Un enfoque gramatical para la extracción de términos índice*

Jesús Vilares Ferro y **Miguel A. Alonso Pardo**
Departamento de Computación, Universidade da Coruña
Campus de Elviña s/n, 15071 - A Coruña
{jvilares,alonso}@udc.es

Resumen: La extracción de los términos que caracterizan un documento es una tarea de vital importancia en el desarrollo de sistemas de Recuperación de Información. En este artículo proponemos la utilización de análisis sintáctico superficial, implementado mediante cascadas de traductores finitos, para la extracción de términos índice complejos en base a una gramática aproximada del español que si bien es incompleta permite obtener adecuadamente las palabras involucradas en las dependencias sintácticas más importantes. La efectividad de los términos extraídos ha sido evaluada en la colección CLEF de textos en español.
Palabras clave: Análisis sintáctico superficial, Recuperación de Información

Abstract: The extraction of the keywords that characterize a document in a given collection is one of the most important components of an Information Retrieval system. In this article, we propose to apply shallow parsing, implemented by means of cascades of finite-state transducers, to extract complex index terms based on an approximated grammar of Spanish. The coverage of the grammar is small but it allows us to extract the words involved in the most relevant syntactic dependencies. The effectiveness of the extracted index terms has been evaluated in the CLEF collection of Spanish texts.
Keywords: Shallow parsing, Information Retrieval

1. Introducción

En un sistema de Recuperación de Información (IR), la correcta extracción de términos índice, tanto en los documentos como en las consultas, es la base para lograr un buen rendimiento. De poco sirve disponer de un mecanismo de emparejamiento y ordenación muy efectivo si no se han extraído y ponderado adecuadamente los términos que capturan el contenido semántico del texto.

En este contexto, la utilización de técnicas de Procesamiento del Lenguaje Natural (NLP) resultaría factible, en particular cuando se trata de procesar documentos escritos en lenguas con unas estructuras morfológicas y sintácticas más complejas que las presentes en inglés, como es el caso del español. A este respecto, se ha estudiado la aplicación de técnicas de NLP que operan a nivel de palabra con el fin de reducir la *variación morfológica* debida a la flexión y derivación (Figuerola et al., 2001; Vilares et al., 2001; Vilares et al., 2002a) así como la *variación léxica* fruto de la sinonimia (Fernández et al., 2002). La utilización de tales técnicas no conlleva un aumento significativo del coste computacional con respecto a técnicas clásicas como el *stemming*, ya que pueden ser implementadas mediante autómatas o traductores de estado finito.

Una vez establecida la viabilidad de las técnicas de NLP a nivel de palabra, el siguiente paso consiste en aplicar técnicas de análisis a nivel de frase con el fin de, por una parte, hacer frente a la *variación sintáctica*, y por otra, obtener términos índice más precisos. Este es el ámbito en que se enmarca el trabajo descrito en este artículo.

Llegados a este punto, nos enfrentamos al problema que supone el alto coste computacional de los analizadores sintácticos. A fin

* Parcialmente financiado por el Ministerio de Ciencia y Tecnología (TIC2000-0370-C02-01, HP2001-0044 y HF2002-81), becas FPU de la Secretaría de Estado de Educación y Universidades, Xunta de Galicia (PGIDT01PXI10506PN, PGIDIT02PXIB30501PR y PGIDIT02SIN01E) y Universidade da Coruña.

de mantener una complejidad lineal en relación con el tamaño del texto a analizar, nos hemos alejado de propuestas que propugnan la realización de un análisis sintáctico completo (Perez-Carballo y Strzalkowski, 2000), optando por aplicar técnicas de *análisis sintáctico superficial*, buscando, además, una mayor robustez. Durante el proceso de análisis, procederemos a extraer, en forma de pares, las dependencias sintácticas presentes en el texto, para ser utilizadas como términos índice.

En base a la teoría de lenguajes formales sabemos que, dada una gramática independiente del contexto y una cadena de entrada, los subárboles sintácticos de altura k generados por un analizador sintáctico pueden ser recreados mediante k capas de traductores finitos: la primera capa obtiene los nodos etiquetados por no terminales correspondiente a la parte izquierda de producciones que sólo contienen terminales en su parte derecha; la segunda capa obtiene aquellos nodos que involucran símbolos terminales y aquellos símbolos no-terminales generados en la capa anterior; y así sucesivamente. Evidentemente, la acotación en la altura de los árboles limita el tipo de construcciones sintácticas que se puede reconocer. Sin embargo, este tipo de análisis sintáctico superficial (Abney, 1997) ha mostrado ser de utilidad en diversos ámbitos de aplicación del NLP, particularmente en el de la Extracción de Información. Su aplicación en IR, no tan estudiada, ha sido ensayada por Xerox (Hull et al., 1997) para el caso del inglés, demostrando su superioridad respecto a aproximaciones clásicas basadas en pares de palabras contiguas.

