Anotación con UMLS de história clínica electrónica en español*

Annotation of Spanish electronic health records with UMLS

Naiara Perez[1], **Montse Cuadros**[1] y **German Rigau**[2]
[1]HSLT Group, Vicomtech-IK4, Donostia-San Sebastián
[2]IXA NLP Group, UPV/EHU, Donostia-San Sebastián
{nperez,mcuadros}@vicomtech.org, german.rigau@ehu.es

Resumen: En este artículo presentamos una aproximación a la anotación de história clínica electrónica en español con conceptos recogidos en la base de terminologías de referencia UMLS. La herramienta se ha evaluado con un corpus paralelo de informes médicos creado para la tarea.
Palabras clave: PLN, anotación semántica, UMLS, HCE

Abstract: This paper presents an approach to annotate Spanish electronic health records with concepts included in UMLS. The tool has been evaluated against a parallel corpus of health records developed for the task.
Keywords: NLP, semantic tagging, UMLS, EHR

1 Introducción

Gracias a la implantación de la histora clínica electrónica (HCE) en los centros de la red sanitaria española, el interés por crear tecnología de procesamiento del lenguaje natural para el español aplicada al dominio médico ha crecido notablemente en los últimos años con la idea de desarrollar, por ejemplo, herramientas de soporte a la decisión, de detección automática de efectos adversos a medicamentos, codificación automática, etcétera.

Una tarea básica es la identificación de conceptos médicos relevantes en los textos clínicos. Existen desde hace tiempo herramientas con este fin que analizan textos en inglés, siendo una de ellas MetaMap (Aronson, 2001), la cual basa su funcionamiento en la base de terminologías clínicas de referencia Unified Medical Language System (UMLS) (Bodenreider, 2004). Para el español, Castro et al. (2010) presentan un planteamiento parecido al de Aronson (2001), y Oronoz et al. (2013) enriquecen Freeling (Padró y Stanilovsky, 2012) con listas terminológicas.

En este artículo presentamos un sistema que aborda la anotación de conceptos médicos, al igual que MetaMap, como un problema de búsqueda de correspondencias léxicas en UMLS y de desambiguación de acepciones. En este caso nos centramos en el subconjunto de fuentes en español del UMLS 2016AA (excepto LOINC®).

* Este trabajo ha sido financiado por Vicomtech-IK4 y el proyecto TUNER TIN2015-65308-C5-1-R (MINECO/FEDER, UE)

2 Descripción de la herramienta

El UMLS agrupa mediante relaciones N a N diferentes expresiones para un mismo concepto con un código identificador único. Tras normalizar y filtrar los términos del subconjunto mencionado, se ha creado un índice Apache Lucene™ que, dado un término de búsqueda, devuelve, entre otros, los códigos de los conceptos más similares. En total se han indexado 538.026 términos para 352.075 conceptos. Este índice es el componente principal del sistema. A continuación se resume el funcionamiento del mismo:

2.1 Búsqueda de correspondencias

Como primer paso del procesamiento, se detectan las abreviaturas en el texto de entrada y se sustituyen por las formas completas correspondientes. Para ello se emplea un detector de abreviaturas basado en reglas (Montoya, 2017) y un diccionario de abreviaturas validado por profesionales sanitarios.

Utilizando IXA pipes (Rodrigo, Bermudez, y Rigau, 2014) el texto expandido se segmenta en tokens y oraciones, y se etiqueta morfosintácticamente. Este etiquetado sirve de apoyo para la extracción de sintagmas nominales. La extracción se basa en heurísti-

cos y busca identificar en una oración todos los sintagmas nominales posibles, también anidados y/o solapados, como candidatos a ser anotados con un concepto UMLS; es decir, dado el texto "lesión en rodilla derecha", se extraen los sintagmas "lesión en rodilla derecha", "lesión en rodilla", "rodilla derecha", "lesión" y "rodilla". Para cada uno de los sintagmas extraídos se generan además variantes utilizando los lemas de los tokens y palabras relacionadas.

Todos los sintagmas y sus variantes se ordenan por subsunción de grande a pequeño para consultar el índice UMLS uno a uno. Cuando se encuentra un resultado válido para un sintagma, se descartan todos los sintagmas que éste subsume. Un resultado válido es el concepto con mayor peso entre los devueltos por el índice que supera cierto umbral, donde el peso se calcula con la métrica de Aronson (2001) adaptada para tener en cuenta la negación (es decir, para penalizar resultados con diferente polaridad al término de búsqueda).

