

Extracción de contextos definitorios en el área de biomedicina

Extraction of Definitional Contexts from Biomedical Corpora

César Aguilar^a, Olga Acosta^b, Gerardo Sierra^c
Sergio Juárez^d, Tomás Infante^b

^aCatólica de Chile - Campus San Joaquín, Santiago de Chile caguilara@uc.cl

^bCognitiva –IBM - Apoquindo No. 5400, Las Condes, Santiago de Chile {oacosta, tinfante}@cognitiva.la

^cGrupo de Ingeniería Lingüística - Instituto de Ingeniería, UNAM - Torre de Ingeniería, CU, Ciudad de México
gsierram@ii.unam.mx

^dFacultad de Estadística e Informática, Universidad Veracruzana - Xalapa, Veracruz, México sejuarez@uv.mx

Resumen: En este proyecto se formula una metodología para extraer contextos definitorios desde corpus de biomedicina en español, con el fin de generar los siguientes productos: (i) un listado de candidatos a términos, (ii) un listado de candidatos a definiciones, y (iii) una taxonomía de términos biomédicos basada en relaciones de hiponimia/hiperonimia. Nuestro método permite crear un sistema capaz de extraer tales contextos, el cual puede verse como un módulo que cubriría las primeras etapas a seguir para construir una ontología basada en información textual.

Palabras clave: Contexto definitorio, término, definición, extracción de información, taxonomía.

Abstract: In this project we formulate a methodology for extracting definitional contexts from corpus of biomedicine in Spanish, in order to generate the following products: (i) a list of candidate terms, (ii) a list of candidates for definitions, and (iii) a taxonomy of biomedical terms relationships based on hyponym/hyperonym. Our methodology allows the creation of a system capable of extracting such contexts, which can be seen as a module that would cover the first steps to follow to build an ontology based on textual information.

Keywords: Definitional Context, Term, Definition, Information extraction, Taxonomy.

1 Introducción

Debido al incremento de información, la biomedicina tiene un gran interés en el desarrollo de herramientas que le ayuden a identificar, extraer y clasificar tal información, con miras a obtener conocimientos relevantes. Por ello, la Biblioteca Nacional de Medicina (NLM) de Estados Unidos desarrolló la base de datos *MedLine*, la cual cuenta con un total aproximado de 21 millones de referencias a artículos provenientes de 4.500 revistas. Para acceder a tal repositorio, se ha implementado el motor de consulta *PubMed*¹, el cual brinda acceso gratuito a sus datos. La implementación de esta clase de recursos ha hecho que la biomedicina mantenga una estrecha relación con el área de procesamiento del lenguaje natural (PLN).

A pesar de los avances logrados por parte

del PLN en la explotación de *MedLine*, (mayoritariamente en inglés, aunque también hay aportes en español), en Chile tales logros han tenido un bajo impacto hasta hoy.

De acuerdo con la *Estrategia Nacional de Salud 2010-2020*², elaborada por el Ministerio de Salud, existen varias limitaciones dentro de los sistemas de gestión de información médica, entre las cuales cabe destacar: (i) ausencia de normativa para la generación y recepción de información; (ii) escasa infraestructura tecnológica; (iii) ausencia de sistemas en línea, y (iv) errores en el traspaso manual de la información.

2 Objetivos

El objetivo de este proyecto es diseñar un sistema de extracción de contextos definitorios (CDs) en el área de biomedicina en español.

¹ Al respecto, véase el siguiente sitio WEB: www.ncbi.nlm.nih.gov/pubmed.

² Al respecto, véase el siguiente sitio WEB: <http://web.minsal.cl/portal/url/item/c4034eddbbc96ca6de0400101640159b8.pdf>.