2. *Análisis sintáctico superficial*

En nuestro sistema hemos utilizado una arquitectura basada en cuatro capas más una capa de preprocesamiento que denominaremos capa 0. Cada una de ellas ha sido implementada mediante traductores finitos, lo cual nos permite mantener una complejidad lineal respecto al tamaño del texto de entrada.

Describiremos a continuación el funcionamiento de cada una de dichas capas. Para ello utilizaremos como notación reglas independientes del contexto, extendidas mediante los operadores clásicos utilizados en la definición de expresiones regulares: ? expresará opcionalidad, * indicará repetición con opcionalidad (0 o más veces), y | separará alternativas. Por otra parte, los identificadores con mayúscula referenciarán conjuntos de términos, tanto preterminales (las etiquetas resultantes del proceso de etiquetación léxica) como elementos de una categoría gramatical dada. Cuando se requiera la presencia de un lema en concreto, éste se indicará empleando la fuente `typewriter`.

2.1. Capa 0

La capa 0 toma como entrada la salida de un etiquetador-lematizador y realiza las siguientes transformaciones de la entrada con el fin de eliminar el ruido generado por ciertas construcciones en las etapas posteriores, en las que se realizará el análisis sintáctico propiamente dicho.

Tratamiento de cifras en formato no numérico. La tarea de identificar aquellas secuencias de palabras que se corresponden con cifras escritas parcial o totalmente en formato no numérico es más compleja de lo que aparenta. Esto se debe a las ambigüedades ligadas a la aparición de la conjunción coordinada *y*, como en *vendí cuarenta y cinco quedaron sin vender*, donde sería preciso un análisis sintáctico completo para determinar que la conjunción no forma parte de la cifra *cuarenta y cinco*. Afortunadamente, la extracción de cifras no es tan importante en los sistemas de IR como en los de Extracción de Información, ya que las consultas expresadas en lenguaje natural raramente involucran términos numéricos, bastando con aplicar heurísticas que resuelvan los casos más frecuentes.

Tratamiento de expresiones de cantidad. Las expresiones del tipo *algo más de dos millones* o *unas dos docenas*, que si bien se refieren a una cifra, establecen una cierta vaguedad en relación al valor de la misma, son tratadas mediante reglas que definen el conjunto de modificadores PMC (`casi`, `cerca de`, `poco más`, `algo menos`, etc.) y *numerales colectivos* NC, sustantivos que representan como unidad un número determinado de elementos (`decena`, `docena`, `centenar`, `millar`, etc.). La siguiente regla permite definir expresiones de cantidad del tipo *poco más de cinco millones*, formadas por un modificador opcional, una cifra, un numeral colectivo y una preposición *de* que sólo será necesario si la expresión tiene función de determinante, pues si funciona como pronombre no acompañará a ningún sustantivo:

$$SNum \rightarrow PMC?\ Cifra\ NC\ \texttt{de}?$$

La siguiente regla nos permite identificar expresiones que denotan una cantidad apro-

ximada de numerales colectivos, como por ejemplo *cientos de miles de*:

$$SNum \rightarrow PMC?\ (\texttt{medio} \mid \texttt{un})?\ \ (NC\ \texttt{de})?$$
$$NC\ (\texttt{y medio})?\ \texttt{de}?$$

La última regla nos permite identificar aquellas expresiones que no contienen numerales colectivos, como por ejemplo *poco más de trecientos treinta*:

$$SNum \ \rightarrow \ PMC\ \ Cifra$$

Simplificación de expresiones verbales. Ciertas expresiones verbales deben considerarse como una unidad para así simplificar el trabajo de las capas superiores. De este modo, la expresión *tener en cuenta*, por ejemplo, debe tomarse como una unidad, sinónima del verbo *considerar*, para evitar que capas posteriores identifiquen *en cuenta* como complemento del verbo, lo que a su vez podría impedir la correcta identificación de otros complementos verbales de interés. El criterio seguido a la hora de identificar estas expresiones es el de considerar sólo aquellas secuencias Verbo-Preposición-Sustantivo especialmente frecuentes, que sean fácilmente sustituibles por un sinónimo verbal o perifrástico y en las que no se pierda información al realizar dicha sustitución.

