Procesamiento del Lenguaje Natural, Revista nº 45, septiembre 2010, pp 77-85 recibido 30-04-10 revisado 20-05-10 aceptado 20-05-10

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

Noa P. Cruz, Manuel J. Maña, Jacinto Mata
Dpto. de Tecnologías de la Información. Universidad de Huelva
Ctra. Huelva - Palos de la Frontera s/n. 21819 La Rábida (Huelva)
{noa.cruz, manuel.mana, jacinto.mata}@dti.uhu.es

Resumen: En este artículo, presentamos un sistema de aprendizaje automático que identifica las expresiones de negación y especulación en textos biomédicos, en concreto, en la colección de documentos BioScope. El objetivo de este trabajo es contrastar la eficiencia de este enfoque centrado en aprendizaje automático con el que se basa en expresiones regulares. Entre los sistemas que siguen este último enfoque, hemos utilizado NegEx por su disponiblidad y popularidad. La evaluación se ha llevado a cabo sobre las tres subcolecciones que forman BioScope: documentos clínicos, artículos científicos y resúmenes de artículos científicos. Los resultados muestran la superioridad del enfoque basado en aprendizaje automático respecto a la utilización de expresiones regulares. En la identificación de expresiones de negación, el sistema propuesto mejora la medida F_1 de NegEx entre un 20 y un 30%, dependiendo de la colección de documentos. En la identificación de la especulación, el sistema propuesto supera la medida F_1 del mejor algoritmo *de línea base* entre un 10 y un 20%.
Palabras clave: Detección de la negación y la especulación, aprendizaje automático, expresiones regulares, biomedicina.

Abstract: In this paper, we present a machine learning system that identify the negation and speculation signals in biomedical texts, in particular, in the BioScope corpus. The objective of this research is to compare the efficiency of this approach focused on machine learning with which it is based on regular expressions. Among the systems that follow the latter approach, we used NegEx because of its availability and popularity. The evaluation has been carried out on the three subcollections which make up Bioscope: clinical documents, scientific papers and abstracts of scientific articles. The results show the superiority of the approach based on machine learning over the use of regular expressions. In the detection of expressions of negation, the proposed system improves the F_1 measure of NegEx by between 20 and 30%, depending on the collection of documents. In the speculation detection, the proposed system outperforms the F_1 measure of the best system between 10 and 20%.
Keywords: Negation and speculation detection, machine learning, biomedicine.

1 Introducción

La medicina está incorporando, cada vez en mayor medida, el soporte de la evidencia clínica en las decisiones de la práctica facultativa habitual. La disponibilidad de ingentes bases de datos de artículos científicos permite obtener esas evidencias. Sin embargo, sus enormes dimensiones también dificultan el acceso a los artículos relevantes. Por otra parte, algunos hospitales disponen de registros electrónicos de las historias clínicas de sus pacientes y otros muchos están procediendo a su digitalización. Esto facilita a los médicos la realización de estudios clínicos que permiten avanzar en la medicina basada en la evidencia. Sin embargo, como en el caso del acceso a la literatura

* Este trabajo ha sido parcialmente financiado por el Ministerio de Ciencia e Innovación, el Plan E del Gobierno español y la Unión Europea con cargo al FEDER (TIN2009-14057-C03-03)

científica los médicos necesitan disponer de herramientas eficientes de acceso a esa información.

De esta preocupación dan buena cuenta el notable número de iniciativas, en forma de congresos competitivos, que se organizan en el dominio de la biomedicina en torno a retos de características muy distintas. Algunos de los más relevantes que se han llevado a cabo o se siguen organizando son: BioCreative, i2b2, BioNLP Shared Task o Genomic TREC Track.

Las herramientas avanzadas de acceso a la información o minería de textos a las que hemos hecho referencia no pueden basarse en una aproximación simple como la *bolsa de palabras*. Es necesario realizar un análisis del texto en mayor profundidad. Este análisis debería incluir la detección de la negación y la especulación. Esto se hace especialmente necesario en el dominio biomédico donde, dependiendo del tipo de documentos, la cantidad de frases especulativas o negadas varía entre el 13 y el 20% (Vincze et al., 2008).

La negación convierte una frase afirmativa en negativa, como por ejemplo: "no se aprecia neumotórax ni fracturas". La especulación se utiliza para expresar que algún hecho no se conoce con certeza: "puede corresponder con un proceso neumónico incipiente".

