UN EXPERIMENTO DE EXTRACCIÓN DE TERMINOLOGÍA UTILIZANDO ALGORITMOS ESTADÍSTICOS SUPERVISADOS

Rogelio Nazar[*]
Teresa Cabré[**]

RESUMEN: Este artículo describe una metodología de extracción automática de candidatos a término basada en técnicas de análisis estadístico de textos. A diferencia de la mayoría de los extractores terminológicos que aparecen en la literatura especializada en el tema, nuestra propuesta no incorpora ningún tipo de conocimiento explícito acerca de la lengua o del dominio de especialidad que se está analizando. Este algoritmo extrae la información que necesita directamente de los datos analizados, gracias a una fase de entrenamiento en que un usuario le "enseña" al sistema ejemplos de unidades terminológicas (una lista de términos validados) y unidades no terminológicas (una colección de texto no especializado). A partir de estos ejemplos, el algoritmo realiza una abstracción que le permite aprender a distinguir nuevas unidades terminológicas en nuevos textos. La evaluación del desempeño de este algoritmo en términos de precisión y cobertura demuestra una calidad suficiente como para ser utilizado en el procesamiento de terminología. La principal ventaja de nuestra propuesta es su gran facilidad de adaptación a nuevas lenguas y dominios de especialidad, lo que la convierte en una herramienta apropiada para lenguas con pocos recursos.

PALABRAS CLAVE: Extracción de Terminología; Análisis Estadístico de Textos Especializados; Aprendizaje Automático; Terminografía Computacional; Conocimiento Implícito.

RESUMO: Este artigo descreve uma metodologia para a extração automática de candidatos a termos baseada em técnicas de análise estatística de textos. Diferentemente da maioria dos extratores de terminologia que aparecem na literatura sobre o assunto, a nossa proposta não integra qualquer conhecimento explícito sobre a língua ou o domínio que está sendo analisado. Este algoritmo extrai as informações diretamente dos dados analisados, por meio de uma fase de treinamento na qual um usuário "ensina" os exemplos de unidades terminológicas (uma lista de termos validados) e unidades não terminológicas (uma coleção de textos não especializados). A partir destes exemplos, o algoritmo realiza uma abstração que permite distinguir novas unidades terminológicas em novos textos. A avaliação de desempenho deste algoritmo em termos de precisão e cobertura demonstra qualidade suficiente para ser útil no processamento de terminologia. a principal vantagem da nossa proposta é a sua fácil adaptação a novas línguas e domínios de especialidade, tornando uma ferramenta adequada para línguas com poucos recursos.

PALAVRAS-CHAVE: Extração de Terminologia; Análise Estatística de Textos Especializados; Aprendizado de Máquina; Terminografia Computacional; Conhecimento Implícito**.**

ABSTRACT: This article describes a methodology for automatic extraction of term candidates based on techniques of statistical analysis of texts. Unlike most terminology extractors reported in the literature, our proposal does not integrate any explicit knowledge about the language or special domain being analyzed. This algorithm extracts the necessary information directly from the analyzed data, through a training phase in which a user trains the system with examples of terminological units (a list of validated terms) and non-terminological units (a collection of non specialized texts, such as press articles). From these examples, the algorithm performs an abstraction that allows to distinguish new terminological units in new texts. an evaluation of the performance of this algorithm in terms of precision and recall shows that it has enough quality to be useful in terminology processing. The main advantage of our proposal is its easy adaptation to new languages and domains of expertise, making it an appropriate tool for languages with scarce resources.

KEYWORDS: Terminology Extraction, Statistical Analysis of Specialized Texts, Machine Learning, Computational Terminography, Implicit Knowledge.

Cómo citar este artículo: Nazar, Rogelio; Cabré, Teresa. Un experimento de extracción de terminología utilizando algoritmos estadísticos supervisados. *Debate Terminológico.* No. 07, Abril 2011.; pp. 36-55

[*] Dr. por la Universidad Pompeu Fabra. Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra. rogelio.nazar@upf.edu

[**] Dra. en Filosofía y Letras. Universitari de Lingüística Aplicada. Universitat Pompeu Fabra. teresa.cabre@upf.edu

1. INTRODUCCIÓN

En este artículo describimos una metodología para la extracción automática de términos (o candidatos a término) basada en algoritmos estadísticos supervisados, e incluimos un ejemplo de su aplicación al análisis de una muestra concreta de texto en el dominio de la medicina en castellano. El método consiste principalmente en las dos fases típicas de todo algoritmo estadístico supervisado: una fase de entrenamiento y una fase de análisis. En la fase de entrenamiento es donde un usuario le "enseña" al sistema ejemplos de términos y ejemplos de texto no especializado. En la fase de análisis, el usuario somete texto especializado nuevo para que el algoritmo extraiga términos aplicando el conocimiento adquirido durante la fase de entrenamiento. Es evidente que no se trata de reconocer en un texto analizado las unidades que se encuentran ya en el diccionario, lo cual sería una tarea distinta a la extracción de terminología y que a veces recibe el nombre de "detección de terminología" (Vivaldi y Araya, 2004). En el caso del extractor terminológico presentado en este artículo, la idea es que el algoritmo pueda reconocer unidades en el texto que no están en el diccionario de entrenamiento pero que tienen una estructura similar a la de las entradas de este diccionario. Lo fundamental de todo el proceso es que el conocimiento necesario para la extracción de los términos a partir del texto analizado es adquirido automáticamente y no hay en ningún momento un usuario o un programador que tenga que proveerle al algoritmo reglas o detalles explícitos acerca de los rasgos que deben tener los términos a extraer, tales como listas de formantes cultos, patrones sintácticos u ontologías.

En la fase de entrenamiento, al sistema se le debe presentar un vocabulario especializado –un fichero de texto con un término por línea– y un corpus de referencia de lengua general, es decir texto no especializado, como una muestra de artículos de prensa. El aprendizaje del sistema se da en tres niveles: léxico, morfológico y sintáctico. En el nivel léxico se compara la frecuencia de las palabras individuales del diccionario, incluyendo cada uno de los componentes de los términos sintagmáticos de la nomenclatura, con la frecuencia de las palabras que se encuentran en el corpus de referencia. En el nivel morfológico se hace lo mismo con las secuencias de caracteres, de distinta extensión, al principio y al final de cada palabra o componente de término. En el nivel sintáctico, finalmente, se extraen los patrones sintácticos más frecuentes en las entradas del diccionario de entrenamiento, con la idea de ponderar mejor aquellos candidatos a término del texto analizado que posean una estructura sintáctica frecuente en el diccionario terminológico.

El resultado del análisis es una lista de candidatos a término extraídos de la muestra analizada, ordenada sobre la base de una ponderación que es la combinación de los coeficientes léxico, morfológico y sintáctico. La idea general es otorgar mayor puntuación a aquellas unidades que muestren elementos que en cualquiera de estos tres niveles sean preponderantes en el diccionario utilizado durante el entrenamiento en relación al corpus de referencia de lenguaje general. Esto quiere decir que el resultado no es categórico: no es el resultado de un proceso de clasificación de las cadenas de texto de la muestra analizada en las clases de "término" y "no término", sino que se postula un continuo en forma de una lista de candidatos ordenados según esta ponderación que el algoritmo otorga. Mientras mayor sea la proporción de unidades terminológicas ocupando las primeras posiciones de esta lista, mejor es la calidad de los resultados. En cuando a la forma de presentar estos resultados, una opción es ordenar todas las unidades en una misma lista y la otra es el desglose por categoría sintáctica según los patrones sintácticos extraídos del diccionario, que son ordenados por frecuencia decreciente del patrón en tal diccionario.

En cuanto al tamaño de la muestra de análisis, éste no es determinante, lo cual quiere decir que se puede trabajar con muestras pequeñas de texto. En el experimento que describimos aquí, nuestro corpus analizado es un artículo científico de una extensión total de 1800 palabras (aproximadamente cinco páginas) seleccionado aleatoriamente del Corpus Técnico del IULA (Vivaldi, 2009). Lo que sí tiene consecuencias es el tamaño del corpus de entrenamiento. Como en todas las técnicas basadas en estadística, cabe esperar que, mientras mayor sea el tamaño del corpus, mejor será el resultado. Sin embargo, y como demostraremos en la evaluación, el aumento de tamaño del corpus de entrenamiento, si bien comporta un incremento en la calidad, éste no es de importancia crítica, lo cual quiere decir que el algoritmo tiene un desempeño aceptable incluso cuando se encuentra pobremente entrenado. Para determinar con exactitud cómo incide el tamaño del corpus de entrenamiento en el desempeño, incluimos

una evaluación del algoritmo con distintos tamaños de corpus.