2.2. Capa 1

La primera capa consta de reglas que contienen únicamente etiquetas y/o lemas en su parte derecha. Con el fin de que las capas siguientes puedan extraer pares de dependencias, nos interesa poder asociar al no terminal del lado izquierdo de cada regla el lema correspondiente al núcleo del sintagma que se esté reconociendo, así como una etiqueta con los rasgos morfosintácticos pertinentes. La notación que utilizaremos para este mecanismo de herencia está inspirada en la empleada en la especificación del conjunto de restricciones en las gramáticas basadas en estructuras de rasgos (Carpenter, 1992), como quedará ilustrado en estas primeras reglas.

La primera regla del analizador nos permite identificar secuencias de adverbios (W), que denominaremos *sintagmas adverbiales* ($SAdv$). Aunque no van a participar en la formación de los pares de dependencia, es necesario identificar este tipo de sintagmas para que las etapas posteriores trabajen correctamente. Consideraremos que el último adverbio constituye el núcleo del sintagma, por lo que su lema y su etiqueta constituirán el lema y la etiqueta del no terminal $SAdv$:

$$SAdv \ \rightarrow \ W^* \ \ W_1 \ \left\{ \begin{array}{l} SAdv.lem \doteq W_1.lem \\ SAdv.etiq \doteq W_1.etiq \end{array} \right.$$

Existen expresiones tales como *de forma rápida*, con un adjetivo (A) como núcleo, que sin embargo equivalen a un adverbio, en este caso *rápidamente*. Estas construcciones son procesadas por la siguiente regla:

$$SAdv \ \rightarrow \ \texttt{de}\ \ (\texttt{forma} \mid \texttt{manera} \mid \texttt{modo})\ \ A \quad \left\{ \begin{array}{l} SAdv.lem \doteq A.lem \\ SAdv.etiq \doteq A.etiq \end{array} \right.$$

El siguiente conjunto de reglas nos permite identificar *grupos verbales de primer nivel* ($GV1$) correspondientes a formas pasivas, tanto simples como compuestas. La primera regla trata las formas compuestas: la etiqueta se toma del verbo (V) auxiliar `haber`, mientras que el lema se toma del verbo principal, que debe ser un participio, al igual que sucede con el verbo auxiliar `ser`. La segunda regla trata las formas simples: la etiqueta se obtiene de la forma del verbo auxiliar `ser`, mientras que el lema se toma del verbo principal, de nuevo un participio.

$$GV1 \ \rightarrow \ V_1 \ \ V_2 \ \ V_3 \ \left\{ \begin{array}{l} GV1.lem \doteq V_3.lem \\ GV1.etiq \doteq V_1.etiq \\ GV1.voz \doteq \text{PAS} \\ V_1.lem \doteq \texttt{haber} \\ V_2.lem \doteq \texttt{ser} \\ V_2.tiempo \doteq \text{PART} \\ V_3.tiempo \doteq \text{PART} \end{array} \right.$$

$$GV1 \ \rightarrow \ V_1 \ \ V_2 \ \left\{ \begin{array}{l} GV1.lem \doteq V_2.lem \\ GV1.etiq \doteq V_1.etiq \\ GV1.voz \doteq \text{PAS} \\ V_1.lem \doteq \texttt{ser} \\ V_2.tiempo \doteq \text{PART} \end{array} \right.$$

Las formas activas compuestas y simples son identificadas, respectivamente, mediante las siguientes reglas:

$$GV1 \ \rightarrow \ V_1 \ \ V_2 \ \left\{ \begin{array}{l} GV1.lem \doteq V_2.lem \\ GV1.etiq \doteq V_1.etiq \\ GV1.voz \doteq \text{ACT} \\ V_1.lem \doteq \texttt{haber} \\ V_2.tiempo \doteq \text{PART} \end{array} \right.$$

$$GV1 \ \rightarrow \ V_1 \ \left\{ \begin{array}{l} GV1.lem \doteq V_1.lem \\ GV1.etiq \doteq V_1.etiq \\ GV1.voz \doteq \text{ACT} \end{array} \right.$$

Es interesante observar que, para garantizar que los tiempos verbales sean identificados correctamente, en caso de que puedan aplicarse varias reglas, tendrá preferencia la que presente una parte derecha más larga.

