Extracción de términos índice mediante cascadas de expresiones regulares*

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

El rendimiento de los sistemas de Recuperación de Información se ve limitado por los fenómenos de variación lingüística presentes en los textos. Las técnicas de Procesamiento de Lenguaje Natural a nivel de palabra han mostrado su utilidad para reducir dicha variación. Proponemos en este artículo extender esta aproximación a la variación a nivel de frase; para ello se indexarán las dependencias sintácticas presentes en los documentos, las cuales son obtenidas por medio de un analizador sintáctico. Para reducir en lo posible el coste computacional asociado al proceso de análisis, hemos optado por emplear un analizador sintáctico superficial basado en cascadas de traductores de estado finito. Si bien este artículo se centra en el caso del español, nuestra aproximación es extensible a otros lenguajes adaptando convenientemente la gramática empleada por el analizador.

Palabras clave: Análisis sintáctico superficial, traductores de estado finito.

1. INTRODUCCIÓN

En trabajos anteriores [18, 19, 5] se ha mostrado la viabilidad de la aplicación de técnicas de Procesamiento del Lenguaje Natural (NLP) para el tratamiento de la *variación lingüística* [3] a nivel de palabra en Recuperación de Información (IR) sobre textos en español. Por otra parte, su utilización no produce un aumento significativo del coste computacional con respecto a técnicas clásicas como el *stemming*, pues pueden ser implementadas mediante autómatas o traductores de estado finito [18].

El siguiente paso lógico consiste en aplicar técnicas de análisis a nivel de frase para, por una parte, manejar la *variación sintáctica* y, por otra, obtener términos índice más precisos. Sin embargo, llegados a este punto, debemos enfrentarnos al alto coste computacional asociado al empleo de analizadores sintácticos. A fin de mantener una complejidad lineal en relación con el tamaño del texto a analizar, nos hemos alejado de aproximaciones que proponen la realización de un análisis sintáctico completo [14], optando por aplicar técnicas de *análisis sintáctico superficial*, buscando, a la vez, una mayor robustez. Frente a otras propuestas que se limitan a tratar los sintagmas nominales [11], nuestro sistema cubre también sus variantes sintácticas y morfosintácticas [10], incluso aquellas que implican el análisis de sintagmas verbales. El objetivo de este proceso de análisis es el de extraer las dependencias sintácticas presentes en el texto, las cuales, una vez normalizadas, constituirán términos índice complejos.

Proponemos, pues, aproximar el análisis sintáctico de gramáticas independientes del contexto mediante cascadas de expresiones regulares. La teoría de lenguajes formales nos dice 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. Evidentemente, la acotación en la altura de los árboles limita el tipo de construcciones sintácticas reconocibles. Sin embargo, este tipo de análisis ya ha demostrado su utilidad en diversos ámbitos de aplicación del NLP, particularmente en Extracción de Información [2, 8]. Su aplicación en IR, no tan estudiada, ha sido ensayada por Xerox [9], para el inglés, en el TREC-5, demostrando su superioridad respecto a aproximaciones clásicas basadas en pares de palabras contiguas.

* 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 and PGIDIT02SIN01E) y Universidade da Coruña.

2. FUNCIONAMIENTO DEL SISTEMA

2.1. Análisis superficial

En nuestro sistema hemos utilizado una arquitectura basada en cinco capas, alimentadas por la salida de un preprocesador-etiquetador [7]. Perfilaremos a continuación las funciones desempeñadas por cada una de dichas capas en el proceso de análisis:

Capa 0: ampliación del preprocesado. Su función es la de tratar cierto tipo de construcciones lingüísticas para minimizar el ruido generado durante su análisis propiamente dicho. Dichas construcciones comprenden:

- *Cifras en formato no numérico.*
- *Expresiones de cantidad.* Construcciones tales como *algo más de dos millones* o *cientos de miles de*, que si bien se refieren a una cifra, establecen una cierta vaguedad en relación al valor de la misma, son identificadas como sintagmas numerales ($SNum$).
- *Expresiones con función verbal.* Ciertas construcciones verbales tales como *tener en cuenta*, deben considerarse como una unidad, en este caso sinónima de *considerar*, para así evitar que capas posteriores cometan errores como el de identificar *en cuenta* como un complemento verbal.

