Extracción de Candidatos a Términos del Dominio Médico a Partir de la categorización automática de Palabras

Koza W.
Grupo Infosur, Universidad Nacional de Rosario, Santa Fe, Argentina

Resumen
Se presenta una serie de experimentos con el objetivo de desarrollar un método automático de extracción terminológica del dominio de la medicina. Uno de los inconvenientes típicos es el cambio constante de la terminología médica, que imposibilita mantener terminologías actualizadas inmediatamente por medios manuales. Se parte de la hipótesis de que las palabras incluidas en los textos médicos que no aparecen en el diccionario fuente del software analizador, denominadas PD, son, en su mayoría, expresiones específicas del dominio médico, por lo cual, una categorización automática de estas ayudaría en la extracción. Primeramente, se intenta deducir a qué categoría pertenecen las PD mediante reglas de formación de palabras (1er. Nivel de análisis) y sintácticas (2do. Nivel de análisis). Luego, se procede a la conformación de sintagmas nominales que involucren PD, para extraerlos como candidatos a términos del dominio. Finalmente, se evalúa la precisión de las categorizaciones y, posteriormente, con el asesoramiento de profesionales del área de medicina, se verifica la posibilidad que tienen los candidatos a términos extraídos de ser promovidos a términos. En trabajo computacional, se utilizan las herramientas Smorph [1] y Módulo Post-Smorph (MPS) [2]. Smorph realiza el análisis morfológico y MPS trabaja sobre gramáticas locales.

Palabras Clave
Terminología Médica, extracción automática, candidatos a término, Smorph, MPS.

Introducción

Se presenta una serie de experimentos realizados con el objetivo de deducir la categoría gramatical de aquellas palabras que no se encuentran en el diccionario fuente de los softwares de análisis automático de textos. Este trabajo se enmarca en el ámbito la lingüística informática, por un lado, y, por otro, en las tareas de minería textual, y forma parte de una serie de tareas tendientes a desarrollar un método de extracción terminológica del dominio de la medicina. A tales efectos, se tomaron en consideración dos tipos de antecedentes: los estudios sobre terminología y extracción de términos, y los de utilización de formalismos y softwares declarativos, en los que la máquina algorítmica está disociada de los datos a utilizar.

Las tareas de extracción de términos poseen un lugar destacado en actividades de extracción y organización del conocimiento. Un término es una unidad léxica caracterizada por una referencia especial dentro de una disciplina [3], y puede estar conformado por una sola palabra (unigrama), por ejemplo ‘asma’, ‘hormona’; o una combinación de ellas (n-gramas), como ser ‘tuberculosis pulmonar’, ‘sistema cardiovascular’ (bigramas); ‘esquema de tratamiento’, ‘estado de enfermedad’ (trigramas), etcétera. Un conjunto de términos constituye la terminología.

Las tareas de extracción de términos suelen enfocarse en dominios específicos, y uno de ellos es el de la medicina. En este caso, la extracción de términos representativos suele destinarse a la elaboración de listas de entradas para diccionarios electrónicos específicos, la creación de base de datos o de ontologías y taxonomías que organizan y especifican el dominio de conocimiento, etcétera. Otra de las aplicaciones apunta a la clasificación textual, es decir que, a partir de la extracción realizada sobre varios textos o informes de

medicina, los especialistas (médicos, enfermeros, técnicos del área e incluso pacientes) puedan reconocer a qué subáreas pertenecen dichos textos.
Uno de los inconvenientes en la identificación automática de términos en este ámbito es el cambio constante de la terminología médica. Esto dificulta de sobremanera mantener terminologías actualizadas inmediatamente por medios manuales, por lo que se hace necesario contar con herramientas que puedan considerar en el análisis aquellos términos nuevos.
Parto de la hipótesis de que las palabras incluidas en los textos médicos que no aparecen en el diccionario fuente del software analizador (se trata de un diccionario estándar), y que son etiquetadas como palabras desconocidas (PD), son, en su mayoría, expresiones específicas del dominio médico, por lo cual, una categorización automática de estas ayudaría a la extracción terminológica. A partir de allí, se elaboran reglas para deducir las PD y las que son categorizadas como nombres, pasan a formar parte de la lista de términos del dominio médico establecida automáticamente. En este caso la lista se divide en unigramas (‘osteoporosis’), bigramas (‘insuficiencia ovárica’) y trigramas (‘síndrome de Cushing’).
El artículo se organiza de la siguiente manera. En primer lugar, se presentan los antecedentes en esta área. En segundo lugar, la metodología. Finalmente, en tercer lugar, la descripción de la tareas a llevar a cabo.

