Identificación de Términos a partir de Enumeraciones Sintagmáticas Nominales: Una Aplicación al Dominio Médico.

Conrado MS[1], Koza W[2], Díaz Labrador J[3], Abaitua Odriozol JK[3], Rezende SO[1], Pardo TAS. [1], Solana Z[2]

[1]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Brasil;
[2]Grupo Infosur – CONICET – Lingüística General I, Facultad de Humanidades y Artes, de la Universidad Nacional de Rosario, Argentina;
[3]Universidad de Deusto, Bilbao-España

Resumen

Partiendo de la hipótesis de que las enumeraciones sintagmáticas nominales (ESN) que se encuentran en los textos médicos se componen de términos específicos del dominio, presentamos un método de reconocimiento de dichas enumeraciones con el objetivo de contribuir a la extracción automática. La metodología se conforma de tres etapas: (i) reconocimiento de enumeraciones sintagmáticas nominales, aquí se utiliza exclusivamente información lingüística, a partir de la cual se elaboran reglas de análisis sintáctico; (ii) extracción automática de los candidatos a términos que se correspondían con unigramas y bigramas, y (iii) evaluación de los candidatos extraídos con el asesoramiento de expertos del área médica.

Los experimentos fueron realizados en el corpus IULA, conformado por textos médicos en español. Los resultados obtenidos fueron alentadores, ya que se logró un 67% y 68% de precisión en las enumeraciones detectadas para unigramas y bigramas respectivamente.

Palabras Clave

Extracción de Términos Médicos, Enumeraciones Sintagmáticas Nominales, Dominio de Medicina.

1. Introducción

Se describe la implantación en máquina realizada con el objetivo de extraer términos propios del dominio médico a partir de la detección automática de enumeraciones sintagmáticas nominales. El presente trabajo se inscribe dentro del área de la minería textual y, específicamente, en la extracción de términos.

Las tareas de extracción de términos poseen un lugar destacado en actividades de extracción y organización del conocimiento. Un término puede ser definido como una unidad léxica caracterizada por una referencia especial dentro de una disciplina [1] y puede estar conformado por una sola palabra (unigrama), como por ejemplo 'hormona', 'asma', o una combinación de ellas (n-gramas), como en el caso de 'tuberculosis pulmonar', 'sistema cardiovascular' (bigramas); 'esquema de tratamiento', 'estado de enfermedad' (trigramas), etcétera. Un conjunto de términos constituye la terminología.

Este tipo de tareas suele enfocarse en dominios específicos. Uno de ellos es el de la medicina. Aquí, la extracción de términos representativos suele constituir el punto de partida para la elaboración de tareas más complejas, como ser listas de entradas para diccionarios electrónicos específicos, creación de base de datos o de ontologías y taxonomías, que organizan y especifican el dominio de conocimiento, entre otras.

Existen diversos trabajos en la literatura que realizan la extracción de términos [2, 3, 4, 5]. En relación con el dominio de medicina, de acuerdo con Castro et al. [6], para el caso del inglés, hay varias investigaciones orientadas al procesamiento de textos y de datos de ese

dominio [7, 8], sin embargo, se encuentran pocas iniciativas para el español [9, 10, 11, 12, 13].
A modo de aporte, el método de detección aquí presentado se basa en información lingüística en la medida en que se ha podido observar que el análisis, y posterior modelización e implantación en máquina, de algunas expresiones sintácticas específicas pueden ayudar a las tareas de extracción de términos. Una de estas expresiones sería la enumeración, que ya ha sido trabajada con el inglés por autores como Nenadić y Ananaidou [14], y que formaría parte, junto con la topicalización, aposición, etcétera, de los recursos a los que recurre un escritor a fin de destacar algún elemento del texto que considera relevante. Por otro lado, en contextos de dominios especializados, tales elementos destacados se corresponderían con términos. A modo de ejemplo, una enumeración detectada en un fragmento del corpus:

> "(…) fases evolutivas del **enfisema pulmonar, bronquitis crónica y asma bronquial**. (...)"

A tales efectos, se desarrolló un método de detección automática de enumeraciones a partir de reglas de reagrupamiento. En las secciones siguientes se presenta una breve descripción de la enumeración sintagmática nominal y la metodología utilizada.