La identificación de la negación o la especulación puede dividirse en dos fases. En la primera, se identificarían las expresiones que indican negación/especulación. En la segunda, se determinaría el alcance de las mismas, es decir, que palabras se ven afectadas por una expresión de negación o especulación. En este trabajo nos centramos en la primera fase.

El resto del artículo está organizado de la siguiente forma. En primer lugar, se presentan los trabajos más relevantes relacionados con la detección de la negación/especulación, tanto los basados en expresiones regulares como en aprendizaje automático. En el apartado 3 se presenta la colección de documentos BioScope, utilizada para el entrenamiento y evaluación del sistema que se propone. En el apartado 4 se describen las principales características del sistema. A continuación se presenta el marco de evaluación y, seguidamente, se discuten los resultados. Se finaliza con las conclusiones y el trabajo futuro.

2 Trabajos relacionados

La detección de la negación y de la especulación ha recibido mucha atención en los últimos años, especialmente en el ámbito de la biomedicina. Prueba de ello son las numerosas investigaciones recientes relacionadas con el tema, las cuales están basadas en dos enfoques diferentes: expresiones regulares (ER) y aprendizaje automático (AA).

Entre las investigaciones basadas en ER hay que destacar la desarrollada por (Chapman et al., 2001). Su algoritmo, NegEx, implementa un conjunto de patrones que indican negación. Filtra las frases que falsamente parecen ser de negación y limita el alcance de éstas. Para ello, se identifican las frases relevantes haciendo uso de UMLS y se aplican las reglas de negación. Aunque el algoritmo es calificado por la propia autora como simple, ha demostrado ser potente en la detección de la negación en resúmenes médicos. Los resultados muestran una precisión del 84,5%, cobertura del 77,8% y especificidad del 94,5%. No obstante, cuando NegEx se aplica sobre un conjunto de documentos pertenecientes a un dominio distinto para el que fue concebido, el rendimiento empeora (Mitchell et al., 2004).

Por su parte, (Mutalik et al., 2001), desarrollan un algoritmo denominado Negfinder. Dicho algoritmo consiste en un analizador léxico, así como un analizador sintáctico basado en una gramática LALR(1). La entrada del algoritmo es un texto médico el cuál es preprocesado para reconocer conceptos UMLS. El sistema muestra una cobertura del 95,7% y una especificidad del 91,8%.

Recientemente, (Elkin et al., 2005) describen un sistema basado en reglas que asigna a los conceptos un nivel de certeza como parte de la generación de un árbol sintáctico en dos fases. Primero, se preprocesa cada frase, dividiéndola en texto y operadores. El texto es analizado sintácticamente y los conceptos son representados en base a la mayor coincidencia dentro de SNOMED-CT para el fragmento de frase analizada. En la segunda fase, el sistema decide si cada concepto es una afirmación positiva, negativa o incierta. La evaluación del sistema muestra una precisión del 91,2%, cobertura del 97,2% y especificidad del 98,8%.

(Huang y Lowe, 2007) proponen un sistema híbrido para detectar negaciones en informes clínicos de radiología. El algoritmo, utiliza ER acompañado de un analizador gramatical. Con

este enfoque, los autores consiguen una precisión del 98,6%, cobertura del 92,6% y especificidad del 99,8%.

Muchos de los trabajos recientes en el campo de la identificación de la negación están basados en el enfoque de AA.

El trabajo de (Averbuch et al., 2004) es un ejemplo de la detección de conceptos negados en la narrativa médica usando técnicas de AA. Dicho algoritmo se basa en la ganancia de información para aprender patrones de negación. Se ha demostrado que este algoritmo es superior a los algoritmos tradicionales de clasificación. En concreto, la cobertura es del 95,45%, la especificidad es del 97,47% y se obtiene una medida F_1 del 99,57%.

(Goldin y Chapman, 2003) describen una extensión de NegEx utilizando técnicas de AA, con el objetivo de distinguir las frases presentes en informes clínicos, donde la palabra "not" niegue una observación médica. Como algoritmos de clasificación se utilizan Naïve Bayes y árboles de decisión. Con este enfoque, se consigue mejorar ligeramente el rendimiento del algoritmo *de línea base*, NegEx.

Por su parte, (Morante, Liekens y Daelemans, 2008) estudian el problema de la detección de la negación en la narrativa médica centrándose en la identificación del alcance. El sistema de AA propuesto por los autores, consiste en dos clasificadores. Un primer clasificador identifica las expresiones de negación presentes en cada frase. Un segundo clasificador, determina qué palabras en cada frase, se ven afectadas por la negación. El sistema muestra, para la colección de resúmenes de artículos científicos, una medida F_1 del 80,99% e identifica correctamente el 50,05% de los alcances.

