C-value aplicado a la extracción de términos multipalabra en documentos técnicos y científicos en español

L. Alberto Barrón
Instituto de Ingeniería, UNAM
alberto@pumas.ii.unam.mx

Gerardo Sierra
Instituto de Ingeniería, UNAM
gsierram@ii.unam.mx

Elio Villaseñor
Facultad de Ciencias, UNAM
elio@www.dynamics.unam.edu

Resumen

La extracción automática de terminologías en documentos de especialidad ha sido explorada ampliamente en idiomas como el inglés, el francés y el alemán, pero no lo ha sido tanto en el español. El algoritmo C-value/NC-value se ha implementado exitosamente para la extracción automática de términos biológicos sobre documentos en inglés. En el presente trabajo se describe la adaptación de la etapa C-value, un método lingüístico y estadístico para la extracción de términos multipalabra en inglés, para la extracción de términos multipalabra, en particular del área de ingeniería, en español.

1. Introducción

En la lingüística, y en particular en la lexicografía y terminología, es necesario revisar textos para extraer los elementos necesarios para su investigación. La extracción de términos automatizada por medios computacionales puede resultar de gran utilidad en su trabajo diario, además de ser útil en el desarrollo de herramientas lexicográficas, en la delimitación de glosarios, en el análisis diacrónico de la lengua (para encontrar el momento en que un término surge en una disciplina, por ejemplo) y para dar soporte al web semántico, entre otras cosas.

En este trabajo se presentan los primeros pasos en la implementación de un algoritmo para la extracción automática de términos multipalabra en textos técnicos y científicos, originalmente para el inglés y que se ha aplicado exitosamente para diferentes idiomas, pero no antes en español.

La entrada para el algoritmo es un corpus textual de carácter científico o técnico y da como resultado una lista con los candidatos a términos obtenidos por medio de una combinación de métodos lingüísticos y estadísticos.

El artículo está estructurado de la siguiente manera. En el segundo apartado se presenta una descripción sobre el estudio de la disciplina conocida como terminología y de su objeto de estudio: el término. En el tercero se ofrece una reseña sobre la etapa C-value del algoritmo C-value/NC-value. En la cuarta parte se presenta la adaptación de C-value para la extracción de términos en español, por lo que se ha optado por distinguir esta adaptación con el nombre C-value-E (C-value, español); el apartado incluye un breve análisis de las construcciones sintácticas de los términos en español, de gran importancia para su extracción automática. El quinto apartado presenta algunos de los resultados que se han obtenido hasta el momento. Finalmente, en el apartado seis se presentan las conclusiones y se describe el trabajo a futuro.

2. El estudio de las terminologías

La terminología es una disciplina que ha florecido en los últimos tiempos, principalmente en regiones inmersas en una fuerte interacción entre lenguas. Existen varias escuelas en el mundo que han destacado sobre las demás; entre ellas vale la pena señalar las escuelas rusa, austriaca, canadiense y catalana. Para Cabré, principal representante de la escuela catalana, la terminología es “la disciplina que se ocupa de los términos especializados” [3], a la vez que la considera como el “conjunto de directrices o principios que rigen la recopilación de términos”.

Ampliando esta definición, Cardero considera que la terminología, como disciplina, se ocupa de “identificar el vocabulario de una especialidad en forma sistemática en una situación comunicativa específica en los textos propios de la especialidad y entre los profesionales del área, analizarlo desde la lingüística y, si es necesario, crearlo entre el especialista y el terminólogo, además de normalizarlo para un funcionamiento concreto con la finalidad de responder a las necesidades de expresión de sus usuarios” [6].

La terminología se considera a la vez como una disciplina integrante de la lingüística y como un campo

multidisciplinario construido a partir de las teorías del conocimiento, la comunicación y el lenguaje [4], teniendo a los términos, las unidades constituyentes de los vocabularios de especialidad, como su objeto de estudio.

Queda ahora precisar qué es un término. Para Dubuc [8], el término es "el elemento constitutivo de cualquier nomenclatura terminológica que esté relacionada con una lengua de especialidad […], es la denominación de un objeto, propio de una determinada área de especialidad". En este sentido, cabe señalar que si un término es sacado de contexto, es decir, del entorno del área de especialidad a la que pertenece, pierde la categoría de término.

