Normalización de términos multipalabra mediante pares de dependencia sintáctica*

Jesús Vilares, Fco. Mario barcala y Miguel A. Alonso
Departamento de Computación, Universidad de La Coruña
Campus de Elviña s/n, 15071 La Coruña
jvilares@mail2.udc.es, barcala@dc.fi.udc.es, alonso@dc.fi.udc.es
http://coleweb.dc.fi.udc.es/

Resumen En este artículo se presentan dos nuevas técnicas para la indexación de textos escritos en español. A nivel de palabra, proponemos la utilización de la morfología derivativa para obtener conjuntos de palabras relacionadas semánticamente. Esta técnica se combina, a nivel de frase, con la utilización de una gramática aproximada, lo que nos permitirá normalizar a una forma base común las variantes sintácticas y morfosintácticas de un término multipalabra. Dichos métodos han sido evaluados sobre un corpus de documentos periodísticos, obteniendo unos resultados que muestran una mejora considerable con respecto a los métodos clásicos de indexación.

1 *Introducción*

En tareas de Recuperación de Información (RI), los documentos son representados en forma de conjuntos de términos índice o palabras clave representativas. Para ello se recurre a operaciones tales como la eliminación de *stopwords* (palabras excesivamente frecuentes y sin significación aparente) o técnicas de *stemming* (que reducen las palabras a su supuesta raíz gramatical). A dicho tipo de operaciones se las denomina *transformaciones de texto*, y proporcionan una *representación lógica* del documento.

En efecto, los sistemas de RI normalizan los documentos antes de su indexación mediante la agrupación conjunta de términos que se refieren a conceptos idénticos o similares, para así decrementar la variedad lingüística de dichos documentos [1, 8]. Desafortunadamente, las técnicas más empleadas carecen habitualmente de base lingüística. Incluso aquellas técnicas que presumiblemente gozan de ella (por ejemplo *stemming*), si bien consiguen resultados muy aceptables para el inglés, se muestran del todo insuficientes cuando se aplican a idiomas más ricos desde el punto de vista morfológico, caso del español. Por ello, las técnicas a aplicar en estos idiomas deberán emplear mayores y mejores recursos lingüísticos, lo que redundará en una mayor complejidad y en un aumento del coste computacional. Llegados a este punto debemos llamar la atención sobre uno de los grandes problemas a los que se enfrentan las técnicas de Procesamiento de Lenguaje Natural (PLN) para el caso del español, la escasa disponibilidad de grandes recursos lingüísticos de libre acceso [1].

En este contexto, proponemos una extensión de las técnicas clásicas de indexación que nos permitirá salvar tales inconvenientes.

2 *Normalización de términos simples*

En inglés, la normalización de términos simples puede llevarse a cabo de forma bastante satisfactoria empleando un *stemmer* [10], una herramienta muy sencilla desde el punto de vista lingüístico y de escaso coste computacional. Esto se debe a que la morfología flexiva del inglés es harto sencilla. No sucede lo mismo en el caso del español, con una morfología flexiva compleja e irregular [13].

Se hace necesario, pues, el empleo de técnicas de PLN, con el consiguiente incremento en complejidad y coste computacional. Como primer paso, el empleo de un lematizador permite solventar los problemas derivados de la flexión de las palabras [5]. Pero es posible ir más allá, utilizando la morfología derivativa para la normalización de términos

* Este trabajo ha sido financiado en parte por el Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (TIC2000-0370-C02-01), los fondos FEDER de la EU (1FD97-0047-C04-02) y la Xunta de Galicia (PGIDT99XI10502B).

[1] Corpora etiquetados, bancos de árboles, diccionarios avanzados, etc.

simples mediante el empleo de familias morfológicas.

2.1 Familias morfológicas

La riqueza léxica y morfológica del español queda reflejada en la gran productividad y flexibilidad que presentan sus mecanismos de formación de palabras, lo que conlleva una morfología derivativa rica y compleja. En concreto, en español el mecanismo preferido para la formación de nuevas palabras es la derivación, en detrimento de la composición.