El artículo se organiza de la siguiente manera: en la sección 2 comentaremos las líneas más importantes en el ámbito de la extracción automática de terminología. En la sección 3 ofreceremos una descripción de la metodología, tanto de la fase de entrenamiento como de la de análisis. En la sección 4 detallaremos los resultados del experimento de extracción. En la sección 5 aportaremos métricas de evaluación para cuantificar la calidad de los resultados. En la sección 6, finalmente, ofreceremos una discusión de los resultados de este experimento y delinearemos algunas posibilidades de trabajo futuro.

2. ANTECEDENTES DE EXTRACCIÓN DE TERMINOLOGÍA

Entre los autores interesados en el estudio de la terminología especializada, podemos encontrar por un lado grupos vinculados a corrientes de lingüística teórica (Sager, 1990; Cabré, 1999; Cabré y Estopà, 2005) y, por otro lado, corrientes vinculadas a la práctica terminográfica y la estandarización de terminología (Wüster, 1979; Arntz y Picht, 1988). En las dos últimas décadas, sin embargo, ambas corrientes comparten el interés por las tecnologías de extracción automática de terminología.

De manera tangencial, aunque con efectos importantes, la necesidad de sistematización para el desarrollo de extractores terminológicos promueve una reflexión teórica sobre las características formales que tienen que tener las unidades del texto para ser consideradas terminológicas, y sobre si realmente existe una manera objetiva de identificarlas. Tampoco es evidente que esta pregunta acerca de las condiciones necesarias y suficientes sea válida si se considera a las unidades fuera de su contexto de utilización. Cabré (1999) considera fundamental el contexto de utilización porque un contexto especializado "activa" el valor de especialidad de una unidad, sin importar que esa misma forma aparezca con un significado no especializado en la lengua cotidiana.

Los trabajos de Kageura y Umino (1996) y Cabré et al. (2001) pueden servir como introducción al estudio de la extracción de terminología. Como en otras tareas de procesamiento de lenguaje natural, existen también aquí dos aproximaciones teóricas fundamentales: por un lado la que incluye el conjunto de estrategias llamadas simbólicas, que implica la utilización de conocimiento explícito acerca de la lengua y del dominio temático sobre el que se trabaja, y por otro lado la aproximación cuantitativa, que no requiere necesariamente este tipo de conocimiento. Existe, además, una tercera vía, la de las estrategias híbridas, que integran la incorporación de conocimiento explícito y métodos estadísticos.

Con respecto a las propuestas orientadas a la incorporación de conocimiento de la lengua analizada, encontramos autores que incorporan patrones morfológicos o sintácticos (Bourigault, 1994; Jacquemin, 1997), listas de formantes cultos (Ananiadou, 1994, Estopà et al., 2000) o conocimiento ontológico del dominio especializado (Vivaldi y Rodríguez, 2006). Con respecto a las técnicas basadas en estadística, encontramos algoritmos que calculan medidas como la asociación entre los componentes de unidades poliléxicas o la forma en que se distribuyen los términos en las colecciones de documentos (Sparck Jones, 1972; Enguehard y Pantera, 1994; Daille, 1994; Frantzi, 1997; Frantzi et al, 2000; Hisamitsu et al., 2000; Pantel y Lin, 2001; Party y Langlais, 2005) o la comparación de la frecuencia de términos en un corpus de referencia de lenguaje general (Drouin, 2003; Maicher, 2004), y tal como hemos dicho, en ambas vertientes se dan distintos grados de hibridación (Justeson y Katz, 1995; Maynard y Ananiadou, 1999; Maynard y Ananiadou, 2000; Vivaldi, 2001; Vivaldi y Rodríguez, 2001; Sheremetyeva, 2009). Algunos de estos autores ofrecen demos en la web para probar sus herramientas[1]. Son pocas las que pueden procesar texto en castellano.

La mayoría los autores revisados hasta el momento parten del supuesto de que un extractor de términos es un algoritmo que hace el vaciado o extracción de términos a partir de un corpus conformado por uno o más documentos pero tratado como una unidad. En este artículo también asumimos este supuesto de partida, de manera tal que la entrada a analizar es un documento o un corpus textual tratado como bloque, es decir, sin particiones internas que podrían dar lugar a, por ejemplo, estudios diacrónicos, tal como hemos hecho en otros trabajos (ver, por ejemplo, Nazar, en prensa). La salida es, por tanto, el

vaciado terminológico en forma de una lista ponderada de candidatos a término.

3. METODOLOGÍA

Como se dijo en la introducción, en esta sección explicamos cada uno de los pasos del entrenamiento y análisis necesarios para la replicación de nuestros experimentos. Naturalmente, un usuario final de nuestro sistema no necesitará participar en estos pasos ni comprender la naturaleza del procedimiento si la finalidad es meramente práctica, ya que concebimos una implementación de nuestro algoritmo como un programa de tipo "caja negra", por tanto, el entrenamiento y el análisis podrán ser supervisados por un operador que ignore por completo el funcionamiento del sistema. A modo de ejemplo, reproduciremos en esta sección los resultados del entrenamiento con términos tomados de un diccionario terminológico concreto, la versión castellana del diccionario de medicina Mosby (2001).

3.1. FASE DE ENTRENAMIENTO

La fase de entrenamiento se compone de una serie de pasos que incluyen tanto el análisis del corpus terminológico (la lista de términos) como del no terminológico o de referencia (el texto no especializado). Este análisis se da, como ya se anticipó, en los niveles léxico (sección 3.1.1.), morfológico (sección 3.1.2) y sintáctico (sección 3.1.3).

3.1.1. Entrenamiento a nivel léxico

En este primer paso el algoritmo calcula la frecuencia de las palabras en el diccionario terminológico de entrenamiento, tanto de las palabras individuales en el caso de las entradas monoléxicas como de la frecuencia de cada una de las palabras que conforman las entradas sintagmáticas. La tabla 1 muestra las primeras posiciones en la lista ordenada por frecuencia decreciente de las palabras encontradas en ese diccionario.

En esta misma etapa del aprendizaje es donde el algoritmo analiza también el corpus de referencia de lengua general. En este caso, como corpus de referencia utilizamos una colección de artículos del diario EL PAÍS[2] de un período de 10 años (1997-2007). Tal como se hizo con el diccionario especializado, la tabla 2 muestra las palabras más frecuentes del corpus de referencia, donde faltan algunos elementos que se eliminaron previamente con una lista de exclusión (palabras como *el*, *la*, *los*, *de*, *que*, etc.).

	componente	frecuencia
1	síndrome	378
2	enfermedad	319
3	prueba	232
4	ácido	167
5	reflejo	162
6	fiebre	141
7	fractura	138
8	clorhidrato	133
9	sistema	113
10	tratamiento	109
...	...	...

Tabla 1. Frecuencia de los componentes en las entradas de un diccionario terminológico

	componente	frecuencia
1	por	843948
2	con	792858
3	no	749325
4	su	652935
5	para	624130
6	ha	380629
7	más	376928
8	lo	361949
9	como	360638
10	pero	263361
...	...	...

Tabla 2: Frecuencia de las palabras de un corpus de texto no especializado.

3.1.2. Entrenamiento a nivel morfológico

La segunda etapa del entrenamiento es el aprendizaje en el nivel morfológico. Esta etapa se plantea de una forma muy similar a la del entrenamiento en el nivel léxico, ya que aquí también estamos comparando la frecuencia de los eventos en el diccionario especializado respecto al corpus de referencia. La diferencia es que en este caso los eventos no son las palabras aisladas sino las secuencias de tres, cuatro y cinco caracteres al principio y al final de las palabras provenientes del diccionario terminológico de entrenamiento por un lado y del corpus de referencia, por otro. La tabla 3, por ejemplo, muestra el resultado del ordenamiento por frecuencia decreciente de todas las secuencias de tres caracteres a principio de palabra en el diccionario terminológico, mientras que la tabla 4, por su parte, muestra la información correspondiente a las secuencias de cuatro caracteres. Además, las tablas 5 y 6 muestran las secuencias de tres y cuatro caracteres, respectivamente, a final de palabra.

rango	componente	frecuencia
1	con-	859
2	int-	667
3	tra-	610
4	pro-	526
5	est-	508
6	hip-	502
7	car-	497
8	par-	492
9	per-	488
10	enf-	546
...	...	...