De manera simplificada, podemos considerar que un término es un sintagma asociado a un concepto, dentro de un área de especialidad, en donde un sintagma es un conjunto de una o más palabras cohesionadas que tienen un concepto asociado, p. ej. *gato hidráulico*.

Hay que notar que un término no está limitado a estar constituido por una sola palabra, como en el caso de las palabras *procesador* o *Internet*, en el área de la computación, sino que puede estar conformado por un sintagma multipalabra. De hecho, las palabras que conforman a un término multipalabra pueden ser también términos, como en el caso de *red*, *red de computadoras* y *red de computadoras inalámbrica*. Dicha dependencia entre sintagmas terminológicos de distintas longitudes es de gran importancia para el algoritmo C-value. En el siguiente apartado se presenta la descripción del algoritmo C-value/NC-value, abordando principalmente su etapa C-value.

3. C-value/NC-value

En un estudio reciente [5] se ha hecho una descripción de algunos de los extractores de términos que se han desarrollado para el inglés, francés y alemán. Varios de ellos se desarrollaron para apoyar en actividades lexicográficas o de traducción. Sin embargo, poco se ha hecho sobre extracción automática de términos en documentos en español.

Los sistemas desarrollados hasta ahora se basan en técnicas lingüísticas, estadísticas o una combinación de ambas. En este trabajo, para la extracción de términos en español, se ha optado por adaptar la etapa C-value del algoritmo C-value/NC-value [9], desarrollado para la extracción de términos biológicos multipalabra en inglés. El algoritmo está dividido en dos etapas, una lingüística y una estadística, las cuales se abordan a continuación.

3.1. Etapa lingüística de C-value

Antes de comenzar de lleno con la descripción de la etapa lingüística de C-value cabe señalar que, para la extracción automática de términos, resulta de gran importancia contar con un texto procesado en formato POST (*part of speech tagging*). Con ello, la detección de candidatos a términos se reduce a la detección de patrones sintácticos (parte lingüística) y al conteo de la frecuencia de aparición de sintagmas en un corpus (parte estadística).

La etapa lingüística es la encargada de obtener sintagmas candidatos a términos con base en el cumplimiento de ciertos patrones lingüísticos. Se divide en tres partes:

1) Etiquetado de partes de la oración (POS) del corpus. Cada una de las palabras en el texto es etiquetada con su categoría gramatical (nombre, adjetivo, preposición, etc.). Esta etapa se llevó a cabo en la Universidad de Montreal por medio del etiquetador de Eric Brill [2].

2) Aplicación del filtro lingüístico sobre el corpus etiquetado. La definición del filtro lingüístico para documentos en inglés, se basa en el hecho de que la mayoría de los términos en esta lengua están compuestos por nombres y adjetivos [11], de manera que el filtro lingüístico acepta ese tipo de sintagmas. Para la extracción de términos biológicos en inglés, han determinado tres sencillas reglas sintácticas:

1) NC+NC,
2) (ADJ | NC)+NC
3) ((ADJ |NC)+|((ADJ | NC)*(NCPREP)?)(ADJ | NC)*)NC

Todo sintagma que cumpla con estos patrones es seleccionado como un candidato a término.

3) Aplicación de la lista de paro. Cuentan con una lista de 229 palabras funcionales y de contenido que, con base en sus estudios, no se espera que se encuentren en términos del área de biología.

En esta etapa se obtienen los candidatos que serán ponderados en la etapa estadística.

3.2. Etapa estadística de C-value

Luego de la aplicación de la etapa lingüística, en la que se ha obtenido un conjunto de sintagmas candidatos, se pasa a la etapa estadística. Se trata básicamente de un conteo de ocurrencias de los sintagmas detectados por el filtro lingüístico; sin embargo, dicho conteo no es trivial, pues toma en cuenta diversos valores que intentan discriminar con mayor precisión los sintagmas que en verdad son términos de los que no lo son.

La etapa estadística es la encargada de medir la probabilidad de que un sintagma sea un término. Para ello, se basa en cuatro valores:

1) La frecuencia total de ocurrencia del sintagma candidato en el corpus.

2) La frecuencia total de ocurrencia del sintagma candidato dentro de candidatos de mayor longitud.

3) El número de ocurrencias de esos candidatos de mayor longitud.

4) La longitud del sintagma candidato.

