Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 49-56 recibido 04-04-2018 revisado 01-05-2018 aceptado 04-05-2018

Construcción de recursos terminológicos médicos para el español: el sistema de extracción de términos CUTEXT y los repositorios de términos biomédicos

Construction of medical terminological resources for Spanish: the CUTEXT term extraction system and biomedical term repositories

Jesús Santamaría[1,2]
Martin Krallinger[1,2]
[1]CNIO (Centro Nacional de Investigaciones Oncológicas)
[2]BSC (Barcelona Supercomputing Center)
jsantamaria@cnio.es
krallinger.martin@gmail.com

Resumen: El uso frecuente de términos médicos motivó la construcción de grandes recursos terminológicos para el inglés, como el Unified Medical Language System (UMLS) o las ontologies Open Biological and Biomedical Ontology (OBO). La construcción exclusivamente manual de recursos terminológicos es en sí misma muy valiosa, pero constituye (1) un proceso laborioso que requiere mucho tiempo, (2) no garantiza que los conceptos o términos incluidos se 'alineen' realmente con el lenguaje médico y los términos que se usan en los documentos clínicos escritos por los profesionales de la salud y (3) requiere actualización constante y revisión debido a los cambios y la aparición de nuevos conceptos biomédicos. En este artículo presentamos una herramienta de extracción de términos médicos multilingüe, llamada CUTEXT (CValue Utilizado para Extraer Términos), un recurso promovido por el Plan de Impulso de las Tecnologías del Lenguaje (Villegas et al., 2017), disponible en: `https://github.com/Med-TL/Plan-TL/tree/master/CUTEXT`
Palabras clave: CValue, términos, extracción automática de términos

Abstract: The heavy use of medical terms motivated the construction of large terminological resources for English, such as the Unified Medical Language System (UMLS) or the Open Biological and Biomedical Ontology (OBO) ontologies. Purely manual construction of terminological resources is by itself very valuable, but constitutes (1) a highly time-consuming process, (2) it does not guarantee that included concepts or terms do actually align with the medical language and terms as they are being used in clinical documents by healthcare professionals and (3) requires constant update and revision due to changes and emergence of new biomedical concepts over time. In this paper we present a multilingual medical term extraction tool, called CUTEXT (Cvalue Used To Extract Terms), a resource promoted by the Spanish National Plan for the Advancement of Language Technology (Villegas et al., 2017), available at: `https://github.com/Med-TL/Plan-TL/tree/master/CUTEXT`
Keywords: CValue, terms, automatic terms extraction

1 Introducción

Una característica común de los textos biomédicos y clínicos es el uso excepcionalmente frecuente de términos técnicos específicos de dominio. Esta característica es compartida por la mayoría de los documentos médicos, independientemente de si están escritos en inglés, español u otros idiomas. Además, en el caso del dominio médico, existe una considerable brecha de recursos terminológicos para todos los demás idiomas en comparación con el inglés. La detección y el procesamiento eficientes de términos técnicos médicos es clave, no solo para enriquecer o incluso ayudar en la construcción automática de recursos terminológicos, sino que también constituye un elemento clave para otras aplicaciones de tecnología del lenguaje médico, incluyendo traducción automática, extracción de información, generación de resúme-

ISSN 1135-5948. DOI 10.26342/2018-61-5
© 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

nes y particularmente sistemas automáticos de codificación de documentos clínicos.

Desde la perspectiva de la extracción de términos, la tarea consiste en identificar dos rasgos principales (Kageura y Umino, 1996): (1) Unicidad (*unithood*): grado de cohesión o estabilidad de las palabras de una locución. (2) Termicidad (*termhood*): grado de especificidad del término con respecto a una disciplina concreta.

Por tanto, si deseamos comunicarnos sin confusión ni malentendidos, debemos ponernos de acuerdo sobre qué términos utilizaremos para representar los conceptos y cómo se deben traducir estos términos a diferentes idiomas (Foo, 2012).

La extracción manual de términos es una tarea larga, repetitiva, y tediosa, por ello, corre el riesgo de ser poco sistemática y subjetiva. Además, es muy costosa en términos económicos y está limitada por la información disponible. Por ello, la extracción automática de términos es una tarea crucial en el procesamiento del lenguaje natural, ya que permite que todo el proceso se lleve a cabo de forma mucho más ágil, eficiente, y económica, tanto en tiempo como en dinero.

