
Mutalik et al. [33] (Negfinder)	0,95	0,91	0,93	-
Huang and Lowe [21]	0,92	0,98	0,95	-

Tab. 4.12: Tabla comparativa de los resultados obtenidos en este trabajo y los resultados de otros trabajos de detección de negaciones en textos médicos. Las primeras tres filas de la tabla corresponden a los resultados del presente trabajo sobre el conjunto de test. NegExMod con el conjunto de triggers *compilado* (conjunto de bigramas y trigrams, triggers provistos por el médico radiólogo y triggers traducidos y adaptados de la versión de NegEx en inglés). Chapman et al. [5] desarrollan NegEx original para inglés, basado en expresiones regulares; Wu et al. [49] desarrollan un método para realizar búsquedas en informes radiológicos basándose en una modificación de NegEx para inglés; Skeppstedt [40] adaptan NegEx a sueco; Costumero et al. [9] adaptan NegEx a español; Mutalik et al. [33] desarrollan un programa para identificar patrones de negaciones y utilizan reglas gramaticales para determinar el alcance de las mismas; Huang and Lowe [21] realizan detección automática de negaciones en informes radiológicos escritos en inglés, combinando expresiones regulares y parsing gramático.

reproducibilidad en el Procesamiento del Lenguaje Natural: métodos de adaptación, recursos y software en IJCAI 2015), se realizaron pruebas utilizando técnicas de *machine learning* para anotar automáticamente los *findings* de los informes en *Affirmed* o *Negated*, usando un clasificador.

Para dichas pruebas, se utiliza un subconjunto de datos de 196 oraciones (extraídas del conjunto de análisis) como entrada para un clasificador *Naive Bayes*[5] (NB).

Se emplea una herramienta de *machine learning* para tareas de NLP llamada MALLET[6] (*MAchine Learning for LanguagE Toolkit*, herramientas de aprendizaje automático para lenguajes). MALLET utiliza el modelo de *Bag-of-Words* (bolsa de palabras) para representar las frases. *Bag-of-Words* define un diccionario, que contiene todo el vocabulario incluido en el conjunto de entrenamiento, en el que se asigna cada palabra una posición única en un vector. Las frases se representan como vectores con la longitud del diccionario. Cada posición, conocida comúnmente como valor característico, tiene un valor correspondiente a la cantidad de veces que aparece la palabra en la oración.

Los resultados de la automatización del proceso de anotación son *accuracy*: 0,93 %, *precision*: 0,91 %, *recall*: 0,89 % y F1: 0,90 %. Este es un resultado prometedor, ya que implica que este proceso, que requiere mucho tiempo, podría automatizarse y podrían obtenerse recursos de procesamiento del lenguaje para español más fácilmente.

[5] Debido a que se busca evaluar la viabilidad del enfoque, se optó por un algoritmo que podría ser visto como un punto de referencia para este tipo de enfoques. Los modelos alternativos podrían mejorar los resultados.

[6] `http://mallet.cs.umass.edu/`

4.5.2. Experimentos sobre el conjunto de test

A partir de los resultados mostrados en la tabla 4.5, de la *precision* obtenida usando el conjunto de bigramas y trigramas se puede deducir que los triggers que contempla verdaderamente se utilizan para denotar negación, sin embargo está incompleto, generando que muchas negaciones no se detecten, por eso se obtiene un *recall* tan bajo. Del conjunto de triggers provistos por el radiólogo se deduce algo similar respecto a que los triggers son realmente frases utilizadas para denotar negaciones en el ámbito radiológico, aunque este conjunto es un poco más completo que el de bigramas y trigramas porque se obtienen más casos de TP (118), y por esto, mayor *recall*. En el conjunto de triggers obtenido de las traducciones, la alta cantidad de casos de FP sugiere que hay triggers que no se utilizan para denotar negaciones, o que es más factible que produzcan falsos positivos. Habría que estudiar si conviene eliminar algunos triggers de este conjunto. Los resultados obtenidos con el conjunto de triggers compilado indican que es el conjunto de triggers más completo porque es el que detecta mayor número de negaciones, lo cual sabemos que es cierto porque contiene a los tres conjuntos anteriores.

Observando la tabla 4.6 se puede decir que los resultados obtenidos utilizando el conjunto de triggers compilado y *genTriggers* resultan muy similares, siendo el conjunto de triggers compilado superior, en particular, en el *recall*. Debido a que esta medida indica la proporción de oraciones que el algoritmo detecta correctamente que están negadas, se infiere que con los triggers del conjunto compilado se detectan mayor número de negaciones correctamente que con los del conjunto *genTriggers*, indicando que el de compilado es más completo que el de *genTriggers*. Sin embargo, con el conjunto *genTriggers* se alcanzan los resultados que se muestran teniendo muchos menos triggers y siendo estos genéricos. Si bien era esperado que compilado diera un mejor resultado, es interesante alcanzar estos valores con *genTriggers* porque es más simple. Con el fin de obtener un conjunto de triggers genérico este resultado es prometedor.

En la tabla 4.7 se puede ver que el enfoque de shallow parsing obtiene los peores resultados, aunque se puso a prueba bajo el supuesto de que usando este método se iba a obtener mejores resultados que PoS-tagging y NegEx en la detección de negaciones, ya que en este método no se consideran ventanas fijas de palabras entre la negación y el término de interés (como ocurre en NegEx) o cada palabra que forma la frase (como se hace en el método PoS-tagging), sino que segmenta la oración en grupos sintácticos relacionados.

Sin embargo, las técnicas de NegEx y PoS-tagging tienen mejores resultados. Uno de los factores que se cree que influyen en estos, es que las oraciones de los informes del conjunto de datos utilizado, son por lo general relativamente cortas. Esto sugiere que probablemente no es necesario utilizar métodos complejos, independientes de la longitud de las oraciones, que no fijan ventanas de palabras, y además que un enfoque sencillo es suficientemente bueno para este conjunto de datos.

El análisis que realiza PoS-tagging a partir de la oración podría ser suficiente para estas frases. En cambio, podría utilizarse shallow parsing para las frases más complejas ya que hace un análisis basado en la estructura de la oración.

4.5.3. Experimentos con otro conjunto de datos

Las tablas 4.8 y 4.9 presentan resultados interesantes. En la primera (4.8) se observa que de las dos adaptaciones de NegEx para español (la de Costumero et al. [9] y NegExMod) sobre

el conjunto de datos de SciElo, NegExMod obtiene mejores *accuracy* y *recall*, pero peores *precision* y F1. De lo que se infiere que para el conjunto de datos de SciElo, la adaptación de Costumero et al. [9] es mejor (si se contempla F1 como criterio de evaluación). Sin embargo, para el conjunto de test construido para este trabajo, utilizando el conjunto de triggers compilado (triggers de bigramas y trigramas, triggers provistos por el médico radiólogo y triggers traducidos y mejorados de NegEx para inglés) NegExMod obtiene resultados mejores en comparación con los de Costumero et al. [9] sobre SciElo.