2. La enumeración

La enumeración es la enunciación sucesiva de un conjunto de elementos que componen un todo. La coma separa los componentes de las enumeraciones siempre que estos no sean complejos y que previamente contengan comas en su expresión, ya que en ese caso debe utilizarse como elemento de separación al punto y coma. Un ejemplo de este tipo de enumeración sería:

(1) [Los alumnos destacados son: Santamaría, Carlos; Taborda, Elena; Varlotta, José Armando.]

En el caso de las enumeraciones completas o exhaustivas, el último elemento debe ir introducido por una conjunción, delante de la cual no debe escribirse coma.
Ejemplos:

(2) [Compró pan, verduras y carne.]
(3) [No compró pan, verduras ni carne.]
(4) [¿Compró pan, verduras o carne?]

Prácticamente no se han encontrado clasificaciones de las distintas enumeraciones dentro del ámbito de la lingüística informática; con excepción de la realizada por Garat Baridón [15]. No obstante la clasificación que este autor presenta se limita a distinguir sólo dos grandes grupos de series: las proposicionales y las simples, que, a su vez, comprenden las series nominales, adjetivales y las que él denomina "otras" y que incluyen las que no se consideran en los casos anteriores.
A tales efectos, en trabajos anteriores [16] se ha establecido una clasificación propia, constituida sobre la base de dos fenómenos lingüísticos:

I- Respecto de la estructura de los componentes de la serie: Se refiere a la relación que existe entre los elementos enumerados (estructuras, características semánticas, etcétera);
II- Respecto del número de los elementos: Aquí se va a remitir acerca del carácter finito o infinito que puede poseer la enumeración.

En I se describen los diferentes elementos que pueden ser enumerados, como por ejemplo sintagmas, denominadas "sintagmáticas" (5 y 6), subordinadas (7), etcétera.

(5) [María es simpática, agradable y muy linda.] (Enumeración sintagmática nominal)
(6) [En el supermercado compré una docena de huevos, un kilo de harina y dos paquetes de acelga.] (Enumeración sintagmática nominal)
(7) [Quiero que me entiendan, que no me juzguen y que me esperen.]

El trabajo de extracción estuvo focalizado en enumeraciones del tipo de (6).
Por otro lado, en I se incluyen las enumeraciones completas, es decir, aquellas que presentan un conjunto cerrado, a las que no se les pueden adicionar nuevos elementos, tal como se muestra en (2), (3) y (4), a los que habría que adicionar los casos de asíndeton (8). No obstante, puede haber enumeraciones que quedan abiertas, como en el caso de las finalizadas con punto suspensivo (9) o "etcétera" (10) y las que incluyen expresiones del tipo "entre otros", "entre otras cosas" (11), etcétera.

(8) [Platero es pequeño, peludo, suave.]
(9) [Los números primos son el dos, el tres, el cinco, el siete…]
(10) [Le gusta la pintura, la música clásica y la poesía, entre otras cosas.]

Puede haber ESN tanto completas como infinitas, por lo que debieron constituirse reglas que contemplaran ambos casos.

3. Metodología

El objetivo fue reconocer las ESN incluidas en los textos y evaluar si los SN enumerados en ellas se correspondían con términos del dominio específico. La metodología, que se muestra en la Figura 1, se conforma de tres etapas: (i) reconocimiento de ESN, aquí se utiliza exclusivamente información lingüística, a partir de la que se elaboran reglas de análisis sintáctico; (ii) extracción automática de los candidatos a términos enumerados en las ESN que se correspondían con unigramas y bigramas, y (iii) evaluación de los candidatos extraídos con el asesoramiento de expertos del área médica.