Definimos una *familia morfológica* como el conjunto de palabras obtenidas a partir de una misma raíz morfológica mediante la aplicación de mecanismos de derivación. Es de esperar que se mantenga una relación semántica entre las palabras de dicho conjunto[2]. A la hora de obtener unos patrones regulares de formación de palabras podemos valernos de las denominadas *reglas de formación* propuestas por la Fonología Generativa y la Gramática Transformacional-Generativa [9]. Aunque dicho paradigma no es completo, supone un avance considerable puesto que permite diseñar un sistema de generación automática de familias morfológicas con un grado aceptable de corrección y completud [11].

Los mecanismos básicos de derivación en español son: *parasíntesis*, *prefijación*, *sufijación apreciativa*, *sufijación no apreciativa* y *derivación regresiva*. La prefijación consiste en la adición de un prefijo a una forma base, la sufijación en la adición de un sufijo y la parasíntesis en la adición simultánea de un prefijo y un sufijo. En el caso de la sufijación, debemos distinguir entre sufijación apreciativa, que altera semánticamente el lexema base de un modo subjetivo emocional pero sin cambiar su categoría gramatical[3], y sufijación no apreciativa, que involucra un cambio fundamental más que marginal en el significado del lexema base, frecuentemente acompañado de un cambio de categoría sintáctica.

En lo referente a la derivación regresiva, se trata de un mecanismo para la formación de sustantivos a partir de verbos. Su característica principal radica en que, en lugar de incrementar el tamaño del lexema base, lo reduce, al añadir tan solo una vocal a la raíz verbal[4].

Un fenómeno importante a tener en cuenta es la existencia de *alomorfos*, variantes de un mismo morfema derivativo [5]. El alomorfo a utilizar en cada caso puede estar determinado por la fonología o venir impuesto por convención o por la etimología.

Tampoco debemos descuidar el análisis de las condiciones fonológicas que dominan el proceso de formación de palabras, puesto que una operación morfológica suele implicar a su vez alteraciones fonológicas en el lexema base, que pueden ser regulares[6] o aparentemente irregulares[7] [9].

2.2 Transformación del texto mediante familias morfológicas

Al abordar la normalización de términos simples mediante familias morfológicas, reemplazaremos cada palabra con contenido por un representante de su familia morfológica fijado a priori. De este modo se utiliza un mismo término índice para representar a todas las palabras pertenecientes a una misma familia morfológica, lo que nos permite trasladar al índice las relaciones semánticas existentes entre dichas palabras.

Hemos comparado la efectividad como operaciones de transformación de texto de la lematización y las familias morfológicas con respecto a la técnica clásica de stemming. Después de probar el funcionamiento de diferentes stemmers diseñados para el español, los mejores resultados se obtuvieron para el stemmer utilizado por Muscat[8], un motor de búsqueda de código abierto; dicho stemmer está basado en el algoritmo de Porter [1]. Sin embargo, los resultados obtenidos son muy pobres, con un grado de corrección general aproximado del 37%. Debemos reseñar que mientras el lematizador muestra un rendimiento uniforme para todas las categorías gramaticales, para los stemmers es variable, obteniendo una corrección del 46% para sustantivos, del 36% para adjetivos y de un 0%

[2]Relaciones del tipo proceso-resultado (por ejemplo *fijación-fijado*), proceso-agente (por ejemplo *inhibición-inhibidor*), etc.

[3]Los sufijos apreciativos pueden subdividirse en diminutivos, aumentativos y peyorativos.

[4]Por ejemplo, de *deteriorar* se obtiene *deterioro*

[5]Por ejemplo, *innecesario*, *imprudente* e *irreal*

[6]Por ejemplo, la derivación de *respondón* a partir de *responder* o la derivación de *invencible* a partir de *vencer*.

[7]Por ejemplo, la derivación de *panadero* (y no de **pandero*) a partir de *pan* o la derivación de *acuatizar* (y no de **agüizar*) a partir de *agua*.

[8]http://open.muscat.com

para verbos[9]. Una ventaja adicional de los lematizadores con respecto a los stemmers es su habilidad para desambiguar en base al contexto de una palabra.