Capa 1: sintagmas adverbiales y grupos verbales de nivel 1. En esta capa son identificados, en primer lugar, los sintagmas adverbiales ($SAdv$), tanto aquéllos con núcleo adverbial, p.ej. *rápidamente*, como ciertas expresiones propiamente no adverbiales pero de función equivalente, p.ej. *de forma rápida*. Por otra parte, son también procesados los grupos verbales de nivel 1 o no perifrásticos ($GV1$), tanto en sus formas simples como compuestas, y tanto en sus formas activas como pasivas.

Capa 2: sintagmas adjetivales y grupos verbales de nivel 2. Los sintagmas adjetivales ($SAdj$) como *azul* o *muy alto* son tratados aquí, al igual que los grupos verbales de nivel 2 o perifrásticos ($GV2$) como *tengo que ir*. 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., los cuales no pueden ser expresados mediante las formas simples o compuestas del verbo.

Capa 3: sintagmas nominales. Para los sintagmas nominales (SN), además de estructuras sencillas como la adjunción de determinantes y adjetivos varios, se han considerado fenómenos más complejos como la existencia de *complementos partitivos* (CP), como *cualquiera de*, para así dar cobertura a estructuras nominales de mayor complejidad tales como *cualquiera de aquellos coches nuevos*.

Capa 4: sintagmas preposicionales. Formados por un sintagma nominal precedido por una preposición, se han considerado tres tipos diferentes en función de dicha preposición de cara a facilitar la posterior extracción de dependencias: los precedidos por la preposición *por* o $SPpor$, los precedidos por *de* o $SPde$, y los restantes, designados por SP.

2.2. Extracción de términos índice

El objetivo final del análisis sintáctico es la extracción de pares de palabras ligadas por relaciones de dependencia sintáctica. Estos *pares de dependencia* son extraídos una vez el análisis ha finalizado, siendo entonces identificadas las funciones sintácticas de los sintagmas reconocidos para, a continuación, identificar los pares considerados:

- Sustantivo-Adjetivo.
- Sustantivo-Complemento nominal. Debido a la ambigüedad propia de la adjunción de sintagmas preposicionales, sólo se considerarán los sintagmas preposicionales en *de*, los $SPde$.
- Sujeto-Verbo predicativo.
- Sujeto-Atributo.
- Verbo predicativo activo-Objeto directo.
- Verbo predicativo pasivo-Agente.
- Verbo predicativo-Complemento circunstancial.
- Sujeto-Complemento circunstancial (sólo con verbos copulativos).

A fin de obtener los términos índice, dichas dependencias son luego normalizadas mediante la aplicación de relaciones morfológicas derivativas, y así dar cobertura a las

variantes morfosintácticas del término original [10, 17] como, por ejemplo, en *cambio del clima* y *cambio climático*.

2.3. Aspectos de implementación

Cada una de las capas de nuestro analizador ha sido implementada mediante traductores de estado finito. A diferencia de otras aproximaciones similares [8], no perseguimos obtener a la salida una versión parentizada de la entrada, con los paréntesis identificando los sintagmas, sino una lista, en forma de pares, de las dependencias del texto. En dichos pares sólo intervienen los núcleos de los sintagmas, por lo que para cada sintagma identificado durante el análisis sólo nos interesa preservar el lema de su núcleo —y así eliminar su flexión— junto con sus rasgos morfosintácticos. Ésta es la razón por la cual los pares Sustantivo-Adjetivo suponen una excepción al proceso general de extracción, pues son ya extraídos en la fase de identificación de sintagmas nominales (capa 3), y no tras la fase de análisis. Esto se debe a que esa información desaparecerá al reducir a su núcleo el sintagma nominal.

