

Explotación computacional del metalenguaje en corpus especializados para la generación de lexicones no convencionales

Carlos Rodríguez Penagos

Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas
Universidad Nacional Autónoma de México (UNAM)
Cto. Escolar S/N, Cd. Universitaria, Coyoacán, México, D. F.
crodriguezp@turing.iimas.unam.mx

Resumen: En este artículo se presentan técnicas de análisis automático de base estadística y simbólica para la detección y procesamiento de metalenguaje en textos altamente técnicos, en varios dominios de especialidad. La extracción selectiva de información metalingüística realizada por el sistema MOP permite la obtención de lexicones complementarios no convencionales como ayuda para el procesamiento del lenguaje natural.

Palabras clave: Terminología, procesamiento del lenguaje natural, conocimiento científico, definiciones, metalenguaje, extracción de información, adquisición léxica, corpus especializados, procesamiento del discurso.

Abstract: This paper presents the application of automatic analysis (of statistical and symbolic nature) for the detection and processing of metalanguage in highly technical texts from various domains. The selective metalinguistic information extraction performed by the MOP system allows compilation of non-conventional lexicons to aid domain-restricted NLP.

Keywords: Terminology, natural language processing, scientific knowledge, definitions, metalanguage, information extraction, lexical acquisition, specialized corpora, discourse processing.

1 Introducción

El crecimiento exponencial de la información científica en formato electrónico presenta un reto a disciplinas como la biomedicina, cuyos especialistas deben manejar una vasta literatura de millones de artículos (el llamado *biobiblioma*). Sin técnicas robustas de procesamiento informático (creación de índices, extractores de terminología, etc.) sería imposible construir bases de conocimiento como el repositorio indexado MedLine de la National Library of Medicine norteamericana, que cada año añade alrededor de 400.000 abstracts de artículos de investigación (Powell *et al.*, 2002).

Este trabajo describe la aplicación de técnicas derivadas de la Extracción de Información para la explotación del metalenguaje presente en los textos de especialidad; también muestra cómo es posible la consolidación de dicha información en bases de conocimiento útiles para el trabajo terminográfico y para sistemas de

procesamiento e inferencia dentro de dominios especializados. El sistema Procesador de Operaciones Metalingüísticas (*Metalinguistic Operation Processor*, o MOP) extrae enunciados metalingüísticos y definiciones de documentos técnicos, utilizando tanto autómatas de estados finitos como algoritmos de aprendizaje automático. Este sistema se diferencia de otros sistemas terminográficos en que no se basa simplemente en regularidades sintácticas, morfológicas o semánticas implícitas de la formación de los términos (Jacquemin, 2001), las estructuras semánticas (Hearst, 1998) o los conceptos (Kageura, 2002), sino que explota la dimensión más discursiva de actos de habla en los cuales los términos son establecidos, modificados o valorados de manera explícita por los propios participantes en la actividad científica.

El sistema crea bases semi-estructuradas de información acerca de la terminología utilizada, llamadas Bases de Información Metalingüística (*Metalinguistic Information Databases*, o MIDs). Además de su utilidad para el PLN, esta tecnología permite la investigación de las dinámicas de la evolución y construcción

consensual del conocimiento científico en las disciplinas modernas. Presentaremos primero una perspectiva general sobre el metalenguaje en contextos especializados, y luego pasaremos a la descripción del sistema y a las evaluaciones del mismo.