2.3. Capa 2

En esta capa se identifican los sintagmas adjetivales y grupos verbales perifrásticos. Un *sintagma adjetival* ($SAdj$) será aquél cuyo núcleo es un adjetivo (A), el cual puede ir precedido por un sintagma adverbial:

$$SAdj \rightarrow SAdv?\ A \begin{cases} SAdj.lem \doteq A.lem \\ SAdj.etiq \doteq A.etiq \end{cases}$$

También se identifican los *grupos verbales de segundo nivel* ($GV2$), incluyendo los grupos verbales perifrásticos. Las *perífrasis verbales* son uniones de dos o más formas verbales que funcionan como una unidad, dotando a la semántica del verbo principal de matices de significado tales como obligación, grado de desarrollo de la acción, etc., que no pueden ser expresados mediante las formas simples o compuestas del verbo. En cuanto a su estructura, las perífrasis están generalmente formadas por un verbo auxiliar conjugado que aporta la flexión, un verbo en forma no personal (infinitivo, gerundio o participio) que aporta el significado principal, y un elemento opcional (preposición o conjunción) de enlace entre ambos.

Las perífrasis de infinitivo se identifican mediante la siguiente regla, que contempla la posibilidad de que los verbos vayan seguidos de un pronombre enclítico (separado previamente de la forma verbal por el etiquetador). Por cuestiones prácticas, sólo se han considerado los nexos más comunes. La etiqueta se hereda del verbo auxiliar, mientras que el lema y la voz se heredan del verbo principal:

$$GV2 \rightarrow GV1_1\ (\texttt{me}|\texttt{te}|\texttt{se})?\ (\texttt{que}|\texttt{de}|\texttt{a})?\ GV1_2 \begin{cases} GV2.lem \doteq GV1_2.lem \\ GV2.etiq \doteq GV1_1.etiq \\ GV2.voz \doteq GV1_2.voz \\ GV1_1.voz \doteq \text{ACT} \\ GV1_2.tiempo \doteq \text{INF} \end{cases}$$

Las perífrasis de gerundio se tratan de modo similar, salvo que en este caso no se permite la presencia de nexos:

$$GV2 \rightarrow GV1_1\ (\texttt{me} \mid \texttt{te} \mid \texttt{se})?\ GV1_2 \begin{cases} GV2.lem \doteq GV1_2.lem \\ GV2.etiq \doteq GV1_1.etiq \\ GV2.voz \doteq GV1_2.voz \\ GV1_1.voz \doteq \text{ACT} \\ GV1_2.tiempo \doteq \text{GER} \end{cases}$$

De forma análoga se procede con las perífrasis de participio:

$$GV2 \rightarrow GV1_1\ (\texttt{me} \mid \texttt{te} \mid \texttt{se})?\ GV1_2 \begin{cases} GV2.lem \doteq GV1_2.lem \\ GV2.etiq \doteq GV1_1.etiq \\ GV2.voz \doteq GV1_2.voz \\ GV1_1.voz \doteq \text{ACT} \\ GV1_2.tiempo \doteq \text{PART} \end{cases}$$

Por último, los grupos verbales de primer nivel que no forman parte de grupos perifrásticos son promocionados a grupos verbales de segundo nivel:

$$GV2 \rightarrow GV1 \begin{cases} GV2.lem \doteq GV1.lem \\ GV2.etiq \doteq GV1.etiq \\ GV2.voz \doteq GV1.voz \end{cases}$$

2.4. Capa 3

En esta capa son procesados los *sintagmas nominales* (SN). En la definición de las reglas para su identificación y procesado, se ha contemplado la posibilidad de que vengan precedidos de un *complemento partitivo* CP (`alguno de`, `ninguno de`, etc.). Por cuestiones prácticas sólo hemos contemplado las fórmulas más frecuentes.

Tras el núcleo del sintagma nominal puede aparecer un modificador en forma de dos sintagmas adjetivales unidos por una conjunción coordinada (Cc), o bien una secuencia de uno, dos y hasta tres sintagmas adjetivales.

$$\begin{aligned} PostModifSAdj &\rightarrow SAdj\ Cc\ SAdj \\ PostModifSAdj &\rightarrow SAdj \\ PostModifSAdj &\rightarrow SAdj\ SAdj \\ PostModifSAdj &\rightarrow SAdj\ SAdj\ SAdj \end{aligned}$$

El núcleo del sintagma nominal estará formado por un nombre común (representado por N), una sigla o un nombre propio; su etiqueta y lema determinarán la etiqueta y lema del sintagma completo. En el caso de la aparición sucesiva de varios candidatos a núcleo, consideraremos que el último es el que realiza esta función. Por otra parte, la etiqueta establecida para el sintagma puede verse modificada en presencia de un complemento partitivo, ya que en este caso, y de cara a establecer concordancias con otros sintagmas, el número del sintagma nominal se corresponderá con el aportado por dicho partitivo. Por ejemplo, se debe decir *"Cualquiera de ellos lo sabe"*, y no **"Cualquiera de ellos lo saben"*.