3. UN EJEMPLO DETALLADO

Para ilustrar mejor el funcionamiento de nuestro sistema, mostraremos la evolución, paso a paso, de la frase:

Docenas de niños muy alegres han tenido que aprender hoy en el colegio una lección de historia

En primer lugar, la oración es desambiguada y etiquetada por nuestro preprocesador-etiquetador. La salida inicial, en forma de ternas *forma*-etiqueta-*lema*, es transformada al formato requerido por el analizador, *lema*-etiqueta-*no_terminal*, donde inicialmente el no terminal está conformado por la categoría gramatical del término. De este modo se obtiene la entrada al analizador:

[*docenas* NCFP docena] [*de* X de]
[*niño* NCMP N] [*muy* WQ W]
[*alegre* AQFP A] [*haber* V3PRI V]
[*tener* VPMS V] [*que* Cs Cs]
[*aprender* VRI V] [*hoy* WI W] [*en* P P]
[*el* DAMS DA] [*colegio* NCMS N]
[*un* DAFS DA] [*lección* NCFS N]
[*de* P P] [*historia* NCFS N]

Es conveniente indicar que los corchetes han sido añadidos expresamente durante la redacción del presente artículo para facilitar la lectura del ejemplo.

3.1. La fase de análisis

Capa 0: En primer lugar, se amplía el preprocesado, permitiendo identificar la expresión numeral *docenas de*, definida por la secuencia de ternas [*docena* NCFP N] [*de* P P]. Ésta es reducida al sintagma numeral [*docena&de* Cifra $SNum$]:

[*docena&de* Cifra $SNum$] [*niño* NCMP N]
[*muy* WQ W] [*alegre* AQFP A]
[*haber* V3PRI V] [*tener* VPMS V]
[*que* Cs Cs] [*aprender* VRI V]
[*hoy* WI W] [*en* P P] [*el* DAMS DA]
[*colegio* NCMS N] [*un* DAFS DA]
[*lección* NCFS N] [*de* P P]
[*historia* NCFS N]

Adviértase que el carácter *&* es empleado a la hora de concatenar palabras.

Capa 1: Son identificados los sintagmas adverbiales ($SAdv$) *muy*, definido por la terna de entrada [*muy* WQ W], y *hoy*, definido por [*hoy* WI W]. Una vez reducidos, se obtienen las ternas de salida [*muy* WQ $SAdv$] y [*hoy* WI $SAdv$], respectivamente.

Al mismo tiempo son identificados dos grupos verbales de primer nivel ($GV1$). El primero de ellos, *han tenido*, dado por [*haber* V3PRI V] [*tener* VPMS V], es reducido a [*tener* V3PRI $GV1$], mientras que el segundo, *aprender*, dado por la entrada [*aprender* VRI V], es transformado en [*aprender* VRI $GV1$]:

[*docena&de* Cifra $SNum$] [*niño* NCMP N]
[*muy* WQ $SAdv$] [*alegre* AQFP A]
[*tener* V3PRI $GV1$] [*que* Cs Cs]
[*aprender* VRI $GV1$] [*hoy* WI $SAdv$]
[*en* P P] [*el* DAMS DA]
[*colegio* NCMS N] [*un* DAFS DA]
[*lección* NCFS N] [*de* P P]
[*historia* NCFS N]

Capa 2: Los sintagmas adjetivales ($SAdj$) y grupos verbales de segundo nivel ($GV2$) del texto son ahora procesados. El único sintagma adjetival de nuestro ejemplo es *muy alegres*, dado por [*muy* WQ $SAdv$] [*alegre* AQFP A], y que a su vez es reducido a

[*alegre* AQFP *SAdj*]. Recordemos que *muy* había sido ya reducido a un *SAdv* en la capa anterior.