1. Acerca de la terminología médica y las tareas de extracción automática

Se denomina “términos” a los elementos léxicos utilizados en un ámbito temáticamente restringido para denominar un concepto [4], con lo cual, una correcta identificación de estos es clave para acceder al conocimiento que se transmite en los textos. Generalmente, son los sintagmas nominales los que se corresponden con los términos [5], a tales efectos, en el presente proyecto, la extracción estará focalizada en ellos.
En lo que atañe al área de la medicina, se observa que la terminología es extremadamente cambiante y compleja, por lo que la identificación de términos en este dominio se ha convertido en uno de los principales tópicos de investigación, tanto en el procesamiento del lenguaje natural, como así también en las comunidades biomédicas [6]. Términos tales como nombres de genes, proteínas, organismos, drogas, componentes químicos, etcétera, son los medios que se utilizan en medicina para identificar conceptos del dominio.
Existen diversos trabajos en la literatura que realizan la extracción de términos [7,8,9,10]. En relación con el dominio de medicina, de acuerdo con Castro [11], para el caso del inglés, hay varias investigaciones orientadas al procesamiento de textos y de datos de ese dominio [12,13], sin embargo, se encuentran pocas iniciativas para el español [14,15,16,17,18].
El principal inconveniente para una identificación automática exitosa es que las palabras y los términos comparten la misma estructura superficial [4], otros problemas se derivan de las variaciones léxicas, los casos de sinonimia (cuando un concepto está representado por varios términos) o de homonimia (cuando un término tiene varios significados). Por otro lado, cabe mencionar el cambio constante de la terminología médica; algunos términos aparecen en un período muy corto y, a la vez, se crean nuevos casi a diario. Esto hace imposible mantener terminologías actualizadas inmediatamente por medios manuales. Otro de los problemas, señalado por Krauthammer y Nenadić [6], remite a la falta de convenciones firmes en la nomenclatura. Si bien existen algunas directrices para ciertos tipos de entidades médicas, estas no imponen restricciones a los términos del dominio. Tuason et al. [19], por ejemplo, mencionan que las causas de las falencias en sus

experimentos de extracción se debieron principalmente a variaciones de "puntuación" ('bmp-4', 'bmp4'), uso de diferentes tipos de numerales ('syt4', 'syt iv') y diferentes transcripciones de las letras del alfabeto griego ('igα', 'ig alpha').
Ante la imposibilidad de mantener diccionarios totalmente completos, Aït-Mokthar y Rodrigo Mateos [20] señalan, en la descripción de la herramienta Smorph [1], que, en algunos casos, se puede describir la categoría de una palabra desconocida (PD) a partir de su terminación morfológica. Por ejemplo, toda cadena de caracteres terminadas en '-ción' es un nombre femenino singular, o toda cadena terminada en '-ó' es un verbo flexivo en pretérito perfecto simple.
A tales efectos, en primer lugar, se intentará deducir a qué categoría pertenecen las PD mediante reglas de formación de palabras (1er. Nivel de análisis) y reglas sintácticas (2do. Nivel de análisis). En segundo lugar, se procederá a la conformación de sintagmas nominales que involucren PD, para luego extraerlos. Finalmente, en tercer lugar, se evaluará la precisión de las categorizaciones y, posteriormente, con el asesoramiento de profesionales del área de medicina, se verificará la posibilidad que tienen los candidatos a términos extraídos de ser promovidos a términos.