2 El metalenguaje en lenguajes formales y naturales

Desde la óptica de la matemática y la lógica, un metalenguaje constituye la fundamentación del lenguaje-objeto de los sistemas formales. En sistemas semióticos de base no axiomática como la lengua se ha reconocido también un rol fundacional. Aunque el lenguaje cotidiano podría verse sincrónicamente como un sistema completo, con la mayor parte del léxico adquirido en una etapa relativamente temprana de la vida, la comunicación especializada depende crucialmente de un inventario léxico en constante cambio, con la introducción permanente de términos y conceptos que refleja los cambios en el estado del arte de una disciplina.¹

Mantener al día el conocimiento de dominios altamente fluidos es sumamente costoso. Se ha dicho, por ejemplo (Boguraev y Levin, 1993), que una de las limitaciones de las bases de datos lexicográficas como “repositorios” es su incapacidad para representar la productividad y el carácter abierto del léxico. Esta limitante motivó la idea de Boguraev y Pustejovsky (1996) de extraer el lexicón a partir de los propios textos, así como el proyecto ACQUILEX de la década de los noventa (Copestake y Vossen, 1993). Más recientemente, podemos mencionar los trabajos de Pearson (1998), Rebeyrolle y Péry-Woodley (1998), Rodríguez (1999), Cuoto y Crispino (1999), Meyer (2001), Klavans y Muresan (2001), y Sierra y Alarcón (2002) como esfuerzos similares tendientes a entender y procesar automáticamente este tipo de fragmentos textuales especializados, que Meyer ha llamado “contextos ricos en conocimiento”.

Algunos ejemplos de estos contextos son los siguientes:

- (1) *In 1965 the term soliton was coined to describe waves with this remarkable behaviour.*

¹ Ives Gentilhomme (1994) ha utilizado la metáfora de los sistemas metaestables de la química para describir esta dinámica.

- (2) *En este entramado se puede llegar a distinguir sucesivas expresiones de lo heredado, aunque el término fenotipo se suele reservar para la expresión final.*

El control terminológico se da en segmentos textuales cognitivamente importantes, de los cuales las familiares definiciones analíticas y lexicográficas forman un subconjunto limitado fuera de diccionarios y de textos didácticos. Algunos de estos actos de habla singulares difícilmente pueden ser homologados como definiciones en un sentido estricto o lógico, ya que no establecen de una manera completa el lugar de la unidad léxica en el sistema conceptual o terminológico del área; con frecuencia se limitan a señalar un rasgo adicional, una condición de uso del término o una relación semántica específica con otro. Su importancia reside en constituir instancias discursivas que atestiguan el proceso de negociación, cambio y consolidación de un terminología, y a través de ella del conocimiento especializado.

3 El sistema MOP para la extracción de bases de información metalingüística

El trabajo preliminar llevado a cabo para desarrollar el sistema MOP estudió empíricamente el metalenguaje en lenguajes de especialidad en inglés, enfocándose en el establecimiento, modificación y negociación de una terminología común entre los grupos de especialistas de cada área, lo que Clark (1998) ha llamado los Léxicos Comunitarios. Este estudio se basó en un análisis tanto manual como automático de varios corpus con un alto grado de tecnicidad, y provenientes de diversos dominios. Después de una exploración inicial de una colección de documentos especializados (mayoritariamente artículos de Journals de Sociología) y la revisión de algunos inventarios de marcadores léxicos de la definición, se hicieron búsquedas de dichos patrones en segmentos escritos especializados del British National Corpus, y se compiló así un corpus general de más de 11,000 oraciones potencialmente metalingüísticas (el corpus EMO, *Explicit Metalinguistic Operations*). Durante el análisis del extenso corpus descrito anteriormente se marcó de manera manual si una oración era metalingüística o no (5.407 lo eran, un 49,6% del total). Por otro lado se compiló una serie de archivos que se detallarán más adelante (entre otros, documentos

sociología, histología, biomedicina) que fueron también marcados manualmente para ser usados como estándares de evaluación y validación de los resultados generales del procesamiento.