Para diseñar este sistema, se toma en cuenta la metodología para extraer CDs desde corpus planteada por Sierra *et al.* (2008), y Sierra (2009), la cual considera los siguientes aspectos:

- El análisis lingüístico respecto a la estructuración de un CD.
- La implementación de un sistema de búsqueda, el cual genere como *output*: (a) un conjunto de candidatos a términos, (b) un conjunto de candidatos a CDs, clasificados según el tipo de definición asociada, y (c) una posible taxonomía de términos jerarquizados conforme a relaciones léxicas identificables entre ellos.
- El uso de métodos probabilísticos para evaluar la eficacia del sistema.

3 Descripción del proyecto

Definimos un CD como fragmentos textuales que contengan términos y definiciones ligadas por predicaciones verbales (Sierra *et al.*, 2008; Sierra, 2009). Un ejemplo es el siguiente:

Generalmente ^{marcador discursivo}, [la **célula** ^{Término/Marcador tipográfico}] [puede definirse como ^{Frase predicativa}] [una porción de protoplasma individualizado, dotado de núcleo y de una membrana plasmática, que nace, crece, se reproduce y muere ^{Definición}]

En este ejemplo, se pueden reconocer los componentes básicos de un CD: un término, una definición y una frase predicativa que opera como conector entre las unidades anteriores. A estas unidades se añaden marcadores discursivos (el adverbio *generalmente*) o marcadores tipográficos (signos ortográficos o tipos de fuente).

Una forma de representar lo anterior es el esquema de la figura 1, en donde se muestra la configuración de un CD en torno a sus componentes básicos (término, frase predicativa y definición), así como opcionales (marcadores discursivos y tipográficos).

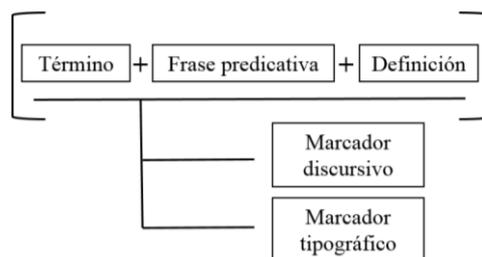


Figura 1: Estructura de un CD

Un aspecto relevante a observar aquí es que las definiciones asumen rasgos específicos para expresar el concepto asociado a un término, dependiendo del tipo de verbo que aparece en la frase predicativa. Estos rasgos tienen que ver con la formulación de los dos componentes básicos: género próximo y diferencia específica. Tomando en cuenta esto, Sierra *et al.* (2008) proponen 4 tipos de definiciones, derivables del modelo analítico:

- **Definición analítica o aristotélica:** se da cuando el género próximo y la diferencia específica aparece de manera explícita dentro de una definición.
- **Definición sinonímica:** se da cuando en una definición se hace explícito el género próximo, estableciendo una equivalencia conceptual con el término que es definido.
- **Definición funcional:** se da cuando se hace explícita la diferencia específica, ofreciendo una definición de un término a partir de su uso o aplicación en una situación dada.
- **Definición extensional:** se da cuando se hace explícita la diferencia específica, presentando una definición que enumera los componentes que conforman un objeto representado por el término a definir. Esta enumeración de componentes sigue un orden basado en relaciones que van de un todo hacia las partes, o de las partes hacia el todo.

Estos 4 tipos brindan la posibilidad de establecer una taxonomía que se puede reforzar a partir de la corroboración de relaciones léxicas existentes en CDs, a saber: hponimia/hiperonimia, meronimia y sinonimia. Conforme a la propuesta de Buitelaar, Cimiano y Magnani (2005), un CD contiene tanto unidades conceptuales (términos y definiciones)

como relaciones léxicas y semánticas, de acuerdo con este esquema:

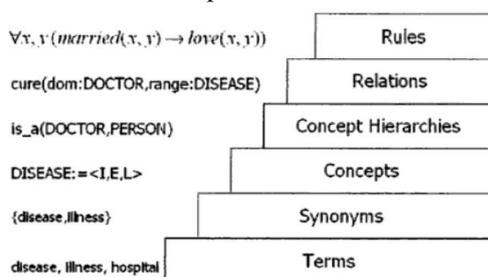


Figura 2: Esquema de desarrollo de una ontología según Buitelaar, Cimiano y Magnini