La extracción automática de términos es una tarea relevante que puede ser útil en una amplia gama de tareas, como el aprendizaje ontológico, la traducción asistida y automática, la construcción de tesauros, la clasificación, la indexación, la recuperación de información, así como en la minería de textos y en la extracción de resúmenes.

Sin embargo, el proceso de extracción automática de términos no es una tarea trivial. El cambio constante en la terminología hace necesario que las herramientas sean capaces de detectar dichos términos nuevos, así como sus posibles variaciones. Las tareas de extracción suelen ser específicas, y por tanto, deben adaptarse a los requerimientos y particularidades propias de cada una de ellas. Según Ananiadou y Nenadić (2006) se pueden distinguir cinco variaciones terminológicas: ortográfica, morfológica, léxica, estructural, acrónimos y abreviaturas. A las que podemos añadir también las producidas por sinonimia y homonimia. Por último, existe una falta de convenciones firmes en la nomenclatura. Se han creado directrices pero no se imponen restricciones, así, junto con los términos denominados "bien formados" existen otros ad-hoc, que son problemáticos para los sistemas automáticos de identificación de términos.

Para hacer frente a estos problemas, los sistemas de extracción automática de términos se suelen clasificar en cuatro grandes grupos no excluyentes: (1) Sistemas basados en **características internas**: suelen atender a la ortografía, como mayúsculas, uso de dígitos, caracteres griegos, etc. (2) Sistemas que aprovechan **pistas morfológicas**: atienden sobre todo a afijos específicos y formantes cultos (principalmente griegos y latinos). (3) Sistemas que aprovechan la información procedente del **análisis sintáctico**: atienden a la estructura gramatical, para extraer términos procedentes de los sintagmas nominales, verbales, y preposicionales. (4) Sistemas basados en **medidas estadísticas** para promover candidatos a términos: basados en frecuencias, *log-likelihood*, *logDice*, *l-value*, *c-value*, etc. Como veremos más adelante, CUTEXT aprovecha tanto la información lingüística (3), como la estadística (4). CUTEXT permite el reconocimiento automático de términos técnicos, de una o varias palabras, a partir de documentos médicos, admitiendo varios formatos de entrada. Este sistema permite a los usuarios sin conocimientos técnicos identificar fácilmente términos médicos detectados en grandes corpus clínicos, y así facilitar la construcción de diccionarios clínicos, índices o glosarios médicos, con un énfasis particular en aplicaciones de reconocimiento de conceptos clínicos. Este artículo resume las principales características de CUTEXT, así como los resultados que se obtienen al aplicarlo a diferentes corpus, incluidos textos biomédicos en inglés, literatura médica española y textos clínicos en español.

En el apartado 2, se revisa el estado del arte, en el 3 se describen las características principales de CUTEXT, en el 4 se muestran los corpus, la colección que se ha utilizado durante el uso real con CUTEXT, y, por último, en el apartado 5 se resumen las conclusiones obtenidas así como las líneas de trabajo futuro.

2 Trabajo Relacionado

Las aproximaciones utilizadas en el estado del arte para la extracción automática de términos se pueden dividir, siguiendo a Krauthammer y Nenadic (2004), en cuatro tipos:

1. **Basadas en diccionario**: utilizan listas de palabras, de *stop-words* (sin con-

tenido semántico), ontologías, glosarios, y tesauros del dominio. Se utilizan como "filtro lingüístico" para eliminar palabras y reconocer términos. Tiene la ventaja de ser simple y eficiente, pero suele ser incompleta, además de no estar disponible en todos los dominios e idiomas.

2. **Basadas en reglas**: En este enfoque se emplean patrones y conocimiento gramatical para la detección de términos. Es uno de los métodos más empleados desde los años 90, sin embargo, posee el mismo inconveniente que el anterior, a saber, no está extendido igualmente en todos los dominios e idiomas.