(Morante y Daelemans, 2009a), presentan una mejora del sistema anteriormente descrito. En este caso, utilizan cuatro clasificadores para identificar el alcance de las expresiones de negación. En primer lugar, tres clasificadores predicen si cada palabra en una frase es la primera de una secuencia de alcance, la última o bien no pertenece al alcance de la expresión. Un cuarto clasificador utiliza la predicción de los tres clasificadores anteriores para predecir el alcance. El conjunto de documentos usados para la experimentación es más extenso ya que se utiliza la colección completa de documentos BioScope (Vince et al., 2008). Para los documentos clínicos, la medida F_1 en el caso de la negación es del 84,2% mientras que se identifican correctamente el 70,75% de los alcances. Para los artículos científicos, el sistema obtiene una medida F_1 del 70,94% y el 41% de los alcances es correctamente identificado. Para los resúmenes de artículos científicos, la medida F_1 es del 82,6% y el porcentaje de alcances correctamente clasificados es del 66,07%.

Esta investigación es ampliada por estos autores incluyendo la detección de la especulación y su alcance en (Morante y Daelemans, 2009b). Se muestra así, como el mismo enfoque puede ser aplicado tanto a la negación como a la especulación. La medida F_1 en la detección de la especulación para los documentos clínicos es del 38,16% y el 26,21% de los alcances son correctamente identificados. Para los artículos científicos, la medida F_1 es del 59,66% y se identifican correctamente el 35,92% de los alcances. Por último, para la colección de resúmenes de artículos científicos, el sistema obtiene una medida F_1 del 78,54% y el 65,55% de los alcances son correctamente clasificados.

3 La colección de documentos BioScope

La colección de documentos que hemos utilizado para entrenar y evaluar nuestro sistema es el proporcionado en la colección de documentos BioScope. Cabe destacar lo novedoso y reciente de la recopilación de documentos. Novedoso porque se trata del primer corpus con anotaciones acerca de palabras clave negativas o que indican incertidumbre así como el alcance de éstas. Reciente porque este trabajo se presentó en junio de 2008 en el BioNLP (Columbus, Ohio).

La colección está formada por tres tipos de documentos:

- Documentos clínicos. Representa la mayor parte de la colección ya que consiste en 1954 documentos. En concreto, informes de radiología, cada uno de los cuáles contiene las partes de historia clínica e impresiones del radiólogo. Esta misma colección de informes, fue usada para la tarea de etiquetar con códigos ICD-9-CM cada uno de los documentos (Pestian et al., 2007), organizado por el Centro de Medicina Computacional en Cincinnati (Ohio) en 2007.

- Artículos científicos. 5 artículos de "FlyBase" y 4 artículos de "BMC Bioinformatics". Estos documentos son útiles para evaluar la clasificación de la negación y/o especulación, ya que las distintas partes de un artículo muestran diferentes propiedades en el uso de frases negadas o especulativas.
- Resúmenes de artículos científicos. En concreto, 1273 resúmenes científicos obtenidos del corpus Genia (Collier et al., 1999). Este tipo de documentos son el principal objetivo de muchas aplicaciones de Minería de Textos debido a su accesibilidad pública.

La colección de documentos la componen en total más de 20.000 sentencias junto con sus anotaciones. Este tamaño es considerado por los autores lo suficientemente grande para servir de recopilación estándar de evaluación para la negación/especulación en el ámbito biomédico.

La Tabla 1 resume las principales características de las tres subcolecciones. Las filas 6 y 7 de la tabla muestran el porcentaje de frases negadas y especulativas que aparecen en cada subcolección. En artículos y resúmenes de artículos científicos, el número de frases especulativas es mayor que el número de frases negadas. Esto se debe a que la ambigüedad en este tipo de documentos es mayor. El conjunto de documentos clínicos, por su parte, es el más denso, respecto al número de palabras, en lo que a número de expresiones de negación y especulación se refiere.

	Clínicos	Artículos	Resúmenes
Documentos	1954	9	1273
Frases	6383	2670	11871
Palabras	42495	62794	301975
Expresiones de negación	872	378	1757
Expresiones especulativas	1137	682	2694
Frases de negación	13,55%	12,69%	13,45%
Frases especulativas	13,39%	19,43%	17,69%

Tabla 1: Estadísticas de las tres subcolecciones de documentos BioScope

Tal y como se muestra en el ejemplo (1), en la colección BioScope, cada frase es anotada con información de la especulación y la negación así como el alcance asociado a éstas.