En la segunda tabla (4.9), se ve que la implementación de NegEx, utilizando el conjunto de triggers específicos para el conjunto de datos sobre el que será probado (NegExMod con el conjunto de triggers compilado que contiene triggers específicos del ámbito radiológico) obtiene mejores resultados que cuando se aplica sobre un conjunto de datos de otro ámbito. Sin embargo, con el conjunto de triggers más genérico (genTriggers) se obtienen resultados similares al de Costumero et al. [9] cuando se aplica sobre SciElo, y similar a NegExMod cuando se aplica sobre el conjunto de test (proveniente de informes radiológicos), aunque los conjuntos de triggers específicos generan resultados superiores.

Por lo tanto, para obtener el mejor resultado dentro de un tipo de datos específico, la mejor opción podría ser utilizar un conjunto de triggers específico del dominio, pero para cuando se tienen conjuntos de datos de distintos tipos y se quieren resultados óptimos para todos los tipos de datos, el conjunto de triggers genérico podría ser el más apropiado.

Finalmente, de la tabla 4.10 se ve que que las alternativas a NegEx producen resultados peores (con F1 más bajo). Esto sugiere que los patrones propuestos posiblemente necesiten ser refinados, o que deben estudiarse otro tipo de patrones para conjuntos de datos de otro tipo.

4.5.4. Comparaciones con otros trabajos

En las tablas 4.11 y 4.12 puede verse que los enfoques y algoritmos propuestos en este trabajo, superan los resultados obtenidos por otros trabajos basados en NegEx tanto para inglés como para las adaptaciones a otros idiomas, y son similares a los resultados obtenidos por trabajos que usan otras técnicas, para idiomas distintos del español. Esto parece ser un buen indicador para los resultados obtenidos, sin embargo, habrá que considerar los errores y las limitaciones de los algoritmos propuestos para poder concluir que tan buenos son estos resultados. Sin embargo, es difícil comparar los resultados ya que son distintos conjuntos de datos y en distintos idiomas.

4.5.5. Análisis de errores

Una de las principales razones por las que se generan falsos positivos en todos los algoritmos es porque en la oración aparece una negación pero no está negando al *finding* (este es un problema de determinación del alcance de la negación). Esto ocurre en los tres algoritmos, pero por diferentes causas. En NegEx 25 oraciones de los 31 FP (81 %) tienen este problema. En el enfoque de PoS-tagging, 24 de 31 (77 %) fallan por esto mismo y en shallow parsing, 15 de 19 (79 %).

Una de las causas de este problema se debe a falta de puntuación en las oraciones. A pesar de haber corregido manualmente problemas de puntuación en el conjunto de test, quedaron algunos problemas. Hay oraciones que en realidad corresponden a más de una, y el *finding* aparece en lo que sería una de esas oraciones, y la negación en la otra. Pero la falta de puntuación hace que el *finding* quede dentro del alcance de la negación, haciendo

difícil la distinción aún para las personas. En NegEx 7 errores (23 %) se deben a falta de puntuación, en PoS-tagging 4 (13 %), mientras que en shallow parsing 5 (26 %).

Cuando no se debe a problemas de puntuación, en NegEx ocurre que se niega otro finding primero, por ejemplo con el trigger *sin*, pero luego termina el alcance de esa negación, a veces con una coma, a veces con la palabra *con*, lo que acentúa más la correcta elección de ese trigger en el conjunto *genTriggers*. Por ejemplo, *sin finding1, con finding2*, donde el *finding* de interés es el *finding2*. Esto sucede 8 veces (26 %). En PoS-tagging ocurre lo mismo, siendo 7 de los errores (23 %) debido a esta causa. Este problema no ocurre con shallow parsing.

Otra de las causas se debe a que se niega una característica o cualidad, y la negación de esa cualidad corresponde a un *finding* (el *finding* en cuestión). Los tres algoritmos fallan por este motivo.

También es motivo de falla cuando hay algo de la forma *no VERBO por finding*. Generalmente ocurren expresiones del tipo *no se logra visualizar por presencia de finding*. Esto ocurre 7 veces para NegEx (23 %), 5 para PoS-tagging (16 %), y 6 para shallow parsing (32 %), en los tres algoritmos, son casi las mismas oraciones en las que fallan por esto.

Este motivo de falla sugiere que debería incorporarse *por* como trigger con etiqueta PSEU, y que habría que incorporar patrones para los algoritmos que usan PoS-tagging y shallow parsing, que contemplen estos casos.

Las otras cuestiones que motivan la existencia de falsos positivos tienen menor incidencia, siendo una de ellas la aparición de doble negación en la oración. Todos los algoritmos fallan cuando hay doble negación, como por ejemplo en la frase *No se logra descartar... finding*[7]. Todos los algoritmos marcan como negada esta oración y deberían detectarla como afirmada. Desde NegEx se podría incorporar el trigger *no se logra descartar* con la etiqueta PSEU que corresponde a las dobles negaciones, y se solucionaría para ese caso particular. Para los otros enfoques es más difícil, porque habría que encontrar alguna forma de detectar palabras que indiquen negación como *descartar*, y deberían detectar si esas palabras a la vez están negadas, pero no hay una clase de palabras para ese caso, con lo cual, no hay forma de detectar las doble negaciones de ese tipo. Sin embargo hay una sola oración en el corpus que tiene este problema. Posiblemente no sea tan común que los médicos en el ámbito radiológico en español utilicen doble negaciones.

Otra cuestión que motiva la existencia de falsos positivos es la doble aparición del *finding* en la oración: El criterio desarrollado para los anotadores establecía que debían anotar la oración en base a la primer aparición del *finding*. Sin embargo, los algoritmos detectan ambas apariciones del *finding*, si alguna la detectan como negada, marcan la oración como negada.

Por último, se encontraron dos oraciones mal anotadas en el Gold Standard, lo cual generó falsos positivos.

Respecto a los falsos negativos, tanto en los enfoques de NegEx como el de PoS-tagging, hay muy pocos casos (5 y 6 respectivamente), y se deben a oraciones mal escritas, a falta de puntuación, o a que si bien se menciona al *finding* como presente, está acompañado de cuantificaciones que podrían indicar que el *finding* es despreciable, pero para tener certeza de esto, es necesario tener conocimiento del dominio.

[7] Los puntos suspensivos (...) denotan que pueden haber palabras entre medio.

Pero shallow parsing presenta otros problemas, además de los mencionados para NegEx y PoS-tagging. El principal problema (17 de 25 FN, 68 %) se debe a oraciones en las que hay negaciones que tienen el siguiente patrón: *sin ... de finding*, por ejemplo, *sin evidencia de derrame*, siendo el *derrame* el *finding*. Este problema se podría solucionar agregando un patrón que contemple este caso. En los otros algoritmos este error no se produce debido a que el alcance de la negación se evalúa en función de la cantidad de palabras que hay entre la negación y el *finding* (y la presencia de otros triggers en el caso de NegEx) y además en PoS-tagging se evalúa que estén las palabras que conforman el patrón (negación, verbo, etc).