Tabla 3: Frecuencia de las secuencias de tres caracteres a principio de palabra en el diccionario terminológico

rango	componente	frecuencia
1	enfe-	435
2	inte-	414
3	sínd-	380
4	hipe-	269
5	cond-	256
6	comp-	248
7	prue-	239
8	cont-	238
9	neur-	234
10	anti-	230
...	...	...

Tabla 4: Frecuencia de las secuencias de cuatro caracteres a principio de palabra en el diccionario terminológico

rango	componente	frecuencia
1	-ión	3993
2	-ico	2143
3	-ica	1830
4	-sis	1465
5	-ina	1154
6	-dad	998
7	-nte	919
8	-lar	840
9	-tis	831
10	-nto	748
...	...	...

Tabla 5: Frecuencia de las secuencias de tres caracteres a final de palabra en el diccionario terminológico

rango	componente	frecuencia
1	-ción	3266
2	-itis	785
3	-osis	776
4	-tico	708
5	-ento	695
6	-idad	640
7	-tica	637
8	-ular	633
9	-ncia	623
10	-sión	592
...	...	...

Tabla 6: Frecuencia de las secuencias de cuatro caracteres a final de palabra en el diccionario terminológico

Incluimos sólo 10 ejemplos por tabla por cuestiones de espacio, y además sólo mostramos secuencias de tres y cuatro caracteres cuando en realidad el entrenamiento de nuestro algoritmo también incluye la frecuencia de las secuencias de 5 caracteres, tanto a principio como a final de palabra. Sin embargo, un examen a simple vista del resto de los datos confirma que esta metodología nos permite capturar la morfología típica del dominio, con formas de prefijo altamente productivas en medicina como, por ejemplo: *radi-*, *peri-*, *tran-*, *hipo-*, *para-*, *refl-*, *hemo-*, *arte-*, *célu-*, *auto-*, *fibr-*, *card-*, *psic-*, *inmu-*, *micr-*, *poli-*, *dent-*, *oste-*, *arti-*, etc. Lo mismo puede decirse de las secuencias de letras a final de palabra. En el caso de las secuencias de tres caracteres, encontramos formas típicas como *-oma*, *-cia*, *-ria*, *-nal*, *-ral*, *-ado*, *-ato*, *-ura*, *-sia*, *-rio*, *-mía*, *-smo*, *-oso*, *-gía*, etc. De la misma forma, con las secuencias de cuatro letras a final de palabra encontramos ejemplos como *-ente*, *-omía*, *-ante*, *-rome*, *-ismo*, *-edad*, *-nico*, *-isis*, *-emia*, *-inal*, *-esis*, *-oide*, *-afía*, *-geno*, etc.

Tal como hemos indicado más arriba, y de la misma forma que con el entrenamiento a nivel del léxico, aquí el algoritmo aplica el mismo análisis de afijos al corpus de referencia. Es decir que calcula la frecuencia de las secuencias de 3, 4 y 5 caracteres a principio y final de palabra, aunque, de nuevo por cuestiones de espacio, en la tabla 7 mostramos sólo las frecuencias de las secuencias de cuatro letras a principio de palabra.

rango	componente	frecuencia
1	cont-	1387
2	inte-	1375
3	comp-	1213
4	cons-	1145
5	desc-	1053
6	anti-	947
7	desa-	881
8	desp-	850
9	auto-	801
10	reco-	727
...	...	...

Tabla 7: Frecuencia de las secuencias iniciales de cuatro letras en el corpus de texto no especializado.

3.1.3. Entrenamiento a nivel sintáctico

En el entrenamiento a nivel sintáctico, el algoritmo calcula la frecuencia de las secuencias de categorías gramaticales que se encuentran en las entradas del diccionario terminológico. Este algoritmo

es entrenado con un diccionario terminológico del área con el objeto de identificar las secuencias de categorías gramaticales que son frecuentes en las entradas de ese diccionario y elaborar así un modelo sintáctico de los términos. Este modelo luego permite segmentar un texto que se somete a análisis, identificando aquellas secuencias que podrían ser terminológicas desde el punto de vista sintáctico. Con este método el algoritmo aprende que la categoría *nombre* o las secuencias como *nombre+adjetivo* o *nombre+de+nombre* son muy frecuentes en las entradas del diccionario y, por lo tanto, si encuentra esas secuencias en el texto las privilegiará como candidatos a término.

En el caso del aprendizaje a nivel sintáctico, no es necesaria la utilización del corpus de referencia, debido a que los patrones sintácticos pueden ser tan productivos en terminología especializada como en texto general. Es importante tener en cuenta que la restricción sintáctica actúa como filtro, es decir que sólo se considerarán como candidatos a término aquellas secuencias de palabras del texto analizado que configuren alguna de las estructuras sintácticas frecuentes en el diccionario (el umbral de frecuencia es arbitrario).

Creemos que utilizar conocimiento sintáctico sobre las unidades terminológicas no contradice nuestra afirmación de que el algoritmo es independiente de lengua, ya que, como se explica en los párrafos anteriores, la lista de los patrones sintácticos válidos no forma parte del diseño del algoritmo y toda la cadena de procesamiento que concebimos está planteada como un problema estadístico. Como hemos dicho, el algoritmo aprende estos patrones de manera automática sobre la base de la frecuencia que tienen en el diccionario de entrenamiento. Naturalmente, para poder contabilizar las secuencias de categorías morfosintácticas necesitamos una herramienta de etiquetado morfo-sintáctico. Se puede discutir si esto representa una forma de conocimiento explícito de una lengua. Para evitar este problema hemos elegido el programa *TreeTagger* del IMS de la Universidad de Stuttgart (Schmidt, 1994), una herramienta de etiquetado morfosintáctico que tiene una filosofía muy parecida a la nuestra, ya que es un programa que en principio puede ser utilizado en cualquier lengua en la que sea entrenado. Esta herramienta resulta más apropiada para un enfoque independiente de lengua como el de este artículo, y es la razón por la cual no utilizamos otras herramientas como *Freeling* (Padró et al., 2010), que es multilingüe pero no independiente de lengua. *TreeTagger* resulta conveniente porque ya está entrenado para la mayoría de las lenguas europeas, por lo tanto es posible aplicarlo a nuestra cadena de proceso si se trabaja en estas lenguas. También es posible entrenarlo en una lengua nueva, pero ello implica una tarea algo laboriosa ya que se necesita una gran cantidad de texto previamente etiquetado para poder obtener buenos resultados. De cualquier manera, la elección por uno u otro etiquetador no es de una importancia vital para lo que se pretende evaluar en este artículo. Nuestra metodología asume que el etiquetado morfosintáctico es un problema resuelto en la mayoría de las lenguas, siendo prácticamente un paso de rutina en las aplicaciones de procesamiento del lenguaje natural. Alternativamente, también prevemos una implementación de nuestro extractor sin el etiquetado morfosintáctico, para aquellos usuarios que no dispongan de tiempo para entrenar el POS-tagger en la lengua en la que trabajan. Lógicamente, sin el etiquetado esperamos obtener resultados de calidad inferior.

De forma concisa, la función del etiquetador morfosintáctico es la de identificar, en cada una de las entradas del diccionario terminológico, la lista de las secuencias de categorías morfosintácticas más frecuentes, de la cual se muestra un fragmento en la tabla 8. En esa tabla, el código *N* hace referencia a la categoría gramatical *nombre*, *J* a *adjetivo*, *P* a *preposición* y *A* a *artículo*. De esta forma, el código *NJ* hace referencia al patrón *nombre-adjetivo*, *NPN* al patrón *nombre-preposición-nombre*, etc. Para simplificar la lectura, los códigos de las categorías de la tabla 8 muestran una sola letra por categoría. Sin embargo, el código contiene una letra más, que especifica mejor cada categoría. Por ejemplo, nombre común (*N5*) respecto a nombre propio (*N4*), o verbo en participio (*VC*) por oposición a verbo en indicativo (*VD*), etc. Con independencia de cuáles sean y cómo se llamen o codifiquen las categorías gramaticales que tenga una lengua, con este sistema quedarán registradas sus combinaciones más frecuentes.

rango	patrón	frecuencia
1	N	11848
2	NJ	8362
3	NPN	3546
4	J	1928
5	NJJ	801
6	NPNJ	611
7	NN	572
8	V	442
9	NV	371
10	NPAN	331
…	…	…

Tabla 8: Frecuencia de patrones sintácticos en las entradas del diccionario Mosby

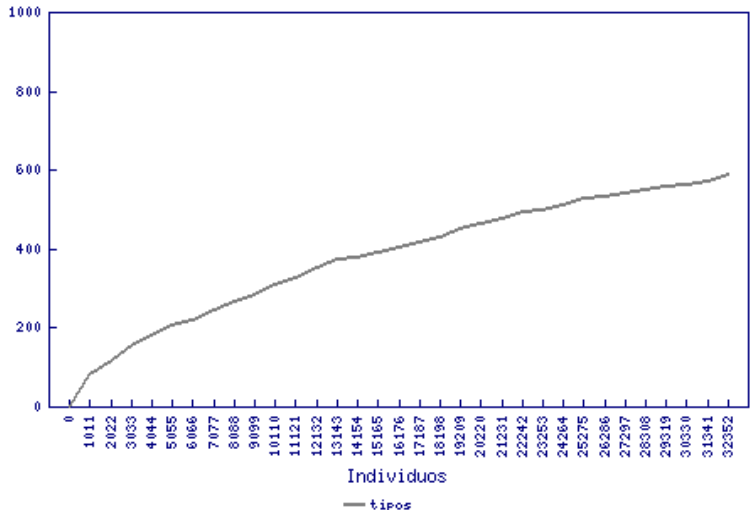

Figura 1: Crecimiento de la diversidad de patrones sintácticos (eje Y) a distintos tamaños de muestra en las entradas del diccionario de medicina Mosby (eje X).