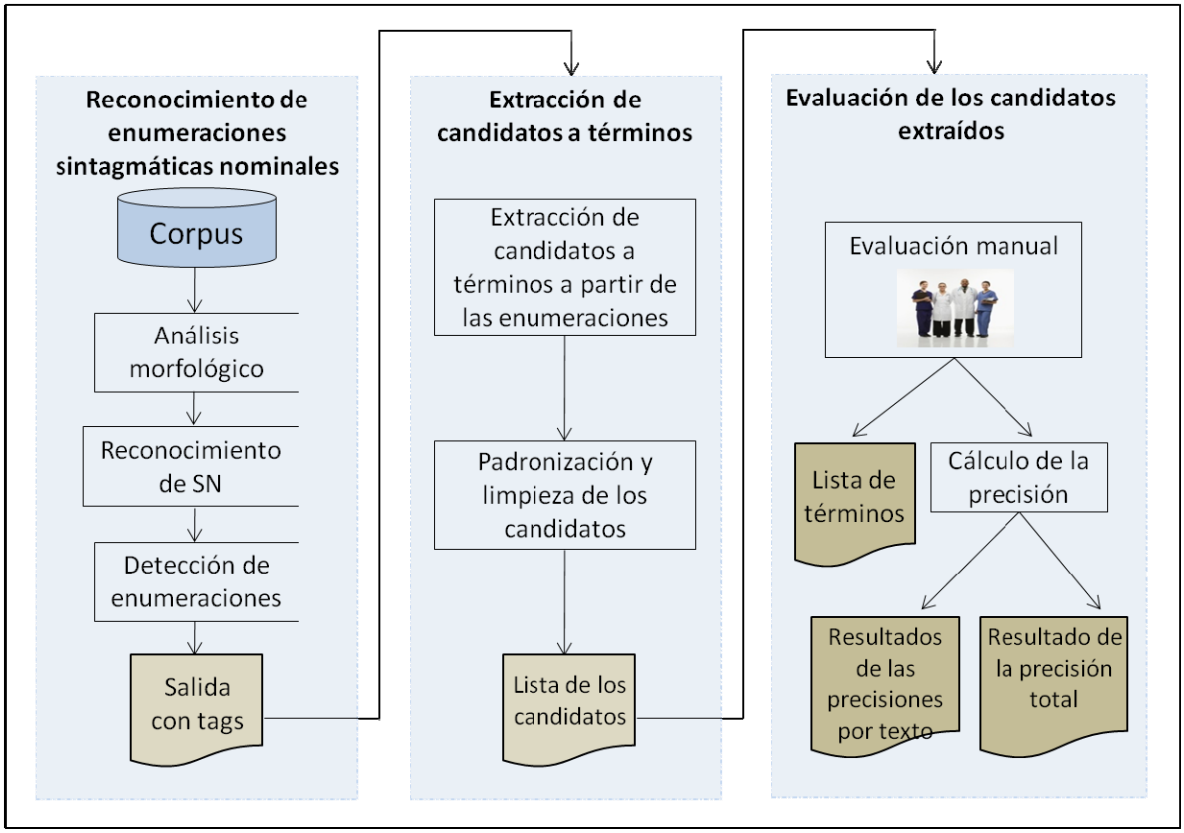

Figura 1. Extracción de términos de dominio especifico a partir del reconocimiento de enumeraciones sintagmáticas nominales.

La primera etapa corresponde al **reconocimiento de ESN**. Aquí se pasa el corpus por el analizador sintáctico a fin de obtener el *análisis morfológico* de las palabras y el reconocimiento de los signos de puntuación. Para ello, se utilizó es software Smorph [17], un analizador y generador textual que, en una única etapa, realiza la delimitación previa de los segmentos a considerar y el análisis morfológico. El programa le asigna a cada segmento (por lo general, una palabra) una etiqueta morfosintáctica con la información del lema correspondiente y los rasgos previamente declarados por el usuario. Para este experimento, se contó con la adaptación que realizó el Grupo Infosur[1] para la lengua española. A continuación, se presenta un ejemplo del análisis de la herramienta. Dado el siguiente fragmento textual:

"(...) presión arterial, frecuencia cardíaca y frecuencia respiratoria (...)".

Smorph dividirá la expresión en la lista de elementos mostrada en la Tabla 1.

[1] Sitio electrónico: http://www.infosurrevista.com.ar

‘presión’.
[‘presión’, ‘EMS’,’nom’, ‘GEN’,’fem’, ‘NUM’,’sg’].
‘arterial’.
[‘arterial’, ‘EMS’,’adj’, ‘GEN’,’_’, ‘NUM’,’sg’].
‘,’.
[‘cc’, ‘EMS’,’coma’].
‘frecuencia’.
[‘frecuencia’, ‘EMS’,’nom’, ‘GEN’,’fem’, ‘NUM’,’sg’].
‘cardíaca’.
[‘cardíaco’, ‘EMS’,’adj’, ‘GEN’,’fem’, ‘NUM’,’sg’].
‘y’.
[‘y’, ‘EMS’,’cop’].
‘frecuencia’.
[‘frecuencia’, ‘EMS’,’nom’, ‘GEN’,’fem’, ‘NUM’,’sg’].
‘respiratoria’.
[‘respiratorio’, ‘EMS’,’adj’, ‘GEN’,’fem’, ‘NUM’,’sg’].

