Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 177-180 recibido 28-03-2016 revisado 11-04-2016 aceptado 20-04-2016

EasyLecto: Un sistema de simplificación léxica de efectos adversos presentes en prospectos de fármacos en español

EasyLecto: A lexical simplification system for adverse drug effects in Spanish patient information leaflets

Luis Núñez-Gómez, Isabel Segura-Bedmar, Paloma Martínez
Universidad Carlos III de Madrid
Av. Universidad, 30, 28913
lununezg@pa.uc3m.es,{*isegura*, *pmf*} @inf.uc3m.es

Resumen: Presentamos EasyLecto, un sistema de simplificación léxica de efectos adversos presentes en prospectos de fármacos en español. El método de simplificación léxica que utiliza EasyLecto se basa en la frecuencia de las palabras para determinar los sinónimos más simples. Este sistema propone los mejores sinónimos y definiciones, obtenidos a partir de los recursos MedlinePlus y MedDRA. El sistema puede ayudar a los lectores con bajo nivel de alfabetización, con problemas cognitivos o discapacidad a la hora de entender los efectos adversos presentes en prospectos de fármacos.
Palabras clave: Simplificación léxica, Procesamiento de Lenguaje Natural

Abstract: We introduce EasyLecto, a lexical simplification system of adverse effects in patient information leaflet in Spanish. The method of lexical simplification using EasyLecto is based on the frequency of words to determine the simplest synonyms. This system proposes the best synonyms and meanings, obtained from the MedlinePlus and MedDRA resources, representing their benefit for readers with low literacy, with cognitive problems or handicapped in understanding the adverse drug effects in patient information leaflet in Spanish.
Keywords: Lexical simplification, Natural Language Processing

1 Introdución

La simplificación automática de textos tiene como objetivo transformar un texto complejo en uno sencillo de leer y de entender. Es una tarea en Procesamiento de Lenguaje Natural (PLN) que en los últimos años ha crecido considerablemente (Abrahamsson et al., 2014) y que puede beneficiar a distintos grupos de usuarios como las personas mayores, personas con bajo nivel de alfabetización, personas con discapacidad lectora o incluso personas que están aprendiendo un idioma. Para simplificar un texto es aconsejable aplicar las pautas de lectura fácil[1], que pueden resumirse en: uso de lenguaje directo y sencillo, empleo de voz activa, expresar una idea por oración, evitar el uso de jerga y tecnicismos así como abreviaturas, estructurar el texto de forma clara y coherente, y usar palabras que representen un único concepto. En la simplificación automática de textos las oraciones complejas se dividen en oraciones más simples y el vocabulario complejo se reemplaza por un vocabulario más fácil de entender. La aplicación de técnicas de simplificación de textos puede ser beneficiosa para disminuir la complejidad de los textos en dominios tan diversos como la administración electrónica, la enseñanza o la salud, entre muchos otros. En este artículo nos centramos en el dominio de la salud y en particular en la simplificación de prospectos de medicamentos. Los prospectos de fármacos no son fáciles de comprender ya que contienen un gran número de términos técnicos del ámbito médico y porque sus oraciones suelen contener un gran número de estructuras gramaticales complejas (Davis et al., 2006). Por lo general, las personas desconocen la terminología médica y cuando tratan de entender los efectos adversos de un fármaco no logran hacerlo y terminan quitando importancia. Por esta razón es importante reducir la complejidad de los prospectos y de esta manera contribuir a evitar un uso inadecuado y peligroso de los medicamentos.