Otro caso de falla que solo ocurre en shallow parsing ocurre cuando en la oración, el verbo que se utiliza cuando se niega al *finding* (cuando se trata del patrón "no ... verbo ... *finding*"), aparece en la forma de gerundio compuesto. Por ejemplo, *no observandose lesiones*, siendo *lesiones* el *finding*. Para estas formas, el árbol de shallow parsing obtenido es distinto al que se obtiene con un verbo conjugado, por lo tanto, el algoritmo no detecta estos casos. Sin embargo, el PoS-tag del gerundio, denota que pertenece a la categoría de verbos, por lo tanto, en PoS-tagging, no produce error, y en NegEx depende de que haya un trigger específico para ese tipo de gerundios.

Otro error ocurre cuando el *finding* se menciona en el paciente como presente, pero es un *finding* ya conocido, y no es el motivo de la visita del paciente al médico. Estos casos, los anotadores debían marcarlos como *Doubt*, pero, detectó una oración del conjunto de datos que fue mal anotada. Los algoritmos propuestos no contemplan el caso en que el *finding* se refiera a la historia del paciente y los reporta como *Affirmed*. Esto es un error de anotación.

La limitación de scope de shallow parsing genera menos falsos positivos (el algoritmo más conservador) pero más falsos negativos. Es posible que para obtener mejoras con el algoritmo de shallow parsing, se requiera tener patrones más específicos, por ejemplo, teniendo en cuenta los casos de falsos negativos, incorporando el siguiente patrón: *sin ... de finding*, y teniendo en cuenta los falsos positivos, podría incorporarse el patrón *no ... verbo ... por finding* tanto en PoS-tagging, como en shallow parsing. Sin embargo, tener patrones más específicos requiere tener mayor conocimiento del conjunto de datos con la posibilidad de hacerlo muy adecuado a este dominio (radiológico) y a este conjunto en particular, pero poco adecuado a otros. También requiere mayor conocimiento lingüístico.

4.5.6. Limitaciones

Una de las limitaciones que presenta este trabajo es la técnica utilizada para el etiquetado de *findings* en los informes radiológicos. Esta técnica presenta errores en algunos casos, marcando como *findings* términos que no representan verdaderos *findings*. Esto provoca ruido en el proceso de anotación, además genera casos que no tienen sentido. De todas formas, estos errores no afectan al trabajo, debido a que las oraciones que tenían *findings* mal etiquetados fueron descartadas del conjunto datos.

Otra limitación se presenta en el criterio de anotación propuesto, en donde los anotadores podían anotar oraciones como *Probable*, sin embargo, los algoritmos no consideran estos casos, por lo que se transformaron en *Affirmed*. Esta decisión puede impactar en los resultados. Además, el criterio establecía que debían anotar la oración según la primer aparición del *finding* en la frase (si es que aparecía más de una vez). Pero los algoritmos anotan como *Negated* la oración si cualquiera de las apariciones del *finding* está negada. Otra cuestión relacionada con el criterio de anotación es que los anotadores no debían

considerar como presente un *finding* histórico (que no fuera el motivo de la visita actual del paciente). Sin embargo, ninguno de los enfoques propuestos tienen la capacidad de distinguir si un *finding* mencionado en un informe corresponde al historial del paciente o está relacionado con la visita actual del paciente. A pesar de esto, se cree que los criterios establecidos son los adecuados para la tarea.

También resulta una limitación importante no contar con un conjunto de datos de radiología en español anotados, debido a que construirlo es una tarea difícil que requiere mucho trabajo y realizar un análisis profundo, prueba y error para poder establecer un criterio de anotación óptimo para el conjunto de datos que se tiene y la tarea a realizar, además el criterio debe ser claro y sin ambigüedades para los anotadores. Por otro lado, el proceso de anotación es tedioso, los informes pueden ser ambiguos, y en algunos casos se requiere conocimiento del dominio para poder dar una sentencia sobre la oración que se está anotando. Además, contar con un especialista en el dominio para que realice esta tarea es costoso.

No tener otros algoritmos para español contra los cuales comparar, es otra de las limitaciones.

Otra gran limitación está en los algoritmos sintácticos, debido a que es necesario emplear mucho tiempo para ejecutar el análisis sintáctico sobre todo el conjunto de datos.

5. CONCLUSIONES

En este trabajo se desarrollaron tres métodos distintos para la detección de negaciones de términos de interés (*findings*) en informes de radiología escritos en español. La posibilidad de conocer el alcance de las negaciones, permitiría detectar informes que son importantes para un experto en radiología y posibilitaría estructurar la información permitiendo su análisis.

La evaluación de los métodos se centra en dos aspectos: obtener alta performance (utilizando como medida de evaluación F1) en la detección de negaciones, y la generalidad de los enfoques para distintos tipos de datos. Se elaboró un conjunto de datos para utilizarlo como Gold Standard y evaluar las propuestas.

Uno de los métodos es una adaptación para español del algoritmo NegEx. Este utiliza expresiones regulares. Se distingue de otras adaptaciones por modificaciones hechas al algoritmo específicas para el idioma español, y el conjunto de triggers usado, que contiene algunos particulares del dominio de radiología. Además se estudió un subconjunto de triggers que fuera genérico respecto del tipo de datos. En esta técnica, el alcance de la negación sobre el término de interés solo se considera parcialmente, según una ventana de palabras de tamaño fijo.

Las otras dos propuestas elaboran reglas basadas en información sintáctica de las oraciones para determinar el alcance de la negación respecto de los términos de interés. La técnica basada en shallow parsing se distingue de la que utiliza PoS-tagging en que la primera toma en cuenta la estructura de la oración, mientras que la segunda, no. Los PoS-tags y el orden de las palabras en frases que contienen negaciones permiten elaborar reglas basadas en ellos. Pero estas reglas dependen de cada palabra de las oraciones. Es difícil modelar todas las formas que pueden tener las oraciones con términos de interés negados.

De los enfoques propuestos se esperaba que, debido a la complejidad y análisis del alcance de las negaciones que se obtiene a través de shallow parsing, esta alternativa fuera la que diera los mejores resultados, y que la adaptación de NegEx, debido a su sencillez y la ventana fija de palabras que delimita el alcance de las negaciones fuera la propuesta de peores resultados. Sin embargo, sucedió lo contrario. Tampoco se esperaba que la propuesta basada en PoS-tagging y la versión de la adaptación de NegEx específica para radiología obtuvieran resultados parecidos. Una hipótesis es que esto se debe al tipo de oraciones de los informes con los que se trabajó (son informes de radiología, escritos por médicos, que contienen abreviaturas no usuales y errores de tipografía, no tienen lenguaje formal, son de longitud corta y de formato no estructurado). Se debe realizar un análisis con frases más complejas y largas para evaluar que ocurre es esos casos.