El sistema MOP (Figura 1) realiza una extracción selectiva de enunciados metalingüísticos y definiciones en documentos especializados, utilizando tanto autómatas de estados finitos como algoritmos de aprendizaje automático. Una vez seleccionados los enunciados, el sistema realiza un procesamiento de base lingüística para crear bases semiestructuradas de información terminológica llamadas Bases de Información Metalingüística (*Metalinguistic Information Databases*, o MIDs).²

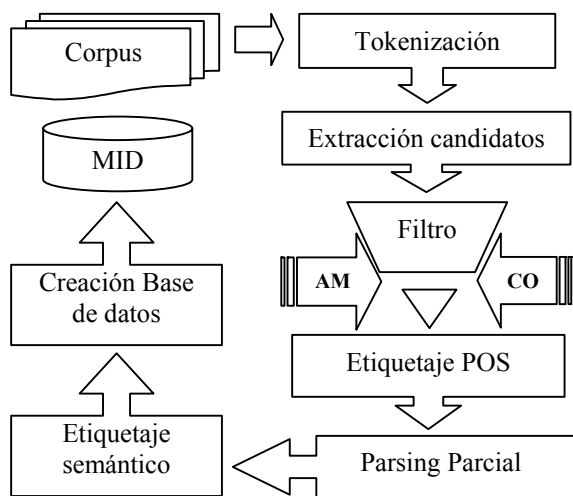


Figura 1. Arquitectura del sistema MOP (AM: Aprendizaje-Máquina; CO: Colocaciones)

3.1 Dos estrategias de extracción

Un problema difícil para los sistemas que utilizan indicadores léxicos es el de controlar la cantidad de “ruido” que se introduce si no se realiza una desambiguación efectiva. Para la extracción inicial de metalenguaje se utilizaron un total de 44 patrones léxico-ortográficos que demostraron ser indicadores fiables de condición metalingüística en nuestra exploración inicial del corpus EMO, como comillas, “known as”, “called”, “termed”, “denote”, “coined”, “defined as”, etc. Una vez que se obtuvieron un conjunto de oraciones candidatas mediante cascadas de expresiones

² El sistema fue codificado en Python, utilizando la plataforma de desarrollo Natural Language Toolkit (<http://nltk.sf.net>).

regulares sobre los textos especializados de evaluación, se experimentó con dos técnicas diferentes para descartar instancias no metalingüísticas como en (4), o conservar instancias claramente pertinentes como (3), en la que se define el término *unacknowledged shame*, (ambas obtenidas a partir de una expresión regular de un mismo lema en negritas).

(3) *Since the shame that was elicited by the coding procedure was seldom explicitly mentioned by the patient or the therapist, Lewis **called** it unacknowledged shame.*

(4) *It was Lewis who **called** attention to emotional elements in what until then had been construed as a perceptual phenomenon.*

Primero se realizó un análisis estadístico de las colocaciones atestiguadas en el corpus EMO, algunas de las cuales permiten la desambiguación entre usos verbales y nominales, como en la siguiente muestra de patrones exclorios, de manera que en el ejemplo (4) la aparición de “attention” después del marcador “called” descartaría su uso metalingüístico.

colocación previa	colocación posterior
in, duty, personal, conference, local, next, the, their, house, anonymous, phone, telephone, system ...	out, someone, charges, before, charge, back, contact, for, upon, to, into, off, 911, by...

Tabla A. Muestra de colocaciones exclorios para el lema “called”

Para cada patrón de extracción se utilizaron como criterio excluyente las colocaciones atestiguadas más de 2 veces en el corpus en instancias marcadas como no metalingüísticas, y que a su vez no se encontraran atestiguadas ni una sola vez en los ejemplos efectivamente metalingüísticos. Es decir, definimos el conjunto E de patrones excluyentes como $\{x: x \in N-S \ \& \ n(x) \geq 2\} \cup A$, donde N es el conjunto de colocaciones atestiguadas en oraciones NO metalingüísticas, S el de las colocaciones en oraciones metalingüísticas, y A es un conjunto de colocaciones adicionales obtenidas de diccionarios de verbos frasales y otras fuentes complementarias.