(1) <xcope id="X6.2.2"><cue type="negation" ref="X6.2.2">No</cue> focal consolidation to <xcope id="X6.2.1"><cue type="speculation" ref="X6.2.1">suggest</cue>pneumonia </xcope></xcope>

Durante el proceso de anotación, los autores siguen una estrategia min-max. Cuando se marcan las expresiones se sigue una estrategia minimalista, marcando como expresión, la mínima unidad que exprese especulación o negación. A la hora de marcar el alcance, éste se extiende a la mayor unidad sintáctica posible.

4 Aprendizaje automático para la detección de las expresiones de negación y especulación

4.1 Descripción del sistema

La arquitectura del sistema puede dividirse en dos partes, una primera parte dedicada al entrenamiento y construcción del modelo de clasificación y una segunda fase de test. Ambas fases han sido implementadas usando métodos supervisados de AA, entrenadas y evaluadas por separado para cada una de las subcolecciones que forman la colección de documentos BioScope.

La Figura 1 muestra la arquitectura del sistema en la fase de entrenamiento. Como puede observarse, el conjunto de documentos usados para el entrenamiento es preprocesado con objeto de obtener una representación válida para el algoritmo de clasificación. En esta representación, cada instancia es una palabra del texto junto con un conjunto de atributos asociados a la misma. Por último y mediante el uso del algoritmo de clasificación correspondiente, se obtiene el modelo de clasificación final.

En la figura 2, se muestra la arquitectura del sistema en la fase de test, en la que se evalúa el rendimiento del sistema haciendo uso del modelo de clasificación generado en la etapa de entrenamiento. En esta fase, el clasificador decide si las palabras de cada frase se encuentran al principio, en el interior o fuera de una expresión de negación/especulación. Esto permite al sistema detectar expresiones complejas formadas por más de una palabra.

En ambas fases, utilizamos la versión del algoritmo C4.5 implementado en Weka (Witten y Frank, 2005). También experimentamos con

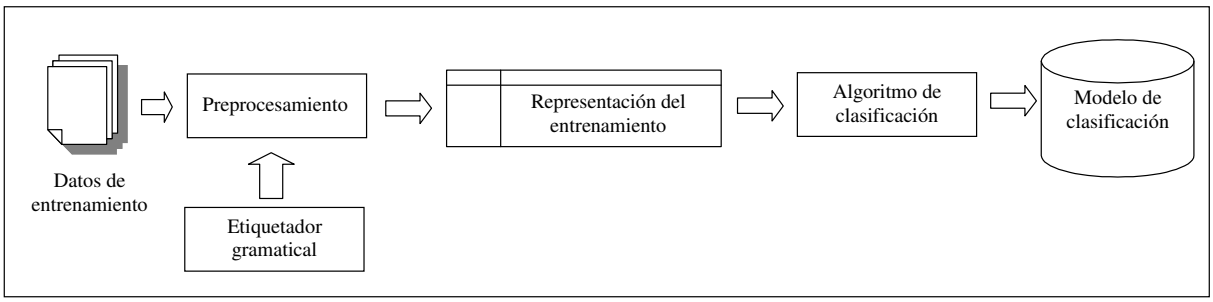

Figura 1: Arquitectura del sistema en la fase de entrenamiento

el algoritmo Naïve Bayes pero mostró peor rendimiento en general.

El algoritmo C4.5 (Quinlan, 1993), genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente de la misma forma que ID3 (Quinlan, 1986), usando el concepto de entropía. El algoritmo ha sido parametrizado usando un factor de confianza del 50%, el cuál influye en el tamaño y capacidad de predicción del árbol construido.

Por su parte, Weka, es un conocido software escrito en Java, robusto y disponible libremente bajo la licencia pública general de GNU que permite aplicar, analizar y evaluar, sobre cualquier conjunto de datos del usuario, las técnicas más relevantes de análisis de datos, principalmente las provenientes del AA.

Éste es un problema de clases no equilibradas, en el que los algoritmos de clasificación tienen una tendencia hacia la clase mayoritaria. Las técnicas de *Selección de instancias*, aplicadas sobre el conjunto de datos de entrenamiento, permiten resolver este problema mediante un *incremento* de las clases minoritarias y una *reducción* de la clase mayoritaria, usando para ello, una estrategia aleatoria. En nuestro caso, para cada una de las colecciones, hemos experimentado con las técnicas de *Resample* y *SpreadSubsample* implementadas en Weka así como con combinaciones entre ambas técnicas.