Opcionalmente, antes del núcleo pueden aparecer uno o más determinantes (D) y un sintagma adjetival. La aparición de postmodificadores adjetivales es también opcional,

dando lugar, finalmente, a la regla:

$$\begin{array}{rcl} SN & \rightarrow & CP? \\ & & D^* \ \ (SAdj \mid Cifra \mid SNumeral)? \\ & & (N \mid Sigla \mid Propio)^* \\ & & (N \mid Sigla \mid Propio)_1 \\ & & PostModifSAdj? \\ & & \left\{ \begin{array}{l} SN.lem \doteq ()_1.lem \\ SN.etiq \doteq ()_1.etiq \\ SN.num \doteq CP.num \end{array} \right. \end{array}$$

2.5. Capa 4

Por último, la capa 4 se encarga de la identificación de los *sintagmas preposicionales* (SP, $SPde$, $SPpor$), aquéllos formados por un sintagma nominal (SN) precedido de una preposición (P). Para facilitar la extracción de dependencias, nos interesará distinguir del resto los sintagmas que comienzan por las preposiciones *de* y *por*, dando lugar a las siguientes reglas:

$$SPde \rightarrow P \ \ SN \left\{ \begin{array}{l} P.lem \doteq \texttt{de} \\ SP.lem \doteq SN.lem \\ SP.etiq \doteq SN.etiq \end{array} \right.$$

$$SPpor \rightarrow P \ \ SN \left\{ \begin{array}{l} P.lem \doteq \texttt{por} \\ SP.lem \doteq SN.lem \\ SP.etiq \doteq SN.etiq \end{array} \right.$$

$$SP \rightarrow P \ \ SN \left\{ \begin{array}{l} SP.lem \doteq SN.lem \\ SP.etiq \doteq SN.etiq \end{array} \right.$$

3. Extracción de dependencias

El objetivo final del análisis sintáctico es la extracción de pares de palabras ligadas por relaciones de dependencia sintáctica. Dicho proceso se desarrolla en dos fases: una primera fase de *identificación de funciones sintácticas* de los sintagmas identificados durante el proceso de análisis, y una segunda fase de *extracción de dependencias* propiamente dicha.

En la etapa de *identificación de funciones sintácticas*, y debido a la superficialidad del análisis realizado, nos tenemos que enfrentar a diversas limitaciones, entre las que destaca el problema de establecer los límites de cada oración. Tomando como hipótesis de trabajo que para cada núcleo verbal existe una oración asociada, y en base a consideraciones prácticas, damos por terminada una oración cuando se alcanza uno de los siguientes elementos: *signos de puntuación*, sin limitarnos sólo al punto de cierre de las oraciones, pues tampoco podremos retomar con garantías el análisis de una frase cuando éste es interrumpido por un inciso entre comas, o por una cita entre comillas; *relativos*, al separar una cláusula subordinada de la oración principal; *conjunciones*, debido a las ambigüedades sintácticas que introducen respecto a cuáles son las partes de la oración que conectan; y *los grupos verbales de segundo nivel ($GV2$) cuyo núcleo es una forma personal* cuando no existe ningún otro límite de oración entre dicho grupo verbal y el anterior, ya que por norma general la aparición de un verbo implica que estamos ante una nueva oración.

Esta limitación del ámbito en el cual trabaja el extractor de dependencias no representa un gran obstáculo en el contexto para el cual se ha diseñado, la extracción de términos índice para un sistema de IR, ya que lo que se persigue no es tanto la exhaustividad como la fiabilidad de las dependencias obtenidas.

En última instancia, lo que se persigue es identificar oraciones que sigan alguno de los siguientes criterios de formación:

- Sujeto activo + grupo verbal predicativo activo + complemento directo.
- Sujeto activo + grupo verbal copulativo + atributo.
- Sujeto pasivo + grupo verbal predicativo pasivo + complemento agente.

Evidentemente, estas estructuras ideales raramente aparecen en estado puro, ya que lo habitual es encontrarse con sintagmas diversos entre el sujeto y el grupo verbal y entre el grupo verbal y sus complementos.

Las funciones sintácticas identificadas, junto con los criterios empleados para ello, son las siguientes:

Complemento nominal. Debido a la ambigüedad en la adjunción de sintagmas preposicionales en cuanto a si nos encontramos realmente ante un complemento nominal o bien ante un complemento verbal, sólo hemos considerado el caso de los sintagmas preposicionales introducidos por *de*, los $SPde$, por ser altamente fiables. En consecuencia, cuando nos encontremos con un $SPde$ que siga inmediatamente a un sintagma nominal o preposicional, será etiquetado como complemento nominal.