Como sucede frecuentemente en el caso de las reglas elaboradas manualmente algunos casos resultan altamente dependientes de dominio (e.g. “esophageal coins”), lo que

reduce la portabilidad del sistema. Este hecho, y el costo general de desarrollo de una aproximación manual de escritura de reglas, han hecho que se experimente con la compilación automática de diccionarios de patrones mediante técnicas de aprendizaje automático (Riloff y Jones, 1999), para las cuales se han reportado recientemente resultados equiparables al tratamiento manual (Chieu, Ng y Lee, 2003). Este hecho nos condujo a probar el uso de algoritmos de aprendizaje con características bien descritas en la literatura, y que pudiesen funcionar con un conjunto de datos relativamente escaso. Se implementaron algoritmos Bayes ingenuo y Entropía Máxima (este último con dos modalidades de entrenamiento reiterativo). El corpus de entrenamiento fueron las casi 11.000 oraciones del corpus EMO. Se utilizaron como vectores para cada algoritmo 1, 2 y 3 lexemas antes y después del marcador, en consonancia con nuestra hipótesis de que el contexto gramatical del marcador era importante. En esas posiciones se probó tanto la categoría gramatical de los lexemas (la etiqueta POS),³ como la forma ortográfica de las palabras.

En los siguientes ejemplos de vectores a partir de una oración obtenida mediante el lema “calls” (“... creates what Croft calls a description constraint ...”) se generaron estructuras de datos etiquetadas a partir de los elementos en 3 posiciones antes y después del marcador, ya sea con etiquetas POS o mediante las formas ortográficas:

(‘VB WP NNP’, ‘calls’, ‘DT NN NN’)/‘YES’@[102].

(‘creates what Croft’, ‘calls’, ‘a description constraint’)/‘YES’@[102].

Los clasificadores así entrenados se aplicaron al las oraciones obtenidas por la extracción inicial para determinar si se trataba de casos genuinamente metalingüísticos, o no.

3.2 Generación de MIDs

Una vez que se obtuvo un conjunto confiable de operaciones metalingüísticas, se procedió a realizar un procesamiento lingüístico más o menos estándar. Tras un análisis morfo-sintáctico adaptado a contextos terminológicos que identificó posibles bloques nominales, se realizó un etiquetaje semántico que asignó

ciertos roles a argumentos de la oración articulados alrededor del marcador metalingüístico. A continuación se aplicaron reglas basadas en marcos predicativos (*semantic frames*) obtenidos a partir de análisis de corpus y la consulta en FrameNet (Baker *et al.*, 1998), con el objetivo de crear los registros de las MIDs, las bases de datos que son el objetivo final del sistema. Las *Metalinguistic Information Databases* son archivos XML con registros para los tres elementos constitutivos de las operaciones metalingüísticas explícitas:

(a) el **término** o las unidades terminológicas en condición autonómica a las cuales se refiere predicación; (b) la **información** de naturaleza semántico-pragmática que se está aportando para dicha unidad (segmentos informativos), y (c) los **marcadores/operadores** que permiten vincular los dos elementos anteriores y marcar la naturaleza metalingüística del fragmento textual.

El siguiente ejemplo es el resultado del procesamiento mediante el sistema de un pequeño fragmento de MedLine.

Autonym:	postsynaptic density-95
Info:	a protein
Marker: <i>known as</i>	The NMDA receptor can bind a protein known as postsynaptic density-95 (PSD-95), which may regulate the localization of and/or signalling by the receptor.
Autonym:	decoy receptor 3 (DcR3)
Info:	a soluble decoy receptor
Marker: <i>termed</i>	Here we report the discovery of a soluble decoy receptor, termed decoy receptor 3 (DcR3), that binds to FasL and inhibits FasL-induced apoptosis.
Autonym:	Zfp106
Info:	The H3a gene
Marker: <i>called</i>	The H3a gene, now called Zfp106, encodes a 1888-amino acid protein with three zinc fingers and a beta-transducin domain consistent with DNA/protein binding.

Tabla B. Fragmento de una MID.

Debido a que se buscó limitar la complejidad del procesamiento lingüístico no se incluyó en esta versión del sistema ni un parser sintáctico exhaustivo ni un módulo de resolución de anáfora o coreferencia, por lo que algunas de las entradas recogen como información “semántica” tan sólo el pronombre que apunta al referente del término.