Este esquema describe un proceso escalonado para construir una ontología con información textual. En el primer escalón se concibe un proceso de extracción de términos (en nuestro caso, éstos se reconocen en CDs previamente delimitados). Después se hace una búsqueda de sinónimos ligados a los términos extraídos, ya sea a partir de la consulta de diccionarios, o ya sea empleando *WordNet* (Fellbaum, 1998).

En el siguiente escalón se pasa a formalizar la información conceptual asociada al término definido (en nuestro caso, tal información está contenida en CDs), conforme a una jerarquía léxica basada en relaciones de hiponimia/hiperonimia, meronimia y sinonimia. Luego, en un plano de representación conceptual, se pueden construir marcos semánticos como los que propone *FrameNet* (Baker, Fillmore y Cronin, 2003). Finalmente, el último escalón representa la formulación de axiomas, tomando en cuenta la información léxica y semántica obtenida.

4 Avances

Lo que mostramos aquí es el desarrollo de un método de extracción de términos, así como un método para detectar relaciones de hiponimia/hiperonimia y meronimia.

4.1 Implementación de un método de extracción de términos

Hemos implementado un método de extracción de términos, el cual aplica un contraste entre corpus, usando 4 medidas para asignar relevancia a palabras que ocurren tanto en el corpus de dominio como en un corpus de lengua general. Tales medidas son: razón *log-likelihood*; (ii) diferencia de rangos aplicada por Kit y Liu (2008); (iii) razón de frecuencia

relativa, y la aproximación a la distribución binomial mediante el uso de la distribución normal estándar, planteada por Drouin (2003).

Los resultados obtenidos muestran un desempeño mejor de las medidas diferencia de rangos y razón de frecuencias relativas. Igualmente, nuestro método es útil para asignar relevancia a palabras de un dominio, ya que el vocabulario estrechamente relacionado con un dominio tendrá mayor probabilidad de ocurrencia en éste, que en un corpus general. Para más detalles véase a Acosta, Aguilar e Infante (2015).

4.2 Identificación de relaciones de hiponimia/hiperonimia y meronimia

En el caso de relaciones de hiponimia/hiperonimia, se ha desarrollado un método híbrido que utiliza un *chunker* capaz de reconocer sintagmas nominales cuyo modificador es un adjetivo relacional (p.e.: *cáncer pancreático*, *infección sanguínea*, *médula ósea*, etc.). Para obtener esta clase de candidatos, se filtran todos aquellos sintagmas que no contengan esta clase de adjetivos, considerando una heurística reportada en Acosta, Sierra y Aguilar (2015).

Estos experimentos de extracción han sido útiles para desarrollar la arquitectura de un sistema prototipo, el cual se muestra en la figura 3. De acuerdo con tal esquema, el *input* es un corpus especializado, el cual se etiqueta con anotado morfosintáctico (o POST), para luego hacer una identificación de segmentos que contengan candidatos a CDs. Una vez localizados estos candidatos, se pasa a identificar aquellos sintagmas nominales que posean o bien un adjetivo relacional, o bien un sintagma prepositivo introducido por *de*.

Tras localizar estos candidatos, se contrastan sus frecuencias de aparición tanto en el corpus de dominio como en un corpus general, con miras a determinar su grado de unicidad (ing.: *unithood*), junto con su valor terminológico (ing.: *termhood*). Al final, el output esperado se compondrá de: (i) un conjunto de candidatos a CDs, un conjunto de candidatos a términos, y (iii), un conjunto de posibles hipónimos e hiperónimos (en el caso que se busque esta relación), o en su defecto candidatos a merónimos.