La técnica de *Resample* produce una muestra aleatoria del conjunto de datos usando

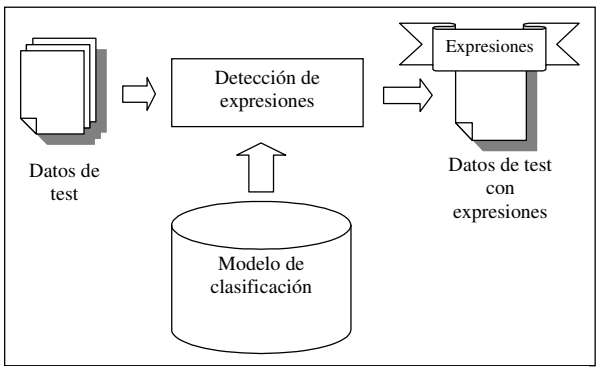

Figura 2: Arquitectura del sistema en la fase de test

muestreo con reemplazamiento. Cabe destacar que se ha utilizado la versión supervisada de esta técnica. El parámetro más importante a definir es la distribución de las clases, el cuál puede tomar valores entre 0 y 1, donde 1 correspondería a una muestra con distribución de clases totalmente uniforme.

La técnica de *SpreadSubsample* produce una muestra aleatoria del conjunto de datos. Al igual que la técnica de *Resample*, se ha utilizado en su versión supervisada. En este caso, el parámetro más importante es la relación en la muestra entre la clase mayoritaria y la/s clase/s minoritaria/s.

Para la colección de documentos clínicos, el uso de estas técnicas ha demostrado no ser efectivo. Esto se debe tanto a las características de los documentos como a que se trata de la colección más densa en expresiones de negación/especulación, por lo que el desequilibrio es menor que en el resto de colecciones.

La técnica utilizada en el caso de los resúmenes de artículos científicos, ha sido *Resample* ya que es con la que se obtienen los mejores resultados. Tras pruebas con distintos valores del parámetro de distribución de clases, el más eficiente ha demostrado ser 0,15. Otro problema que presenta esta colección, además del problema del desequilibrio de clases, es su gran tamaño. Por ello, al aplicar la técnica de *Selección de instancias* se redujo el conjunto de datos de entrenamiento al 10%.

Por último, para la colección de artículos científicos, los mejores resultados se obtienen utilizando la técnica de *SpreadSubsample*, con una relación entre la clase mayoritaria y las clases minoritarias de 500 a 1. En este caso además, se añade a la colección utilizada para el entrenamiento, las negaciones y especulaciones presentes en la colección de resúmenes de artículos científicos ya que ambas colecciones presentan características similares. Esta técnica ha resultado ser muy efectiva especialmente en el caso de la especulación, donde se consigue un incremento en la medida F1 de 0,217.

4.2 Atributos utilizados

El conjunto de documentos BioScope usado en nuestra experimentación está organizado en frases. Por ello, decidimos trabajar a nivel de frase, lo que implica considerar las palabras de cada frase de forma independiente.

Cada palabra del texto ha sido representada mediante un conjunto de atributos que fueron seleccionados, de un conjunto inicial, aplicando las técnicas de *InformationGain* y *ChiSquared* implementadas en Weka. Este conjunto final de atributos está formado por las siguientes 13 características:

- Token: Lema, categoría gramatical, atributo booleano que indica si una palabra está al comienzo de una frase o no, atributo booleano que indica si una palabra está al comienzo de un documento o no así como una etiqueta que puede tomar los valores BN,IN,BE,IE,O según si una palabra se encuentra al principio de una expresión de negación, dentro de una expresión de negación, al principio de una expresión de especulación, dentro de una expresión de especulación o fuera de la expresión.
- Contexto del token: Lema, categoría gramatical, atributo booleano que indica si una palabra está al comienzo de una frase o no, atributo booleano que indica si una palabra está al comienzo de un documento o no. Todas estas características para la primera palabra a la izquierda y la primera palabra a la derecha de la palabra que se está analizando.

Las técnicas de selección de atributos muestran que la característica que aporta más información es el lema de la palabra seguida del lema de las palabras anterior y posterior a ésta.

Para la subcolección de resúmenes de artículos científicos se añaden a las anteriores dos atributos más, ya que con ello se mejora el rendimiento del sistema. Dichos atributos son el bigrama, formado por la palabra en análisis y la palabra inmediatamente posterior a ésta y una característica booleana que será cierta si la palabra que se está analizando contiene el prefijo "in" o "un" o falso en caso contrario.Por último, destacar que la categoría gramatical de las palabras se ha obtenido usando el etiquetador gramatical desarrollado por la Universidad de Stanford. Se caracteriza por ser una implementación Java del etiquetador gramatical probabilístico basado en Máxima Entropía descrito por (Kristina Toutanova et al. 2000).