³ Se utilizó un etiquetador tipo Brill implementado en el MIT por Hugo Liu.

4 Evaluación del sistema

La evaluación del sistema de extracción automática se realizó en base a tres corpus de referencia (que fueron marcados manualmente): A) un conjunto de 19 artículos de revistas académicas de sociología (5581 oraciones), B) un manual de histología clínica (5146 oraciones) y C) un conjunto de aproximadamente 200 abstracts de biomedicina del MedLine (1403 oraciones). Se utilizaron los parámetros estándares de Extracción de Información, como la cobertura (*recall*) o la precisión (*precision*), y se utilizó la medida F-measure con un factor beta de 1 para balancear ambos parámetros.

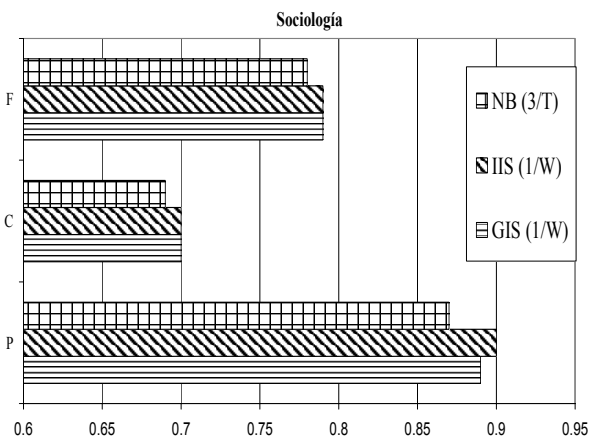
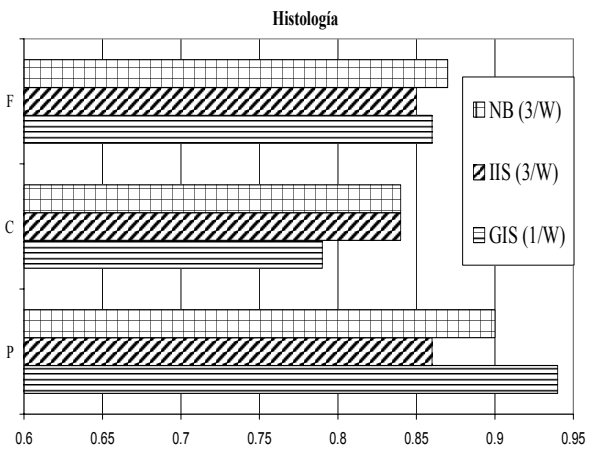
La primera evaluación permitió comparar los dos métodos de filtraje implementados. Por un lado, el método basado en colocaciones dio resultados muy satisfactorios, como lo muestra la siguiente tabla (C) para dos de nuestros archivos de prueba:

Dominio	Líneas extraídas	Filtradas (%)	Precisión	Cobertura
Sociología	143	14 (9,8)	0,97	0,79
Histología	37	5 (13,5)	0,94	0,81

Tabla C. Métricas de la tarea de extracción

Aunque las evaluaciones del método basado en colocaciones reforzaron nuestra hipótesis de que el entorno gramatical de los marcadores era determinante para el filtraje de las oraciones, las pruebas con aprendizaje automático no validaron la intuición de que un contexto gramatical más restrictivo (usando las etiquetas de POS de los elementos aledaños) o un mayor número de posiciones antes y después de los marcadores generarían mejores resultados. La precisión y cobertura con algoritmos de aprendizaje automático fueron igualmente buenas, pero no fueron concluyentes por lo que respecta a superioridades fundamentales de uno u otro respecto a tipo de algoritmo o al conjunto de rasgos vectores para la clasificación. Los métodos estocásticos y los de compilación manual de colocaciones presentaron resultados muy similares, aunque el aprendizaje automático es más promisorio respecto a las posibilidades de portabilidad a otros dominios. Por otra parte, se pudo comprobar que los patrones de extracción obtenidos mediante el trabajo de corpus son bastante fiables, y presentan una independencia relativa respecto al dominio (aunque esto no descarta que

algunas disciplinas tengan protocolos muy específicos para la predicación metalingüística, y se deba hacer un proceso de compilación adicional).