Basado en esos cuatro parámetros, C-Value determina la posibilidad de que un sintagma candidato sea en verdad un término mediante la siguiente fórmula.

$$C\text{-}value = \begin{cases} log_2|a| * f(a) \\ log_2|a| * \left(f(a) - \frac{1}{P(T_a)} \sum f(b) \; b \in T_a \right) \end{cases}$$

La primera parte de la fórmula se usa sólo en los candidatos de mayor longitud *N* en el corpus, porque evidentemente no se han hallado candidatos de mayor longitud que los pudieran contener. Para los candidatos de longitud $l \leq N-1$ se aplica la segunda parte de la fórmula. A continuación la descripción de los elementos de la fórmula señalada:

a sintagma candidato.

$|a|$ longitud del sintagma candidato *a.*

f(a) frecuencia de ocurrencia del sintagma *a* en el corpus.

T_a conjunto de candidatos de mayor longitud que contienen a *a.*

$P(T_a)$ número de esos candidatos (tipos).

$\Sigma f(b)$ ocurrencia total de *a* como subcadena de cualquier sintagma candidato *b* tal que $|a|<|b|$.

Es así como C-value genera una lista ordenada con base en la posibilidad de que un sintagma candidato sea en verdad un término.

Con el algoritmo C-value/NC-value aplicado al inglés, se obtienen candidatos del tipo *optic nerve*, *carcinoma basal cell* y *fibrous tissue*, todos ellos términos correctos. Sin embargo, se extraen también sintagmas como *plane of section* o *stump of optic nerve*, que no son en realidad términos.

4. C-value para el español: C-value-E

La construcción sintáctica de términos en inglés es totalmente distinta a la de los términos en español, razón por la cual el algoritmo C-value requiere ser adaptado para explotarse en documentos en este idioma. A dicha adaptación la nombraremos C-value-E.

A continuación se aborda el C-value-E para la extracción de términos en español en documentos del área de ingeniería.

4.1. Construcción de términos multipalabra en el español

Para este experimento se ha utilizado un corpus recopilado por el Grupo de Ingeniería Lingüística de la UNAM, compuesto por documentos del área de ingeniería en español de México: el Corpus Lingüístico de Ingeniería (CLI) [10], por lo que los patrones sintácticos obtenidos son de términos de ingeniería en español de México.

Se ha encontrado que si bien existen intersecciones en los conjuntos de patrones de diversas áreas, éstas no son suficientes para hacer búsquedas generales sobre documentos de cualquier área con las mismas reglas, por lo que la etapa lingüística no es solo dependiente de la lengua, sino que también lo es, en menor grado, del área de especialidad tratada.

Los términos multipalabra en español mexicano, y en particular los del área de ingeniería, tienen patrones lingüísticos distintos a los del inglés. Los patrones que con mayor frecuencia presentan los términos en español son *nombre*, *nombre-adjetivo*, *nombre-nombre*, *nombre-preposición-nombre* y combinaciones más complejas entre ellos. Ejemplos de dichos patrones pueden ser observados en la tabla 1.

Tabla 1. Patrones de términos en ingeniería

Término	***Patrón sintáctico***
motor	N
tracción independiente	N Adj
flecha motriz	N N
palanca de velocidades	N Prep N
palanca de cambios de velocidades	N Prep N Prep N
tiempo efectivo de trabajo	N Adj Prep N
motocultor de alto despeje	N Prep Adj N

Tomando como base los patrones hallados con una extracción manual, se han diseñado las siguientes cuatro reglas:

1) <NC|NP>$^+$<PREP><CARD>*<NP|NC|ADJ>$^+$

2) <NC><NEG>*<ADJ|Vladj|NC|NP>

3) <NC|NP>$^+$ [<ADJ><PREP><NC>]*

4) <NC><PREP>]$^+$<NC>[[<ADJ><CC|PREP>]* <ADJ>]*

Dichas reglas se han implementado por medio de un autómata finito programado con la ayuda del Natural Language Tool Kit [1], encontrando buenos y malos candidatos como los mostrados en la tabla 2.

Tabla 2. Comparativa de candidatos

Buenos candidatos	*Malos candidatos*
análisis estructural	Ciencias de Kiev
analizador de espectros	colt I14
ángulo de fase	cuenta Ias
carga gradual	e5.6 oats
desplazamiento de placas	forma directa
Ciudad Universitaria	E1 modelo
comportamiento estático y dinámico	densidades DOYLAXAM15
condición de apoyo físico	estudios futuros
condiciones de apoyo	FIOCO Me
configuración gráfica	frecuencias bajas
curva carga desplazamiento	modo teórico

Es importante señalar que la lista de paro contemplada para la extracción de términos en inglés no es necesaria, en principio, en la extracción de términos en español, debido a que las palabras funcionales que son descartadas previamente en inglés suelen integrar buena parte de los términos multipalabra en español.