Sujeto. El sintagma nominal (SN) más próximo que antecede a un grupo verbal ($GV2$) se toma como su sujeto. Adicionalmente, consideraremos que carecen de sujeto aquellos grupos verbales cuyo núcleo es una

forma no personal (infinitivo, gerundio o participio).

Atributo. En presencia de un verbo copulativo, identificaremos como atributo aquel $SAdj$ no ligado, núcleo del SN o $SPde$ más próximo que sigue al grupo verbal.

Objeto directo. El SN más próximo que aparece después de un $GV2$ predicativo activo se considera su complemento directo.

Agente. El $SPpor$ más próximo que sigue a un $GV2$ predicativo pasivo es considerado su complemento agente.

Complemento circunstancial. Debido al citado problema de adjunción de sintagmas preposicionales,hemos optado por un criterio conservador a la hora de identificar los complementos circunstanciales del verbo. Nuestro objetivo es minimizar el ruido introducido por complementos incorrectamente identificados. Por ello, consideraremos como complemento circunstancial sólo a aquel sintagma preposicional posterior al verbo, más próximo a él, y anterior a todo complemento verbal o atributo previamente identificado.

Una vez identificadas las funciones sintácticas de los sintagmas, la siguiente fase consiste en la *extracción de las dependencias sintácticas* existentes entre estos. Para ello, se crean los pares formados por:

- Un sustantivo y cada uno de los adjetivos que lo modifican. Mientras que el resto de las dependencias son extraídas tras finalizar el proceso de análisis, estas dependencias se extraen en la fase de identificación de sintagmas nominales de la capa 3. Esto se debe a que son dependencias internas al sintagma nominal, y, de no extraerlas entonces, dicha información se perdería una vez el sintagma quedase reducido a su núcleo.
- Un sustantivo y el núcleo de su complemento nominal.
- El núcleo del sujeto y el verbo predicativo.
- El núcleo del sujeto y el del atributo. Los verbos atributivos se consideran meros elementos copulativos, por lo que la dependencia se establece directamente entre el sujeto y el atributo.
- Un verbo activo y el núcleo de su objeto directo.
- Un verbo pasivo y el núcleo de su complemento agente.
- Un verbo predicativo y el núcleo de su complemento circunstancial.
- El núcleo del sujeto y el del complemento circunstancial de su verbo, únicamente en caso de que el verbo sea atributivo, dado su especial comportamiento.

Una vez extraídos y normalizados, los pares de dependencias constituirán los términos índice. En nuestro caso se ha aplicado un esquema de normalización basado en la utilización de relaciones morfológicas para mejorar el tratamiento de la variación sintáctica. El objetivo es dar cobertura a la aparición de variantes morfosintácticas del término original (Vilares et al., 2002b).

4. Evaluación

Nuestra aproximación ha sido probada sobre el corpus monolíngüe para español correspondiente a las ediciones 2001 y 2002 del CLEF (Peters, 2002), compuesto por noticias de la Agencia EFE pertenecientes a 1994, totalizando 215.738 documentos SGML. Las 100 consultas empleadas, de la 41 a la 140, constan de tres campos: un breve *título*, una somera *descripción* en una frase del tema, y una *narrativa* de mayor complejidad especificando los criterios de relevancia. En nuestros experimentos se han empleado los tres campos, dándole doble relevancia al *título* respecto a los otros por ser, en última instancia, el campo que resume la semántica básica de la consulta. Los documentos han sido indexados por medio del motor de indexación vectorial SMART (Buckley, 1985), empleando un esquema de pesos `atn-ntc`.

Experimentos previos (Vilares et al., 2002a) apuntan a la lematización como el mejor punto de inicio para el desarrollo de métodos de normalización basados en NLP que hagan frente a niveles de variación lingüística más complejos. Por ello tomaremos como punto de referencia la indexación de términos simples lematizados.