Figuras 2 y 3. Mejores resultados por dominio, para algoritmos de aprendizaje automático⁴

Con el corpus de sociología (Figura 3), el algoritmo de Entropía Máxima que utilizó formas ortográficas de una palabra antes y después del marcador presentó precisión de 0,9 y cobertura de 0,7 mientras que los mejores resultados para el corpus de histología (Figura 2) se lograron con una red bayesiana que utilizó 3 formas ortográficas antes y después del marcador (P=0,9 /C=0,84). Por lo que toca a la evaluación de la Extracción de Información

⁴ En las figuras, NB significa Bayes ingenuo; IIS: Entropía Máxima con entrenamiento *Improved IS*; GIS: Entropía Máxima con entrenamiento *Generalized IS*. Entre paréntesis: (posiciones a la derecha e izquierda del marcador / W: formas ortográficas o T: categoría gramatical); Las columnas corresponden a F: F-Measure; C: Cobertura y P: Precisión).

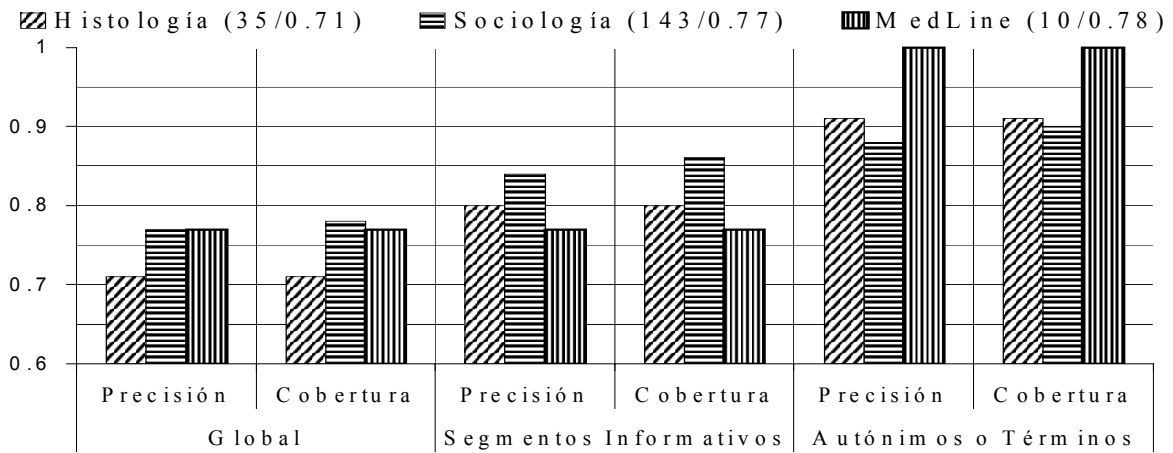


Figura 4. Métricas para 3 corpus
(# de registros/F-Measure global)

propia mente dicha,⁵ se obtuvieron valores comparables o superiores a los de otros sistemas similares por ejemplo el sistema DEFINDER (P=0,8 / C=0,75; Klavans y Muresan, 2001). Sin embargo, es necesario hacer notar que mientras DEFINDER trabaja con definiciones de documentos lexicográficos y didácticos en los que existe una gran regularidad en la redacción, MOP se enfrenta a una amplia casuística en documentos de investigación con características muy diferentes. Por otro lado, las comparaciones con sistemas tradicionales de Extracción de Información son poco ilustrativas debido al orden de complejidad de la tarea de extraer tan solo tres campos a una base de datos, frente a algunos sistemas que extraen al menos una docena de campos heterogéneos.