En cuanto a los grupos verbales *GV*2, nuestro texto únicamente contiene el grupo perifrástico *han tenido que aprender*, cuyas ternas integrantes [*tener* V3PRI *GV*1] [*que* Cs *Cs*] [*aprender* VRI *GV*1] son reducidas a la terna de salida [*aprender* V3PRI *GV*2]:

[*docena&de* Cifra *SNum*] [*niño* NCMP *N*]
[*alegre* AQFP *SAdj*] [*aprender* V3PRI *GV*2]
[*hoy* WI *SAdv*] [*en* P *P*] [*el* DAMS *DA*]
[*colegio* NCMS *N*] [*un* DAFS *DA*]
[*lección* NCFS *N*] [*de* P *P*]
[*historia* NCFS *N*]

Capa 3: En esta fase son identificados los sintagmas nominales (*SN*). El primero de los sintagmas encontrados es *docenas de niños muy alegres*, cuyos componentes han sido repetidamente reducidos hasta [*niño* NCMP *N*] [*alegre* AQFP *SAdj*], y que son ahora transformados en [*niño* NCMP *SN*].

El segundo sintagma nominal del ejemplo es *el colegio*, definido por [*el* DAMS *DA*] [*colegio* NCMS *N*], y que es reducido a [*colegio* NCMS *SN*].

El siguiente sintagma identificado es *una lección*, [*un* DAFS *DA*] [*lección* NCFS *N*], devuelto a la salida como [*lección* NCFS *SN*].

El último sintagma es *historia*, dado por la terna [*historia* NCFS *N*] y que es ahora reducido a [*historia* NCFS *SN*]:

[*niño* NCMP *SN*] [*aprender* V3PRI *GV*2]
[*hoy* WI *SAdv*] [*en* P *P*]
[*colegio* NCMS *SN*] [*lección* NCFS *SN*]
[*de* P *P*] [*historia* NCFS *SN*]

Sólo uno se estos sintagmas nominales, *docenas de niños muy alegres*, contiene dependencias *sustantivo-adjetivo* (*ADJ*) a extraer:

ADJ (*niño* NCMP, *alegre* AQFP)

Nótese que el tipo de la dependencia, junto con los lemas y etiquetas de sus componentes, principal y modificador, se denotan mediante:

tipo(*lema-pral* etiq-pral, *lema-mod* etiq-mod)

Capa 4: El último paso consiste en identificar los sintagmas preposicionales, en este caso dos. El primero de ellos es un *SP*, *en el colegio*, representado por la secuencia [*en* P *P*] [*colegio* NCMS *SN*], y que es transformado en [*colegio* NCMS *SP*]. El segundo sintagma es un *SPde*, *de historia*, dado por [*de* P *P*] [*historia* NCFS *SN*] y que es ahora reducido a [*historia* NCFS *SPde*]:

[*niño* NCMP *SN*] [*aprender* V3PRI *GV*2]
[*hoy* WI *SAdv*] [*colegio* NCMS *SP*]
[*lección* NCFS *SN*] [*historia* NCFS *SPde*]

3.2. La fase de extracción de dependencias

En primer lugar, son identificadas las funciones sintácticas de los sintagmas obtenidos durante el análisis para, a continuación, extraer sus dependencias asociadas, salvo en el caso de las dependencias internas a los sintagmas nominales, que han sido ya extraídas durante el análisis.

Mostramos a continuación, sobre la salida del analizador, las funciones sintácticas de los sintagmas identificados, pudiendo comprobar que hemos identificado un sujeto activo (*SUJact*), un grupo verbal predicativo activo (*Vact*), su complemento circunstancial (*CC*), su objeto directo (*OD*), y el complemento nominal de éste último (*CN*):

[*niño*	NCMP	*SN*]	-- ⟨	*SUJact* ⟩
[*aprender*	V3PRI	*GV*2]	-- ⟨	*Vact* ⟩
[*hoy*	WI	*SAdv*]	-- ⟨	⟩
[*colegio*	NCMS	*SP*]	-- ⟨	*CC* ⟩
[*lección*	NCFS	*SN*]	-- ⟨	*OD* ⟩
[*historia*	NCFS	*SPde*]	-- ⟨	*CN* ⟩

Seguidamente, extraemos sus dependencias asociadas:

SUJ (*aprender* V3PRI, *niño* NCMP)
CC (*aprender* V3PRI, *colegio* NCMS)
OD (*aprender* V3PRI, *lección* NCFS)
CN (*lección* NCFS, *historia* NCFS)

4. EVALUACIÓN DEL SISTEMA

Tres son las técnicas de indexación comparadas a la hora de evaluar el comportamiento de nuestro sistema:

Lematización (*lem*). Indexación de las palabras con contenido del texto vía preprocesado y lematización. Experimentos anteriores con diferentes esquemas de peso y diferentes modelos de recuperación [20, 16, 13] apuntan a la lematización como el mejor punto de

Tabla 1: Resultados para todas las dependencias (*ds*): 5548 docs. relevantes esperados

	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

inicio para el desarrollo de métodos de normalización basados en NLP que hagan frente a niveles de variación lingüística más complejos. Por ello tomaremos sus resultados como punto de referencia.

Pares de dependencia sintáctica (*ds*). Empleo combinado de términos simples lematizados y términos complejos obtenidos a partir de la normalización de las dependencias sintácticas extraídas del texto.

Pares de sintagmas nominales (*sn*). Como el anterior, si bien restringiéndose a las dependencias entre un núcleo nominal y sus adjetivos (ADJ), y a aquéllas entre un núcleo nominal y su complemento nominal (CN).

Para estos experimentos hemos empleado el corpus monolingüe español de las ediciones 2001 y 2002 del CLEF [13], compuesto por noticias de la Agencia EFE pertenecientes a 1994 que totalizan 215.738 documentos con un espacio en disco de 509 MBs. Las 100 consultas empleadas, de la 41 a la 140, constan de tres campos: un breve *título*, una somera *descripción* del tema en una frase, y una *narrativa* algo más compleja especificando los criterios de relevancia. Los tres campos han sido empleados, si bien dándole doble relevancia al *título* respecto a los otros campos por ser, en última instancia, el campo que resume la semántica básica de la consulta.

Los documentos fueron indexados por medio del motor de indexación vectorial SMART [4], empleando un esquema de pesos `atn-ntc` [15]. Sin embargo, podría haberse empleado cualquier otro motor de indexación de texto en su lugar, ya que el proceso de normalización y obtención de términos índice se realiza previamente al proceso de indexación propiamente dicho. Debe tenerse en cuenta, por otra parte, que cada motor seguiría comportándose de acuerdo a sus propias características: paradigma de recuperación, esquema de pesos, algoritmo de ordenación, etc. [20]

Llegados a este punto debemos precisar que, si bien los experimentos aquí recogidos han sido realizados siguiendo el procedimiento establecido por el CLEF, éstos no pueden ser considerados *oficiales* al no haber sido evaluados por los correctores de la organización.

La Tabla 1 recoge el primer conjunto de resultados considerado, aquél obtenido empleando la totalidad de dependencias sintácticas extraídas (*ds*). La primera columna presenta los resultados para la lematización (*lem*), nuestra referencia. Las siguientes columnas, *dsx*, recogen los resultados para el método *ds* conforme la ponderación de pesos entre términos simples y complejos, x a 1, evoluciona. La penúltima columna, *opt*, 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*. Por su parte, 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 pa-

Tabla 2: Resultados para sintagmas nominales (*sn*): 5548 docs. relevantes esperados