Con el objeto de cuantificar la diversidad de los patrones morfosintácticos que se puede encontrar en el diccionario de entrenamiento, la figura 1 expresa de forma gráfica la cantidad de patrones sintácticos distintos (eje Y) a distintos tamaños de muestra (eje X), con las entradas del diccionario ordenadas de forma aleatoria. Se entiende que los individuos en el eje X no son ya las entradas en sí sino sus correspondientes patrones sintácticos, es decir, no ya *fiebre amarilla* sino NJ (nombre+adjetivo). De esta forma podemos comprobar que existe una gran diversidad de patrones con una curva que no tiende a estabilizarse. Cabría esperar que con una muestra de términos tan grande la curva se vuelva asintótica, lo cual significaría que, si bien hay una gran cantidad de patrones distintos, cada vez encontraríamos menos patrones nuevos. Sin embargo, gracias a la ley de Zipf, esta diversidad no es un problema a los fines de elaborar un modelo sintáctico, ya que la mitad de la muestra está representada por patrones que ocurren una sola vez (hapax legomena) y dos veces (dis legomena), lo cual los hace irrelevantes para el modelo. La comprobación empírica es sencilla: si ordenamos por frecuencia decreciente los patrones de distintas muestras aleatorias del mismo diccionario, comprobaremos que los primeros puestos en la tabla de frecuencia corresponden siempre a los mismos patrones.

3.2. FASE DE ANÁLISIS

Una vez que el proceso de entrenamiento ha concluido, el sistema ya puede ser utilizado en el análisis de textos en la lengua y el dominio de especialidad para los cuales el algoritmo ha sido entrenado. Las fases del análisis son las siguientes: preparación y etiquetado morfosintáctico de la muestra a analizar (sección 3.2.1); preparación de las primeras listas brutas de candidatos a término por patrón sintáctico; eliminación de falsos candidatos utilizando el corpus de referencia (sección 3.2.2.) y, finalmente, la aplicación de coeficientes a cada candidato utilizando la información almacenada durante el entrenamiento (sección 3.2.3), con la consiguiente ponderación final de los candidatos.

3.2.1. Selección y preparación del texto a analizar

Para los intereses de la extracción automática de terminología, en este artículo partimos del supuesto de que el texto analizado se encuentra en formato de texto plano y que no tenemos acceso, por tanto, a ningún tipo de información adicional que podría ayudar en la detección. Nos referimos a marcas de tipo estructural que nos permitirían identificar términos que se encuentren en títulos o que tengan algún tipo de marca tipográfica, como subrayado, negrita, cursiva, paréntesis o comillas. Consideramos que todo este tipo de recursos, que llamaríamos de "información extrínseca", es demasiado *ad hoc* y por tanto no es de interés para nuestro objetivo, que es el de proponer una metodología de análisis de la mayor generalidad posible. Nuestra muestra de análisis, por tanto, es texto plano tal como el fragmento que se muestra en la tabla 9. El primero de los pasos de análisis sobre esta muestra de texto es el etiquetado morfosintáctico con el POS-tagger. Un fragmento del resultado de este proceso se encuentra en la tabla 10.

TRASTORNOS DE LA MOTILIDAD *Los trastornos de la función motora sobrepasan probablemente a todos los demás síntomas neurológicos en frecuencia e importancia , por razones que no son dificiles de discernir.* *La porción más importante del sistema nervioso humano está destinada a movilizar el cuerpo en el espacio y a mover las diferentes partes del cuerpo, unas con relación a otras.* *...*

Tabla 9: Fragmento de una muestra de texto a analizar.

TRASTORNOS/trastorno/N5-MP DE/de/P LA/el/AFS MOTILIDAD/motilidad/N5-FS
Los/el/AMP trastornos/trastorno/N5-MP de/de/P la/el/AFS función/función/N5-FS motora/motor/JQ--FS sobrepasan/sobrepasar/VDR3P-probablemente/probable/D a/a/P todos/todo/EN--66 los/el/AMP demás/demás/EN--66 síntomas/síntoma/N5-MP neurológicos/neurológico/JQ--MP en/en/P frecuencia/frecuencia/N5-FS e/y/C importancia/importancia/N5-FS ,/=/Z por/por/P razones/razón/N5-FP que/que/RR---66 no/no/D son/ser/VDR3P-difíciles/difícil/JQ--6P de/de/P discernir/discernir/VI---- ./=/Z
La/el/AFS porción/porción/N5-FS más/más/D importante/importante/JQ--6S de/de/P l/el/AMS sistema/sistema/N5-MS nervioso/nervioso/JQ--MS humano/humano/JQ--MS está/estar/VDR3S- destinada/destinar/VC--SF a/a/P movilizar/movilizar/VI---- el/el/AMS cuerpo/cuerpo/N5-MS en/en/P el/el/AMS espacio/espacio/N5-MS y/y/C a/a/P mover/mover/VI---- las/el/AFP diferentes/diferente/JQ--6P partes/parte/N5-6P de/de/P l/el/AMS cuerpo/cuerpo/N5-MS ,/=/Z unas/uno/E6--FP con/con/P relación/relación/N5-FS a/a/P otras/otro/EN--66 ./=/Z
…

Tabla 10: Fragmento de una muestra de texto a analizar después del etiquetado morfosintáctico.

3.2.2. Extracción por patrón sintáctico

Como advertimos ya en la introducción, este algoritmo no es un clasificador binario de unidades léxicas en las categorías de *término* y *no término*, sino una herramienta de ordenación de candidatos a término según una combinación de coeficientes estadísticos. Sin embargo, como un paso previo necesario para este ordenamiento de las unidades terminológicas, se debe hacer una selección de las unidades que conformarán la muestra sometida a análisis atendiendo sólo a aquellas que tengan alguna probabilidad de ser interesantes desde un punto de vista terminológico. Esta muestra, como listado de términos, será ordenada por los distintos coeficientes que se presentarán en la sección 3.2.3.

Con el texto ya etiquetado, el paso siguiente es la constitución de listas de palabras o secuencias de palabras del texto que cumplan con aquellos patrones sintácticos frecuentes encontrados en el texto analizado. Dos restricciones adicionales a estos patrones se explican a continuación. Como se verá, la segunda de ellas implica un grado de error, ya que asume que los términos de la disciplina no pueden tener la misma forma de una palabra muy común en la lengua general, lo cual, naturalmente, no es cierto.

1. No incluir un candidato si es, o, en el caso de ser un término sintagmático, si tiene como primer o último componente un miembro de las *n* palabras más frecuentes del corpus de referencia (tales como *el, los, las, al, la, de,* etc.).
2. No incluir un candidato si es o si tiene como primer o último componente una unidad cuya relación (fórmula 1 en sección 3.2.3) entre su frecuencia observada y la frecuencia esperada supera un umbral arbitrario *b*.

En estos experimentos, $b = 0.05$, pero en una implementación este valor puede funcionar como un parámetro de entrada. Incrementar o disminuir este valor tendrá como efecto que el algoritmo sea más o menos tolerante con aquellas unidades que si bien tienen una frecuencia importante en el corpus de especialidad, también la tienen en el corpus de referencia.

Un aspecto a ser destacado es que para eliminar un candidato no tenemos en cuenta información relacionada con su frecuencia en el texto analizado. No sería apropiado ignorar las unidades que tengan baja frecuencia, ya que puede existir una gran cantidad de términos especializados entre los hapax y dis

legomena.