La Figura 4 muestra los resultados de la evaluación de esta etapa en los tres ámbitos temáticos diferentes de los documentos de evaluación, para cada tipo de campo extraído, y global para todos los registros de la base de datos. En general, esta primera aproximación a una comparación entre dominios de especialidad permite suponer que el sistema MOP es aplicable a un gran abanico de áreas

⁵ Se utilizó para la evaluación general sólo el método de colocaciones en la extracción inicial, por ser el mejor entendido. También, se estableció que si al menos las dos terceras partes de la cadena obtenida para cada campo coincidían con el campo en el estándar de evaluación, el registro se tomaría como correcto, para compensar pequeños errores en los algoritmos de procesamiento de sintagmas preposicionales y de acrónimos.

temáticas, y es susceptible de adaptación a lenguas diferentes del inglés. Se observó que pese a las obvias diferencias idiomáticas y culturales, los protocolos de negociación metalingüística están altamente normalizados al interior de cada lengua, tanto respecto a los patrones léxicos y paralingüísticos que los marcan como respecto a sus funciones discursivas.

5 Hacia la exploración del meta-lenguaje en el biobiblioma: conclusiones y perspectivas

Sistemas como MOP que permiten explorar una dimensión metalingüística poco explotada pero epistemológicamente rica del discurso especializado abren posibilidades muy interesantes de aplicación. Sin embargo, es importante hacer notar que las MIDs generadas por el sistema no son en pleno sentido recursos terminológicos acabados, sino bases semi-estructuradas a mitad de camino entre los corpora textuales y las bases de datos terminológicas. Una característica de estos recursos es que a diferencia de los diccionarios y lexicones, la información que ofrecen no contiene los defaults usuales de los diccionarios, sino que precisamente recogen las excepciones, las innovaciones, los cambios, y en general los datos que los hablantes consideran relevantes discursivamente porque o no pertenecen a la competencia general supuesta en el especialista, o no es posible derivarlos inferencialmente de la información a disposición de los hablantes en una

comunicación técnica. Esto hace que sea posible considerarlos lexicones con información no convencional que pueden servir a motores de inferencia a los cuales los lexicones más regulares no pueden proporcionar información cuando se procesan enunciados que contienen usos novedosos o altamente restringidos de una unidad léxica. La información léxica más convencional puede sobrepasarse mediante información discursiva o pragmática más específica, si el contexto lo requiere (Lascarides y Copestake, 1995). Evidentemente, se trataría de complementar esos recursos, no de substituirlos.

Las MIDs también tienen como objetivo informar el trabajo terminográfico ofreciendo neología e información semi-procesada para su validación manual, o como materia prima para un refinamiento automático posterior que permita obtener datos estructurados; por ejemplo, uno de los registros de la Figura 1, informa que el *decoy receptor 3* es un subtipo de receptor soluble, y que puede abreviarse mediante *DcR3*. A través de una tipificación semántica como esta, las MIDs podrían ayudar a la revisión y actualización de ontologías compiladas manualmente.

Las posibilidades del análisis cuantitativo de terminologías y conocimiento disciplinario están siendo exploradas en la actualidad por el autor mediante la constitución de varios corpus considerablemente más vastos que los empleados en la etapa de desarrollo y pruebas. Se han compilado otros corpus de artículos de revistas académicas de biomedicina y de sociología con al menos 5 millones de palabras (incluyendo puntuación), para los cuales se conocen sus índices de citación. Su procesamiento se encuentra en una etapa preliminar, pero la precisión lograda hasta el momento con el método de colocaciones es excelente (entre 0,94 y 0,95) en ambos casos, aunque la cobertura no ha sido determinada aún con exactitud.

Consideramos que la utilización de estos corpus más extensos hará posible investigar algunas cuestiones poco exploradas, como saber si el uso de metalenguaje juega algún papel en la aceptación o relevancia de los artículos, si tiene dinámicas diferentes según la disciplina, etc. El procesamiento de un segmento más amplio del *biobiblioma* permitirá establecer, por ejemplo, qué tan amplia es la cobertura de los lexicones MeSH y Specialist (con más de 2,5 millones de términos) del Instituto Nacional de Medicina de

los Estados Unidos, respecto a los MIDs de términos extraídos con el sistema MOP directamente a partir de los textos.