De los resultados de las evaluaciones de las adaptaciones de NegEx a español (Costumero et al. [9] con los triggers usados por los autores, y la versión desarrollada en este trabajo específica para el dominio de radiología[1]) se deduce que el algoritmo de Costumero et al. [9] es mejor para sus datos, mientras que la versión presentada en este trabajo es mejor para el conjunto del dominio de radiología construido para este trabajo. Es decir

[1] NegExMod con el conjunto de triggers compilado (conjunto de triggers que contiene a los triggers obtenidos de bigramas y trigramas, los provistos por el médico radiólogo y los traducidos y adaptados de la versión para inglés de NegEx).

que ambos algoritmos están adecuados al dominio del conjunto de datos utilizado. Se deduce también que la versión de NegEx genérica desarrollada podría servir para sacar del dominio al algoritmo y usarlo en otros dominios.

También se observó que los resultados de evaluar la versión de NegEx específica para el dominio de radiología y la técnica basada en PoS-tagging sobre el conjunto de datos utilizado por Costumero et al. [9] dejan de ser parejos. Se deduce que la técnica basada en PoS-tagging está más adecuada al dominio de radiología que la adaptación de NegEx.

A partir de todo esto se cree que la técnica basada en NegEx es sencilla y rápida de aplicar, pero debido a la ventana de palabras fija, da mejores resultados sobre oraciones no muy largas. La versión desarrollada específicamente para el dominio de radiología obtiene muy buena performance en dicho dominio, pero ésta se ve degrada en otros dominios. La versión genérica desarrollada de este algoritmo si bien no presenta resultados tan buenos como la versión específica para el dominio radiológico, son resultados bastantes buenos para distintos dominios, y también comparado con otras adaptaciones de NegEx para otros idiomas. Además, debido a que el método que utiliza PoS-tags depende del orden de todas las palabras de cada oración, se cree que se consiguen mejores resultados con oraciones cortas. Por último, se entiende que la propuesta con shallow parsing es la más lenta de todas en aplicarse. Además requiere más reglas y más específicas para dar buenos resultados. Sin embargo, como considera la estructura de las oraciones, se estima que podría ser útil para oraciones más largas o complejas. Por ejemplo, si un informe que contiene una oración con dos términos negados y un tercer término de interés no negado, entonces un radiólogo querría tener este informe marcado como importante (hay un término de interés). Detectar el alcance de las negaciones permitiría conocer cuáles de los términos de interés se encuentran dentro del alcance de la negación. En casos como estos, la técnica con shallow parsing resultaría útil.

De estas deducciones se concluye que la versión genérica de la adaptación de NegEx para español tiene la ventaja de que es fácil de utilizarlo sobre un conjunto de datos de otro dominio, pues una vez que se tradujeron los triggers no requiere realizar ningún proceso adicional, sino que puede ser usado directamente. En cambio, los algoritmos basados en técnicas que usan información sintáctica requieren de mucho análisis (de oraciones y patrones) y trabajo manual para usarlos sobre un conjunto de datos, y se deduce de los resultados obtenidos que para usarlos sobre otro de dominio distinto hay que volver a realizar el proceso.

Debido a que hay pocos recursos de NLP para el idioma español, y las herramientas para este idioma están menos avanzadas que para otros como el inglés, el desarrollo de los distintos enfoques para español, en particular para el dominio de radiología, presenta un desafío. Para la realización de este trabajo no se encontraron corpus de datos estándares, por lo cual se optó por construir un Gold Standard para poder evaluar los resultados. Pero la tarea de anotación es difícil principalmente por dos motivos: establecer un criterio de anotación que sea óptimo según el conjunto de datos que se tiene y la tarea a realizar requiere de conocimiento del corpus de datos, de la tarea específica, además debe ser claro y sin ambigüedades para los anotadores, por lo que requiere hacer un análisis profundo, prueba y error, en primer lugar. Por otro lado, el proceso de anotación es tedioso, los informes pueden ser ambiguos, y en algunos casos se requiere conocimiento del dominio para poder dar una sentencia sobre el informe que se está anotando. Además, contar con un especialista en el dominio para que realice esta tarea es costoso. Por estas razones se creó un conjunto de datos, y fue dividido en subconjuntos para ser anotado por tres anota-

dores: un médico radiólogo y dos anotadores no expertos en este dominio. Algunos de los subconjuntos fueron anotados por dos anotadores para poder evaluar el IRA, obteniéndose resultados altos. Una conclusión que se desprende de esto es que este tipo de informes en casos que resulten de complejidad similar a estos, posiblemente puede ser anotado por personas no especializadas en este dominio, y obtenerse así un conjunto de datos anotados de buena calidad, lo cual es un resultado importante dada la escasez de recursos para el idioma. Además, los resultados de las pruebas para evaluar la viabilidad de la automatización del proceso de anotación son prometedores (*accuracy*: 0,93 %, *precision*: 0,91 %, *recall*: 0,89 % y F1: 0,90 %) ya que sugieren que este proceso, que es tedioso y requiere mucho tiempo, podría automatizarse y podrían obtenerse recursos de procesamiento del lenguaje para el idioma español más fácilmente.

5.1. Trabajo futuro

En base a los resultados y discusiones presentados en las secciones anteriores se considera evaluar nuevos triggers y conjuntos de triggers para la adaptación de NegEx a español propuesta en este trabajo y modificar las etiquetas asignadas a algunos de los triggers, por ejemplo incorporar el trigger *por* con etiqueta PSEU, estudiar como reducir el conjunto de triggers de traducciones para disminuir los falsos positivos, analizar el trigger *con* en los distintos conjuntos. También se considera incorporar nuevos patrones para PoS-tagging, creando y variando ventanas que determinen el alcance de la negación, agregar nuevos patrones para shallow parsing (más específicos para el conjunto de datos, y evaluarlos en conjuntos de otros dominios), analizar como funciona este enfoque con oraciones más largas, ejecutar en paralelo algunos de los procesos realizados en los algoritmos sintácticos para disminuir el tiempo de ejecución de los mismos e incorporar las sugerencias que se exponen en la documentación de las herramientas utilizadas para optimizar el tiempo de procesamiento en las operaciones de análisis sintáctico. También se evalúa la posibilidad de desarrollar un híbrido entre las distintas alternativas, con el objetivo de mejorar la performance. Asimismo, se planea incorporar la detección de incertidumbre en textos del dominio médico en español. Para el algoritmo de NegEx esto se puede lograr agregando triggers, para los algoritmos sintácticos, se requiere evaluar y construir patrones de incertidumbre.

Se considera seguir trabajando en la propuesta de anotación automática usando técnicas de *machine learning* utilizando otros clasificadores más sofisticados que el de *Naive Bayes*, y evaluarlo con el conjunto de datos de test, que resulta de mejor calidad que el de análisis.

Además se proyecta evaluar los enfoques y algoritmos propuestos utilizando el corpus de BioScope [45]. La ventaja de este corpus es que es público y gratuito. La desventaja es que habría que rearmar el trabajo para inglés. Tener resultados con este corpus permitiría que se puedan comparar otros algoritmos que utilicen otros enfoques o los mismos (pero para otros idiomas, por ejemplo) con los presentados en este trabajo.

A su vez, se podría considerar evaluar enfoques alternativos, con análisis de dependencias y *machine learning*.

Por otro lado, se pretende analizar el uso de otras ontologías para mejorar la detección de *findings*.