La tabla 1 recoge los resultados obtenidos. La primera columna presenta los resultados para la lematización de términos simples (lem), nuestro referente. Las siguientes columnas, dsx, recogen los resultados obtenidos conjugando términos simples lematizados y términos complejos basados en dependencias sintácticas (ds), conforme la ponderación de pesos entre términos simples y complejos, x a 1, evoluciona. La penúltima columna, opt,

	lem	***ds1***	***ds2***	***ds3***	***ds4***	***ds5***	***ds6***	***ds7***	***ds8***	*opt*	Δ
Documentos devueltos	99k	99k	99k	99k	99k	99k	99k	99k	99k	--	--
Relevantes devueltos	5220	5214	5250	**5252**	**5252**	5248	5249	5244	5242	5252	32
R-precision	.5131	.4806	.5041	.5137	.5175	.5174	.5200	**.5203**	.5197	.5203	.0072
Precisión no interpolada	.5380	.5085	.5368	.5440	.5461	.5462	.5464	**.5472**	.5463	.5472	.0092
Precisión por documento	.5924	.5489	.5860	.5974	.6013	.6025	**.6028**	.6026	.6020	.6028	.0104
Precisión a 5 docs.	.6747	.6525	**.6909**	.6869	.6848	.6788	.6808	.6828	.6808	.6909	.0162
Precisión a 10 docs.	.6010	.5859	.6091	.6192	**.6202**	.6192	.6192	.6172	.6152	.6202	.0192
Precisión a 15 docs.	.5623	.5441	.5690	.5737	.5778	**.5791**	**.5791**	.5764	.5758	.5791	.0168
Precisión a 20 docs.	.5374	.5040	.5298	.5328	.5354	.5343	.5384	**.5394**	.5384	.5394	.0020
Precisión a 30 docs.	.4825	.4549	.4778	.4852	.4892	.4886	.4882	**.4896**	**.4896**	.4896	.0071
Precisión a 100 docs.	.3067	.2873	.3017	.3070	.3084	**.3095**	.3087	.3089	.3083	.3095	.0028
Precisión a 200 docs.	.2051	.1959	.2033	.2057	.2062	.2063	**.2067**	**.2067**	.2065	.2067	.0016
Precisión a 500 docs.	.0997	.0980	.0997	.1001	.1004	**.1005**	**.1005**	**.1005**	**.1005**	.1005	.0008
Precisión a 1000 docs.	.0527	.0527	.0530	**.0531**	**.0531**	.0530	.0530	.0530	.0529	.0531	.0004

Tabla 1: Experimentos con el corpus CLEF

está integrada por los mejores resultados obtenidos para *ds* en cada parámetro considerado, los cuales están señalados en negrita. Finalmente la columna Δ presenta la mejora de *opt* respecto a *lem*. El rendimiento del sistema para cada uno de estos métodos se ha medido en base a los parámetros recogidos en cada fila: número de documentos devueltos, número de documentos relevantes devueltos (5548 esperados), *R-precision*, precisión media no interpolada para todos los documentos relevantes, precisión media por documento para todos los documentos relevantes, y precisión a los N documentos devueltos.

Como puede apreciarse en la columna *ds1*, el empleo directo de dependencias sintácticas como términos índice ha producido una disminución general del rendimiento del sistema a todos los niveles. Tras estudiar el comportamiento del sistema para las diferentes consultas, se llegó a la conclusión de que el problema residía en una sobreponderación de los pesos de los términos complejos, mucho menos frecuentes que los términos simples y, por tanto, con un peso asignado mucho mayor. Esto se ha traducido en una creciente inestabilidad del sistema, en tanto que al producirse correspondencias no deseadas de términos complejos con documentos no relevantes, su puntuación asignada aumenta considerablemente, disparando su nivel de relevancia. Por otra parte, y debido a la misma causa, cuando se producen correspondencias de términos complejos con documentos sí relevantes, se produce una clara mejora de los resultados respecto al empleo de términos simples. De acuerdo con esto podría argumentarse que deberían esperarse, cuando menos, unos resultados similares a los obtenidos para los términos simples. Sin embargo las correspondencias de términos complejos son mucho menos frecuentes que las de términos simples, por lo que las correspondencias casuales, de producirse, son muchísimo más perjudiciales que para el caso de los términos simples, cuyo impacto tiende a diluirse debido al efecto de las restantes correspondencias.

Debemos, por tanto, corregir esa sobrevaloración de los términos índice complejos, y así minimizar el efecto negativo de las correspondencias no deseadas. Para ello se corrigió el factor de ponderación entre los pesos de los términos índice simples y complejos, disminuyendo el alto grado de relevancia otorgado inicialmente a los términos complejos, tal y como se muestra en las restantes columnas *dsx*. Como puede apreciarse, la mejora de los resultados es inmediata, sobre todo a nivel de precisión hasta los primeros 15 documentos devueltos, y en cuanto a número de documentos relevantes devueltos (pasamos de 5220 en *lem* y 5214 en *ds1* a 5250 en *ds2*).