Tabla 1. Salida ejemplo del programa Smorph[2].

Una vez obtenido el análisis morfológico y el reconocimiento de los signos de puntuación, se toma como input el output de Smorph para que la segunda herramienta identifique las ESN. En este caso, se recurrió al parser Módulo Post Smorph (MPS) [18], que aplica un conjunto de reglas de reagrupamiento que permite, primero, obtener los SN y, segundo, las ESN. He aquí ejemplo de las dos reglas:

- Artículo + Nombre + Adjetivo = SN (Ejemplo: ‘la gastritis crónica’)

Notación específica de MPS:

```
art+adj+nom da snn%
S1 [L1, 'TDET', 'art']
S2 [L2, 'EMS','nom']
S3 [L3, 'EMS','adj']
--> S1+S2+S3 [L1+L2+L3, 'EMS', 'SN' ].
%la gastritis crónica%
```

- SN + coma + SN + Copulativo + SN = ESN (Ejemplo: ‘presión arterial, frecuencia cardíaca y frecuencia respiratoria’)

Notación específica de MPS:

```
SN+coma+SN+cop+SN da ESN%
S1 [L1, 'EMS', 'SN']
S2 [L2, 'EMS','coma']
S3 [L3, 'EMS','SN']
```

[2] Referencias: *EMS*: Etiqueta Morfosintáctica, *nom*: nombre, *GEN*: genero, *fem*: femenino, *NUM*: número, *sg*: singular y *cop*: copulativo.

S4 [L4, 'EMS','cop']
S5 [L5, 'EMS','SN']

--> S1+S2+S3+S4+S5 [L1+L2+L3+L4+L5, 'EMS', 'ESN'].
% presión arterial, frecuencia cardíaca y frecuencia respiratoria%

De esta manera, hemos desarrollado un conjunto de reglas que modelan los casos posibles de ESN, estas se describen a continuación:

- (SN + coma) ≥ 1 + SN + conjunción + SN = ESN
- (SN + coma) ≥ 2 + punto suspensivos = ESN
- (SN + coma) ≥ 2 + '*etcétera*' = ESN
- (SN + coma) ≥ 2 + '*entre otras cosas*' = ESN
- (SN + coma) ≥ 3 = ESN (para los casos de asíndeton)

A partir de las reglas constituidas a partir de la modelización, se pudieron detectar las ESN incluidas en los textos que conformaban el corpus, a continuación dos ejemplos:

> **[Ejemplo 1]** "(...) tal como la hipertensión, isquemia coronaria, diabetes mellitus, asma bronquial, etcétera (...)"
>
> **[Ejemplo 2]** "(...) fases evolutivas del enfisema pulmonar, bronquitis crónica y asma bronquial. (...)"

[Ejemplo 1]

'tal'. ['tal', 'EMS', 'sadjn']. 'como'. ['como', 'EMS', 'adv']. 'la hipertensión , isquemia coronaria , diabetes mellitus , asma bronquial , etcétera'.
['la hipertensión cc isquemia coronario cc diabetes mellitus cc asma bronquial etcétera', 'EMS', 'ESN'].

[Ejemplo 2]

'fases evolutivas'. ['fase evolutivo', 'EMS', 'SN']. 'del'.
['del', 'EMS', 'contr']. 'enfisema pulmonar , bronquitis crónica y asma bronquial'.
['enfisema pulmonar cc bronquitis crónica y asma bronquial', 'EMS', 'ESN'].

Tabla 2. Ejemplos de enumeraciones detectadas.

A partir de las ESN reconocidas, **se extraen automáticamente los SN en calidad de candidatos a términos** (segunda etapa de la metodología). Dichos candidatos son *estandarizados* a partir de la eliminación de los numerales o cualquier símbolo que pudiera perjudicar la calidad del candidato a término. Asimismo, también se eliminan las *stopwords* de las extremidades de los candidatos.

Finalmente, se realiza la tercera etapa de la metodología, que es la **evaluación de los candidatos extraídos**. Cada uno de ellos es analizado por expertos del dominio del corpus con el objetivo de verificar si el candidato es un verdadero término del dominio. Esa evaluación es computada utilizando la medida de precisión en relación con cada texto (*precisión por texto*), que es la precisión media obtenida considerando el número de textos del corpus, y en relación al corpus como un todo (*precisión total*). Entonces, se graban los verdaderos términos en listas, aquí llamadas de *lista de términos*.
Si bien los experimentos fueron hechos sobre corpus en español, vale aclarar que la metodología puede ser adaptada a otras lenguas cambiando la información lingüística de los analizadores utilizados.