2. Metodología

La presente investigación se basa, principalmente, en "deducir" la categoría de las palabras que no se encuentran en el diccionario fuente de los softwares de análisis lingüístico. Para ello, se tomarán en consideración los estudios de formación de palabras [21] y la relación entre morfología y terminología [22], como así también, los análisis de conformación de sintagmas.
Para el trabajo informático, se recurrirá a las herramientas Smorph [1] y Módulo Post Smorph (MPS) [2]. El primero permite analizar morfológicamente la cadena de caracteres, dando como salida la asignación categorial y morfológica correspondiente a cada ocurrencia de acuerdo con los rasgos declarados. MPS, por su parte, tiene como input la salida de Smorph y, a partir de reglas de recomposición, descomposición y correspondencia declaradas por el usuario, analiza la cadena de lemas resultante del análisis morfológico.
Las fuentes declarativas de Smorph están constituidas por 5 archivos: (i) ascii.txt: contiene los códigos ascii específicos tales como los separadores de oración y de párrafo; (ii) rasgos.txt: incluye etiquetas de rasgos morfológicos a aplicar en el análisis de las cadenas de caracteres con sus posibles valores (ej.: EMS 'nombre', 'verbo'; Género: 'masculino', 'femenino', etcétera); (iii) term.txt: carga las diferentes terminaciones que cada lema puede presentar en su derivación morfológica (ej.: -o, -a, -os, -as); (iv) entradas.txt: es el listado de lemas y modelos correspondientes de derivación (ej. casar v1), y (v) modelos.txt: define las clases de acuerdo con los parámetros de concatenación regular de cadenas a partir de las entradas y las terminaciones (ej.: modelo v1: raíz + terminaciones de la 1ª conjugación regular + rasgos).
Las fuentes declarativas de MPS, en cambio, están constituidas por un único tipo de archivo, rcm.txt, que incluye un listado de reglas que especifican cadenas posibles de lemas con una sintaxis informatizada. Las reglas pueden ser de tres tipos: (i) de reagrupamiento: D + N = SN; (ii) de descomposición: Contracc = P + D, y (iii) de correspondencia: Art = D.
Se utilizará el archivo entradas.txt elaborado por Infosur, y la correspondiente modelización desarrollada por el mismo equipo para las formas flexivas.

El proceso de reconocimiento de PD y posterior extracción de candidatos a términos se compone de las siguientes etapas:

- **Etapa I:** Análisis morfológico y reconocimiento de los signos de puntuación por medio de Smorph. Aquí se les asignará a las palabras desconocidas la etiqueta 'PD';
- **Etapa II:** Modificación del archivo term.txt mediante la asignación de terminaciones distinguidas con su correspondiente clasificación morfológica (ej.: ción pd/nom/fem/sg', '-ciones pd/nom/fem/pl'). Teniendo en cuenta el género textual del corpus, solo se incluirán las terminaciones distinguidas de los verbos de la tercera persona singular y plural, formados a partir de los sufijos '-ar', '-ear', '-ecer', '-ificar', '-izar', del presente y el pretérito del indicativo y el subjuntivo, el participio y el gerundio. No se cargará la terminación del infinitivo para evitar etiquetados erróneos (ej.: 'vascular' puede ser etiquetado como verbo). Posteriormente, se volverá a pasar el corpus por Smorph a fin de obtener las categorías que se ajusten a dichas terminaciones. También en esta etapa se considerará la posibilidad de que la PD sea un nombre propio o una sigla a partir de si presenta o no caracteres en mayúscula;
- **Etapa III:** Creación y aplicación de reglas sintácticas que permitan deducir la categoría de las PD. Aquí se hará hincapié en la estructura del sintagma nominal (SN) (Ej.: Det + PD + Adj = SN/ART+NOM+ADJ);
- **Etapa IV:** Extracción de los SN que involucran PD, en calidad de candidatos a términos. Aquí los términos serán simplificados con la técnica de stemming [23], que consiste en reducir las palabras a sus formas no flexivas y no derivativas;
- **Etapa V:** Evaluación de las categorizaciones y de los candidatos a términos extraídos.

He aquí un esquema de las etapas de trabajo.