3.2.3. Cálculo de coeficientes para cada candidato

Como se anticipó ya en las secciones anteriores, los coeficientes que se calculan para los candidatos son: un coeficiente de ponderación léxica, uno de ponderación morfológica y uno de ponderación sintáctica. Los distintos coeficientes se combinan según la media que es la que da la puntuación final a cada candidato.

$$w(j) = f_o(j) / (f_e(j) + s)$$

j = unidad analizada
$f_o(j)$ = frecuencia relativa observada de j
$f_e(j)$ = frecuencia relativa esperada de j
s = coeficiente de alisado (~0.001)

(1)

La fórmula 1 es válida para explicar la mayoría de los coeficientes aplicados. Sin importar la naturaleza del fenómeno observado (léxico o morfológico) lo que hacemos en definitiva es poner en relación la frecuencia observada, que es la frecuencia de determinado evento en el corpus de entrenamiento, con la frecuencia esperada, la frecuencia en el corpus de referencia de lengua general que teníamos almacenada durante la etapa de entrenamiento. Para evitar que la frecuencia esperada sea nula, aplicamos un coeficiente de alisado consistente en sumar a la frecuencia esperada un valor mínimo *s*. Este valor mínimo es el que corresponde a un elemento de baja frecuencia relativa, y el propósito fundamental es determinar que un evento que no ha ocurrido antes tiene aproximadamente la misma probabilidad de ocurrir que uno que sí ha ocurrido antes pero con poca frecuencia. Tal como el parámetro *b*, recién introducido en la sección 3.2.2, este valor puede ser definido como un parámetro por el usuario. Ambos parámetros están relacionados, puesto que reducir el valor para el alisado tendrá como efecto una mayor sensibilidad del algoritmo a las unidades de menor frecuencia del corpus de referencia, es decir penalizará más aquellas unidades del texto analizado que estén registradas en el corpus de referencia incluso con baja frecuencia. Incrementar el valor tendría el efecto opuesto, es decir que penalizará sólo aquellas unidades del texto analizado que tengan una alta frecuencia en el corpus de referencia.

En caso de que los candidatos analizados sean términos sintagmáticos, la fórmula 1 se aplica a cada uno de los componentes del término y luego se calcula la media.

Para cada candidato a término j en cada uno de los aspectos evaluados se obtiene una matriz inicial con cuatro valores: 1) la media de los valores del coeficiente léxico (con la proporción obtenida a nivel de forma y a nivel de lema), que denotamos $l(j)$, 2) la media de los valores obtenidos con el coeficiente morfológico (a partir de las secuencias de tres, cuatro y cinco caracteres a principio y final de palabra), denotada $m(j)$, 3) el coeficiente sintáctico que, como vimos en la sección 3.1.3., viene dado por la frecuencia de la estructura sintáctica del término en el diccionario de entrenamiento, denotada $s(j)$, a los que se añade, finalmente, 4) la frecuencia observada de j en el texto analizado, denotada $f(j)$. La puntuación final $w(j)$ de un candidato es el producto (Fórmula 2) de estos cuatro valores.

$$w(j) = l(j) \cdot m(j) \cdot s(j) \cdot f(j) \quad (2)$$

Candidatos	Frec	Forma	Lema	Forma Ref	Lema Ref	Forma/ Ref Ratio	Lema/ Ref Ratio	Ratio Media	...	Total
músculos	14	0.040	0.098	0.001	0.001	26.162	66.010	46.086	...	9920
función	11	0.040	0.040	0.008	0.008	4.696	4.696	4.696	...	85
movimiento	11	0.0512	0.051	0.011	0.011	4.500	4.500	4.500	...	89
parálisis	**9**	**0.175**	**0.175**	**0.001**	**0.001**	**101.798**	**101.798**	**101.798**	**...**	**11552**
acción	8	0.069	0.069	0.016	0.016	4.230	4.230	4.230	...	76
neuronas	8	0.003	0.01	0.001	0.001	2.684	13.322	8.003	...	288
fibras	7	0.019	0.067	0.001	0.001	16.711	46.623	31.667	...	2940
médula	6	0.040	0.040	0.001	0.001	26.028	26.028	26.028	...	1289
trastornos	6	0.027	0.215	0.001	0.001	17.042	157.718	87.380	...	638
corteza	5	0.024	0.024	0.001	0.001	21.839	21.839	21.839	...	118

Tabla 11: Ejemplos de coeficientes aplicados a la lista palabras de la muestra ordenadas por frecuencia.

A los fines ilustrativos, la tabla 11 muestra un ejemplo de los valores que obtienen las palabras de la muestra analizada según algunos de los coeficientes analizados. Se trata de los 10 nombres simples más frecuentes en la muestra, ordenados por frecuencia decreciente. Así, esta tabla nos muestra, además de la frecuencia observada en la muestra analizada, la frecuencia relativa de esa forma en el diccionario de entrenamiento (columna *Forma*), la frecuencia relativa de la lematización de esa forma (columna *Lema*) y, a continuación, los mismos valores pero en el corpus de referencia de lengua general (columnas *Forma Ref* y *Lema Ref*). Las dos columnas siguientes indican la proporción entre los valores del diccionario de entrenamiento y del corpus de referencia (columnas *Forma/Ref Ratio* y *Lema/Ref Ratio*) y la siguiente la media de esos dos valores (columna *Ratio Media*). Por comodidad, omitimos varios de los coeficientes, particularmente los morfológicos, porque el principio es el mismo y se aplican de manera idéntica. Lo importante es ver cómo la puntuación total que obtienen (columna *Total*) difiere del valor dado por la frecuencia, lo cual era uno de los objetivos del experimento. Así, vemos que unidades muy frecuentes pero poco informativas, como *función*, *movimiento* o *acción* son menos ponderadas que otras tanto o menos frecuentes pero más terminológicas, como *parálisis*, *músculos* o *fibras.*

4. RESULTADOS

Esta sección expone los resultados del algoritmo para la muestra analizada. El resultado consiste, por un lado, en una única lista en la que se ordenan los candidatos de acuerdo a la puntuación final que reciben. Tal como se indicó en la fórmula 1, la puntuación final se calcula combinando las puntuaciones que cada candidato obtiene después del análisis léxico, sintáctico y morfológico. Todas las unidades de la muestra sintácticamente aptas aparecen en este listado (en este caso, algo menos de mil unidades). Como es habitual, la idea es que las unidades más terminológicas aparezcan al principio de la lista y las menos al final. Los valores exactos de acierto serán dados en la sección 5, sin embargo, en el primer examen manual de esta lista de 1000 unidades encontramos que alrededor de un cuarto de ellas son unidades terminológicas y, efectivamente, la mayoría de estas unidades se encuentran en las primeras posiciones de la lista de candidatos, un fragmento de la cual se muestra en la tabla 12. Esta tabla exhibe los valores obtenidos para los coeficientes definidos en la sección anterior, y representa la primera de las listas, la que fusiona todos los candidatos sin tener en cuenta su estructura sintáctica. En las tablas posteriores, sin embargo, mostramos además las primeras posiciones en las tablas desglosadas por patrón sintáctico, ordenadas por frecuencia decreciente según el diccionario terminológico (tabla 8). Encontramos el patrón *nombre* en la tabla 13, el patrón *nombre-adjetivo* en la tabla 14, *nombre-preposición-nombre* en la tabla 15, *adjetivo* en la tabla 16, *nombre-adjetivo-adjetivo* en la tabla 17, *nombre-preposición-nombre-adjetivo*, en la tabla 18 y, finalmente, el patrón *nombre-nombre*, en la tabla 19. Naturalmente, existen varios patrones sintácticos más, que no incluimos.

Rango	Candidato	Lema	Frec	Coef Sintax	Coef Lexic	Coef Morf	Total
1	parálisis	parálisis	9	1.00	101.79	12.59	11552.78
2	músculos	músculo	14	1.00	46.08	15.36	9920.42
3	enfermedad	enfermedad	4	1.00	95.30	20.63	7869.79
4	células	célula	3	1.00	49.43	32.96	4889.44
5	nervio	nervio	2	1.00	140.57	13.85	3897.85
6	músculo	músculo	3	1.00	66.01	16.61	3292.35
7	fibras	fibra	7	1.00	31.66	13.25	2940.81
8	célula	célula	1	1.00	59.37	35.30	2096.93
9	fibra	fibra	2	1.00	46.62	14.92	1392.89
10	médula	médula	6	1.00	26.02	8.245	1289.30
11	nervios	nervio	2	1.00	74.59	7.913	1182.16
12	cerebral	cerebral	8	0.16	73.79	12.23	1176.74
13	fibras musculares	fibra muscular	5	0.70	38.61	8.465	1154.82
14	cisura	cisura	1	1.00	48.61	20.86	1014.85
15	ganglios	ganglio	1	1.00	64.40	15.54	1001.67
...	...	...	...	...	...	...	...