[1] www.lecturafacil.es/es

ISSN 1135-5948

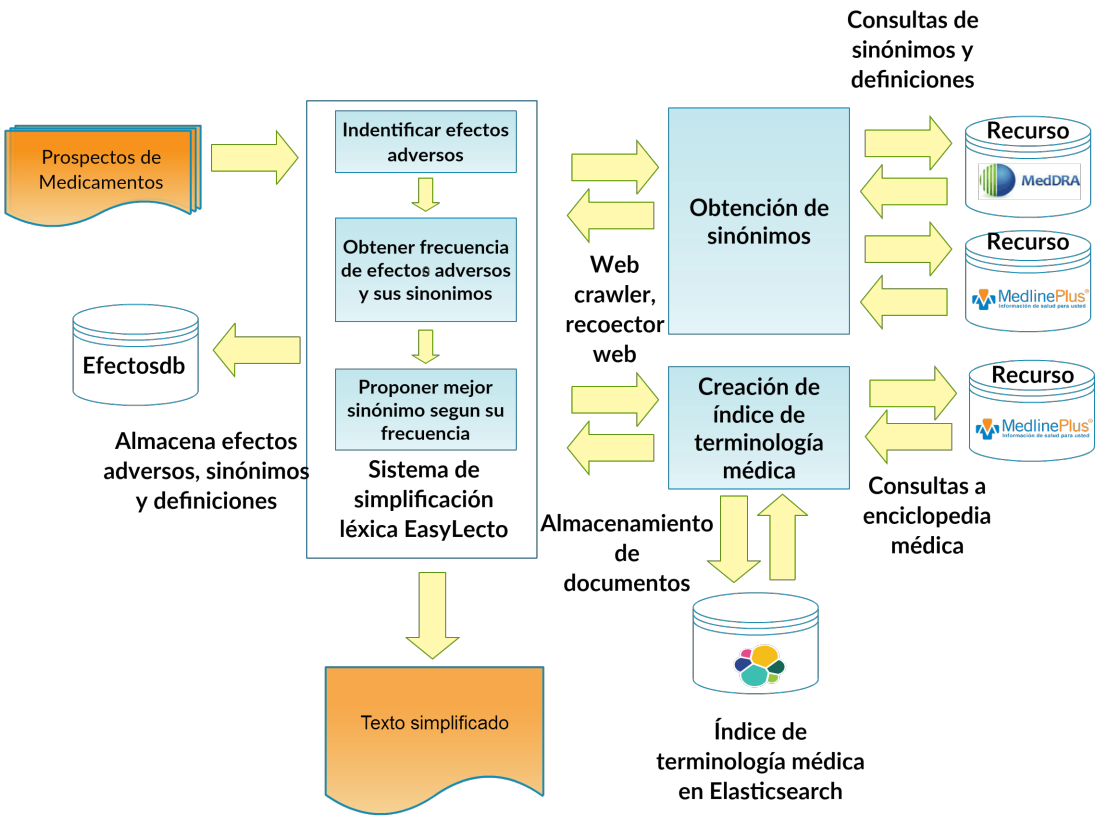

Figura 1: Esquema del proceso de simplificación léxica de EasyLecto.

Aunque existen herramientas de PLN basadas en técnicas de simplificación léxica que permiten reducir la complejidad de documentos, muy pocas están orientadas al español y al dominio de la salud (Bott, Saggion, y Mille, 2012; Grigonytea et al., 2014).

En las siguientes secciones se presenta la herramienta EasyLecto, un sistema de simplificación léxica de los efectos adversos presentes en los prospectos de fármacos en español. La herramienta combina recursos terminológicos para obtener los sinónimos de los efectos adversos y el cálculo de sus frecuencias en una colección de textos recopilada de la enciclopedia online MedLinePlus[2]. Nuestra principal hipótesis es que el sinónimo más frecuente debería ser el más simple. En el siguiente apartado, se describe con más detalle el método de simplificación propuesto.

2 *Proceso de simplificación Léxica de EasyLecto*

El sistema EasyLecto propone los mejores sinónimos para cada efecto adverso presente en un prospecto de medicamentos. El método de simplificación léxica que utiliza EasyLecto es la frecuencia de términos para determinar el sinónimo más simple; siendo el más frecuente. Figura 1 muestra la arquitectura del sistema EasyLecto.