	lem	***sn1***	***sn2***	***sn3***	***sn4***	***sn5***	***sn6***	***sn7***	***sn8***	*opt*	Δ
Documentos devueltos	99k	99k	99k	99k	99k	99k	99k	99k	99k	--	--
Relevantes devueltos	5220	5202	5241	5247	**5248**	5247	5244	5242	5241	5248	28
R-precision	.5131	.4792	.5036	.5121	.5133	.5145	**.5157**	.5156	.5149	.5157	.0026
Precisión no interpolada	.5380	.5093	.5344	.5408	.5432	.5437	.5441	.5439	**.5443**	.5443	.0063
Precisión por documento	.5924	.5518	.5855	.5956	.5993	.6005	**.6008**	.6007	.6003	.6008	.0084
Precisión a 5 docs.	.6747	.6505	**.6869**	.6747	.6788	.6727	.6727	.6727	.6768	.6869	.0122
Precisión a 10 docs.	.6010	.5768	.5990	.6081	.6111	**.6141**	.6131	.6101	.6111	.6141	.0131
Precisión a 15 docs.	.5623	.5394	.5643	.5690	**.5744**	**.5744**	**.5744**	.5737	.5724	.5744	.0121
Precisión a 20 docs.	.5374	.5020	.5237	.5308	.5323	.5338	.5369	**.5389**	**.5389**	.5389	.0015
Precisión a 30 docs.	.4825	.4519	.4744	.4855	**.4892**	.4879	.4869	.4872	.4875	.4892	.0067
Precisión a 100 docs.	.3067	.2891	.3013	.3045	.3071	**.3083**	.3076	.3077	.3077	.3083	.0016
Precisión a 200 docs.	.2051	.1961	.2036	.2057	.2064	.2065	.2063	**.2067**	.2064	.2067	.0016
Precisión a 500 docs.	.0997	.0981	.0998	.1001	.1005	.1004	.1004	.1004	**.1005**	.1005	.0008
Precisión a 1000 docs.	.0527	.0525	.0529	**.0530**	**.0530**	**.0530**	**.0530**	.0529	.0529	.0530	.0003

ra 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 produjo una disminución general del rendimiento del sistema a todos los niveles. Tras analizar el comportamiento del sistema para las diferentes consultas, se pudo concluir que el problema residía en una sobreponderación de los pesos de los términos complejos respecto a los simples, al ser aquéllos mucho menos frecuentes y, por tanto, con un peso asignado mucho mayor. Esto se tradujo en una creciente inestabilidad del sistema, en tanto que al producirse correspondencias no deseadas de términos complejos con documentos no relevantes, su puntuación asignada aumentaba considerablemente, disparando su nivel de relevancia. Por otra parte, y debido a la misma causa, al producirse correspondencias de términos complejos con documentos sí relevantes, se producía una clara mejora de los resultados respecto al empleo de términos simples.

De acuerdo con esto se podría argumentar 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, con lo que las correspondencias casuales, de producirse, son muchísimo más perjudiciales que en el caso de los términos simples, cuyo impacto tiende a diluirse debido al efecto de las restantes correspondencias.

Podemos afirmar, pues, que esta primera aproximación amplificaba notablemente el ruido introducido por falsas correspondencias. Debíamos corregir, por tanto, esa sobrevaloración de los términos complejos, para 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, decrementando el excesivo grado de relevancia otorgado a los términos complejos en un inicio, tal y como se muestra en las restantes columnas *dsx*. Como puede apreciarse, la mejora de los resultados es inmediata, sobre todo en cuanto a la precisión hasta los 15 primeros documentos devueltos, y en cuanto a la cobertura (pasamos de 5220 relevantes devueltos 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* es el que obtiene unos mejores resultados, además de la mejor cobertura, con 5252 documentos relevantes devueltos. Sin embargo, los mejores resultados en cuanto a medidas de rendimiento globales se obtienen con un factor de ponderación mayor, en *ds7*. Mención aparte merece el caso de *ds2*, que obtiene la precisión más alta en los documentos iniciales y la segunda mejor cobertura, si bien a costa de sacrificar rendimiento en otros aspectos.

Los resultados obtenidos restringiéndonos a los sintagmas nominales (*sn*), recogidos en la Tabla 2, son similares en cuanto a su interpretación, si bien las mejoras en los resul-

tados respecto a *lem* son menores, al no contemplar las dependencias integradas por grupos verbales. Se produce también una mayor dispersión de los resultados, sobre todo en el caso de la ordenación de documentos devueltos, siendo más difícil decidir qué factores de ponderación se muestran superiores. Debemos destacar el hecho de que, en contra de lo que inicialmente se pueda suponer, el coste computacional de nuestra aproximación es el mismo, puesto que los sintagmas nominales y preposicionales son identificados en las capas 3 y 4, respectivamente, mientras que los grupos verbales, necesarios para establecer las dependencias restantes, ya han sido procesados en las capas previas.