Si comparamos los stemmers con las familias morfológicas, obtenemos que el stemmer de Muscat es capaz de identificar el 27% de las familias generadas. De ellas, el 95% son familias con un único lema, un 3% de dos lemas y menos de un 2% de tres lemas.

Con respecto al coste computacional, las familias morfológicas, al haber sido generadas a priori, no influyen en el coste final de indexación y consulta. El coste de ejecución de un stemmer es lineal en la longitud de la palabra, siendo el del lematizador sólo ligeramente superior debido al proceso previo de desambiguación. Dicho coste será también lineal respecto a la longitud de la palabra, pero cúbico respecto al tamaño del conjunto de etiquetas. Sin embargo, tal y como se detalla en la sección 3.3, nos basta con conocer la categoría gramatical de la palabra, por lo que el conjunto de etiquetas es muy pequeño, lo que supone un aumento mínimo en el coste computacional.

3 Normalización de términos multipalabra

En el ámbito de la recuperación de información se denomina *término multipalabra* a aquel término que contiene dos o más palabras con contenido (sustantivos, verbos y adjetivos)[10]. En la literatura se describen varios métodos para su obtención. Uno de los más utilizados es el denominado *simplificación del texto* [6]: en una primera fase, se realiza un stemming de las palabras individuales, y se procede a eliminar las stopwords; posteriormente se extraen y normalizan los términos empleando para ello patrones [3], técnicas estadísticas [4], etc. Existe, pues, una clara falta de base lingüística en dichas operaciones[11], lo que redunda frecuentemente en simplificaciones erróneas. Sin embargo es el método más sencillo y menos costoso.

Por otra parte, existen otros métodos de sólida base lingüística que realizan un *análisis sintáctico* del texto mediante un analizador sintáctico, el cual devuelve a su salida un conjunto de árboles sintácticos que denotan relaciones de dependencia entre las palabras involucradas. De este modo, estructuras con relaciones de dependencia similares pueden ser normalizadas a una misma forma.

A medio camino estaría la *correspondencia de patrones*, que se basa en la hipótesis de que las partes más informativas del texto siguen unas construcciones sintácticas bastante bien definidas que se pueden aproximar mediante patrones [7].

En el presente trabajo proponemos una aproximación que conjuga estos dos últimos métodos y que se basa en la indexación de sintagmas nominales y de sus *variantes sintácticas y morfosintácticas* [6].

Una variante sintáctica o morfosintáctica de un término multipalabra es una frase perteneciente al texto, tal que:

- Las variantes sintácticas son producto de la flexión de palabras individuales y de la modificación de la estructura sintáctica del término original. Por ejemplo, *medidas de longitud y tiempo* es una variante de *medida de tiempo*.
- Las variantes morfosintácticas difieren de las anteriores en que al menos una de las palabras con contenido del término original se transforma en otra que deriva de la misma raíz morfológica. Por ejemplo, *medición del contenido* es una variante de *medir el contenido*.
- La variante puede ser sustituida por el término original del que deriva en lo que respecta al acceso a la información.

Desde el punto de vista morfológico, las variantes sintácticas hacen referencia a la morfología flexiva, mientras que las morfosintácticas entran además en el ámbito de la morfología derivativa. En lo referente a la sintaxis, las variantes sintácticas tienen un campo de actuación mucho más restringido, el sintagma nominal, mientras que las variantes morfosintácticas lo amplían a prácticamente toda la oración, incluyendo los verbos y sus objetos[12].

A continuación estudiaremos los mecanismos involucrados en la obtención de variantes sintácticas y morfosintácticas.

[9] Ello se debe a la complejidad del paradigma verbal del español, que no es abarcado en profundidad por ningún stemmer.

[10] Por ejemplo, *el perro grande del vecino* o *la casa de mis padres*.

[11] Por ejemplo, algunas stopwords tales como artículos y preposiciones son componentes clave de la estructura sintáctica.

[12] Consideremos por ejemplo las frases *comida para perros* y *los perros comen*.