El trabajo presentado aquí ha mostrado la factibilidad de implementar sistemas robustos para el procesamiento automático del metalenguaje en textos no estructurados. El sistema MOP puede refinarse subsecuentemente mediante la adición de módulos de análisis sintáctico profundo, de resolución de anáfora y de post-procesamiento de las MIDs para crear recursos prácticos para el procesamiento del lenguaje especializado. Tecnologías del lenguaje como la extracción de información comienzan a mostrar un grado de madurez considerable que permitirán en un futuro no muy lejano el salto del laboratorio al trabajo diario de los especialistas.

Bibliografía

- Baker, C., Fillmore, Ch., y Lowe, J. (1998) The Berkeley FrameNet project. *Proc. COLING-ACL*, Montreal, Canada.
- Boguraev, B. y Levin, B. 1993. Models for Lexical Knowledge Bases, *Semantics and the Lexicon*, Kluwer, Dordrecht.
- Boguraev, B. y Pustejovsky, J. 1996. Issues in Text-based Lexicon Acquisition, *Corpus Processing for Lexical Acquisition*, The MIT Press..
- Cuouto, J., Crispino et al. 1999. Estructuración de Índices Gramaticales y Léxicos para la Extracción y Recuperación de Información. *Procesamiento del Lenguaje Natural*. SEPLN. No. 25
- Chieu, H., Ng, H., y Lee, Y. 2003. Closing the Gap: Learning-Based Information Extraction Rivaling Knowledge-Engineering Methods. *Proc. 41st ACL*. Sapporo, Japan.
- Gentilhomme, Y. 1994. L'éclatement du signifié dans les discours technoscientifiques. *Cahiers de Lexicologie*, Vol. LXIV-1, p.5-53.
- Hearst, M. 1998. Automated discovery of wordnet relations. *WordNet: An Electronic Lexical Database*. The MIT Press.
- Jacquemin, Ch. 2001. *Spotting and discovering terms through NLP*. The MIT Press.
- Klavans, J. y S. Muresan. 2001. Evaluation of the DEFINDER System for Fully Automatic Glossary Construction. *Proc. American*

*Medical Informatics Association
Symposium.*

- Lascarides, A. y Copestake A. 1995. The Pragmatics of Word Meaning, *Proc. AAAI Spring Symposium: Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity and Generativity*, Stanford CA.
- Meyer, I. 2001. Extracting knowledge-rich contexts for terminography. *Recent Advances in Computational Terminology*. John Benjamins.
- Pearson, J. 1998. *Terms in Context*. John Benjamins. Vol. 1, Amsterdam
- Powell et al. 2002. Tracking Meaning Over Time in the UMLS Metathesaurus, *Biomedical Informatics. Proc. Annual Symposium of the American Medical Informatics Association*; San Antonio, TX.
- Rebeyrolle, J. y Péry-Woodley, M-P 1998. Repérage d'objets textuels fonctionnels pour le filtrage d'information: le cas de la définition. *Rifra '98 Tunez*.
- Riloff, E. and Jones, R. 1999. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. *Proc. National Conference on Artificial Intelligence (AAAI-99)*.
- Rodríguez, C. 1999. Explicit Metalinguistic Operations in specialized discourse: The construction of lexical meaning in theoretic science. *Terminology and Knowledge Engineering TKE'99*, Innsbruck, Austria.
- Sierra G., y Alarcón R. 2002. Identification of recurrent patterns to extract definitory contexts. *CICLing 2002*. Ciudad de México, México.
- Vossen, P. and Copestake, A. 1993. Untangling Definition Structure into Knowledge Representation. *Inheritance, Defaults and the Lexicon*. Cambridge University Press.