5 Evaluación

Para medir la efectividad, es decir, para tener una estimación de la "bondad" del sistema, se han utilizado como medidas la precisión y la cobertura.

$$\text{Precision (P)} = \frac{\text{Tokens correctamente negados por el sistema}}{\text{Tokens negados por el sistema}}$$

La precisión es denominada por otros autores como Chapman, *sensibilidad*.

$$\text{Cobertura (R)} = \frac{\text{Tokens correctamente negados por el sistema}}{\text{Tokens negados en la subcolección}}$$

Esta medida es llamada por otros autores *valor de predicción positivo* (PPV).

Con objeto de obtener un único valor que permita comparar los distintos experimentos, se utiliza la medida F de (Van Rijsbergen, 1979). En nuestro caso, hemos utilizado la medida F que equilibra precisión y cobertura. Se conoce como F_1 y se calcula como $\frac{2PR}{P+R}$, siendo P la precisión y R la cobertura.

6 Resultados y discusión

El objetivo de nuestro sistema es identificar las expresiones de negación y especulación presentes en la colección de documentos BioScope. Los resultados de la experimentación se han obtenido entrenando y evaluando cada subcolección por separado. En concreto, se ha dividido aleatoriamente cada subcolección en 3 partes, de las cuáles 2/3 se han utilizado para entrenar el sistema y el 1/3 restante se ha utilizado para llevar a cabo la evaluación.

Los resultados obtenidos por nuestro sistema, basado en AA, se han comparado con distintos sistemas basados en ER con el fin de contrastar la eficiencia de ambas técnicas sobre el mismo conjunto de datos.

En primer lugar, se han utilizado dos algoritmos *de línea base*. Una primera *línea base* se ha construido etiquetando directamente

como expresiones de negación o especulación las 2 expresiones de negación/especulación más frecuentes en cada conjunto de datos de entrenamiento. En la segunda *línea base* usamos las 10 expresiones más frecuentes, excepto en el caso de la negación para la colección de documentos clínicos, donde usamos las 8 expresiones más frecuentes. Esto es debido a que el resto de expresiones tienen frecuencia de aparición igual a 1.

Además, los resultados se comparan con los obtenidos por NegEx. Éste es el algoritmo de referencia en lo que a la identificación de la negación se refiere. De hecho, ha sido recientemente incluido en la herramienta Metamap Transfer Information (Aronson, 2001). Ambas evaluaciones no son directamente comparables ya que NegEx realiza una evaluación a nivel de concepto UMLS. No obstante, nos da una idea de lo eficiente que es nuestro sistema respecto a un sistema basado en ER tan popular como NegEx. La comparación con NegEx para el caso de la detección de la especulación no ha podido llevarse a cabo puesto que este sistema no ha sido diseñado para identificar este tipo de expresiones.

A la salida del clasificador se aplica un algoritmo de postprocesamiento. El procedimiento que lleva a cabo es el siguiente: cuando la expresión está formada por una sola palabra, el algoritmo cambia la clase de las palabras clasificadas como dentro de una expresión, por al principio de la expresión. Cuando la expresión está formada por más de una palabra y la clasificación de éstas no concuerda en tipo, se consulta las clasificaciones realizadas para dicha expresión hasta el momento y se sustituye la clase por aquella con mayor frecuencia de aparición. En caso de que aún no haya sido clasificada, se modifica su tipo por especulación ya que en todas las subcolecciones, el número de expresiones de especulación es mayor que el número de expresiones de negación. Sólo se muestran los resultados fruto de aplicar el algoritmo de postprocesamiento en aquellos casos dónde éste mejora los resultados obtenidos sin postprocesamiento.

La Tabla 2, muestra los resultados obtenidos para la colección de documentos clínicos. En el caso de la negación, la primera *línea base* ya obtiene unos resultados bastante buenos. Esto se debe a que las dos expresiones más frecuentes de negación representan el 88,1% de las expresiones de negación presentes en la colección de datos de entrenamiento.