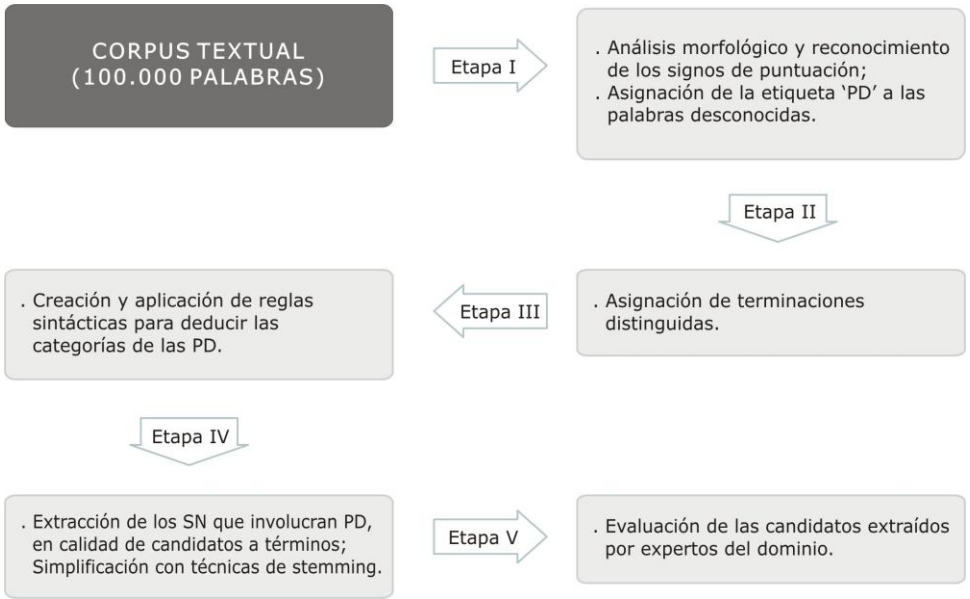

Figura 1

A continuación se ejemplificará con un breve texto del dominio médico.

3. Experimentación

En las etapas preliminares, se tomó un texto base para desarrollar las reglas iniciales. Aquí se presenta un fragmento de este.

> Enfermedades Vasculares
> En esta categoría se encuentran patologías como el accidente cerebrovascular hemorrágico, el accidente cerebrovascular isquémico, la hemorragia subaracnoidea, las malformaciones arteriovenosas, etc., también pueden incluirse, dentro de este grupo, los casos de aneurisma. Todos ellos tienen en común que son eventos de gran severidad y que las consecuencias neurológicas del retraso en el diagnóstico y el tratamiento pueden ser hasta fatales. Dentro de los métodos de diagnóstico que se utilizan en estas patologías están: el TAC, la resonancia magnética, la angiografía cerebral y otros. El tratamiento de estos problemas depende de cada caso, aunque últimamente la mayoría de estos puede ser diagnosticado y tratado a través de la terapia endovascular. La figura de la izquierda muestra una hemorragia intraparenquimatosa.
> (Mora, "Enfermedades del cerebro", texto adaptado)

En la primera pasada, Smorph señaló como palabras desconocidas las siguientes cadenas: 'isquémico', 'subaracnoidea', 'ateriovenosas', 'aneurisma', 'TAC', 'angiografía', 'endovascular' e 'intraparenquimatosa'. Luego, se adicionó en el archivo term.txt, las terminaciones distinguidas. A continuación, ejemplo de algunas de ellas:

```
grafía        pd/nom/fem/sg .
oso           pd/adj/masc/sg .
mente         pd/adv .
izó           pd/v/pret/ind/3p/sg .
```

Una vez cargadas, el texto se volvió a pasar por Smorph y se logró clasificar: 'subaracnoidea' (pd/adj/fem/sg), 'arteriovenosas' (pd/adj/fem/pl), 'angiografía' (pd/nom/fem/sg) e 'intraparenquimatosa' (pd/adj/fem/sg). Se ilustra con el etiquetado de 'angiografía':

```
'angiografía'.
[ 'angiografía', 'EMS','pd', 'EMS','nom', 'GEN','fem',
'NUM','sg'].
```

En segundo lugar, se estableció que todas las palabras que estuvieran completamente en mayúsculas fueran etiquetadas como siglas y las que comenzaran en mayúscula sin estar al inicio de la cláusula fueran nombres propios. Así, se logró clasificar correctamente 'TAC' (pd/abrev).

Vale aclarar que las PD que quedaron no eran verbos (conjugados, participios o gerundios), adverbios (los que no están terminados en '-mente', fueron cargados en el diccionario de Smorph), preposiciones, artículos, pronombres, etcétera (todos estos fueron cargados previamente). Asimismo, cabe destacar que, con este procedimiento, se reduce significativamente la ambigüedad, por ejemplo, los artículos seguidos de una PD no se confunden con pronombres ('la', 'las', 'lo', 'los') ya que dicha PD no puede ser un verbo conjugado (por su terminación), pero tampoco un infinitivo, ya que solo permite un

pronombre clítico. Quizá el único riesgo que se podría correr sería confundir un verbo en infinitivo con un nombre en construcciones como 'el cantar de Rolando', por ejemplo. No obstante, a los efectos del presente trabajo, que consiste en la extracción de SN, la opción seguiría siendo válida, puesto que se trata de un SN con núcleo en infinitivo.
Posteriormente, se procedió a la clasificación de PD a partir del contexto sintáctico. Se tomaron en consideración a las palabras y a los signos de puntuación que rodeaban a las PD para la creación de reglas de reagrupamiento con MPS. He aquí unos ejemplos:

- Artículo + PD + Adjetivo = SN_PD/ART+NOM+ADJ;
- Nombre + Preposición + PD + Signo de puntuación = SN_PD/NOM+ PREP+NOM;
- Preposición + Artículo + PD = SP_PD/PRE+ART+NOM.