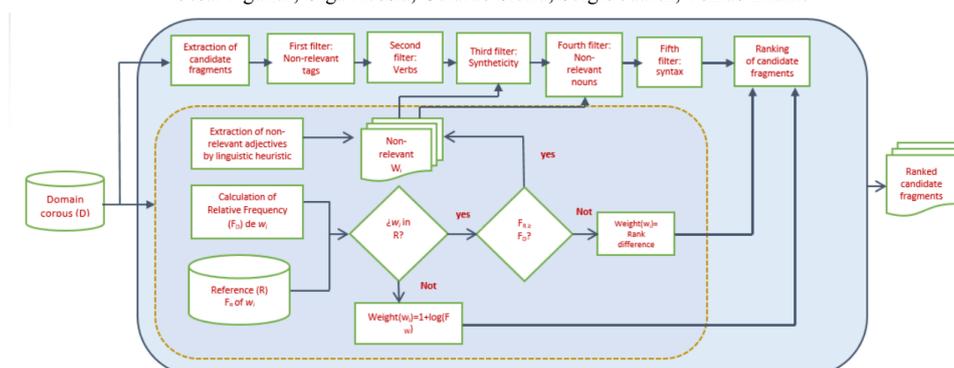


Figura 3: Arquitectura para un prototipo de extractor de CDs

y Tecnológica (CONICYT), del Gobierno de Chile. Número de proyecto: 11130565.

5 Trabajo a desarrollar

Los puntos por cubrir son:

- Concretar un sistema definitivo para la extracción de CDs, considerando una clasificación conforme a los tipos de definiciones descritos.
- Continuar con los procesos de evaluación, incluyendo una revisión manual para verificar la calidad de los resultados obtenidos.
- Generar una taxonomía jerarquizada, basada en las relaciones detectadas, la cual integraría un método para generar sinónimos a través de un proceso de permutación, p. e.: dar como sinónimo de *infección ocular* (nombre + adjetivo relacional) *infección del ojo* (nombre + sintagma prepositivo precedido por *de*).

6 Colaboraciones

Este proyecto cuenta con la colaboración de cuatro instancias:

- El Grupo de Procesamiento del Lenguaje Natural, de la Facultad de Letras de la Pontificia Universidad Católica de Chile.
- El equipo de lingüística computacional de *Cognitiva Latinoamérica* (<http://cognitiva.la>).
- El Grupo de Ingeniería Lingüística, del Instituto de Ingeniería de la UNAM, México.
- La Facultad de Estadística e Informática de la Universidad Veracruzana, México.

Agradecimientos

Este proyecto ha sido patrocinado por la Comisión Nacional de Investigación Científica

Bibliografía

- Acosta, O., Aguilar, C., e Infante, T. 2015. Reconocimiento de términos en español mediante la aplicación de un enfoque de comparación entre corpus, *Linguamática*, 7(2): 19-34.
- Acosta, O., Sierra, G., y Aguilar, C. 2015. Extracting definitional contexts in Spanish through the identification of hyponymy-hyperonymy relations. En Žižka, J., y Dařena, F. (eds.) *Modern Computational Models of Semantic Discovery in Natural Language*. IGI Global, Hershey, Penn, USA: 48-70.
- Buitelaar, P., Cimiano, P. y Magnini, B. 2005. *Ontology learning from text*. IOS Press, Amsterdam.
- Baker, C., Fillmore, Ch., y Cronin, B. 2003. The structure of the FrameNet database. *International Journal of Lexicography* 16(3): 281-296.
- Drouin, P. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1): 99-115.
- Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.
- Kit, Ch., y Liu, Y. 2008. Measuring mono-word termhood by rank difference via corpus comparison. *Terminology*, 14(2): 204-229.
- Sierra, G. 2009. Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos. *Linguamática*, 1(2): 13-37.
- Sierra, G., Alarcón, R., Aguilar, C., y Bach, C. 2008. Definitional verbal patterns for semantic relation extraction. *Terminology* 14(1):74-98.