La detección de nombres propios en español y su aplicación en recuperación de información

Applying proper noun detection in Spanish for information retrieval

Ángel F. Zazo, Carlos G. Figuerola y José Luis Alonso Berrocal

Grupo de investigación en REcuperación de INformación Avanzada (REINA), Facultad de Traducción y Documentación, Universidad de Salamanca, C/ Francisco de Vitoria, 6-16, 37008 - Salamanca (España), {afzazo | figue | berrocal }@usal.es, http://reina.usal.es

Resumen

En este trabajo se describe un método automático para detectar y extraer nombres propios de una extensa colección de documentos en español, con el objetivo de determinar si tal proceso puede aplicarse para mejorar los resultados de recuperación de información, y bajo qué condiciones. La incorporación de mayor información en el proceso de indización, como es en este caso, permite obtener mejores representaciones de los documentos, y por tanto, ello debiera proporcionar mejores resultados en la recuperación. Esto que parece evidente, no parece estar tan claro cuando se analizan los resultados experimentales, al menos en tareas clásicas de recuperación de información. Hemos realizado gran cantidad de experimentos considerando nombres propios simples y compuestos, con diferente valor de ponderación de los nombres propios respecto del resto de palabras indizadas, incluso considerando en la indización los nombres propios compuestos y los simples que los forman, y otras combinaciones. En todos los experimentos hemos comprobado que la detección de nombres propios no mejora los resultados de recuperación, a pesar de que la indización sí se mejora. Los resultados son peores cuando se detectan nombres propios compuestos, debido fundamentalmente a que se introducen variaciones en los nombres de las mismas entidades, si bien, el efecto se reduce al considerar también como términos índice los simples que los forman.

Palabras clave: Indización. Nombres propios. Recuperación de información.

Abstract

In this work an automatic method for proper noun detection in Spanish documents is presented. The objective is to check if it can be applied to improve the retrieval performance. A priori we assume that an indexing process that incorporates more information of the document also provides better retrieval results for classical information retrieval. But our results show that this is not true. A lot of tests were carried out to obtain the best performance for all situations: single proper nouns, compound proper nouns, compound proper nouns plus single proper nouns, different weighting schema for proper nouns, etc. The results were discouraging, as the retrieval performance was deteriorated in all the tests. The worst case is detecting compound proper nouns. The effect is less dramatic if the single nouns of the compound ones are considered.

Keywords: Indexing. Proper nouns. Information retrieval.

1. Introducción

Los mejores modelos de Recuperación de Información (RI) representan los documentos con términos índice, empleando para ello las palabras que aparecen en los mismos. Ahora bien, no todas las palabras de un documento son igual de representativas de su contenido. Tanto si el proceso se realiza de manera manual como automática, interesa obtener aquellas palabras que caractericen mejor el contenido de cada uno de los documentos.

La indización puede verse enriquecida si se detectan ciertas palabras de mayor valor semántico que el resto. Para este cometido son de gran interés las palabras que actúan gramaticalmente como nombres, y especialmente aquéllas que son nombres propios. Ello es más importante cuanto más dependiente de esta característica sea el documento. Un claro ejemplo

son las noticias de prensa. Si analizamos las noticias de cualquier diario escrito vemos que frecuentemente contienen varios nombres propios, junto, claro está, con otras palabras. Intuitivamente podemos determinar el contenido de una noticia sabiendo los nombres propios que contiene, y alguna otra palabra de importancia. Por ejemplo, si en una noticia aparecen los nombres propios *Mariano Rajoy* y *José Luis Rodríguez Zapatero* (o simplemente *Rajoy* y *Zapatero*), rápidamente intuimos que se el documento trata de política, sobre todo si además aparece la palabra *elecciones*.

En ese sentido, a priori parece claro que si incorporamos mayor información en el proceso de indización, por ejemplo, detectando automáticamente nombres propios, obtendremos mejores representaciones de los documentos, y por tanto, ello debiera proporcionar mejores resultados en la recuperación. Esto que parece evidente, no parece estar tan claro cuando se analizan los resultados experimentales, al menos en tareas clásicas de recuperación. Es cierto que en otras tareas, como la extracción de información, la detección automática de nombres propios es un aspecto fundamental, y se consiguen muy buenos resultados. A este extremo véanse las conferencias MUC (*Message Understanding Conferences*) (Voorhees, 2007).