3. **Basadas en estadística y aprendizaje automático**: Las técnicas estadísticas tratan de determinar lo característico de una palabra (o lema) en un corpus específico con respecto a su frecuencia en un corpus general. Es decir, intentan saber qué términos son sobreutilizados o infrautilizados en el corpus utilizado en comparación con su frecuencia en un corpus de referencia.

4. **Híbridas**: Combinan dos o más de las aproximaciones anteriores. CUTEXT, como veremos, es un enfoque híbrido que combina la aproximación lingüística con la estadística.

En los últimos años se han desarrollado varias herramientas de extracción automática de términos, sin embargo, la mayoría de ellas son dependientes del idioma: en portugués - ExATOlp (Lopes et al., 2009), en español-vasco - Elexbi (Gurrutxaga et al., 2006), en español-alemán - Autoterm (Haller, 2008), en árabe (Boulaknadel, Daille, y Aboutajdine, 2008), en esloveno e inglés - Luiz (Vintar, 2010), en inglés e italiano - KX (Pianta y Tonelli, 2010), o en inglés y alemán (Ramm et al., 2018).

Otros trabajos relacionados se pueden encontrar en (Koza Orellana, 2015), así como en (Barrón-Cedeno et al., 2009).

Algunas herramientas se han adaptado a dominios especificos como por ejemplo, TermExtractor (Sclano y Velardi, 2007), TerMine (Frantzi, Ananiadou, y Mima, 2000) o BioYaTeA (Golik et al., 2013). La herramienta TermSuite, se desarrolló durante el proyecto europeo *Terminology Extraction, Translation Tools and Comparable Corpora* (TTC). Este proyecto se centró en la adquisición automática o semiautomática de terminologías alineadas bilingues para la traducción asistida y automática.

Por último, citamos algunas aplicaciones, en español, en las que se ha utilizado la extracción automática de términos: En Castro et al. (2010) utilizan la extracción automática de términos para la detección de conceptos en notas clínicas y su posterior asociación, o mapeo, a la ontología *Systematized Nomenclature of Medicine — Clinical Terms* (SNOMED–CT). Vivaldi y Rodríguez (2010) utilizan un sistema de extracción de términos en un corpus biomédico, de tal forma que, una vez encontrado un candidato a término, se intenta encontrar una página en la Wikipedia que se corresponda con dicho candidato, después se encuentran todas las categorías de la Wikipedia asociadas a dicha página, y por último, se explora la Wikipedia siguiendo recursivamente todos los enlaces de categorías encontrados, para expandir y enriquecer la frontera del dominio. Por último, Moreno-Sandoval y Campillos-Llanos (2013) utilizan un sistema de extracción de términos para elaborar un corpus compuesto por textos biomédicos en español, árabe, y japonés.

3 Descripción de CUTEXT

Las características principales de CUTEXT son las siguientes:

- Está implementado en java, por lo que es multiplataforma. Se ha probado bajo Windows y Linux.
- Es multilingüe: Se ha probado en inglés, castellano, catalán, y gallego. Se puede adaptar fácilmente a otros idiomas con tan solo cambiar el fichero de texto de configuración de etiquetas léxicas.
- Puede utilizar como etiquetador a TreeTagger[1] o a GeniaTagger[2].
- Los documentos que admite pueden ser en texto plano, o en pdf.
- Permite tanto interfaz gráfica como por consola (modo texto).

[1] `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`. Permite etiquetar textos en más de 23 idiomas.

[2] `http://www.nactem.ac.uk/GENIA/tagger/`. A diferencia de Treetagger, sólo etiqueta textos en inglés, pero a cambio tiene la ventaja de haber sido adaptado específicamente para textos biomédicos

- Permite seleccionar el idioma, el etiquetador, los umbrales de frecuencia y de c-value, y la entrada del documento o documentos.
- La salida se proporciona en texto plano, en formato JSON, y en BioC[3] (Comeau et al., 2013). Se incluyen los tiempos parciales y el tiempo total.