Tabla 12: Fragmento de la lista de los candidatos a término extraídos

Rango	Candidato	Lema	Frec	Coef Lexic	Coef Morf	Total
1	parálisis	parálisis	9	101.79	12.59	11552.78
2	músculos	músculo	14	46.08	15.36	9920.42
3	enfermedad	enfermedad	4	95.30	20.63	7869.79
4	células	célula	3	49.43	32.96	4889.44
5	nervio	nervio	2	140.57	13.85	3897.85
6	músculo	músculo	3	66.01	16.61	3292.35
7	fibras	fibra	7	31.66	13.25	2940.81
8	reflejo	reflejo	1	237.76	9.72	376.76
9	célula	célula	1	59.37	35.30	2096.93
...	...	...	...	...	...	...

Tabla 13: Candidatos por patrón *N* (nombre), con frecuencia relativa 1.00

Rango	Candidato	Lema	Frec	Coef Lexic	Coef Morf	Total
1	fibras musculares	fibra muscular	5	38.61	8.46	1154.82
2	nervio periférico	nervio periférico	2	77.79	8.30	912.63
3	corteza cerebral	corteza cerebral	4	47.81	6.65	900.10
4	médula espinal	médula espinal	4	40.32	5.75	656.30
5	célula nerviosa	célula nervioso	1	42.32	21.81	651.96
6	tronco cerebral	tronco cerebral	2	53.61	7.40	560.83
7	enfermedad primaria	enfermedad primario	1	63.01	11.30	503.21
8	fibra muscular	fibra muscular	1	67.41	10.20	485.98
9	células nerviosas	célula nervioso	1	31.11	20.14	442.43
10	sistema nervioso	sistema nervioso	4	15.82	8.79	393.41
11	células motoras	célula motor	1	28.72	17.04	345.71
12	ganglios basales	ganglio basal	1	49.82	8.95	315.25
13	músculos extensores	músculo extensor	2	27.47	7.99	310.64
14	parálisis motoras	parálisis motor	1	54.90	6.86	266.31
15	atrofia muscular	atrofia muscular	1	68.74	4.98	242.40
...	...	...	...	...	...	...

Table 14: Candidatos por patrón *NJ* (nombre-adjetivo), frecuencia relativa 0.705

Rango	Candidato	Lema	Frec	Coef Lexic	Coef Morf	Total
1	tipo de reflejo	tipo de reflejo	1	53.18	6.62	105.61
2	cisura de rolando	cisura de rolando	1	19.32	8.07	46.77
3	sistema de fibras	sistema de fibra	1	13.98	8.38	35.13
4	número de trastornos	número de trastorno	1	29.65	1.95	17.40
5	atrofia de denervación	atrofia de denervación	1	17.09	2.89	14.84
6	trastornos de coordinación	trastorno de coordinación	1	30.34	1.56	14.27
7	tipos de trastornos	tipo de trastorno	1	29.48	1.34	11.92
8	cuerpo en posición	cuerpo en posición	1	5.58	4.97	8.32
9	potencial de acción	potencial de acción	1	8.75	2.22	5.85
10	nombre de fibrilación	nombre de fibrilación	1	3.46	5.14	5.35
11	guía de estímulos	guía de estímulo	1	3.77	1.70	1.93
12	ejecución de movimientos	ejecución de movimiento	1	2.29	1.70	1.17
13	tipo de movimiento	tipo de movimiento	1	1.84	1.66	0.92
14	coordinación de agonistas	coordinación de agonista	1	2.42	1.25	0.92
15	anormalidades de coordinación	anormalidad de coordinación	1	1.52	1.76	0.81
...	...	...	...	...	...	...

Tabla 15: Candidatos por patrón *NPN* (nombre-preposición-nombre), frecuencia relativa 0.299

Rango	Candidato	Lema	Frec	Coef Lexic	Coef Morf	Total
1	cerebral	cerebral	8	73.79	12.23	1176.74
2	reflejo	reflejo	1	237.76	9.72	376.76
3	muscular	muscular	2	88.20	5.48	157.62
4	nervioso	nervioso	5	22.35	8.26	150.44
5	motora	motor	17	27.20	1.91	145.20
6	musculares	muscular	5	45.55	3.67	136.55
7	espinal	espinal	4	54.61	3.26	116.44
8	reflejos	reflejo	2	86.61	2.38	67.47
9	piramidal	piramidal	3	25.05	3.78	46.38
10	fibrosos	fibroso	1	16.47	14.19	38.06
11	nerviosa	nervioso	1	25.28	8.31	34.27
12	nerviosas	nervioso	2	12.78	7.31	30.50
13	neurales	neural	2	15.29	5.32	26.57
14	motoras	motor	16	8.017	1.12	23.69
15	neuromuscular	neuromuscular	1	21.98	6.31	22.63
16	basales	basal	1	35.23	2.36	13.64
17	periférico	periférico	2	15.01	2.74	13.47
...	...	...	...	...	...	...

Tabla 16: Candidatos por patrón *J* (adjetivo), frecuencia relativa 0.162

Rango	Candidato	Lema	Frec	Coef Lexic	Coef Morf	Total
1	célula nerviosa motora	célula nervioso motor	1	37.28	15.18	38.29
2	fibras musculares individuales	fibra muscular individual	2	26.81	5.79	21.05
3	atrofia muscular progresiva	atrofia muscular progresivo	1	49.79	3.68	12.43
4	sistema nervioso humano	sistema nervioso humano	2	12.32	6.24	10.42
5	corteza cerebral contigua	corteza cerebral contiguo	1	32.16	4.60	10.04
6	nervios espinales anteriores	nervio espinal anterior	1	36.39	3.47	8.58
7	fibras nerviosas motoras	fibra nervioso motor	1	17.49	7.23	8.56
8	neurona motora inferior	neurona motor inferior	1	18.81	4.08	5.20
9	neuronas motoras inferiores	neurona motor inferior	2	8.22	2.58	2.89
10	neuronas motoras espinales	neurona motor espinal	1	15.924	2.30	2.49
...	...	...	...	...	...	...

Tabla 17: Candidatos por patrón *NJJ* (nombre-adjetivo-adjetivo), frecuencia relativa 0.067

Rango	Candidato	Lema	Frec	Coef Lexic	Coef Morf	Total
1	número de trastornos funcionales	número de trastorno funcional	1	28.30	1.69	2.48
2	cuerpo en posición vertical	cuerpo en posición vertical	1	8.94	4.16	1.92
3	tipos de trastornos motores	tipo de trastorno motor	1	23.09	1.29	1.54
4	potencial de acción bifásico	potencial de acción bifásico	1	7.47	2.54	0.98
5	guía de estímulos sensoriales	guía de estímulo sensorial	1	4.61	1.59	0.38
6	nombre de haz corticoespinal	nombre de haz corticoespinal	1	0.87	1.80	0.08
...	...	...	...	...	...	...

Tabla 18: Candidatos por patrón *NPNJ* (nombre-preposición-nombre-adjetivo), frecuencia relativa 0.051

Rango	Candidato	Lema	Frec	Coef Lexic	Coef Morf	Total
1	término parálisis	término parálisis	1	51.56	7.26	18.11
2	músculos agonistas	músculo agonista	1	24.86	7.82	9.40
3	núcleos caudal	núcleo caudal	1	7.68	6.30	2.34
4	movimiento agonista	movimiento agonista	1	5.39	1.06	0.27
5	término plejía	término plejía	1	1.16	2.85	0.16
...	...	...	...	...	...	...