El índice de terminología médica se crea utilizando la herramienta Elasticsearch [3], un gestor de base de datos distribuido y orientado a documentos, el cual permite almacenar gran cantidad de información, facilita las consultas mediante JSON [4] y está basado en tecnología Apache Lucene [5]. El proceso de creación del índice empieza con un algoritmo web-crawler que utiliza jsoup[6], una biblioteca en java que permite extraer y manipular datos de la web. El algoritmo recupera los artículos de la enciclopedia de MedlinePlus de forma alfabética; la enciclopedia incluye una gran cantidad de artículos acerca de temas de salud como enfermedades, exámenes médicos, síntomas, lesiones y procedimientos quirúrgicos. Una vez descargado, cada artículo de MedLinePlus se representa como un objeto

[2]https://www.nlm.nih.gov/medlineplus/spanish/

[3]https://www.elastic.co

[4]http://www.json.org

[5]https://lucene.apache.org/core/

[6]http://jsoup.org/

JSON, donde además de almacenar el contenido de su página web, también se guarda un conjunto de sinónimos del concepto descrito en el artículo (dichos sinónimos están identificados con los metadatos "Otros nombres" y "Nombres alternativos" en la página web del artículo) y la primera oración del texto del artículo, que será considerada como la definición del concepto descrito en el artículo. De esta forma, además de obtener un índice de los términos y sus frecuencias en MedLinePlus, también se ha construido de forma automática un posible diccionario de términos médicos y sus sinónimos. En total se almacenaron un total de 6900 documentos en el índice de terminología médica (ver Figura 1).

EasyLecto utiliza un sistema de reconocimiento de entidades, basado en diccionarios, para identificar las menciones de efectos adversos descritos en un prospecto. El lector puede encontrar información más detallada sobre dicho sistema en (Segura-Bedmar et al., 2015).

Para cada uno de los efectos adversos detectados, se obtiene un conjunto de sinónimos y definiciones a partir de los recursos MedDRA y MedlinePlus. En el caso de MedLinePlus, en concreto se utiliza la información almacenada en la base de datos de ElasticSearch, descrita en el apartado anterior. MedDRA [7], un tesauro multilingüe médico con información sobre productos médicos y con información sobre los efectos de los medicamentos. Desde el punto de vista de EasyLecto, la principal ventaja de MedDRA es que los efectos están agrupados en conjuntos de sinónimos.

La técnica de simplificación léxica que utiliza EasyLecto, como ya se ha mencionado, es la frecuencia de los sinónimos en el índice de terminología médica creado a partir de la colección MedLinePlus. Nuestra hipótesis es que el sinónimo más frecuente será el más sencillo para sustituir cada efecto adverso.

Después de recuperar los sinónimos de los recursos, en el subproceso "obtener frecuencia de efectos adversos y sus sinónimos", se consulta el índice de terminología médica; para recuperar la frecuencia de los efectos adversos y de sus sinónimos candidatos. El algoritmo realiza consultas mediante JSON a todos los documentos de nuestro índice. En el subproceso "proponer mejor sinónimo según su frecuencia", y una vez obtenidas las frecuencias de los efectos adversos y sus sinónimos candidatos, se propone el sinónimo más simple. En caso que el efecto adverso es más frecuente que todos los sinónimos candidatos, el efecto adverso no será reemplazado por otro sinónimo. Para mejorar los tiempos de ejecución, una vez procesado un documento, EasyLecto almacena sus efectos adversos y sus mejores sinónimos en la base de datos "Efectosdb", como se puede ver en la Figura 1. De esta forma, cada vez que se simplifique, EasyLecto en primer lugar buscará los efectos y sus mejores sinónimos en la base de datos, y sólo en el caso de no existir, buscará sus sinónimos y sus frecuencias en los recursos MedDRA y en el índice de terminología médica.