CUTEXT tomó como punto de partida un extractor de alto impacto denominado TerMine (Frantzi, Ananiadou, y Mima, 2000), que permite extraer términos en textos escritos en inglés. Como se ha mencionado anteriormente, y siguiendo a TerMine, contiene dos filtros: uno lingüístico y otro estadístico. El filtro lingüístico se compone de expresiones regulares, y una lista de *stop–words*, y es dependiente del idioma. Por tanto, para añadir un nuevo idioma a CUTEXT, también habría que proporcionarle las expresiones regulares, (el filtro lingüístico), para dicho idioma. Para el inglés el filtro está dividido en tres: El primero tiene en cuenta sólo *nombres*, el segundo *nombres* y *adjetivos*, y el tercero *nombres*, *adjetivos* y *preposiciones*. Por otro lado, para castellano, catalán, y gallego es un poco más sofisticado. Está dividido en dos subfiltros, denominados *cerrado* y *abierto*: El filtro cerrado detecta tres patrones: (1) *nombre* seguido de *adjetivo*, (2) *nombre* seguido de *de* seguido de *adjetivo*, y (3) *nombre*. El filtro abierto está basado en cinco expresiones regulares que contienen *nombres*, *palabras extranjeras*, preposición *de*, *acrónimos*, y *adjetivos*. Al ser dependiente del idioma y del ámbito, se puede cambiar. Hemos visto, que los *verbos* no están incluidos. Sin embargo, para ciertas tareas sería deseable incluirlos. Por ejemplo, en un corpus en el que es más frecuente la expresión *"se descarta cáncer"* que *"cáncer descartado"*, los filtros actuales dejarían pasar el segundo pero no el primero, y sería deseable que ambos pasasen el filtro, por lo que se podría incluir una nueva expresión regular que tuviese en cuenta este *tipo* de verbos. Por otra parte, el filtro estadístico es independiente del idioma, y combina las siguientes tres métricas:
1. La frecuencia y longitud del término candidato.
2. La frecuencia del término candidato como parte de otros términos candidatos más largos.
3. El número de estos términos candidatos más largos.
En concreto, el $C-value$ que se asigna a un término candidato a, viene dado por la siguiente expresión:

$$C{-}value(a) = log_2|a|(f(a) - \frac{1}{\#(T_a)} \sum_{b \in T_a} f(b))$$

Donde,
a es el término candidato
$f(a)$ es la frecuencia del término candidato
T_a es el conjunto de términos candidatos que contienen a a (*candidatos mayores*)
$\#(T_a)$ es el número de términos de los *candidatos mayores*
$\sum_{b \in T_a} f(b)$ es la frecuencia total en la que a aparece en el conjunto de los *candidatos mayores*

El diagrama de CUTEXT se muestra en la figura 1.

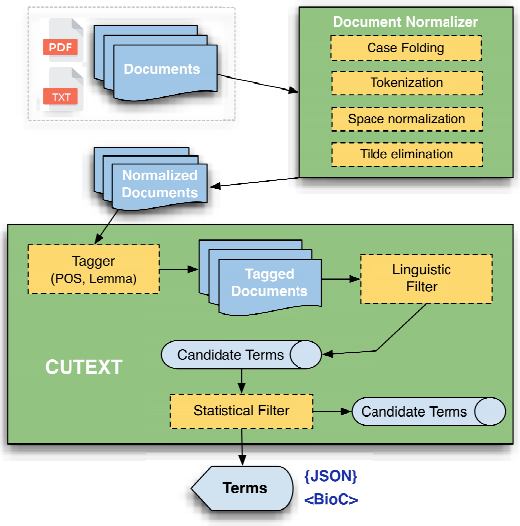

Figura 1: Diagrama de CUTEXT

El proceso denominado *Normalizer* se encarga de normalizar el texto procedente de los diversos documentos, y está separado de CUTEXT, es decir, no está incluido como parte íntegra en él. Básicamente lo que hace este proceso es: (1) convierte todo el documento (o documentos) a minúsculas, (2) separa el texto en tokens[4], (3) asigna un espacio entre

[3] Es un formato interoperable para textos biomédicos, basado en XML. Utilizado principalmente para el intercambio y el almacenamiento de datos de forma sencilla.