Es la lista con los candidatos obtenidos por la etapa lingüística la que se utiliza para realizar el cálculo numérico de C-value. En el siguiente apartado se da un ejemplo de cálculo de C-value para una lista restringida de candidatos a términos.

4.2. Etapa estadística de C-value

Para simplificar la explicación del cálculo de C-value, se ha optado por una lista restringida de candidatos a términos, la que se muestra en la tabla 3. En ella se pueden observar los sintagmas candidatos, longitud y frecuencia total de aparición en el corpus de entrada. Cabe señalar que para la determinación del valor de la frecuencia de aparición de cada candidato, sólo se toman en cuenta aquellas ocurrencias del candidato en las que no se encuentra contenido dentro de otro de mayor longitud.

Tabla 3. Candidatos para C-value

Candidato (a)	*long(a)*	*frec(a)*
teoría de placas	3	3
teoría de análisis	3	3
teoría de la elasticidad lineal	5	2
teoría de la elasticidad	4	4

Suponiendo que ésta es la lista de candidatos, se presenta a continuación el cálculo de C-value para algunos de ellos. Primero, se calcula el C-value para el sintagma "*a"* de longitud mayor *N*, es decir, *a*="teoría de la elasticidad lineal", que tiene cinco palabras, por lo que su longitud es |*a*|=5.

$$C\text{-}value = log_2|a| * f(a)$$
$$C\text{-}value = log_2(5) * 2$$
$$C\text{-}value = 3.21$$

Se continúa con los de longitud $l \leq N-1$, como es el caso de *a*= "teoría de la elasticidad" con longitud |*a*| =4.

$$C\text{-}value = log_2|a| * \left(f(a) - \frac{1}{P(T_a)} \sum f(b) \; b \in T_a \right)$$
$$C\text{-}value = log_2(4) * \left(4 - \frac{1}{1} \sum f(b) \; b \in T_a \right)$$
$$C\text{-}value = log_2(4) * (4 - (1 * 2))$$
$$C\text{-}value = log_2(4) * (4 - 2)$$
$$C\text{-}value = log_2(4) * 2$$
$$C\text{-}value = 2.77$$

Siguiendo este procedimiento, se obtiene una lista ordenada con base en el C-value como la que se muestra en la tabla 4.

Tabla 4. Lista ordenada con C-value

Candidato (a)	*C-value (a)*
teoría de placas	3.29
teoría de análisis	3.29
teoría de la elasticidad lineal	3.21
teoría de la elasticidad	2.77

Así, la posibilidad de que los candidatos *teoría de placas* y *teoría de análisis* sean verdaderos términos es mayor.

C-value se basa en la consideración de que la cantidad de palabras y frecuencia de ocurrencia de un sintagma en un corpus es un factor positivo para determinar si se trata o no un término. Por otro lado, si

dicho candidato se encuentra en otro de mayor longitud, es probable que se trate de una versión simplificada del mismo, por lo que demerita la posibilidad de tratarse de un verdadero término (como en el caso de *red neuronal* y *red neuronal artificial*).

Sin embargo, el denominador de la parte negativa de la fórmula $P(T_a)$ indica que entre más candidatos distintos de longitud mayor a la del candidato analizado existan, éste tiene mayor independencia con respecto a ellos, por lo que es más probable que por sí mismo sea un candidato; por ejemplo, se puede tratar de distintas variedades de un elemento y todos ser términos (como en el caso de *red de computadoras*, *red de computadoras inalámbrica* y *red de computadoras LAN*).

5. Resultados preliminares

Para realizar las primeras pruebas con el algoritmo C-value-E se utilizó un subconjunto de documentos del corpus CLI al azar. Se optó por el documento "Comportamiento estático y dinámico de placas de vidrio", una tesis de posgrado de José A. Noriega y Luis Ferrer, de la Facultad de Ingeniería de la UNAM. El documento tiene un total de 5,722 palabras.