Por medio de dichas reglas, se pudo reconocer 'aneurisma' e 'isquémico'. En 'aneurisma', se observa que el término tiene a su izquierda un artículo más un nombre ('los casos') seguido de una preposición ('de') y, a su derecha un punto. Por ende, se estableció que se trataba del término del SP adjunto del SN. Con respecto a 'isquémico', este está precedido de un nombre y un adjetivo; debido a que una expresión del tipo 'nombre + adjetivo + nombre' es agramatical, la única categoría posible para 'isquémico' es la de adjetivo.
Quedó por resolver el caso de 'endovascular', que no se pudo clasificar dado que no poseía una terminación distinguida y, a su vez, los elementos que la rodeaban no eran suficientes. La expresión del tipo "artículo + nombre + PD" no permitió deducir la categoría de la palabra, ya que la PD se ajustaba a cualquiera de las siguientes tres estructuras:

- Artículo + Nombre + Adjetivo (el caso del ejemplo, 'la terapia endovascular');
- Artículo + Nombre + Nombre ('las células madres');
- Artículo + Nombre + Infinitivo (—vio a— 'la paciente mejorar').

Se plantea, en este tipo de problemas, la posibilidad de postergar secuencias como esas y continuar con el reconocimiento de otras, para luego retomar las primeras con reglas de ejecución secundaria, en donde se aprovechen las secuencias previamente analizadas.
Lo que resulta evidente, no obstante, es la estrecha relación entre la PD y el nombre que la precede, sea esta de nombre-complemento, palabra compuesta o sujeto-verbo. A tales efectos, y para utilizarla momentáneamente, se creó una regla de reagrupamiento en la que las expresiones correspondientes a la estructura "artículo + nombre + PD" fueran etiquetadas como 'SN_PD'.
Una vez realizada la categorización, se necesitó constituir reglas de reconocimientos de SN que incluyeran PD. Por el momento, he tomado las que elaboré en mi tesis doctoral para el reconocimiento de SN, en las que introduje algunas modificaciones pertinentes (cambiar la regla 'artículo + nombre = SN' por 'artículo + PD = SN_PD', por ejemplo). No obstante, van a requerirse reglas más específicas, que contemplen fenómenos que podrían dificultar la clasificación, como por ejemplo un inciso.
Los SN_PD detectados fueron simplificados mediante técnicas de steeming y de esta forma se obtuvieron los siguientes candidatos a términos: 'accidente cerebrovascular isquémico', 'hemorragia subaracnoidea', 'malformación arteriovenosa', 'caso de aneurisma', 'TAC', 'angiografía cerebral', 'terapia endovascular' y 'hemorragia intraparenquimatosa'. De todos ellos, 7 fueron corroborados como términos específicos del dominio de medicina, por profesionales del área. El inconveniente surgió con 'caso de aneurisma', que no fue

considerado término, pero, en dicha expresión, señalaron elementos que sí eran términos por separado: 'caso' y 'aneurisma'.

4. Organización del trabajo y dificultades

La experimentación se llevará a cabo como se ha ejemplificado, sobre un corpus que actualmente se está construyendo. Hasta ahora, se reunieron textos de medicina que suman un total de 1.000.000 de palabras y que están siendo revisados manualmente por médicos a fin de obtener listas de referencias de los términos allí incluidos.

Las dificultades que se han presentado hasta el momento tienen que ver con PD que se encuentran solas, es decir, que no están rodeadas por otros elementos que permitan decir su categoría, o bien, cuando hay una combinación de ellas. Otro de los inconvenientes son los nombres propios que, en algunos casos, pueden ser términos ('Alzheimer'). Por último, hay que mencionar los errores de ortografía cometidos por los autores de los textos.

A tales efectos, será necesario recurrir a técnicas de evaluación automática de la calidad de los candidatos a términos extraídos, para ello, no se descarta para recurrir a métodos estadísticos de evaluación, entre otros.

Referencias

[1] AÏT MOKTHAR, S. (1998) SMORPH: Guide d'utilisation. Rapport technique. Universidad Blaise Pascal/GRIL. Clermont-Fd.