[4] Cada signo de puntuación lo considera un token, que se separa del resto. Así si, como es habitual, hay una palabra seguida de una coma – "por tanto," – la palabra y la coma se separan en dos tokens – "por

los tokens, y, por último, (4) elimina las tildes (si existen). Los documentos, ahora normalizados, se entregan a CUTEXT. El etiquetador (*Tagger*), se encarga de asignar a cada palabra su etiqueta léxica y su lema. Esta información es utilizada por el filtro lingüístico (*Linguistic Filter*) para obtener los términos candidatos, a partir de las reglas asignadas a priori. Los términos candidatos, sin ningún tipo de estadística salvo su frecuencia, se almacenan (*Terms*), y se entregan al filtro estadístico (*Statistical Filter*) que asigna el denominado C-Value. Estos términos, son almacenados de nuevo internamente (*Terms*), y también se muestran en los diferentes formatos (*Terms*) al usuario final.

3.1 Modo Texto – Opciones

En modo texto (por línea de comandos), CUTEXT ofrece más opciones que en modo gráfico. En concreto, en la tabla 1, se muestran las diferentes opciones que permite, así como el tipo para cada opción, y su valor por defecto.

Nombre	Tipo	Por defecto
-help	no aplica	no aplica
-displayon	boolean	true
-postagger	string	TreeTagger
-language	string	Spanish
-frecT	integer	0
-cvalueT	double	0.0
-bioc	boolean	false
-convert	boolean	true
-withoutcvalue	boolean	false
-incremental	boolean	false

Tabla 1: Opciones de CUTEXT en modo texto.

Todas las opciones son autoexplicativas, salvo, quizá, *-displayon*, *-convert*, *-withoutcvalue*, e *-incremental*.

-displayon: Se utiliza para mostrar por la salida estándar el proceso de ejecución, así como el tiempo que tarda en cada una de las fases. Por defecto, es *true*, es decir que sí se mostrará. Si se pone a false no presentará nada por la salida estándar.

-convert: Convierte a minúsculas el texto de entrada. Conviene tener en cuenta que en general, TreeTagger etiqueta las palabras escritas en mayúsculas, como nombres propios, por lo que si se selecciona este etiquetador (con la opción *-postagger*), conviene también poner este parámetro a true (que es su valor por defecto, al igual que el valor por defecto de *-postagger* es TreeTagger).

-withoutcvalue: Si se pone a *true* entonces CUTEXT ejecutará solamente el filtro lingüístico. Esto es útil en las aplicaciones donde es importante la rapidez, y no tanto el valor de c-value, (sólo con la frecuencia nos es suficiente), ya que CUTEXT, lógicamente, tardará menos en obtener los términos.

-incremental: Si se pone a *true* entonces CUTEXT ejecuta cada línea del fichero de entrada por separado. Es decir, para cada línea del fichero de entrada se ejecutan todas y cada una de las fases, tratándose, por tanto, cada línea como si fuese el corpus completo.

La versión más rápida de CUTEXT se da cuando los parámetros *-incremental* y *-withoutcvalue* están ambos a *true*.

4 Corpus Generados

CUTEXT se ha probado en diversos corpus, tanto *de juguete* como reales. Ha sido, como es lógico, con los corpus reales donde nos hemos dado cuenta de que para hacerlo práctico teníamos que hacerlo más eficiente, acelerando su ejecución. A partir de la salida generada por CUTEXT, se han obtenido diferentes corpus terminológicos. En las siguientes subsecciones explicaremos los corpus utilizados así como los obtenidos por CUTEXT. En concreto: (1) corpus Genia, (2) corpus de la tarea Biomedical Abbreviation Recognition and Resolution (BARR), (3) corpus de la Base de datos para la Investigación Farmacoepidemiológica en Atención Primaria (BIFAP).

4.1 Corpus Genia

El corpus Genia[5], está formado por resúmenes en inglés anotados, tomados de la base de datos MEDLINE de la Biblioteca Nacional de Medicina. Están anotadas un subconjunto de las sustancias y de las estructuras subcelulares de proteínas, basadas en un modelo de datos (denominada ontología GENIA) del dominio biológico, en formato XML (GPML). La version 3.0x consta de 2000 resúmenes. Los resúmenes básicos se seleccionan de los

tanto , " –. Esto evita que el etiquetador considere como token palabras seguidas por comas, como "tanto,".