3 Funcionamiento del sistema EasyLecto

El sistema de simplificación léxica EasyLecto, permite identificar y simplificar los efectos adversos en un prospecto. A modo de ejemplo, al abrir el prospecto "Ceftriaxona", si el usuario necesita saber cuál es el mejor sinónimo para el efecto adverso "tromboflebitis", basta que el usuario haga clic sobre dicho efecto adverso y de manera inmediata se abrirá una ventana emergente. La Figura 2 muestra la salida que proporciona la herramienta EasyLecto para el término tromboflebitis. El mejor sinónimo que propone EasyLecto en el recurso MedlinePlus para el efecto adverso "tromboflebitis" es *"trombosis venosa profunda"*, sin embargo, MedDRA propone como mejor sinónimo el mismo término. La definición que ofrece MedlinePlus para este sinónimo es *"la trombosis venosa profunda o tvp, es un coágulo sanguíneo que se forma en una vena profunda en el cuerpo"*, como se puede ver en la Figura 2. Como MedDRA sólo propone sinónimos, nuestro sistema EasyLecto toma como posible definición el término MedDRA de mayor longitud.

4 Conclusiones y trabajo futuro

En este trabajo se presentó EasyLecto, el primer sistema de simplificación léxica de efectos adversos en prospectos de fármacos en español. Una demo del sistema está disponible en `http://163.117.129.251:8080/EasyLecto`. En la actualidad, se está creando un corpus gold-standard anotado con efectos y sus sinónimos, que nos permitirá dar

[7] http://www.meddra.org/

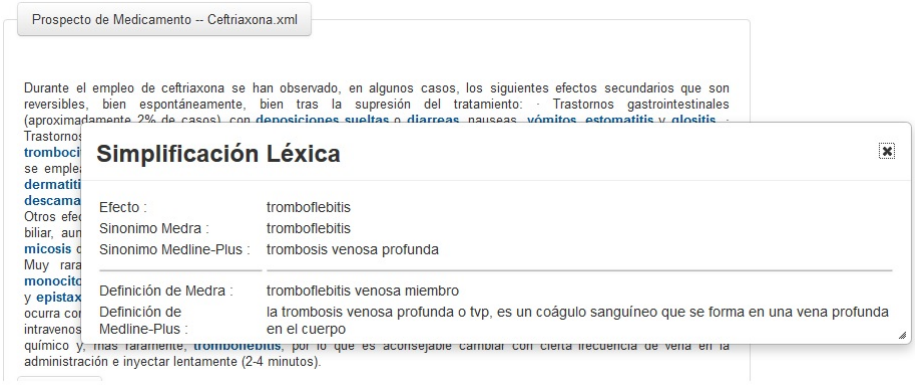

Figura 2: Simplificación léxica de los efectos adversos.

una evaluación cuantitativa de los resultados de la herramienta EasyLecto. En un futuro se pretende integrar otros recursos como BabelNet[8] y abordar la simplificación de otros conceptos médicos (pruebas médicas, tratamientos, enfermedades, etc). También se explorarán otras técnicas para la selección del mejor sinónimo, como por ejemplo el uso de modelos de vectores de palabras.

Agradecimientos

Este trabajo ha sido financiado por el proyecto eGovernAbility-Access (TIN2014-52665-C2-2-R).

Bibliografía

Abrahamsson, E., T. Forni, M. Skeppstedt, y M. Kvist. 2014. Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language. *Proceedings of The 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, páginas 57–65.

Bott, S., H. Saggion, y S. Mille. 2012. Text simplification tools for spanish. En N. C. C. Chair) K. Choukri T. Declerck M. U. Doğan B. Maegaard J. Mariani A. Moreno J. Odijk, y S. Piperidis, editores, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may.

Davis, T. C., M. S. Wolf, P. F. Bass, J. A. Thompson, H. H. Tilson, M. Neuberger, y R. M. Parker. 2006. Literacy and misunderstanding prescription drug labels. *Annals of Internal Medicine*, 145(12):887–894.

Grigonytea, G., M. Kvistbc, S. Velupillaib, y M. Wiréna. 2014. Improving readability of swedish electronic health records through lexical simplification: First results. *EACL 2014*, páginas 74–83.

Segura-Bedmar, I., P. Martínez, R. Revert, y J. Moreno-Schneider. 2015. Exploring spanish health social media for detecting drug effects. *BMC medical informatics and decision making*, 15 Suppl 2(Suppl 2):S6.

[8]http://babelnet.org/