6. APÉNDICE

6.1. Glosario

1. *Bioscope*:

   - Corpus público de acceso gratuito, que consiste de textos médicos y biológicos. En el corpus, cada frase se anota con información acerca de la negación y la especulación. La anotación indica los límites del alcance y las palabras clave. La colección está formada por tres tipos de documentos: documentos clínicos (informes de radiología en formato no estructurado), artículos científicos y resúmenes de artículos científicos [45].

2. *Chunk*:

   - Unidad textual de tokens de palabras adyacentes.

3. *Chunking*:

   - Segmentar en una secuencia no estructurada de *chunks*, que muestran las relaciones que sostienen entre sus palabras internas [15].

4. *Discharge summaries:*

   - Informes clínicos elaborados por un médico u otro profesional de la salud como conclusión de una hospitalización o una serie de tratamientos, en él se escriben los principales síntomas por los que el paciente realizó la visita médica, los resultados de diagnóstico, la terapia administrada y la respuesta del paciente a la misma, y recomendaciones de aprobación[1].

5. EAGLES:

   - EAGLES es una iniciativa de la Comisión Europea, financiado dentro del programa *Investigación e Ingeniería lingüística*, que tiene como objetivo acelerar la provisión de estándares para recursos lingüísticos a gran escala (como corpus de textos, léxicos computacionales y corpus orales), medios de manipulación de esos conocimientos (a través de formalismos lingüística computacional y diversas herramientas de software) y medios de apreciación y evaluación de los recursos, las herramientas y los productos [16].

6. *Finding:*

   - Una observación clínicamente significativa, por lo general utilizada en relación con lo que se encuentra en el examen físico o análisis de laboratorio[2].

[1] Mosby's Medical Dictionary, 8th edition, 2009. Retrieved April 18 2016 from `http://medical-dictionary.thefreedictionary.com/discharge+summary`.

[2] Farlex Partner Medical Dictionary, 2012. Retrieved April 18 2016 from `http://medical-dictionary.thefreedictionary.com/finding`.

- Una observación acerca de un estado de enfermedad particular, por lo general en relación con las pruebas y exámenes de laboratorio físicos.
- Una conclusión que se extrae de un examen, estudio o experimento[3].

7. *FreeLing*:
   - Biblioteca que provee funciones de análisis del lenguaje, como *Morphological Analysis,* Part of Speech tagging *(PoS-tagging), Named Entity Recognition, Shallow Parsing*, entre otros, para varios idiomas, entre ellos, español (cuyo diccionario está basado en EuroWordNet [46], una ontología léxica multilingüe) [36].
8. *Gold Standard Annotation* (*Gold Standard*):
   - Es un corpus de confianza que se utiliza para el entrenamiento y evaluación significativa de algoritmos que usan anotaciones. Estas colecciones se llaman *Gold Standard Corpora* (GSC o GS, corpus estándar de oro). La construcción de un GS es una tarea laboriosa y requiere mucho tiempo de proceso. Además el tamaño, la calidad y sobre todo la disponibilidad de GS de tareas específicas, influyen directamente en el desarrollo de algoritmos de procesamiento de lenguaje natural basados en aprendizaje automático [48].
9. *Health record*:
   - Es una amplia recopilación de información tradicionalmente colocado en el expediente médico, que también abarca aspectos de la salud del paciente física, mental y social, que no necesariamente se relacionan directamente con la condición bajo tratamiento[4].
10. Indexar:
    - Registrar ordenadamente datos e informaciones para elaborar su índice.
11. Lema:
    - Conjunto de formas léxicas que tienen la misma raíz y el mismo sentido [23].
12. Lematización:
    - Proceso que agrupa las distintas formas de una palabra (como por ejemplo *apareció, aparece, aparecerá* al lema de la palabra, también conocido como *forma canónica* de la palabra (para el ejemplo, *aparecer*) [4].
13. *Machine learning*:
    - Aprendizaje automático.
14. MeSH:

[3] Mosby's Medical Dictionary, 8th edition, 2009. Retrieved April 18 2016 from `http://medical-dictionary.thefreedictionary.com/finding`.

[4] Medical Dictionary for the Health Professions and Nursing, 2012. Retrieved April 18 2016 from `http://medical-dictionary.thefreedictionary.com/health+record`.

- Es un vocabulario terminológico controlado de la Biblioteca Nacional de Medicina (NLM) utilizado para publicaciones de artículos y libros de ciencia. Proporciona una terminología organizada jerárquicamente para la indexación y catalogación de la información biomédica, de distintas bases de datos de la NLM[5].

15. Modalidad epistémica:

    - La expresión del grado de certeza o duda que el emisor muestra con respecto a la verdad de la proposición contenida en su enunciado[6].

16. Morbilidad:

    - Cantidad de personas que enferman en un lugar y un período de tiempo determinados en relación con el total de la población.

17. NegEx:

    - Algoritmo basado en expresiones regulares para detección de negaciones en textos médicos escritos en inglés, desarrollado por Chapman et al. [5]. Fue adaptado a distintos idiomas.

18. Ontología:

    - Especificaciones formales explícitas de los términos en el dominio y las relaciones entre ellos [18].
    - Una ontología define un vocabulario común para los investigadores que necesitan compartir información en un dominio. Incluye definiciones interpretables por máquinas de los conceptos básicos en el dominio y las relaciones entre ellos [34].
    - En ciencias de la computación y de la información, una ontología es una definición formal de tipos, propiedades, y relaciones entre entidades que realmente o fundamentalmente existen para un dominio de discusión en particular. Una ontología cataloga las variables necesitadas para algún conjunto de computación y establece las relaciones entre ellos. En los campos de la inteligencia artificial, ingeniería de sistemas, ingeniería de software, informática biomédica y arquitectura de la información, entre otros, se crean ontologías para limitar la complejidad y para organizar la información. Las ontología terminológicas, también conocidas como ontologías lingüísticas, especifican los términos que son usados para representar conocimiento en el universo de discurso. Suelen usarse para unificar vocabulario en un dominio determinado (contenido léxico y no semántico)[7].
    - Las ontologías formales se pueden distinguir de las ontologías terminológicas o léxicas. A diferencia de las ontologías formales, que se centran en los axiomas y las relaciones entre los conceptos definidos lógicamente, las ontologías léxicas tienen que ver con las diversas formas en que los conceptos se pueden instanciar

[5] http://www.ncbi.nlm.nih.gov/mesh
[6] http://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/diccionario/modalidad.htm
[7] https://es.wikipedia.org/w/index.php?title=Ontolog%C3%ADa_(inform%C3%A1tica)

en el lenguaje. El Sistema de Organización Simple del Conocimiento (SKOS, *Simple Knowledge Organization System*), proporciona un mecanismo para la representación de información léxica utilizando vocabularios RDF OWL y que se pueden utilizar para representar ontologías léxicas multilingües. Al igual que SKOS, LexInfo es un modelo RDF que facilita el modelado de léxicos multilingües. Sin embargo, a diferencia de SKOS, LexInfo proporciona los medios para modelar la información lingüística asociada con elementos léxicos (por ejemplo, parte de discurso y la información morfológica) [7].