5. CONCLUSIONES Y TRABAJO FUTURO

En este artículo hemos planteado la utilización de dependencias sintácticas como términos índice complejos para la Recuperación de Información en español, con el objetivo de tratar los problemas derivados de la variación lingüística de origen sintáctico y morfosintáctico y, de este modo, obtener resultados más precisos.

Para extraer dichas dependencias hemos desarrollado un analizador sintáctico superficial del español basado en cascadas de expresiones regulares, lo que nos permite abordar el procesamiento de grandes colecciones de forma ágil y robusta.

Los resultados obtenidos nos permiten ser optimistas respecto a nuestro planteamiento, ya que la mejoras en la precisión de los documentos devueltos es prácticamente inmediata a pesar de la simplicidad de nuestra aproximación. Es de esperar, por tanto, que técnicas más refinadas nos permitan mejorar estos resultados.

Podemos pensar en dos vías principales de desarrollo para nuestro trabajo futuro. La primera de ellas vendría encaminada a la mejora de nuestro analizador, especialmente respecto al clásico problema de la adjunción de sintagmas preposicionales, donde la aplicación de *restricciones de selección* generadas automáticamente [6] sobre los propios textos podría mejorar la capacidad de desambiguación sintáctica del sistema. El objetivo, aumentar en lo posible la cobertura del sistema sin perjudicar la precisión.

El segundo campo de mejora es el de cómo emplear la información sintáctica disponible de un modo más eficaz. El almacenamiento de términos simples y complejos en índices separados para su tratamiento independiente y el empleo de técnicas de *fusión de datos* [21] para la combinación de los resultados obtenidos podrían facilitar dicha tarea.

Sería también interesante comparar nuestros resultados con los obtenidos mediante otras aproximaciones sintácticas como el empleo de colocaciones generadas a partir de bases de datos léxico-semánticas [1], o mediante la conjugación de técnicas pseudo-sintácticas basadas en la localidad [12] con nuestros mecanismos de normalización basados en NLP.

BIBLIOGRAFÍA

[1] ALONSO, Margarita, y SANROMÁN, Begoña (2000): "Construcción de una base de datos de colocaciones léxicas", en *Procesamiento del Lenguaje Natural*, nº24, pp. 97–98.

[2] AONE, Chinatsu, HALVERSON, Lauren, HAMPTON, Tom, y RAMOS-SANTACRUZ, Mila (1998): "SRA: Description of the IE^2 system used for MUC-7", en *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

[3] ARAMPATZIS, Avi T., VAN DER WEIDE, Theo P., KOSTER, Cornelis H.A., y VAN BOMMEL, Patrick (2000): "Linguistically motivated information retrieval", en *Encyclopedia of Library and Information Science*, vol. 69. Marcel Dekker, Inc., Nueva York y Basel.

[4] BUCKLEY, Chris (1985): "Implementation of the SMART information retrieval system". Technical report, Department of Computer Science, Cornell University. Fuentes de SMART disponibles (13.06.2003) en `ftp://ftp.cs.cornell.edu/pub/smart`

[5] FERNÁNDEZ, Santiago, GRAÑA, Jorge, y SOBRINO, Alejandro (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)*, pp. 31–37, León, España.

[6] GAMALLO, Pablo, AGUSTINI, Alexandre, y LOPES, Gabriel P. (2001): "Selections restrictions acquisition from corpora", en *Lecture Notes in Artificial Intelligence*, vol. 2258, pp. 30–43. Springer-Verlag, Berlín/Heidelberg/Nueva York.

[7] GRAÑA, Jorge, BARCALA, Fco. Mario, y VILARES, Jesús (2002): "Formal methods of tokenization for part-of-speech tagging", en *Lecture Notes in Computer Science*, vol. 2276, pp. 240–249. Springer-Verlag, Berlín/Heidelberg/Nueva York.