Tabla 19: Candidatos por patrón *NN* (nombre-nombre), frecuencia relativa 0.048

El examen manual revela rápidamente que algunos patrones ofrecen más ruido que otros. Particularmente, el patrón *NPN* en la Tabla 15 es uno de las que más ruido produce, con candidatos que son claramente falsos como, por ejemplo, *tipo de reflejo*, *sistema de fibras*, *tipos de trastornos*, *nombre de fibrilación*, etc. También produce una proporción alta de errores el patrón *NPNJ*, produciendo falsos candidatos de forma muy similar al anterior: *número de trastornos funcionales*, *cuerpo en posición vertical*, *nombre de haz corticoespinal*, etc. También resultan ruidosos, aunque en menor proporción, otros patrones como el *NN*, con falsos positivos como *término parálisis* o *término plejía*. Más allá de estos errores en patrones puntuales, lo que es más importante en las tablas es la puntuación final que recibe cada candidato a término, ya que es ese valor el que da la ubicación en el ordenamiento general de los términos. Así, por ejemplo, un candidato erróneo como *nombre de fibrilación* obtiene una puntuación total bastante baja (5.35) que lo ubica en posiciones relativamente bajas en la lista ordenada final, concretamente en la posición 209 de todos los candidatos sintácticamente aptos, que es justamente en la parte intermedia de la lista, donde comienza a haber cada vez menos unidades terminológicas y más ruido. Esta puntuación final es la que permite dar juego a la extracción de candidatos buscando el equilibrio entre precisión y cobertura: si queremos mayor precisión -en detrimento de la cobertura-, vamos a elegir sólo aquellos *n* candidatos que obtengan la mejor puntuación.

5. EVALUACIÓN

En la sección anterior analizamos algunos ejemplos de candidatos a término extraídos por nuestro algoritmo. En esta sección, en cambio, vamos a cuantificar la calidad del desempeño del algoritmo utilizando las medidas de precisión y cobertura, muy frecuentemente utilizadas en los trabajos de extracción terminológica.

En la figura 2 podemos comprobar el desempeño en términos de precisión y cobertura de tres instancias de nuestro algoritmo sobre la misma muestra de texto pero con tres muestras aleatorias de corpus de entrenamiento de tamaño distinto: una primera instancia con una lista de 2.500 términos, una segunda instancia con una lista de 5.000 términos, y finalmente una tercera con algo más de 32.000 entradas. Como anticipábamos en la introducción, si bien es verdad que al aumento en el tamaño del corpus de entrenamiento corresponde un incremento en la calidad de los resultados, este impacto no es dramático, y por tanto las curvas de precisión y cobertura de las tres instancias del algoritmo son prácticamente idénticas. La instancia entrenada con el diccionario entero mantiene un 80% de precisión al 60% de cobertura, mientras que las otras dos apenas superan el 70% de precisión en ese punto de cobertura. Sin embargo, este es el único punto de la figura en que las diferencias son tan palpables. En cualquier otro punto las curvas son similares y prácticamente se confunden.

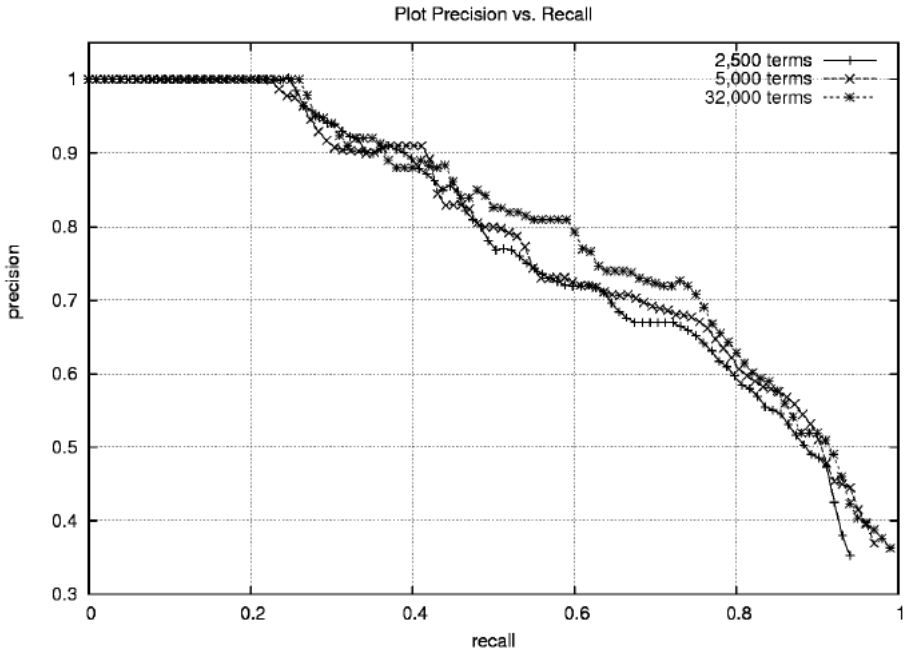

Figura 2: Evaluación del algoritmo en términos de precisión (eje Y) y cobertura (eje X) a distintos tamaños de muestra del corpus de entrenamiento

6. CONCLUSIONES Y TRABAJO FUTURO

En este artículo hemos planteado un método estadístico para la extracción automática de terminología independiente de lengua y de dominio, lo cual representa una contribución de cierta importancia para este campo. Contrariamente a muchos trabajos previos sobre el tema, nuestra metodología no implica la incorporación de conocimiento explícito de la lengua analizada o del dominio sobre el que se está trabajando. En lugar de incorporar esta información, lo que haría muy costoso adaptar el sistema a nuevos contextos de utilización, hemos dotado al sistema de una cierta capacidad de aprendizaje, capacidad que le otorga facilidad de readaptación a nuevas lenguas y dominios y convierte nuestra propuesta en una herramienta útil para ser aplicada en lenguas o en dominios de especialidad que no gozan de amplios recursos lingüísticos o terminográficos. Cabe destacar también, como otra de las cualidades del método, su poca complejidad computacional, lo cual lo hace especialmente apropiado para una implementación en forma de aplicación web, ya que en este entorno conviene que las operaciones no tarden más de unos pocos segundos en realizarse. Naturalmente, el tiempo de respuesta estará en función del tamaño de la muestra analizada, pero lo importante es que el crecimiento del coste computacional es en este caso lineal con respecto al tamaño de la muestra de texto analizada.

En cuanto a las limitaciones de la estrategia, la más evidente es que la calidad de los resultados se verá afectada en buena medida por la calidad de la muestra utilizada para el entrenamiento del algoritmo. La calidad del resultado también dependerá de la selección del dominio a analizar, puesto que no todos los dominios hacen un uso característico de la morfología, como es el caso de los términos de medicina. En un dominio como la lingüística, por ejemplo, tanto la morfología como el léxico se parece mucho más al vocabulario general de la lengua, lo cual presenta un problema para una estrategia como la que presentamos en este artículo. Otra de las limitaciones del algoritmo es la falta de tratamiento de los nombres propios, ya que estos deberían ser identificados y separados del texto a analizar con anterioridad al tratamiento con fines de extracción terminológica. Una de las razones por las que no entramos en esta temática ha sido considerar que se trata de una tarea distinta y que requiere recursos diferentes a los que se han incorporado a nuestra metodología. Creemos, con todo, que esta propuesta se podría integrar con otros procedimientos para la extracción de nombres propios y tenemos, además, algunas ideas al respecto que en principio son también independientes de lengua, tales como medir la probabilidad de que los candidatos aparezcan con mayúsculo inicial, tanto en el texto analizado como en un corpus de referencia. Este recurso, naturalmente,

será problemático en lenguas como el alemán, ya que el uso de las mayúsculas iniciales sigue una normativa distinta.

En relación a líneas de trabajo futuro, el paso inmediato es evaluar esta metodología en diversas lenguas y dominios temáticos para obtener una comparación del desempeño del extractor en distintos casos. Para ello, se está implementando este algoritmo en un servidor web para que quede a disposición de la comunidad y pueda ser entrenado y aplicado en distintas lenguas y dominios. Existe ya una versión demo en la web, aunque de manera algo embrionaria y en una dirección provisional[3]. Posteriormente tenemos previsto incorporar este extractor como un módulo más en el software Terminus (Cabré et al, 2010). Estamos trabajando además en una versión más refinada de este sistema combinándolo con otros que basan la extracción de terminología en grafos de coocurrencia léxica (Nazar, 2010). Este cambio implicaría una gran potenciación de la cobertura de nuestro algoritmo para el reconocimiento de terminología, aunque con un coste computacional más alto. La principal ventaja de los grafos de coocurrencia es que permiten caracterizar de un modo bastante aproximado los fenómenos de polisemia y homonimia, que en el presente artículo se encuentran infrarrepresentados. Esto es un problema porque si un término tiene la misma forma que una palabra de uso muy frecuente en la lengua cotidiana, entonces existe la posibilidad de que tal término no sea registrado por este sistema, lo cual no ocurriría con los grafos de coocurrencia ya que estos permiten caracterizar una unidad léxica no sólo por su forma sino por las unidades que suelen encontrarse en sus contextos de aparición.