19. *Parts-of-speech*:
    - Clases de equivalencia en las que se pueden agrupar las palabras. Por ejemplo, sustantivo, verbo, adjetivo, preposición, adverbio, conjunción, etc.
20. *Part-of-Speech tagging* (*PoS-tagging*, etiquetado gramatical):
    - Proceso de asignación automática de *parts-of-speech* u otro marcador de clase léxica a cada palabra en un corpus.
21. SciElo:
    - Biblioteca virtual formada por una colección de revistas científicas. Esta plataforma virtual proporciona acceso completo a una colección de revistas, así como al texto completo de los artículos [35].
22. *Shallow parsing* (análisis sintáctico superficial):
    - Asignar una estructura sintáctica parcial a oraciones. Este análisis identifica los componentes, pero no especifica su estructura interna, ni su papel en la oración principal.
23. Sintagma:
    - Un sintagma es un grupo de palabras que forman un constituyente sintáctico. Dentro de un sintagma hay una palabra fundamental que recibe el nombre de núcleo sintáctico y es la que aporta las características básicas para la formación de ese grupo. Este núcleo será también el responsable de darle nombre al sintagma. Por ejemplo, si el núcleo de un sintagma es un verbo, estaremos frente a un grupo verbal[8].
24. *Tagging*:
    - Proceso de etiquetado automático.
25. Tokenización:
    - Segmentación de tokens (pueden ser oraciones, palabras, etc).
26. *Triggers*:
    - Frases que indican negación (también llamados *cues* o *lexicons*).

[8] `http://definicion.de/sintagma/`

6.2. Abreviaturas

ACL: *Association for Computational Linguistics* (Asociación para Lingüística Computacional)

BioNLP: *Biomedical Natural Language Processing* (Procesamiento del Lenguaje Natural en Biomedicina)

EAGLES: *Expert Advisory Group on Language Engineering Standards* (Grupo Asesor de Expertos en Normas de Ingeniería del Lenguaje)

FN: *False Negative* (Falso Negativo)

FP: *False Positive* (Falso Positivo)

GS: *Gold Standard*

ICD10: *International Classification of Diseases* (Clasificación Internacional de Enfermedades)

IRA: *Inter Rater Agreement* (Grado de Acuerdo entre Anotadores)

IJCAI: *International Joint Conference on Artificial Intelligence* (Conferencia Internacional Conjunta sobre Inteligencia Artificial)

LM: *Language Model* (Modelo del Lenguaje)

MALLET: *MAchine Learning for LanguagE Toolkit* (Aprendizaje Automático para Herramientas de Idiomas)

MeSH: *Medical Subject Headings* (Encabezados de Temas Médicos)

NB: *Naive Bayes*

NLM: *National Library of Medicine* (Biblioteca Nacional de Medicina)

NLP: *Natural Language Processing* (Procesamiento del Lenguaje Natural)

NLTK: *Natural Language ToolKit* (Conjunto de Herramientas del Lenguaje Natural)

OMS: Organización Mundial de la Salud

PoS: *Parts-of-Speech*

PoS-tagging: *Parts-of-Speech tagging*

SciElo: *Scientific Electronic Library Online*

TN: *True Negative* (Verdadero Negativo)

TP: *True Positive* (Verdadero Positivo)

UMLS: *Unified Medical Language System* (Sistema Unificado de Lenguaje Médico)

6.3. Cantidad de apariciones de los triggers en español

6.3.1. Frecuencia de triggers del conjunto compilado

Trigger	Cantidad de apariciones
sin	142
no	75
no se observan	42
no se observa	38
ni	36
podria corresponder	20
no se visualizan	18
no se visualiza	17
disminuido	13

Tab. 6.1: Cantidad de apariciones del conjunto de triggers compilado (traducciones de triggers de inglés, triggers obtenidos a partir de bigramas y trigramas y los triggers provistos por un radiológo) en el corpus de test. Los triggers que no se muestran en la tabla es porque aparecen 5 o menos veces en el corpus o no aparecen.

6.3.2. Frecuencia de triggers del conjunto traducciones

Trigger	Apariciones
no	226
sin	152
disminuido	13
no se detecta	3
sin signos de	3
descartar	2
salvo	2
secundaria a	1
no puede	1

Tab. 6.2: Cantidad de apariciones del conjunto de triggers traducciones (traducciones de triggers de la versión en inglés de NegEx) en el conjunto de test. Los triggers que no se presentan en la tabla es porque no se utilizan en el corpus.

6.3.3. Frecuencia de triggers del conjunto *genTriggers*

Trigger	Apariciones
con	391
en	284
no	230
y	222
sin	154
:	58
ni	36
si	6

Tab. 6.3: Cantidad de apariciones del conjunto de triggers genérico (*genTriggers*) en el conjunto de prueba. Los triggers que no se presentan en la tabla es porque no se utilizan en el corpus.

6.4. Árboles de shallow parsing

Las siguientes imágenes ilustran la salida de texto obtenida al aplicar la función de shallow parsing de FreeLing a las oraciones de ejemplo para definir el árbol de los patrones utilizados en la técnica basada en shallow parsing. En las ilustraciones se eliminaron algunos símbolos, para que el árbol se entienda con mayor claridad. Los símbolos eliminados, solo sirven a efectos de parsear la salida e interpretar el árbol más fácilmente por un programa.

6.4.1. Patrón 1

```
S
  neg
    (No no RN -)
  grup-verb
    morfema-verbal
      (se se P00CN000 -)
    grup-verb
      verb
        (detecto detectar VMIP1S0 -)
  sn
    grup-nom-ms
      w-ms
        (FINDING finding NP00000 -)
      s-a-ms
        a-ms
          (ureteral ureteral AQ0CS0 -)
  F-term
    (. . Fp -)
```

Fig. 6.1: Esquema que muestra el árbol obtenido en la salida de texto al aplicar la función de shallow parsing en la oración *No se detecto* ***dilatacion*** *ureteral.* cuyo *finding* fue reemplazado por la etiqueta FINDING.

6.4.2. Patrón 2

```
S
  sn
    grup-nom-ms
      w-ms
        (HIGADO higado NP00000 -)
  F-no-c
    (: : Fd -)
  sadv
    (ligeramente ligeramente RG -)
  sn
    grup-nom-ms
      n-ms
        (heterogeneo heterogeneo NCMS000 -)
  grup-sp
    prep
      (en en SPS00 -)
    sn
      grup-nom-fs
        n-fs
          (forma forma NCFS000 -)
        s-a-fs
          a-fs
            (difusa difuso AQ0FS0 -)
  (, , Fc -)
  grup-sp
    prep
      (sin sin SPS00 -)
    sn
      grup-nom-mp
        w-mp
          (FINDING finding NP00000 -)
        s-a-mp
          a-mp
            (focales focal AQ0CP0 -)
  F-term
    (. . Fp -)
```

Fig. 6.2: Esquema que muestra el árbol obtenido en la salida de texto al aplicar la función de shallow parsing en la oración *Higado: ligeramente heterogeneo en forma difusa, sin* ***lesiones*** *focales.* cuyo *finding* fue reemplazado por la etiqueta FINDING.