	Negación			Especulación		
	Precisión	Cobertura	F_1	Precisión	Cobertura	F_1
Línea base 1	**98,1**	85,9	91,6	**97,2**	33,6	50,0
Línea base 2	96,5	**98,4**	**97,4**	94,9	70,5	80,9
NegEx	63,9	67,4	65,6	–	–	–
Nuestro sistema	96,5	98,0	97,3	92,8	**93,4**	**93,1***

* Resultado obtenido tras aplicar el algoritmo de postprocesamiento

Tabla 2: Resultados para la colección de documentos clínicos

	Negación			Especulación		
	Precisión	Cobertura	F_1	Precisión	Cobertura	F_1
Línea base 1	81,3	64,8	72,1	100,0	8,5	15,7
Línea base 2	**92,5**	46,1	85,8	55,0	39,7	46,1
NegEx	66,3	67,5	66,9	–	–	–
Nuestro sistema	77,9	**98,1**	**86,8**	**59,8**	**78,2**	**67,7**

Tabla 3: Resultados para la colección de artículos científicos

	Negación			Especulación		
	Precisión	Cobertura	F_1	Precisión	Cobertura	F_1
Línea base 1	88,3	68,9	77,4	100,0	32,4	48,9
Línea base 2	**89,1**	91,0	90,1	57,8	64,1	60,8
NegEx	72,8	65,5	69,0	–	–	–
Nuestro sistema	85,1	**96,2**	**90,3***	**73,2**	**89,8**	**80,6***

* Resultado obtenido tras aplicar el algoritmo de postprocesamiento

Tabla 4: Resultados para la colección de resúmenes de artículos científicos

Esto no ocurre con la especulación, donde, como muestra la última columna, el rendimiento es menor. En este caso, las dos expresiones especulativas más frecuentes representan sólo el 31,9% del total.

La segunda *línea base*, muestra cómo, con una lista mayor de expresiones, es posible mejorar los valores obtenidos por la primera *línea base*. NegEx, obtiene los peores resultados en el caso de la negación. Esto puede deberse a que el sistema no fue especialmente diseñado para trabajar con informes de radiología. Para la negación, nuestro sistema y los algoritmos *de línea base* obtienen resultados similares, siendo este resultado comparable a los que obtendría un humano realizando la misma tarea. Esto se debe a las características de la colección, donde casi un 90% de las expresiones, engloban el total de expresiones de este tipo presentes en el texto. Esto no ocurre con la especulación, donde la detección se hace más difícil debido a que las expresiones de especulación no se concentran en un pequeño número de expresiones, como ocurría con la negación. Por tanto, en este caso, parece más adecuado un enfoque basado en AA, como muestra la importante diferencia entre nuestro sistema y los algoritmos *de línea base.*

La Tabla 3 muestra los resultados para la colección de artículos científicos. Los resultados empeoran respecto a la colección de documentos clínicos. Esto se debe a que la ambigüedad en este tipo de documentos es mayor y por lo tanto la detección se hace más complicada. Para la negación, la diferencia entre los algoritmos *de línea base* es significativa, por lo que no basta con tener en cuenta sólo las dos expresiones más frecuentes. NegEx, obtiene unos resultados mediocres y no consigue superar tampoco en este caso a ninguno de los algoritmos *de línea base*. La diferencia entre nuestro sistema y la segunda *línea base*, como muestra la cuarta columna, no es importante. No obstante, esta diferencia se hace mayor en el caso de la especulación, donde los resultados obtenidos por nuestro sistema son bastante mejores que los obtenidos por la segunda *línea base*, tal y como se muestra en la última columna. No obstante, ambos resultados son mejorables. Una vez más, con una colección de documentos de distinta naturaleza, se demuestra cómo un enfoque basado en AA permite obtener mejor rendimiento que un enfoque basado en ER.

La Tabla 4 muestra los resultados de los distintos sistemas en el caso de la colección de resúmenes de artículos científicos. Las características de esta colección son similares a la de los artículos científicos, aunque el desequilibrio de clases en esta última es mayor. Esto se verá reflejado especialmente en el caso de la especulación, donde el desequilibrio se hace más patente. Para la negación, nuestro sistema y la segunda *línea base* obtienen los mejores resultados, siendo inexistente la diferencia entre ambos. NegEx, por su parte, vuelve a obtener los peores resultados. En el caso de la especulación, nuestro sistema obtiene los mejores resultados, siendo éstos buenos, más aún si tenemos en cuenta que tan sólo se usa un 10% de la colección para el entrenamiento. Esta diferencia es importante respecto a los algoritmos *de línea base*, como puede observarse en la última columna de la tabla.