Finalmente, como trabajo futuro a largo plazo, pensamos integrar este sistema de extracción de terminología en una cadena de trabajo que incorpora otras operaciones de procesamiento del lenguaje natural. Más específicamente, la extracción de equivalencias entre términos de distintas lenguas y el establecimiento de relaciones conceptuales entre los términos extraídos.

NOTAS

1 Algunos de los extractores que pueden ser probados en línea son los siguientes:
TermExtractor (Sclano; Velardi, 2007): http://lcl2.uniroma1.it/termextractor/
Termostat (Drouin, 2003): http://idefix.ling.umontreal.ca/~drouinp/termostat_web/
TerMine (Frantzi et al., 2000): http://www.nactem.ac.uk/software/termine/cgi-bin/termine_cvalue.cgi
YATE (Vivaldi, 2001): http://igraine.upf.edu/cgi-bin/Yate-on-the-Web/yotwMain.pl
Translated Labs (T-Labs): http://labs.translated.net/terminology-extraction/ o la API Term Extraction de Yahoo: http://developer.yahoo.com/search/content/V1/termExtraction.html

2 Este corpus puede descargarse de Internet en la dirección del diario: http://www.elpais.es

3 La dirección de esta interfaz es http://melot.upf.edu/terminal/ [Con acceso: 15-02-2011].

BIBLIOGRAFÍA

Ananiadou, S. (1994). "A Methodology for Automatic Term Recognition". Coling 1994, 15th International Conference on Computational Linguistics, Kyoto, Japan

Arntz, R ; Picht, H. (1989/1995). "Introducción a la terminología". Madrid, Fundación Germán Sánchez Ruipérez

Bourigault, D. (1994). "LEXTER, un Logiciel d'EXtraction de TERminologie. Application à l'acquisition des connaissances à partir de textes". Ph. D. Thesis. Paris: École des Hautes Études en Sciences Sociales

Cabré, T. (1999). "La Terminología: Representación y Comunicación". Barcelona, Institut Universitari de Lingüística Aplicada

Cabré, T.; Estopá, R. (2005). “Unidades de conocimiento especializado, caracterización y tipología”. En Cabré, M. T.; Bach, C. (eds.). Coneixement, llenguatge i discurs especialitzat, pp. 69–94. Barcelona: Institut Universitari de Lingüística Aplicada

Cabré, M. T.; Estopà, R. ; Vivaldi, J. (2001). “Automatic Term Detection: A review of Current Systems”. En: Bourigault, D., Jacquemin, C. y L'Homme, M-C. (eds.) Recent Advances in Computational Terminology. Amsterdam: John Benjamins, pp. 1–28

Cabré, T; Montane, A.; Nazar, R.; Reus, R. (en prensa). “Estació Terminus: a Web Application for Terminology and Corpus Management”. Proceedings of TKE 2010 Conference (Terminology and Knowledge Engineering) Dublin 12-13 August, 2010

Daille, B. (1994). “Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques”. Thèse de Doctorat en Informatique Fondamentale. Université Paris 7

Drouin, P. (2003). “Term extraction using non-technical corpora as a point of leverage”. Terminology, vol. 9, no. 1, pp. 99–117

Enguehard, C. ; Pantera, L. (1994). “Automatic Natural Acquisition of a Terminology”. Journal of Quantitative Linguistics, vol. 2, no. 1, pp. 27–32

Estopà, R., J. Vivaldi ; T. Cabré. (2000). “Use of Greek and Latin forms for term detection”. Proceedings of the Second International Conference on Language Recources and Evaluation (LREC 2000), pp. 51–56. Athens, Greece

Frantzi, K.T. (1997). “Incorporating context information for extraction of terms”. Proceedings of the Association for Computational Linguistics (ACL/EACL), 501–503. Madrid, Spain

Frantzi, K., Ananiadou, S. ; Mima, H. (2000). “Automatic recognition of multi-word terms”. International Journal of Digital Libraries vol. 3, no. 2, pp.117– 132

Hisamitsu T., Niwa, Y,. Nishioka, S., Sakurai, H., Imaichi, O., Iwayama, M. ; Takano, A. (2000). “Term Extraction Using A New Measure of Term Representativeness”. Proceedings of the Second International Conference on Language Recources and Evaluation (LREC 2000). Workshop Proceedings on: Terminology Resources and Computation, 13–20. May 29, Athens, Greece

Jacquemin, C. (1997). “Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. m´moire d’habilitation ` diriger des recherches en informatique

Justeson J.; Katz, S. (1995). “Technical terminology: some linguistic properties and an algorithm for identification in text”. Natural Language Engineering, vol.1, no. 1, pp. 9–27

Kageura, K. ; Umino, B. (1996). “Methods of Automatic Term Recognition”. Terminology, vol. 3, no. 2, pp. 259–290

Maicher, L. (2004). “Subject Identification in Topic Maps in Theory and Practice”. En Tolksdorf, R., Eckstein, R. (eds.) Berliner XML Tage 2004. XML-Clearinghouse, Berlin, pp. 301–307

Maynard, D. ; Ananiadou, S. (1999). “Identifying contextual information for multi-word term extraction”. TKE ‘99: Terminology and Knowledge Engineering, 212–221. Vienna: TermNet

Maynard, D. ; Ananiadou, S. (2000b). “TRUCKS: A Model for Automatic Multi-Word Term Recognition”. Journal of Natural Language Processing, vol. 8, no. 1, pp. 101–125

Mosby (2001). “Diccionario Mosby de Medicina, Enfermería y Ciencias de la salud”. Quinta edición. Harcourt: Madrid. Versión en lengua española de la 5.ª edición de la obra original en inglés: Mosby's

Medical, Nursing, and Allied Health Dictionary Copyright ©MCMXCVIII by Mosby Year Book, Inc.

Nazar, R. (2010). “Estudio diacrónico de la terminología especializada utilizando métodos cuantitativos: ejemplo de aplicación a un corpus de lingüística aplicada”, Revista Signos, Valparíso, Chile

Nazar, R. (2010). “A Quantitative Approach to Concept Analysis”. Ph.D. Thesis, Universitat Pompeu Fabra

Pantel, P.; Lin, D. (2001). “A Statistical Corpus-Based Term Extractor”. Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence. London, UK, Springer-Verlag, pp. 36–46

Patry, A.; Langlais, P. (2005). “Corpus-Based Terminology Extraction”. Proceedings of the 7th International Conference on Terminology and Knowledge Engineering, Copenhagen, Danemark, pp. 313–321

Sager, J. (1990). “A Practical Course in Terminology Processing”. Amsterdam/Philadelphia: John Benjamins

Schmid, H. (1994). “Probabilistic Part-of-Speech Tagging Using Decision Trees”. Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK, pp. 44–49

Sclano, F. ; Velardi, P. (2007). “TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities”. 9th Conf. on Terminology and Artificial Intelligence TIA 2007, Sophia Antinopolis

Sheremetyeva, S. (2009). “On Extracting Multiword NP Terminology for MT”. Proceedings of the EAMT Conference. Barcelona, Spain, May 13-15

Sparck Jones, K. (1972). “A statistical interpretation of term specificity and its application in retrieval”. Journal of Documentation, vol. 28, no. 1, pp. 11–21

Vivaldi, J. (2001). “Extracción de candidatos a término mediante combinación de estrategias heterogéneas”. PhD Thesis, IULA, Universitat Pompeu Fabra

Vivaldi, J. (2009). “Corpus and exploitation tool: IULACT and bwanaNet”. A survey on corpus-based research = Panorama de investigaciones basadas en corpus [Actas del I Congreso Internacional de Lingüística de Corpus (CICL-09), 7-9 Mayo 2009, Universidad de Murcia]. Murcia: Asociación Española de Lingüística del Corpus, pp. 224–239

Vivaldi, J.; Rodríguez, H. (2001). “Improving term extraction by combining different techniques”. Terminology, vol. 7, no. 1, pp. 31–48

Vivaldi, J; Araya, R. (2004) “Mercedes: detección de términos en el CT-IULA”. Seminario IULATerm, Universidad Pompeu Fabra, 14/07/2004

Vivaldi, J.; Rodríguez, H. (2006). “Ontologías y Extracción de Términos” , en La terminología en el siglo XXI: contribución a la cultura de la paz, la diversidad y la sostenibilidad: Actas del IX Simposio Iberoamericano de Terminología RITERM04. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra, pp. 441–454

Wüster, E. (1979/1998). “Introducción a la teoría general de la terminología y a la lexicografía”. Institut Universitari de Lingüística Aplicada, Barcelona.