[5] `http://www.geniaproject.org/`

resultados de búsqueda con las palabras clave (términos MeSH) *Human, Blood Cells and Transcription Factors*. El corpus está anotado con seis niveles de información lingüística y semántica: (1) Anotación léxica, (2) Constituyentes (anotación sintáctica), (3) Anotación de conceptos Genia, (4) Anotación de eventos, (5) Anotación de relaciones, (6) Anotación de correferencia.

En concreto, la anotación de términos cubre la identificación de entidades biológicas físicas, así como otros términos importantes. Para este corpus, CUTEXT generó un recurso terminológico de 38.903 términos. Los términos con los valores más altos de c-value se corresponden con términos biológicos, por lo que el recurso obtenido por CUTEXT puede valer en aquellas tareas en las que se necesiten dichos términos, como validación, mapeo, etc. Como se ha explicado anteriormente, el corpus Genia contiene anotación de conceptos Genia, por lo que se utilizó esta para su análisis comparativo. La medida F_1, obtenida por CUTEXT utilizando a TreeTagger como etiquetador léxico fue de un $F_1 = 52,6\,\%$, mientras que al utilizar como etiquetador a GeniaTagger obtuvo una medida ligeramente superior $F_1 = 53,8\,\%$. Esto es lógico, ya que GeniaTagger es un etiquetador de dominio biomédico. Esta medida es bastante superior a la obtenida por la versión web de TerMine[6], que fue de tan sólo de $F_1 = 39,7\,\%$.

4.2 Corpus BARR

BARR (Villegas et al., 2018) fue una tarea que se propuso dentro de IBEREVAL 2017[7]. Se trataba de reconocer abreviaturas y su expansión, así como relacionar ambas. Fue particularmente interesante, ya que algunos resúmenes se transcribieron manualmente, algo que se asemeja a las características de preprocesamiento que se encuentran en documentos clínicos. Para llevarla a cabo, se liberaron el corpus BARR anotado manualmente (Gold Standard), y una colección de documentos de resúmenes de artículos médicos escritos en español (la mayor colección unificada existente de la que tenemos noticia), distribuida a través de un acuerdo especial con el editor Elsevier. En concreto, el número total de tokens en los resúmenes (*abstracts*) es de 36.981.968, y el número total de tokens en los títulos de los *abstracts* es de 2.359.516.

Para este corpus, CUTEXT obtuvo 975.963 términos, cubriendo la mayoría de las abreviaturas y de sus formas largas[8]. Por tanto, el corpus terminológico obtenido, que se corresponde con términos biomédicos que incluyen abreviaturas, puede ser útil en tareas no sólo biomédicas sino propias de reconocimiento y extracción de abreviaturas.

4.3 Corpus BIFAP

La base de datos de BIFAP[9] está informatizada con registros médicos de Atención Primaria (AP) para la realización de estudios farmacoepidemiológicos, perteneciente a la Agencia Española de Medicamentos y Productos Sanitarios[10] (AEMPS), y cuenta con la colaboración de Comunidades Autónomas y el apoyo de las principales sociedades científicas implicadas. BIFAP incluye la información registrada por 5.752 médicos de familia y pediatras de AP del Sistema Nacional de Salud, integrando información de 7.890.485 historias clínicas anonimizadas. El 27 de marzo de 2015, la AEMPS pone a disposición de investigadores del ámbito público la base de datos BIFAP[11], para la investigación con medicamentos. Es en este ámbito en el que se ha elaborado un acuerdo de colaboración entre BIFAP y el Centro Nacional de Investigaciones Oncológicas (CNIO) dentro del marco del Plan de Impulso de las Tecnologías del Lenguaje (PlanTL).

Para este corpus, CUTEXT ha obtenido 1.440.306 términos. El corpus generado, no sólo es válido para la tarea en la que hemos colaborado[12], sino también en aquellas otras relacionadas con casos clínicos, ya que los términos extraídos pertenecen a este ámbito. Los términos extraídos, no se han podido evaluar, ya que BIFAP no tiene ningún gold-standard a este respecto. Sin embargo, sí los hemos examinado manualmente, encon-

[6] `http://www.nactem.ac.uk/software/termine/`

[7] `http://cabrillo.lsi.uned.es/nlp/IberEval-2017/index.php`

[8] No incluimos aquí la medida F, ya que no tiene sentido, porque CUTEXT extrae todo tipo de términos, no solamente abreviaturas y formas largas, y sin embargo el gold-standard está compuesto sólo por aquéllas.