6.4.3. Patrón 3

```
S
  sn
    grup-nom-ms
      w-ms
        (VIABILIAR viabiliar NP00000 -)
  grup-verb
    verb
      (intra intra VMIP3S0 -)
  coord
    (y y CC -)
  s-adj
    s-a-fs
      a-fs
        (extrahepatica extrahepatica AQ0FS0 -)
  F-no-c
    (: : Fd -)
  neg
    (no no RN -)
  sn
    grup-nom-ms
      w-ms
        (FINDING finding NP00000 -)
  F-term
    (. . Fp -)
```

Fig. 6.3: Esquema que muestra el árbol obtenido en la salida de texto al aplicar la función de shallow parsing en la oración *VIA BILIAR intra y extrahepatica: no* ***dilatada****.* cuyo *finding* fue reemplazado por la etiqueta FINDING.

6.4.4. Patrón 4

```
S
  neg
    (No no RN -)
  grup-verb
    morfema-verbal
      (se se P00CN000 -)
    grup-verb
      verb
        (detectaron detectar VMIS3P0 -)
  sn
    grup-nom-ms
      w-ms
        (FINDING finding NP00000 -)
  coord
    (ni ni CC -)
  sn
    grup-nom-ms
      w-ms
        (FINDING finding NP00000 -)
      s-a-ms
        a-ms
          (libre libre AQ0CS0 -)
  F-term
    (. . Fp -)
```

Fig. 6.4: Esquema que muestra el árbol obtenido en la salida de texto al aplicar la función de shallow parsing en la oración *No se detectaron* ***adenomegalias*** *ni* ***liquido*** *libre.* cuyo *finding* fue reemplazado por la etiqueta FINDING.

6.5. Performance de trabajos previos en detección de negaciones

Paper	Algoritmo	Resultados (%)				Corpus
		Recall	Precision	F1	Accuracy	
NegEx Chapman et al. [5]	Usando expresiones regulares, determina si un término UMLS está negado (si está a 6 tokens de distancia de una negación)	77,84	84,49	81,02	-	*Findings* de *discharge summaries*. Conjunto de test: dos grupos de 500 oraciones cada uno, uno con oraciones con triggers y el otro con oraciones sin triggers
NegEx for multiple languages Chapman et al. [7]	Se traduce el léxico de NegEx a Sueco, Francés y Alemán, además se modela una representación del léxico	-	-	-	-	-
Negation and uncertainty detection (RadReportMiner) Wu et al. [49]	Se detectan findings negados o inciertos para excluir de resultados de búsqueda los informes correspondientes	72,00	81,00	76,24	-	Informes radiológicos. Conjuntos de test: 464 informes
NegEx for Swedish Skeppstedt [40]	Se traduce los términos de NegEx a Sueco y se expande la lista de triggers	81,90	75,20	78,41	-	*Health records*. Conjunto de test: dos grupos uno con triggers y el otro sin triggers de 558 oraciones y 342 oraciones respectivamente
NegEx for Spanish Costumero et al. [9]	Se traduce los términos de NegEx a Español y se expande la lista de triggers	55,70	49,47	52,38	83,37	Informes de SciElo. Conjunto de test: 500 informes con 422 oraciones distintas y 267 condiciones clínicas únicas
Determine negation, experiencer and temporal status (ConText) Harkema et al. [19]	Se extiende NegEx agregando triggers para determinar experimentador y temporalidad respecto del finding evaluándose en distintos tipos de informes ampliando la definición de scope de los triggers	-	-	-	-	-

Paper	Algoritmo	Resultados (%)				Corpus
		Recall	Precision	F1	Accuracy	
Use negation detection to augment concept indexing of medial documents (Negfinder) Mutalik et al. [33]	Se usan reglas para reconocer patrones de negaciones	95,74	91,84	93,75	-	Documentos médicos de variedad de especialidades. Conjunto de test: dos grupos uno de 60 documentos(30 *discharge summaries* y 30 *surgical notes*), y el otro de 10 documentos. (Se muestran los resultados del segundo grupo)
Hybrid approach to automated negation detection in radiology reports Huang and Lowe [21]	Se usan triggers y se construyen reglas gramáticas manualmente usando información de *Part of Speech* para detectar negaciones en informes radiológicos	92,60	98,60	95,51	-	Informes radiológicos. Conjunto de test: 120 informes

Tab. 6.4: Performance de las adaptaciones de NegEx.(*)Los resultados presentados corresponden al conjunto de oraciones o informes que contienen triggers.

Bibliografía

[1] Andrés Antolino Ibáñez and Celia Chaín Navarro. Identidad cultural en Internet: la difusión del Instituto Cervantes y sus homólogos europeos. *Arbor*, 189(760):a023, 2013.

[2] Alan R Aronson and François-Michel Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[3] Alan R Aronson, Thomas C Rindflesch, and Allen C Browne. Exploiting a Large Thesaurus for Information Retrieval. In *RIAO*, volume 94, pages 197–216, 1994.

[4] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly, Beijing, 2009. ISBN 978-0-596-51649-9. URL `http://www.nltk.org/book`.

[5] Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*, 34(5):301–310, 2001.

[6] Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. Evaluation of Negation Phrases in Narrative Clinical Reports. In *Proceedings of AMIA, American Medical Informatics Association Annual Symposium*, page 105, Washington, DC, USA, 2001.

[7] Wendy W. Chapman, Dieter Hillert, Sumithra Velupillai, Maria Kvist, Maria Skeppstedt, Brian E. Chapman, Mike Conway, Melissa Tharp, Danielle L. Mowery, and Louise Deléger. Extending the NegEx Lexicon for Multiple Languages. In *Proceedings of the 14th World Congress on Medical and Health Informatics*, pages 677–681, Copenhagen, Denmark, 2013.

[8] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and psychological measurement*, 20(1), 1960.

[9] Roberto Costumero, Federico Lopez, Consuelo Gonzalo-Martin, Marta Millan, and Ernestina Menasalvas. An Approach to Detect Negation on Medical Documents in Spanish. In *Brain Informatics and Health*, volume 8609, pages 366–375. 2014. ISBN 978-3-319-09890-6.

[10] Viviana Cotik, Dario Filippo, and Jose Castano. An Approach for Automatic Classification of Radiology Reports in Spanish. *Stud Health Technol Inform.*, 216:634–638, 2015.

[11] Noa P Cruz, Maite Taboada, and Ruslan Mitkov. A Machine Learning Approach to Negation and Speculation Detection. 2015.

[12] Noa P. Cruz Díaz, Manuel Jesús Maña López, and Jacinto Mata Vázquez. Aprendizaje Automático Versus Expresiones Regulares en la Detección de la Negación y la Especulación en Biomedicina [Machine Learning versus Regular Expressions in Negation and Speculation Detection in Biomedicine]. *Procesamiento del Lenguaje Natural [Natural Language Processing]*, 45:77–85, 2010.