[8] HOBBS, Jerry R., APPELT, Douglas, BEAR, John, ISRAEL, David, KAMEYAMA, Megumi, STICKEL, Mark, y TYSON, Mabry (1997): "FASTUS: A cascaded finite-state transducer for extracting information from natural-language text", en ROCHE, Enmanuel, y SCHABES, Yves (eds.): *Finite-State Language Processing*. MIT Press, Cambridge, MA, USA.

[9] HULL, David A., GREFENSTETTE, Gregory, SCHULZE, B. Maximilian, GAUSSIER, Eric, SCHUTZE, Hinrich, y PEDERSEN, Jan O. (1997): "Xerox TREC-5 site report: routing, filtering, NLP, and Spanish tracks", en *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*, pp. 167–180.

[10] JACQUEMIN, Christian, y TZOUKERMANN, Evelyne (1999): "NLP for term variant extraction: synergy between morphology, lexicon and syntax", en STRZALKOWSKI, Tomek (ed.): *Natural Language Information Retrieval*, vol. 7 de *Text, Speech and Language Technology*, pp. 25–74. Kluwer Academic Publishers, Dordrecht/Boston/Londres.

[11] KRAAIJ, Wessel, y POHLMANN, Renée (1998): "Comparing the effect of syntactic vs. statistical phrase indexing strategies for Dutch", en *Lecture Notes in Computer Science*, vol. 1513, pp. 605–614. Springer-Verlag, Berlín/Heidelberg/Nueva York.

[12] DE KRETSER, Owen, y MOFFAT, Alistair (1999): "Effective document presentation with a locality-based similarity heuristic", en *Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 113–120, Nueva York. ACM Press.

[13] PETERS, Carol (ed.) (2002): *Results of the CLEF 2002 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2002 Workshop*, Roma, Italia. Web oficial del CLEF (13.06.2003): `http://www.clef-campaign.org`

[14] PEREZ-CARBALLO, Jose, y STRZALKOWSKI, Tomek (2000): "Natural language information retrieval: progress report", en *Information Processing and Management*, 36(1):155–178.

[15] SAVOY, Jacques, LE CALVÉ, Anne, y VRAJITORU, Dana (1997): "Report on the TREC-5 experiment: Data fusion and collection fusion", en *Proceedings of TREC'5*, NIST publication #500-238, pp. 489–502, Gaithersburg, MD.

[16] VILARES, Jesús, ALONSO, Miguel A., RIBADAS, Francisco J., y VILARES, Manuel (2002): "COLE experiments at CLEF 2002 Spanish monolingual track", en [13], pp. 153–160.

[17] VILARES, Jesús, BARCALA, Fco. Mario, y ALONSO Miguel A. (2002): "Using syntactic dependency-pairs conflation to improve retrieval performance in Spanish", en *Lecture Notes in Computer Science*, vol. 2276, pp. 381–390. Springer-Verlag, Berlín/Heidelberg/Nueva York.

[18] VILARES, Jesús BARCALA, Fco. Mario, ALONSO, Miguel A., GRAÑA, Jorge, y VILARES, Manuel (2002): "Practical NLP-based text indexing", en *Lecture Notes in Computer Science*, vol. 2527, pp. 635–644. Springer-Verlag, Berlín/Heidelberg/Nueva York.

[19] VILARES, Jesús, CABRERO, David, y ALONSO Miguel A. (2001): "Applying productive derivational morphology to term indexing of Spanish texts", en *Lecture Notes in Computer Science*, vol. 2004, pp. 336–348. Springer-Verlag, Berlín/Heidelberg/Nueva York.

[20] VILARES, Jesús, VILARES, Manuel, y ALONSO, Miguel A. (2001): "Towards the development of heuristics for automatic query expansion", en *Lecture Notes in Computer Science*, vol. 2113, pp. 887–896. Springer-Verlag, Berlín/Heidelberg/Nueva York.

[21] VOGT, Christopher C., y COTTRELL, Garrison W. (1999): "Fusion via a linear combination of scores", en *Information Retrieval*, 1(3):151–173.