[9] `http://www.bifap.org/`

[10] `https://www.aemps.gob.es/`

[11] `https://www.aemps.gob.es/informa/notasInformativas/laAEMPS/2015/docs/NI-AEMPS_03-2015-jornada-BIFAP-marzo-2015.pdf`

[12] Básicamente, consiste en una tarea de mapeo de un literal nuevo, introducido por un médico, frente a los literales almacenados en su base de datos.

trando que los términos con un alto valor de c-value (en general por encima de 50.0) son términos específicos del dominio (como por ejemplo, *adenocarcinoma de pulmón*, con un c-value de 405.839), mientras que aquellos con un c-value bajo, suelen ser más genéricos o suelen estar mal escritos (como por ejemplo, *ansiedad gerneralizada*).

5 Conclusiones y Trabajo Futuro

En este artículo se ha mostrado que la extracción automática de términos es una tarea crucial en el ámbito del procesamiento del lenguaje natural. Se ha puesto de relieve que, actualmente, los principales extractores automáticos de términos son dependientes del idioma y de la plataforma. Por todo ello, hemos presentado una herramienta multilingüe y multiplataforma, denominada CUTEXT, que permite extraer automáticamente los términos de un corpus, y asignarle un valor (denominado c-value) a cada uno de ellos, que determina su fiabilidad.

Hemos mostrado tres corpus de distinto ámbito generados por CUTEXT, que se pueden utilizar en tareas muy diversas: (a) Corpus compuesto por términos biológicos, (b) corpus compuesto por términos biomédicos que incluyen abreviaturas y sus formas largas, (c) corpus terminológico de casos clínicos.

Resumimos, a continuación, los puntos principales, y más relevantes de CUTEXT, no vistos en otros extractores automáticos:

1. Es multiplataforma, se ha testeado bajo diferentes sistemas operativos (Windows y Linux), y dispone de una interfaz gráfica y textual altamente configurable.

2. Es un sistema abiertocapaz de generar recursos para medicina, a partir de grandes corpus heterogéneos.

3. Es multilingüe: procesa textos biomédicos en castellano, inglés, catalán, y gallego, pero se pueden añadir idiomas de una forma sencilla.

4. Se ha comprobado su utilidad en aplicaciones reales, como la desarrollada para BIFAP.

5. Admite textos en diferentes formatos, y es capaz de generar una salida en tres tipos de formatos diferentes: plano, JSON, y BioC.

Debido a que la principal utilidad de un extractor de términos es el mapeo[13], como primera línea de trabajo futuro seguiremos a Krauthammer y Nenadic (2004), que determinan 3 etapas secuenciales para su realización:
(1) Reconocimiento del término: Permite diferenciar entre términos y no términos. Esta es la salida que proporciona CUTEXT.
(2) Clasificación del término: Consiste en asignar los términos al dominio específico. Es decir, quedarse sólo con los términos pertenecientes al dominio. El objetivo consiste en medir el grado de distintividad de un término en un corpus especializado en contraste con su frecuencia en un corpus general. Las métricas más empleadas son *log-likelihood ratio test*, y *logDice*.
(3) Emparejamiento del término: Vincula los términos con conceptos bien definidos de fuentes de datos referentes, como vocabularios controlados o bases de datos.

También tenemos pensado utilizar CUTEXT para procesar nuevos corpus de textos médicos bilingües (por ejemplo, el recurso denominado MeSpEN (Villegas et al., 2018)).

Agradecimientos

El presente trabajo fue realizado bajo la financiación de la Encomienda MINETAD-CNIO/OTG Sanidad Plan TL y el proyecto H2020 OpenMinted (654021).

Bibliografía

Ananiadou, S. y G. Nenadić. 2006. Automatic terminology management in biomedicine. En S. Ananiadou y J. McNaught, editores, *Text Mining for Biology and Biomedicine*. Artech House, Inc., páginas 67–98.