3.1 Experimentación

Para este caso, la extracción fue focalizada en unigramas (términos compuestos por una palabra) y bigramas (compuestos por dos palabras), quedando para etapas futuras la extracción de trigramas y expresiones mayores. Se utilizó el corpus IULA-UPF conformado por 12 textos del dominio médico en español y que la media de palabras por texto es 8207.
Se aplicó la metodología descripta previamente, análisis morfológico y reconocimiento de los signos de puntuación con Smorph, primero, y reconocimiento de ESN con MPS, segundo, y se extrajeron los SN enumerados en las ESN en calidad de candidatos a términos.
Luego, se realizó la limpieza de los candidatos, quitando de la lista a los candidatos compuestos solo por una letrea o "palabras vacías". Asimismo, también se quitaron las palabras vacías ubicadas en los extremos de los candidatos. Para esta parte de la experimentación, recurrimos a la lista de palabras vacías enumeradas en el Proyecto Snowball[3] y, a la vez, adicionamos otras palabras vacías como ser las formas de los verbos 'poder' y 'deber' y algunos adverbios como 'siempre', llegando a un total de 773 palabras vacías.

3.2 Resultados

Se evaluó la extracción de los candidatos a términos, considerando solo aquellos que eran unigramas y bigramas. El objetivo fue verificar cuáles candidatos a términos pertenecían realmente al dominio de medicina. Esa verificación fue realizada por dos expertos del área médica y para calcular objetivamente los resultados se utilizó la medida de precisión de acuerdo con el trabajo Gelbukh et al. [19]
Para el corpus IULA-UPF, se extrajeron un total de 302 unigramas. De estos, los expertos verificaron que 98 no pertenecían al dominio médico. A tales efectos, la precisión total del corpus fue 68% [(302-98)/302 = 68%].
Para el caso de los bigramas, se extrajeron 172 y 57 no fueron candidatos a términos; por lo que la precisión lograda fue de 67% [(172-57)/172 = 67%].
Además de la precisión total referente al corpus, se calculó las precisiones obtenidas en cada texto del corpus. El número de candidatos extraídos (*NC*) y la precisión por texto (*P.(%)*) son presentadas en las Tablas 3 y 4. La precisión media obtenida para los unigramas es 69% y para los bigramas es 69,3%.

[3] En: http://snowball.tartarus.org/algorithms/spanish/stop.txt

Textos	NC	P. (%)		Textos	NC	P. (%)
1	29	69		7	25	68
2	50	76		8	25	72
3	27	59		9	27	78
4	44	52		10	52	83
5	53	64		11	30	53
6	33	85		12	39	69

Tabla 3. Resultados de los unigramas extraídos

Textos	NC	P. (%)		Textos	NC	P. (%)
1	21	81		7	8	25
2	23	87		8	16	81
3	19	79		9	26	65
4	14	71		10	19	84
5	23	48		11	18	56
6	19	79		12	17	76

Tabla 4. Resultados de los bigramas extraídos

4. Consideraciones finales

Por medio de la extracción realizada, se lograron detectar términos del dominio de medicina. Los resultados son alentadores, pues se alcanzó un 68% y un 67% para unigramas y bigramas respectivamente.

No obstante, hay que señalar que, obviamente, este método no es suficiente para extraer todos los candidatos a términos que se encuentran en los textos, sino solo los que están incluidos en las enumeraciones. Esto implica que la extracción de candidatos a términos a partir del reconocimiento automático de las ESN debe ser combinada con otras técnicas de extracción. Otro inconveniente radicó en la limitación de MPS a cadenas de 100 caracteres como máximo, esto hizo que algunos SN que iniciaban la enumeración no fueran incluidos.

A tales efectos, el trabajo a futuro se organiza en torno a dos ejes:

- Elaborar reglas de reconocimiento de ESN más abarcativas;
- Combinar el método de extracción de candidatos a términos a partir de las ESN con otras técnicas de extracción.

Agradecimientos

Los autores agradecen a Erasmus Mundus, CNPq, FAPESP y CONICET por el apoyo financiero, y a Vivaldi y Rodríguez por facilitarnos el corpus y las listas de referencias utilizados.