[13] Hercules Dalianis, Martin Hassel, and Sumithra Velupillai. The Stockholm EPR Corpus—characteristics and some initial findings. *Proceedings of ISHIMR 2009, Evaluation and implementation of e-health and health information initiatives: International*

perspectives. 14th International Symposium for Health Information Management Research, pages 243–249, 2009.

[14] Peter L Elkin, Steven H Brown, Brent A Bauer, Casey S Husser, William Carruth, Larry R Bergstrom, and Dietlind L Wahner-Roedler. A controlled trial of automated classification of negation from clinical notes. *BMC medical informatics and decision making*, 5(1):13, 2005.

[15] Stefano Federici, Simonetta Montemagni, and Vito Pirrelli. Shallow parsing and text chunking: a view on underspecification in syntax. *Cognitive science research paper-university of Sussex CSRP*, pages 35–44, 1996.

[16] FreeLing User Manual, Tagset for Spanish. User manual for open-source natural language processing library FreeLing — TALP-UPC. URL `https://talp-upc.gitbooks.io/freeling-user-manual/content/tagsets/tagset-es.html`.

[17] C Friedman and G Hripcsak. Natural language processing and its future in medicine. *Academic Medicine*, 1999.

[18] Thomas R Gruber. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220, 1993.

[19] Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. ConText: An Algorithm for Determining Negation, Experiencer, and Temporal Status from Clinical Reports. *Journal of biomedical informatics*, 42(5):839–851, oct 2009. ISSN 1532-0480.

[20] R Hersh William. Information Retrieval: A Health Care Perspective, 1996.

[21] Yang Huang and Henry J Lowe. A novel hybrid approach to automated negation detection in clinical radiology reports. *Journal of the American Medical Informatics Association*, 14(3):304–311, 2007.

[22] Betsy L. Humphreys, Donald A. B. Lindberg, Harold M. Schoolman, and G. Octo Barnett. The Unified Medical Language System: An Informatics Research Collaboration. *Journal of the American Medical Informatics Association: JAMIA*, 5(1):1–11, 1998.

[23] Daniel Jurafsky and James H Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2000.

[24] Emanuele Lapponi, Jesse Read, and Lilja Ovrelid. Representing and resolving negation for sentiment analysis. In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*, pages 687–692. IEEE, 2012.

[25] Xin Li and Dan Roth. Exploring evidence for shallow parsing. In *Proceedings of the 2001 workshop on Computational Natural Language Learning*, volume 7, page 6. Association for Computational Linguistics, 2001.

[26] Edward Loper and Steven Bird. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, volume 1, pages 63–70, Philadelphia, Pennsylvania, 2002.

[27] Mary L. McHugh. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*, 22(3):276–282, Oct 2012.

[28] Saeed Mehrabi, Anand Krishnan, Sunghwan Sohn, Alexandra M Roch, Heidi Schmidt, Joe Kesterson, Chris Beesley, Paul Dexter, C Max Schmidt, Hongfang Liu, et al. DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. *Journal of biomedical informatics*, 54:213–219, 2015.

[29] Roser Morante and Eduardo Blanco. SEM 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 265–274. Association for Computational Linguistics, 2012.

[30] Roser Morante and Walter Daelemans. A Metalearning Approach to Processing the Scope of Negation. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 21–29, Boulder, Colorado, 2009. ISBN 978-1-932432-29-9.

[31] Roser Morante and Caroline Sporleder. Modality and negation: An introduction to the special issue. *Computational linguistics*, 38(2):223–260, 2012.

[32] Marcia Munoz, Vasin Punyakanok, Dan Roth, and Dav Zimak. A learning approach to shallow parsing. *Proceedings of EMNLP-VLC'99*, pages 168–178, 2000.

[33] Pradeep G. Mutalik, Aniruddha Deshpande, and Prakash M. Nadkarni. Use of General-Purpose Negation Detection to Augment Concept Indexing of Medical Documents: A Quantitative Study Using the UMLS. *Journal of the American Medical Informatics Association: JAMIA*, 2001.

[34] Natalya F Noy, Deborah L McGuinness, et al. Ontology development 101: A guide to creating your first ontology, 2001.

[35] Abel Laerte Packer. SciELO-An Electronic Publishing Model for Developing Countries. *ELPUB*, pages 268–279, april 1999. URL `http://www.scielo.org/`.

[36] Lluís Padró and Evgeny Stanilovsky. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May 2012. ELRA.

[37] Thomas C Rindflesch and Alan R Aronson. Ambiguity resolution while mapping free text to the UMLS Metathesaurus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 240. American Medical Informatics Association, 1994.

[38] Lior Rokach, Roni Romano, and Oded Maimon. Negation Recognition in Medical Narrative Reports. *Journal of Information Retrieval*, 11(6):1–50, 2008.

[39] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5): 507–513, 2010.

[40] Maria Skeppstedt. Negation Detection in Swedish Clinical Text: An Adaption of NegEx to Swedish. *Journal of Biomedical Semantics*, 2(Suppl 3):S3, jan 2011. ISSN 2041-1480.

[41] Sunghwan Sohn, Stephen Wu, and Christopher G Chute. Dependency parser-based negation detection in clinical narratives. *AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science*, 2012:1–8, 2012.

[42] Vanesa Stricker, Ignacio Iacobacci, and Viviana Cotik. Negated Findings Detection in Radiology Reports in Spanish: an Adaptation of NegEx to Spanish. In *IJCAI - Workshop on Replicability and Reproducibility in Natural Language Processing: adaptative methods, resources and software*, Buenos Aires, Argentina, 2015.

[43] Özlem Uzuner, Xiaoran Zhang, and Tawanda Sibanda. Machine Learning and Rule-based Approaches to Assertion Classification. *Journal of the American Medical Informatics Association : JAMIA*, 16(1):109–115, 2009. ISSN 1067-5027.

[44] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[45] Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. The BioScope corpus: Biomedical Texts Annotated for Uncertainty, Negation and Their Scopes. *BMC bioinformatics*, 9(Suppl 11):S9, 2008.

[46] Piek Vossen. Introduction to eurowordnet. In *EuroWordNet: A multilingual database with lexical semantic networks*, pages 1–17. Springer, 1998.

[47] James Paul White. UWashington: Negation resolution using machine learning methods. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 335–339. Association for Computational Linguistics, 2012.

[48] Lars Wissler, Mohammed Almashraee, Dagmar Monett Díaz, and Adrian Paschke. The Gold Standard in Corpus Annotation. In *IEEE GSC*, 2014.

[49] Andrew S. Wu, Bao H. Do, Jinsuh Kim, and Daniel L. Rubin. Evaluation of Negation and Uncertainty Detection and Its Impact on Precision and Recall in Search. *Journal of digital imaging*, 24(2):234–242, apr 2011. ISSN 1618-727X.

[50] Qing T Zeng, Sergey Goryachev, Scott Weiss, Margarita Sordo, Shawn N Murphy, and Ross Lazarus. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC medical informatics and decision making*, 6(1):1, 2006.