3.1 Variantes sintácticas

Las variantes sintácticas de un término multipalabra pueden involucrar variaciones en la flexión de las palabras que lo componen y alteraciones sintácticas de tipo:

- *Coordinación:* empleo de las construcciones coordinantes del español (copulativas y disyuntivas) bien en un modificador, bien en el modificado. Por ejemplo, los términos *desarrollo agrícola* y *desarrollo rural* se combinan en *desarrollo agrícola y rural*, que podemos considerar tanto una variante de *desarrollo agrícola* como de *desarrollo rural*.
- *Sustitución:* empleo de modificadores que hacen más específico el término considerado. Por ejemplo, *desarrollo del arte* puede transformarse en *desarrollo tardío del arte* sustituyendo *desarrollo* por *desarrollo tardío*.
- *Sinapsis:* si bien las anteriores construcciones eran binarias, ésta es una construcción unaria donde se cambia la preposición empleada o bien se añade o elimina un determinante. Por ejemplo, al obtener la variante *abono para plantas* a partir de *abono para las plantas*.
- *Permutación:* hace referencia a la permutación de palabras alrededor de un elemento pivote central. Por ejemplo, *saco viejo* y *viejo saco* son permutaciones del mismo término multipalabra.

3.2 Variantes morfosintácticas

Las variantes morfosintácticas de un término multipalabra se clasifican en función de la naturaleza de las transformaciones morfológicas aplicadas a las palabras con contenido de dichos términos. Tenemos:

- Variantes *iso-categóricas*, cuando la derivación morfológica no altera la categoría gramatical de la palabra. Estaremos por tanto transformando un sintagma nominal en otro sintagma nominal, con los dos tipos siguientes de variantes:
  1. *Sustantivo-Sustantivo*, que contemplan relaciones del tipo proceso/resultado[13] y proceso/agente[14].
  2. *Adjetivo-Adjetivo*, que contemplan relaciones de tipo agente/resultado[15].
- Variantes *hetero-categóricas*, cuando sí se produce un cambio en la categoría gramatical de la palabra. Este tipo de variantes no se restringen sólamente al sintagma nominal. En este caso estamos considerando variantes de tipo:
  1. *Sustantivo-Verbo*, variaciones que conllevan cambios semánticos del tipo proceso/resultado[16].
  2. *Sustantivo-Adjetivo*, que modifican el núcleo de un sintagma nominal mediante construcciones adjetivas o preposicionales equivalentes[17].

3.3 Extracción y normalización de términos multipalabra

Las consultas realizadas a un sistema de recuperación de información suelen expresarse en forma de grupos nominales de complejidad diversa. Es por ello que tomaremos los grupos nominales como términos base a partir de los cuales, y mediante la aplicación de las transformaciones correspondientes, se obtendrán sus variantes sintácticas y morfosintácticas, no necesariamente grupos nominales. Todos estos términos multipalabra, tanto los grupos nominales originales como sus variantes, son susceptibles de ser utilizados como términos índice.

Las estructuras básicas de los grupos nominales para el caso del español son:

- Adjetivo-Sustantivo.
- Sustantivo-Adjetivo.
- Sustantivo-Preposición-Sustantivo.
- Sustantivo-Preposición-Determinante-Sustantivo.

Nos interesará, por tanto, identificar tanto dichos sintagmas como sus variantes para así indexarlos. A la hora de extraer del texto dichos términos, emplearemos patrones obtenidos a partir de la estructura sintáctica de los grupos nominales y sus variantes. Para ello

[13] Por ejemplo, *producción artesanal* y *producto artesanal*.

[14] Por ejemplo, *manipulación de las masas* y *manipulador de las masas*.

[15] Por ejemplo, *compuesto ionizador* y *compuesto ionizado*.

[16] Por ejemplo, *recortar los gastos* y *recorte de gastos*.

[17] Por ejemplo, *cambio del clima* y *cambio climático*.

tomaremos como base la siguiente gramática aproximada del español:

$$S \rightarrow NP\ V\ W^{?}\ (NP|PP)^{*} \quad (1)$$
$$NP \rightarrow D^{?}\ AP^{*}\ N\ (AP|PP)^{*} \quad (2)$$
$$AP \rightarrow W^{?}\ A \quad (3)$$
$$PP \rightarrow P\ NP \quad (4)$$

donde D, A, N, W, V y P son las etiquetas que representan respectivamente a determinantes, adjetivos, sustantivos, adverbios, verbos y preposiciones[18]. La motivación de las reglas de la gramática es la siguiente:

(1) representa una oración de estructura *Sujeto-Verbo-Predicado.*

(2) define el sintagma nominal como un sustantivo modificado por adjetivos y sintagmas preposicionales.