2.2 Desambiguación

Cuando dos conceptos tienen el mismo peso para un mismo sintagma, el sistema recurre a UKB (Agirre y Soroa, 2009), una herramienta de desambiguación de sentidos basada en grafos de conocimiento, para elegir como se explica a continuación el concepto semánticamente más cercano al texto de entrada.

Además de las relaciones léxicas término-concepto descritas, UMLS contiene relaciones semánticas concepto-concepto. Así, es posible recrear un grafo de conocimiento en el cual los vértices son conceptos (los cuales son, a su vez, grupos de términos) y las aristas son relaciones entre conceptos. Con esta configuración, se ha creado un grafo a partir del mismo subconjunto de UMLS indexado para la búsqueda de correspondencias léxicas. La herramienta proyecta a partir de este grafo un Vector PageRank Personalizado (PPV) (Haveliwala, 2002) de cada texto a anotar, asignando mayor peso a los vértices representados por los lemas de los unigramas del texto. Como consecuencia, los conceptos más complejos presentes en el texto se activan también por la propagación de los pesos de acuerdo al principio de composicionalidad.

Entonces, la desambiguación de conceptos consiste en elegir aquél con mayor grado de activación en el PPV generado.

3 Evaluación

El sistema se ha evaluado indirectamente calculando el consenso de anotación entre éste y MetaMap, ya que no existe ningún corpus abierto de informes clínicos en castellano anotado a mano con conceptos de UMLS.

El corpus de evaluación consiste en 36 informes médicos paralelos, 18 en castellano (23.003 tokens) y 18 en inglés (21.093 tokens). Los documentos se han traducido manualmente: 10 del castellano al inglés, y 8 (disponibles on-line[1]) del inglés al castellano.

Los documentos en inglés se han anotado con MetaMap y los documentos en castellano se han anotado con nuestro sistema. Para que las anotaciones sean comparables, se ha reducido la base terminológica de MetaMap de manera que ambos sistemas puedan anotar únicamente los mismos 352.075 conceptos.

La micro-media del consenso, medido como kappa de Cohen, es de $0,36$ (no significativo estadísticamente). MetaMap ha identificado un total de 10.955 conceptos mientras que nuestro sistema ha identificado 8.110.

Si bien el resultado óptimo no es un consenso total, y a falta de un análisis de errores exhaustivo, creemos que las causas principales de la discordancia entre ambos sistemas son: *a)* las diferencias en la formulación de conceptos inherentes a la traducción, *b)* los errores en el análisis lingüístico que empeoran la sensibilidad del extractor de sintagmas, *c)* la generación de variantes con palabras relacionadas dando lugar a falsos positivos, y *d)* la menor disponibilidad de términos por concepto en el subconjunto de UMLS utilizado (1,55) respecto al inglés (2,30).

4 Conclusión y Trabajo Futuro

Hemos presentado un sistema de anotación de informes médicos en español con conceptos de UMLS que ha mostrado un consenso aceptable con MetaMap en una pequeña evaluación basada en un corpus paralelo de 18 informes médicos.

Como trabajo futuro se adaptarán las herramientas de tokenización y análisis morfosintáctico a textos clínicos. Además, se trabajará en la desambiguación de conceptos, cuya dificultad recae en la falta de recursos anotados para este dominio.

[1] `https://www.med.unc.edu/medclerk/education/eduactivities/h-p-examples`

Bibliografía

Agirre, E. y A. Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. *Proceedings of EACL 2009*, páginas 33–41.

Aronson, A. R. 2001. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. En *Proceedings of the AMIA Symposium*, páginas 17–21.

Bodenreider, O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(1):D267–D270.

Castro, E., A. Iglesias, P. Martínez, y L. Castaño. 2010. Automatic Identification of Biomedical Concepts in Spanish Language Unstructured Clinical Texts. En *Proceedings of IHI'10*, páginas 751–757.

Haveliwala, T. H. 2002. Topic-sensitive PageRank. En *Proceedings of WWW2002*, páginas 517–526. ACM.

Montoya, I. 2017. Etiquetado de Historiales Medicos Mediante SNOMED CT y CIE-10.

Oronoz, M., A. Casillas, K. Gojenola, y A. Pérez. 2013. Automatic Annotation of Medical Records in Spanish with Disease, Drug and Substance Names. En *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. páginas 536–543.

Padró, L. y E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. En *Proceedings of LREC2012*.

Rodrigo, A., J. Bermudez, y G. Rigau. 2014. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. En *Proceedings of LREC2014*, páginas 3823–3828.