Entre los candidatos que fueron extraídos correctamente están *densidad de masa*, *efecto de resonancia*, *Facultad de Ingeniería*, *frecuencia natural*, *función de transferencia*, *relación de Poisson* y *teoría de placas*. Sin embargo, otros tantos sintagmas, que no son verdaderos términos, fueron extraídos erróneamente, como es el caso de *frecuencias bajas*, *marco de madera*, *número de placas*, *placas números*, *ventanería particular* y *vidrio de fabricación nacional.*

Se resume en la tabla 5 la cantidad de términos extraídos manualmente, así como los extraídos automáticamente de manera correcta e incorrecta, además de la relación de precisión y recuerdo para este primer experimento.

Tabla 5: Precisión y recuerdo

Términos extraídos manualmente	56
Candidatos extraídos automáticamente	82
Verdaderos	39
Falsos	43
Precisión	0.47
Recuerdo	0.69

6. Conclusiones

En este trabajo se ha presentado la adaptación del algoritmo C-value, un método para la extracción de términos biológicos multipalabra en inglés, para la extracción automática de términos multipalabra de ingeniería en español, dando como resultado la versión C-value-E.

La principal modificación se hizo en las reglas sintácticas para el filtrado en la etapa lingüística, las cuales han tenido que ser cambiadas para adaptarse a los patrones sintácticos que presentan los términos en español y en particular los de ingeniería.

La lista de paro, necesaria en la etapa lingüística de la extracción en inglés no es útil para el español debido a que, si bien la inclusión de palabras como preposiciones y artículos es prácticamente nula en los términos en inglés, en español es muy común.

El siguiente paso es la adaptación e integración de la etapa NC-value, con la que se espera obtener mejores resultados de recuerdo y precisión, además de buscar la posibilidad de detectar términos de una sola palabra integrando una etapa de otro método de extracción (debido a que C-value/NC-value no lo considera). Para ello se ha planteado la posibilidad de adaptar la etapa de detección de términos de una palabra de Termostat [7], basada en la comparación de un corpus técnico con uno no técnico.

La realización de este trabajo ha sido posible gracias al patrocinio del Consejo Nacional de Ciencia y Tecnología (Proyecto 46832) y al co-financiamiento del Macro-proyecto "Tecnologías para la Universidad de la Información y la Computación", dentro del marco del Programa Transdisciplinario en Investigación y Desarrollo de la Secretaría de Desarrollo Institucional, UNAM.

7. Referencias

[1] S. Bird, "NLTK-Lite: Efficient Scripting for Natural Language Processing", *Proceedings of the 4th International Conference on Natural Language Processing (ICON)*, New Delhi: Allied Publishers, Kanpur, India, 2005, pp. 11-18.

[2] E. Brill, "Some advances in transformation-based part-of-speech tagging", *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94)*. Seattle, 1994, pp. 722-727.

[3] M.T. Cabré, *La terminología. Teoría, metodología, aplicaciones*, Antártida/Empúries, Barcelona, 1995.

[4] M. T. Cabré, "Una nueva teoría de la terminología: de la denominación a la comunicación", *La terminología. Representación y comunicación*, Universitat Pompeu Fabra, Barcelona, 1999.

[5] M. T. Cabré, R. Estopa, J. Vivaldi, "Automatic term detection. A review of current systems en Natural Language Processing", *Recent Advances in Computational Terminology*, John Benjamins Publishing Company, 2001, pp. 53-87.

[6] A.M. Cardero, "Terminología y lexicografía", *Primeras jornadas de la Red Iberoamericana de Terminología*, Bogotá, Colombia, 1999

[7] P. Drouin, “Term Extraction Using Non-Technical Corpora as a Point of Leverage”, *Terminology*, vol. 9, no 1, 2003, pp. 99-117.
[8] R. Dubuc, *Manuel pratique de terminologie*, Linguatech éditeur, Canadá, 1992.
[9] K. Frantzi, S. Ananiadou y H. Mima, “Automatic recognition of multi-word terms: the C-value/NC-value method”, *Int. J. on Digital Libraries* 3(2), 2000, pp. 115-130.
[10] A. Medina, G. Sierra, G. Garduño, C. Méndez y R. Saldaña, “CLI: An Open Linguistic Corpus for Engineering”. En *IX Congreso Iberoamericano de Inteligencia Artificial (IBERAMIA),* Actas de congreso. Tonantzintla, Puebla, 23 de noviembre. 2004
[11] J.C. Sager, *A Practical Course in Terminology Processing*. John Benjamins, 1990.