(3) permite que los adjetivos sean modificados por adverbios.

(4) representa un sintagma preposicional formado por una preposición y un sintagma nominal.

Algunas aproximaciones similares, como la descrita en [6], optan por un enfoque estático basado en la reutilización de bases de datos terminológicas previamente disponibles, que son incorporadas en un analizador sintáctico lexicalizado. Puesto que este tipo de recursos son difíciles de obtener en el caso del español, en este trabajo optamos por un enfoque dinámico en el que los términos son identificados dinámicamente durante el proceso de indexación sin realizar un análisis sintáctico completo del documento, sino tan solo superficial, y sin utilizar tampoco base de datos terminológica alguna.

El primer paso para la indexación de un texto consiste en identificar los términos índice. Tomando como base los árboles sintácticos correspondientes a los grupos nominales y de acuerdo con la gramática aproximada descrita anteriormente, aplicaremos manualmente las transformaciones descritas en las secciones 3.1 y 3.2. De este modo obtendremos los árboles sintácticos pertenecientes a las variantes sintácticas y morfosintácticas de dichos grupos nominales. El conjunto de árboles obtenidos para los términos multipalabra (sintagmas nominales y sus variantes) se puede dividir en cuatro grandes grupos: *sustantivo modificado por adjetivo*, *sustantivo modificado por un sintagma preposicional*, *verbo-objeto* y *sujeto-verbo.*

Sin embargo, en nuestro enfoque estos árboles no son todavía aplicables para la extracción de términos, sino que son primero aplanados en forma de expresiones regulares basadas en las categorías gramaticales de los *tokens* involucrados. Tomemos el ejemplo mostrado en la figura 1:

1. Partimos de un sintagma nominal cuya estructura sintáctica viene representada por el árbol de la izquierda, con el sustantivo N_1 como núcleo modificado por un sintagma adjectival.
2. Obtenemos una de sus variantes mediante la incorporación de una coordinación en el sintagma adjetival (paso 1).
3. El árbol sintáctico correspondiente a la variante obtenida es aplanado para obtener el patrón que será aplicado sobre el texto etiquetado (paso 2).

Dedemos señalar que en lugar de realizar un análisis sintáctico completo de las frases, se está realizando sólo un análisis sintáctico superficial, lo que nos permite reducir considerablemente el coste de ejecución, factor muy importante a la hora de trabajar con grandes colecciones de documentos.

Una vez que los términos a indexar han sido identificados mediante la correspondencia de patrones, han de ser normalizados, proceso que se realiza en dos fases. Primero, se identifican las dependencias sintácticas entre pares de palabras con contenido dentro del árbol sintáctico del término multipalabra (*pares de dependencia sintáctica*); dichos pares quedarán asociados al patrón correspondiente a ese árbol. A continuación, los términos simples que constituyen dichos pares son normalizados mediante lematización o el empleo de familias morfológicas; los pares resultantes serán los términos índice a indexar.

Las dependencias que podemos encontrar dentro de un término multipalabra pertenecen a tres tipos:

1. *Modificado-modificador*: presentes dentro de los sintagmas nominales. Se obtendrá un par de dependencia por cada uno de los núcleos de los modificadores y

[18] A la hora de construir las variantes se emplearán también conjunciones coordinantes (representadas por C), y signos de puntuación (representados por Q).