Referencias

[1] Sager, J. 1993. *Curso práctico sobre el procesamiento de la terminología*. Madrid: Fundación Germán Sánchez Ruizpérez.

[2] Barrón-Cedeño et al. 2009. "An improved automatic term recognition method for Spanish". En Gelbukh, A. (Ed.) *Computational Linguistics and Intelligent Text Processing*, Vol. 5449 of Lecture Notes in Computer Science. Springer Berlin/Heidelberg.

[3] Bosma, W. y Vossen, P. 2010. "Bootstrapping language neutral term extraction". En Proceedings of the Seventh conference on International Language Resources and Evaluation, Valletta, Malta, may 2010. European Language Resources Association.

[4] Bonin, F. et al. 2010. “A contrastive approach to multi-word extraction from domain specific corpora. En *Proceedings of the Seventh conference on International Language Resources and Evaluation,* Valletta, Malta, may 2010. European Language Resources Association.
[5] Gelbukh, A. et al. 2010. “Automatic term extraction using log-likelihood based comparison with general reference corpus”. En *Proceedings of the Natural language processing and information systems, and 15th international conference on Applications of natural language to information systems, NLDB'10*. Berlin, Heidelberg. Springer-Verlag.
[6] Castro, E. 2010. “Automatic identification of biomedical concepts in Spanish-language unstructured clinical texts”. En *Proceedings of the 1st ACM International Health Informatics Symposium, IHI '10*. New York, NY, USA, 2010. ACM.
[7] Lacoste, J. 2007. “Medical-image retrieval based on knowledge-assisted text and image indexing”. IEEE Trans. Circuits Syst. Video Techn. 17(7):889-900, 2007.
[8] Sanchez, D. et al. “Web-based semantic similarity: An evaluation in the biomedical domain”. Int. J. Software and Informatics.
[9] López, C. et al. 2006. “Gestión terminológica basada en el conocimiento y generación de recursos de información sobre el cáncer: el proyecto Oncoterm”. En *Revista E Salud*.
[10] López, C. et al. 2006. “Terminología basada en el conocimiento para la traducción y la divulgación médicas: el caso de Oncoterm”. En *Panace*. VII.
[11] Vivaldi, J. et al. 2010. “Automatic summarization using terminological and semantic resources”. En *Proceedings of the Seventh conference on International Language Resources and Evaluation,* Valletta, Malta, may 2010. European Language Resources Association.
[12] Vivaldi, J. y Rodríguez, H. 2010. “Using Wikipedia for term extraction in the biomedical domain: First experiences”. En *Procesamiento del Lenguaje Natural*. N°45.
[13] Alarcón, R. 2010. “Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios”. En *Serie Tesis*. Num. 26. Barcelona: IULA.
[14] Nenadic, G. y Ananiadou, S. “Mining semantically related terms from biomedical literature”. ACM Transactions on Asian Language Information Processing, 5(1).
[15] Garat Baridón, D. 2006. *Análisis de superficie basado en puntuación*. Tesis de maestría. Montevideo: PEDECIBA Informática, Instituto de computación – Facultad de Ingeniería Universidad de la República.
[16] Koza, W. 2008. “Análisis automático de textos: Reconocimiento de enumeraciones”. En Manni Héctor (comp.) *Actas del XI Congreso de la Sociedad Argentina de Lingüística*. Santa Fe: Universidad Nacional del Litoral.
[17] Aït Mokthar, S. 1995. *SMORPH: Guide d’utilisation. Rapport technique*, Clermont-Fd. : Universidad Blaise Pascal/GRILL.
[18] Abbaci, D. 1999. *Développment du Module Post-Smorph*. En Memória del DEA de Linguistique et Informatique. Groupe de Recherche dans les Industries de la Langue. Universidad Blaise-Pascal - Clermont-Ferrand, 1999.
[19] Gelbukh, A. et al. 2010. “Automatic term extraction using log-likehood based comparison with general reference corpus”. En Hopfe et al. (Eds.) Proceedings of the Natural Language Processing and Information Systems, and 15th International Conference on Applications of Natural Language to Information Systems, NLDB'10. Berlin, Heidelberg, 2010. Springer-Verlag.

Datos de Contacto:

Walter Koza. GRUPO INFOSUR-CONICET-UNR. Salta 2960 4°B (2000) Rosario, Santa Fe, Argentina. kozawalter@gmail.com