Como suele ocurrir en IR, no puede hablarse de un método mejor en términos absolutos. Desde el punto de vista de la ordenación, *ds4*, en el que se cuadruplica el peso de los términos simples, obtiene los mejores resultados, además de la mejor cobertura (5252 documentos relevantes devueltos). Sin embargo, los mejores resultados en cuanto a medidas de rendimiento globales[1], los obtiene

[1] *R-precision*, precisión media no interpolada para

ds7, con un factor de ponderación mayor.

5. Conclusiones y trabajo futuro

A lo largo de este artículo hemos planteado la utilización de dependencias sintácticas como términos índice complejos, con el objetivo de tratar los problemas derivados de la variación lingüística de origen sintáctico y morfosintáctico y, de este modo, obtener términos más precisos. Para extraer dichas dependencias hemos desarrollado un analizador sintáctico superficial del español basado en cascadas de traductores finitos, lo que nos permite abordar el procesamiento de grandes colecciones de forma robusta a la vez que ágil, pues la complejidad es lineal respecto al tamaño del texto de entrada. Los resultados obtenidos nos permiten ser optimistas respecto a nuestro planteamiento, residiendo nuestro problema en determinar cómo incorporar la información sintáctica extraída por el analizador.

Con respecto al trabajo futuro, esperamos que la aplicación de *restricciones de selección* generadas automáticamente (Gamallo et al., 2001) sobre los propios textos mejore la capacidad de desambiguación sintáctica del sistema, especialmente en el caso de la adjunción de sintagmas preposicionales También estamos estudiando la posibilidad de almacenar términos simples y complejos en índices separados, combinándolos posteriormente mediante técnicas de *fusión de datos* (Vogt y Cottrell, 1999).

Bibliografía

Abney, Steven. 1997. Partial parsing via finite-state cascades. *Natural Language Engineering*, 2(4):337–344.

Buckley, Chris. 1985. Implementation of the SMART information retrieval system. Informe técnico, Department of Computer Science, Cornell University.

Carpenter, Bob. 1992. *The Logic of Typed Feature Structures*. Cambridge University Press, Cambridge/Nueva York/Melbourne.

Fernández, Santiago, Jorge Graña, y Alejandro Sobrino. 2002. A Spanish e-dictionary of synonyms as a fuzzy tool for information retrieval. En *Actas del XI Congreso Español sobre Tecnologías y Lógica Fuzzy (ESTYLF-2002)*, páginas 31–37, León, España.

Figuerola, Carlos G., Raquel Gómez, Angel F. Zazo, y José L. Alonso. 2001. Stemming in Spanish: A first approach to its impact on information retrieval. En C. Peters, editor, *Working notes for the CLEF 2001 workshop*, Darmstadt, Alemania.

Gamallo, Pablo, Alexandre Agustini, y Gabriel P. Lopes. 2001. Selections restrictions acquisition from corpora. En volumen 2258 de *Lecture Notes in Computer Science*. Springer-Verlag, Berlín-Heidelberg-Nueva York, páginas 30–43.

Hull, David A., Gregory Grefenstette, B. Maximilian Schulze, Eric Gaussier, Hinrich Schutze, y Jan O. Pedersen. 1997. Xerox TREC-5 site report: routing, filtering, NLP, and Spanish tracks. En *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*, páginas 167–180.

Perez-Carballo, Jose y Tomek Strzalkowski. 2000. Natural language information retrieval: progress report. *Information Processing and Management*, 36(1):155–178.

Peters, Carol, editor. 2002. *Results of the CLEF 2002 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2002 Workshop*, Rome, Italia.

Vilares, Jesús, Miguel A. Alonso, Francisco J. Ribadas, y Manuel Vilares. 2002a. COLE experiments at CLEF 2002 Spanish monolingual track. En (Peters, 2002), páginas 153–160.

Vilares, Jesús, Fco. Mario Barcala, y Miguel A. Alonso. 2002b. Using syntactic dependency-pairs conflation to improve retrieval performance in Spanish. En volumen 2276 de *Lecture Notes in Computer Science*. Springer-Verlag, Berlín-Heidelberg-Nueva York, páginas 381–390.

Vilares, Jesús, David Cabrero, y Miguel A. Alonso. 2001. Applying productive derivational morphology to term indexing of Spanish texts. En volumen 2004 de *Lecture Notes in Computer Science*. Springer-Verlag, Berlín-Heidelberg-Nueva York, páginas 336–348.

Vogt, C. C. y Garrison W. C. 1999. Fusion via a linear combination of scores. *Information Retrieval*, 1(3):151–173.

todos los documentos relevantes y precisión media por documento para todos los documentos relevantes.