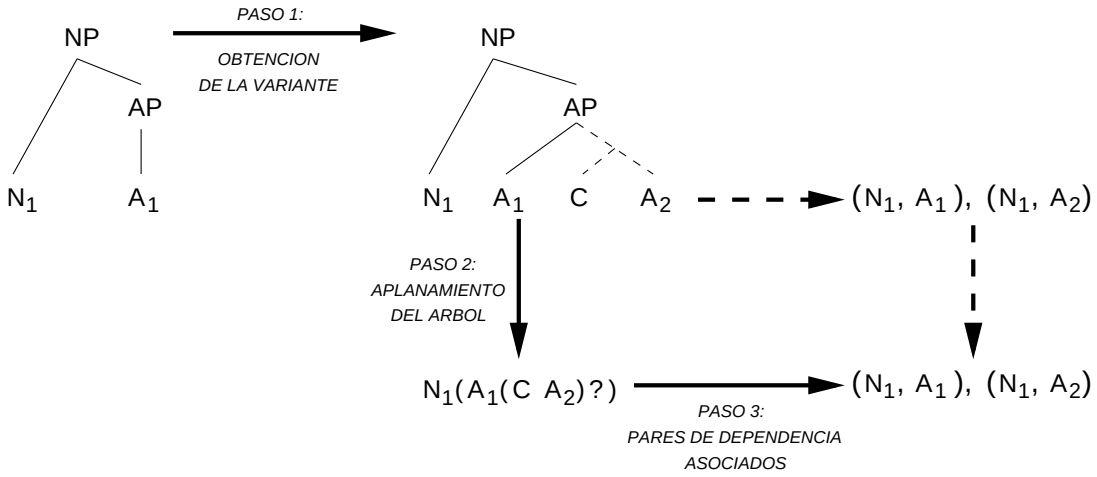

Figura 1: Obtención de los pares de dependencia para la normalización de términos multipalabra

cada uno de los núcleos de sus modificados. Por ejemplo, de *coches y camiones rojos* obtendremos los pares (*coche, rojo*) y (*camión, rojo*).

2. *Sujeto-verbo*: el par principal será el formado por el núcleo del sujeto y el verbo. Por ejemplo, de *los perros comen carne* obtendremos el par (*perro, comer*).
3. *Verbo-objeto*: el par principal será en este caso el formado por el verbo y el sustantivo núcleo del objeto o complemento verbal. Por ejemplo, de *recortar gastos* obtendremos el par (*recortar, gasto*).

En el ejemplo de la figura 1, podemos ver cómo en el paso 3 se procede a la identificación de los pares de dependencia asociados a dicha variante.

En el caso de las variantes sintácticas, las dependencias presentes en el término multipalabra original permanecen siempre en la variante. Sin embargo, esto no ocurre para el caso de las variantes morfosintácticas a menos que utilicemos familias morfológicas en la normalización de los términos simples del par. Por ejemplo, dado el término *recorte de gastos* y su variante morfosintáctica *recortar gastos*, la utilización de la lematización daría lugar a los pares (*recorte, gasto*) y (*recortar, gasto*), respectivamente. En el caso de utilizar familias morfológicas obtendríamos el par (*recorte, gastar*) tanto para el término original como para su variante morfosintáctica[19], obteniendo así un mayor grado de normalización.

4 Evaluación del sistema

Las técnicas descritas en este artículo son independientes del motor de indexación empleado. Esto se debe a que el procesado de los documentos para la obtención de sus términos índice se realiza previamente a la indexación. En consecuencia, el motor recibe, para su indexación, un documento ya transformado. De este modo, cualquier motor de indexación de textos es subceptible de ser utilizado, si bien cada motor se comportará de acuerdo con sus propias características [20].

A la hora de evaluar un sistema de recuperación de información, deben emplearse las medidas de *precisión* P y *cobertura* R, donde:

$$P = \frac{\text{nº de documentos relevantes recuperados}}{\text{número total de documentos recuperados}}$$

$$R = \frac{\text{nº de documentos relevantes recuperados}}{\text{nº total de documentos relevantes}}$$

Hemos estudiado el comportamiento de cinco métodos de indexación diferentes:

pln: texto plano sin stopwords.

lem: normalización de términos simples mediante lematización.

fam: normalización de términos simples mediante familias morfológicas.

[19] En este ejemplo hemos tomado *recorte* como representante de la familia morfológica en la que se encuentran tanto *recortar* como *recorte*, mientras que hemos tomado *gastar* como representante de aquella en la que están *gasto* y *gastar*.