Barrón-Cedeno, A., G. Sierra, P. Drouin, y S. Ananiadou. 2009. An improved automatic term recognition method for spanish. En *International Conference on Intelligent Text Processing and Computational Linguistics*, páginas 125–136. Springer.

Boulaknadel, S., B. Daille, y D. Aboutajdine. 2008. A multi-word term extraction program for arabic language. 01.

[13]Denominado en inglés *mapping*. Consiste en vincular los términos con conceptos bien definidos de fuentes de datos referentes, como vocabularios controlados o bases de datos.

Castro, E., A. Iglesias, P. Martinez, y L. Castaño. 2010. Automatic identification of biomedical concepts in spanish-language unstructured clinical texts. páginas 751–757, 01.

Comeau, D. C., R. Islamaj Doğan, P. Ciccarese, K. B. Cohen, M. Krallinger, F. Leitner, Z. Lu, Y. Peng, F. Rinaldi, M. Torii, y others. 2013. Bioc: a minimalist approach to interoperability for biomedical text processing. *Database*, 2013.

Foo, J. 2012. Computational terminology: Exploring bilingual and monolingual term extraction. 04.

Frantzi, K., S. Ananiadou, y H. Mima. 2000. Automatic recognition of multi-word terms. *International Journal of Digital Libraries*, 3(2):117–132.

Golik, W., R. Bossy, Z. Ratkovic, y C. Nedellec. 2013. Improving term extraction with linguistic analysis in the biomedical domain. *Research in Computing Science*, 70:157–172.

Gurrutxaga, A., X. Saralegi, S. Ugartetxea, y I. Alegria. 2006. Elexbi, a basic tool for bilingual term extraction from spanish-basque parallel corpora.

Haller, J. 2008. Autoterm : Term candidate extraction for technical documentation (spanish/german).

Kageura, K. y B. Umino. 1996. Methods of automatic term recognition - a review. *Terminology. Amsterdam, 1996.*, 3(2):259–289.

Koza Orellana, W. 2015. Proposal for automatic extraction of medical term candidates with linguistic information processing description and evaluation of results. *Alfa: Revista de Linguística (São José do Rio Preto)*, 59(1):113–128.

Krauthammer, M. y G. Nenadic. 2004. Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37(6):512 – 526. Named Entity Recognition in Biomedicine.

Lopes, L., P. Fern, R. Vieira, y G. Fedrizzi. 2009. Exatolp – an automatic tool for term extraction from portuguese language corpora.

Moreno-Sandoval, A. y L. Campillos-Llanos. 2013. Design and annotation of multimedica – a multilingual text corpus of the biomedical domain. *Procedia - Social and Behavioral Sciences*, 95:33 – 39. Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013).

Pianta, E. y S. Tonelli. 2010. Kx: A flexible system for keyphrase extraction. páginas 170–173, 01.

Ramm, A., U. Heid, B. Weissbach, C. Loth, y I. Mingers. 2018. Adapting and evaluating a generic term extraction tool. 03.

Sclano, F. y P. Velardi. 2007. Termextractor: a web application to learn the shared terminology of emergent web communities. páginas 287–290, 01.

Villegas, M., S. de la Peña, A. Intxaurrondo, J. Santamaria, y M. Krallinger. 2017. Esfuerzos para fomentar la minería de textos en biomedicina más allá del inglés: el plan estratégico nacional español para las tecnologías del lenguaje. *Procesamiento del Lenguaje Natural*, 59:141–144.

Villegas, M., A. Intxaurrondo, A. Gonzalez, , M. Marimon, y M. Krallinger. 2018. The mespen resource for english-spanish medical machine translation and terminologies: Census of parallel corpora, glossaries and term translations. En *Proceedings of the LREC 2018 Workshop "MultilingualBIO: Multilingual Biomedical Text Processing*, páginas 32–39.

Vintar, p. 2010. Bilingual term recognition revisited the bag-of-equivalents term alignment approach and its evaluation. 16:141–158, 12.

Vivaldi, J. y H. Rodríguez. 2010. Using wikipedia for term extraction in the biomedical domain: first experiences. *Procesamiento del Lenguaje Natural*, 45(0):251–254.