[2] Abbaci, F. (1999) Développement du Module Post-Smorph. Memoria del DEA de Linguistique et Informatique. Universidad Blaise-Pascal/GRIL. Clermont-Fd.

[3] Sager, J. (1993) Curso práctico sobre el procesamiento de la terminología. Madrid: Fundación Sánchez Ruíz Pérez.

[4] Vivaldi, J. (2011) "Terminología y Wikipedia". Seminario IULATerm. Barcelona: Universitat Pomeu Fabra.

[5] Moreno-Sandoval, A. Terminología y sociedad del conocimiento. 2009.

[6] Krauthammer, M. y Nenadić, G. (2004) "Term identification in the biomedical literature". En J. of Biomedical Informatics.

[7] Barrón-Cedeño et al. (2009) "An improved automatic term recognition method for spanish". In A. Gelbukh, editor, Computational Linguistics and Intelligent Text Processing, volume 5449 of Lecture Notes in Computer Science, pages 125-136. Springer Berlin / Heidelberg.

[8] Bosman, W. and Vossen. P. (2010) Bootstrapping language neutral term extraction. In (N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, Proceedings of the Seventh conference on International Language Resources and Evaluation, Valletta, Malta, may 2010. European Language Resources Association.

[9] Bonin, F. et al. (2010) "A contrastive approach to multi-word extraction from domain-specific corpora". In (N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, Proceedings of the Seventh conference on International Language Resources and Evaluation, Valletta, Malta. European Language Resources Association.

[10] Gelbukh, G. et al. (2010) "Automatic term extraction using log-likelihood based comparison with general reference corpus". In Proceedings of the Natural language processing and information systems, and 15th international conference on Applications of natural language to information systems, NLDB'10, pages 248-255, Berlin, Heidelberg. Springer-Verlag.

[11] Castro, E. (2010) "Automatic identification of biomedical concepts in spanish-language unstructured clinical texts". In Proceedings of the 1st ACM International Health Informatics Symposium, IHI '10, pages 751-757, New York, NY, USA. ACM.

[12] Lacoste, C. et al. Medical-image retrieval based on knowledge-assisted text and image indexing. IEEE Trans. Circuits Syst. Video Techn., 17(7):889-900, 2007.

[13] D. Sánchez, et al. Web-based semantic similarity: An evaluation in the biomedical domain. Int. J. Software and Informatics, 4(1):39-52, 2010.

[14] López Rodríguez, C. et al. Gestión terminológica basada en el conocimiento y generación de recursos de información sobre el cáncer: el proyecto Oncoterm. Revista E Salud, 2(8), 2006.

[15] López Rodríguez, C. et al. Terminología basada en el conocimiento para la traducción y la divulgación médicas: el caso de Oncoterm. Panace, VII (24):228-240. 2006.
[16] Vivaldi, J. and Rodríguez, H. Using wikipedia for term extraction in the biomedical domain: First experiences. Procesamiento del Lenguaje Natural, 45:251-254, 2010.
[17] Vivaldi, J. et al. Automatic summarization using terminological and semantic resources. In (N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, Proceedings of the Seventh conference on International Language Resources and Evaluation, Valletta, Malta, may 2010. European Language Resources Association.
[18] Alarcón, R. Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios. In [CD-ROM], editor, Serie Tesis, number 26. Barcelona: IULA, 2010. ISBN: 13: 978-84-89782-46-4.
[19] Tuason et al. (2004) "Biological Nomenclature: A source of Lexical Knowledge and Ambiguity", en: Proceedings of Pac Symp Biocomput.
[20] Aït-Mokthar, S. y Rodrigo Mateos, J. (1995) "Segmentación y análisis morfológico en español utilizando el sistema Smorph", en SEPLN Revista/'95. Jaén: SEPL.
[21] Lang, M. (2002) Formación de palabras en español. Madrid: Cátedra.
[22] Cabré Castelví, M. (2006) "Morfología y terminología". En Feliú Arquiola (Ed.), La morfología a debate. Jaén: Universidad de Jaén.
[23] Manning, C. et al. Language models for information retrieval. In An Introduction to Information Retrieval, chapter 12. Cambridge University Press, 2008.

Datos de Contacto:
Walter Koza. Grupo Infosur, Universidad Nacional de Rosario. Callao 260 PB Dto. D. E-mail: kozawalter@opendeusto.es