[20] Modelo de indexación, algoritmo de ordenación, etc.

	original	*pln*	*lem*	*fam*	*FNL*	*FNF*
Total	9,780,513	4,526,058	4,625,579	4,625,579	2,666,190	2,666,190
Únicos	154,419	154,071	111,982	105,187	1,210,182	1,036,005

Tabla 1: Características del corpus utilizado

FNL: normalización de términos multipalabra mediante pares de dependencia sintáctica y lematización.

FNF: normalización de términos multipalabra mediante pares de dependencia sintáctica y familias morfológicas.

El corpus de referencia empleado para la evaluación está formado por 21.899 documentos periodísticos (artículos de nacional, internacional, economía, cultura, ...) abarcando la totalidad del año 2000. La longitud media de los documentos es de 447 palabras. Para realizar los experimentos se ha considerado un conjunto de 14 consultas en lenguaje natural de longitud media 7,85 palabras, de las cuales 4,36 son palabras con contenido.

La tabla 1 muestra algunas de las medidas que caracterizan al corpus utilizado. La primera y segunda filas muestran, respectivamente, el número total de términos y el número de términos únicos obtenidos para los documentos indexados, tanto para el texto original como para sus diferentes representaciones normalizadas. Como se puede observar en la primera fila, las técnicas de normalización de términos simples alcanzan una reducción de más del 50% en el número de términos a indexar, mientras que las técnicas de normalización multipalabra alcanzan una reducción de cerca del 75%. Con respecto al número de términos diferentes en el índice, mostrado en la segunda fila, se observa que la reducción resultante de la utilización de stopwords es despreciable, mientras que la utilización de lematización produce una reducción del 27% y la utilización de familias morfológicas proporciona una reducción del 32%, con el consiguiente ahorro de espacio y disminución del tiempo de acceso a los índices. Por su parte, las técnicas de normalización de términos multipalabra incrementan significativamente el tamaño de los índices puesto que cada término de indexación es ahora un elemento compuesto que expresa relaciones sintácticas de dependencia. En este caso, es de destacar que la utilización de familias morfológicas en la normalización de tales elementos complejos reduce significativamente (en un 14%) el número de términos índices con respecto a la lematización.

Los resultados que aquí mostramos han sido obtenidos con SMART[21], motor de indexación basado en el modelo vectorial, y son un breve resumen de los mostrados en [12].

La tabla 2 muestra los valores medios de precisión y cobertura obtenidos en los experimentos. Como se puede observar, los métodos *lem* y *fam*, que aplican técnicas de normalización de términos simples, incrementan considerablemente la cobertura, mientras que *FNL* y *FNF*, que aplican técnicas de normalización de términos multipalabra, incrementan considerablemente la precisión. Es de destacar que el empleo de familias morfológicas por sí solas (*fam*) no plantea mejoras respecto a la lematización (*lem*); sin embargo, su empleo en conjunción con términos multipalabra (*FNF*) produce un aumento más que significativo de la cobertura alcanzada respecto al empleo de la lematización (*FNL*).

En lo referente a la evolución de la precisión respecto a la cobertura, las gráficas de la figura 2 confirman que el peor comportamiento corresponde a *pln* mientras que el mejor se corresponde con *lem* y *FNF*. Para niveles bajos y altos de cobertura, (≤ 0.2, ≥ 0.7) *FNF* es claramente mejor, mientas que para el resto del intervalo *lem* le supera.

Ante estos resultados podemos concluir que si bien el empleo de familias morfológicas no reporta grandes mejoras respecto a la lematización, su empleo conjunto con términos multipalabra incrementa sustancialmente la precisión a la vez que mantiene un nivel de cobertura aceptable.