Por tanto, las comparaciones realizadas con la misma colección de documentos, ponen de manifiesto el hecho de que un enfoque basado en AA resulta más adecuado para problemas de este tipo que los algoritmos basados en ER. Probablemente, NegEx es uno de los algoritmos más utilizados para la identificación de la negación en documentos del ámbito biomédico. Aunque se comporta de una forma homogénea para todos los tipos de documentos, en la negación, los resultados son mejorables, existiendo una diferencia media de entre un 20 y un 30% si lo comparamos con los resultados obtenidos por nuestro sistema.

7 Conclusiones y trabajo futuro

Hemos presentado un sistema basado en AA para la detección de la negación y la especulación en documentos del dominio biomédico. El entrenamiento y la evaluación del sistema se han llevado a cabo utilizando la colección de documentos BioScope. Los resultados muestran la superioridad del enfoque planteado respecto al basado en ER en todas las colecciones que conforman BioScope.

Nuestro trabajo futuro se orientará en dos líneas de investigación. En primer lugar, trataremos de mejorar los resultados de la detección de expresiones de negación y especulación en las dos colecciones en las que hemos obtenido resultados más desfavorables: resúmenes y artículos científicos. Para conseguir esto, estudiaremos la incorporación

de nuevos atributos y experimentaremos con otros algoritmos de clasificación. En segundo lugar, abordaremos la identificación del alcance de la negación y la especulación.

Referencias Bibliográficas

Aronson, AR. 2001. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. En *AMIA Symposium.*

Averbuch, M, T. Karson, B. Ben-Ami, O. Maimon, L. Rokach. 2004. Context-sensitive medical information retrieval. En *Proceedings of the 11th World Congress on Medical Informatics, MEDINFO*, páginas 1-8, San Francisco.

Chapman, WW, W. Bridewell, P. Hanbury, GF. Cooper, BG. Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform.* 34(5):301–10.

Collier, N, HS. Park, N. Ogata, Y. Tateishi, C. Nobata, T. Ohta et al. 1999. The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers. En *Proceedings of the 9th conference on European chapter*, páginas 08.12, Bergen (Noruega).

Elkin, PL, SH. Brown, BA. Bauer, CS. Husser, W. Carruth, LR. Bergstrom, et al. 2005. A controlled trial of automated classification of negation from clinical notes. *BMC Med Inform Decis Mak*. 5(1):13.

Goldin, IM, WW. Chapman. Learning to detect negation with 'Not' in medical texts. 2003. En *Proceedings of the Workshop on Text Analysis and Search for Bioinformatics at the 26th Annual International ACM SIGIR Conference.*

Huang, Y, HJ. Lowe. A novel hybrid approach to automated negation detection in clinical radiology reports. 2007. *J Am Med Inform Assoc*. 14(3):304–311.

Mitchell, KJ, MJ. Becich, JJ. Berman, WW. Chapman, J. Gilbertson, D. Gupta, et al. 2004. Implementation an evaluation of a negation tagger in a pipeline-based system for information extract from pathology reports. *Medinfo.* 11(Pt 1):663-7.

Mutalik, PG, A. Deshpande, PM. Nadkarni. 2001. Use of general purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. *J Am Med Inform Assoc*. 8(6):598–609.

Morante, R, W. Daelemans. 2009. A metalearning approach to processing the scope of negation. En *Proceedings of the 13th Conference on Computational Natural Language Learning*, páginas 21-29, Boulder (Colorado).

Morante, R, W. Daelemans. 2009. Learning the scope of hedge cues in biomedical texts. En *Proceedings of the Workshop on BioNLP*, páginas 28-36, Boulder (Colorado).

Morante, R, A. Liekens, W. Daelemans. 2008. Learning the scope of negation in biomedical texts. En *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, páginas 715-724, Honolulu.

Pestian, J, C. Brew, P. Matykiewicz, DJ. Hovermale, N. Johnson, KB. Cohen et al. 2007. A shared task involving multi-label classification of clinical free text. En *Proceedings of BioNLP*, páginas 97-104, Praga.

Quinlan, JR. 1993. C4.5: *Programs for machine learning.* Morgan Kaufmann Publishers.

Quinlan, JR. Induction of Decision Trees. 1986. *Machine Learning*, 1:81-106.

Toutanova, K, CD. 2000. Manning. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. En *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, páginas 63-70.

Van Rijsbergen, CJ. 1979. *Information Retrieval.*Butterworths-Heinemann, Londres.

Vince, V, G. Szarvas, R. Farkas, G. Móra, J. Csirik. 2009. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. En *Proceedings of BioNLP*, páginas 38-45, Columbus (Ohio).

Witten, IH, E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd ed. Kaufmann M.