5 Conclusiones

Este artículo pretende demostrar la conveniencia de la utilización de técnicas de Procesamiento de Lenguaje Natural en sistemas de Recuperación de Información para idiomas ricos desde el punto de vista morfológico,

[21] `ftp://ftp.cs.cornell.edu/pub/smart/`

	pln	*lem*	*fam*	*FNL*	*FNF*
Precisión	0.17	0.20	0.20	0.31	0.32
Cobertura	0.55	0.63	0.60	0.48	0.56

Tabla 2: Precisión y cobertura media

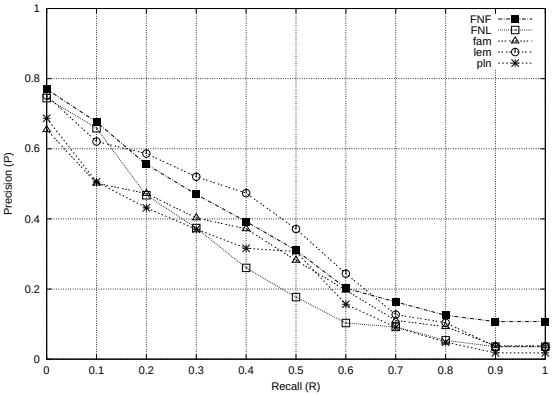

Figura 2: Gráfica de precisión y cobertura

tales como el español. En particular, se han presentado dos nuevas técnicas de normalización que resultan efectivas de cara a reducir la variedad lingüística de los documentos: el empleo de morfología derivativa para tratar términos simples y la extracción de pares de dependencias a partir de una gramática aproximada para tratar términos multipalabra. Frente a otras técnicas similares, basadas en el empleo de analizadores sintácticos y amplias bases de datos terminológicas (confiriéndoles una naturaleza estática), nuestra aproximación es dinámica, pues identifica los términos en tiempo de ejecución, requiriendo además unos recursos lingüísticos mínimos, lo que la hace también adecuada para lenguas minoritarias. Del mismo modo, al tratarse de una aproximación léxica, el incremento del coste computacional es también mínimo, al estar basada en tecnología de estado finito, permitiendo así su aplicación práctica en sistemas reales.

Referencias

[1] R. Baeza-Yates y B. Ribeiro-Neto. 1999. *Modern information retrieval.* Addison-Wesley, Harlow, Inglaterra.

[2] E. Bajo Pérez. 1997. *La derivación nominal en español.* Cuadernos de lengua española. Arco Libros, Madrid.

[3] M. Dillon y A.S. Gray. 1983. FASIT: A fully automatic syntactically based indexing system. *Journal of the American Society for Information Science*, 34(2):99–108.

[4] J.L. Fagan. 1987. Automatic phrase indexing for document retrieval: An examination of syntactic and non-syntactic methods. In *Proceedings of ACM SIGIR'87*, páginas 91–101.

[5] J. Graña. 2000. *Técnicas de Análisis Sintáctico Robusto para la Etiquetación del Lenguaje Natural.* Tesis Doctoral, Universidade da Coruña.

[6] C. Jacquemin y E. Tzoukerman. 1999. NLP for term variant extraction: A synergy of morphology, lexicon and syntax. In T. Strzalkowski, editor, *Natural Language Information Retrieval*, páginas 25–74. Kluwer Academic, Boston.

[7] J.S. Justeson y S.M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27.

[8] G. Kowalski. 1997. *Information retrieval systems: theory and implementation.* Kluwer Academic, Boston.

[9] M. F. Lang. 1990. *Spanish Word Formation: Productive Derivational Morphology in the Modern Lexis.* Croom Helm. Routledge, Londres y Nueva York.

[10] M. Lennon, D.S. Pierce, y P. Willett. 1981. An evaluation of some conflation algorithms. *Journal of Information Science*, 3:177–183.

[11] J. Vilares, D. Cabrero, y M. A. Alonso. 2001. Applying Productive Derivational Morphology to Term Indexing of Spanish Texts, In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volumen 2004 de *LNCS*, páginas 336-348, Springer-Verlag, Berlín-Heidelberg-Nueva York.

[12] J. Vilares, M. Vilares y M. A. Alonso. 2001. Towards the development of heuristics for automatic query expansion. In Database and Expert Systems Applications. A aparecer en *LNCS*. Springer-Verlag, Berlín-Heidelberg-Nueva York.

[13] M. Vilares, J. Graña, y P. Alvariño. 1997. Finite-state morphology and formal verification. In A. Kornai, editor, *Extended Finite State Models of Language*, páginas 37–47. Cambridge University Press.