En este trabajo nos hemos propuesto comprobar, realizando una amplia variedad de experimentos, si la detección de nombres propios aplicada a la recuperación clásica de información obtiene buenos resultados, y bajo qué condiciones. Para ello hemos utilizado una amplia colección de documentos y consultas en español. Se trata de una de las colecciones de prueba utilizada en las famosas conferencias CLEF (*Croos Language Evaluation Forum*) (Peters, 2007). En la siguiente sección se describe la colección de documentos y consultas que hemos empleado en nuestros experimentos, así como el sistema de recuperación de información utilizado. A continuación describimos el proceso que marca automáticamente los nombres propios que aparecen en un texto en español. En la cuarta sección describimos los experimentos realizados, con los resultados obtenidos. Finalizamos con las conclusiones, y el trabajo futuro.

2. La colección de pruebas

Para nuestros experimentos hemos utilizado una colección de 215.718 documentos en español, procedentes de la Agencia EFE, de todas las noticias generadas en el año 1994 (534 MB de información). Esta colección puede ser utilizada para diferentes pruebas en el marco de las conferencias CLEF. Nuestro grupo de investigación lleva participando en estas conferencias desde que se iniciaron en Lisboa en septiembre de 2000.

Cada documento de la colección posee varios campos, tal como puede verse en el ejemplo de la Figura 1. Los más importantes son los campos TITLE y TEXT, si bien, el texto del campo TITLE siempre aparece en mayúsculas, resultando inservible para la detección de nombres propios.

```
<DOC>
<DOCNO>EFE19940101-00004</DOCNO>
<DOCID>EFE19940101-00004</DOCID>
<DATE>19940101</DATE>
<TIME>02.19</TIME>
<SCATE>ECO</SCATE>
<FICHEROS>94F.JPG</FICHEROS>
<DESTINO>ICX MUN EXG</DESTINO>
<CATEGORY>ECONOMIA</CATEGORY>
<CLAVE>DP2406</CLAVE>
<NUM>82</NUM>
<PRIORIDAD>U</PRIORIDAD>
<TITLE> CHINA-TASA CAMBIO
NUEVO CAMBIO UNICO Y FLOTANTE PARA EL YUAN
</TITLE>
<TEXT> Pekín, 1 ene (EFE).- China sustituyó hoy, sábado, su doble sistema de cambio monetario y fijó su primera y única tasa flotante para el yuan, que pasó a tener un valor de 8,7 por dólar, informó la agencia de prensa Nueva China. La tasa de cambio oficial china era hasta ayer de 5,8 yuanes por dólar para residentes extranjeros y turistas, y una tasa flotante fijada en función de la oferta y la demanda, sobre el mercado Swap, para las empresas. EFE
FMR
01/01/02-19/94
</TEXT>
</DOC>
```

Figura 1. *Un documento de la colección.*

En cada campaña de CLEF se proporciona una batería de unas 50 preguntas. Hemos tomado las de la campaña de 2002, dado que fue el año en el que más grupos participaron, y por tanto, los juicios de relevancia fueron los más acertados. Cada consulta posee tres campos (véase el ejemplo de la Figura 2): ES-title (*título*), ES-desc (*descripción*) y ES-narr (*narrativa*), con la intención de que los investigadores experimenten sus modelos de recuperación considerando consultas de diferente longitud. Esta ha sido también nuestra intención y hemos realizado los experimentos con el campo *título* y con el cam-

po *descripción*, es decir, consultas cortas y consultas medianamente largas.

<top>
<num> C112 </num>
<ES-title> Pulp Fiction </ES-title>
<ES-desc> ¿Qué premio ganó la película "Pulp Fiction", dirigida por Quentin Tarantino, en el Festival de Cine de Cannes? </ES-desc>
<ES-narr> El film Pulp Fiction, con John Travolta, recibió un premio en el festival de Cannes. Los documentos relevantes deben dar el nombre del premio asociado con esa cinta. </ES-narr>
</top>

Figura 2. *Una consulta de la colección.*

2.1. El motor de recuperación

Nuestro motor de recuperación utiliza el conocido modelo vectorial (Salton, 1975), empleando lenguaje natural para la entrada de consultas. En el modelo vectorial, cada documento de la colección se representa por un vector de *m* componentes, que se corresponden con el conjunto de los *m* términos índice de la colección. La consulta se representa igualmente por un vector con el mismo número de elementos. Cada elemento del vector posee un peso, que expresa la importancia del término correspondiente para describir el documento o la consulta. El sistema utiliza el producto escalar de los vectores que representan a documentos y consulta para obtener un ranking de documentos ordenados por similitud.

Existen multitud de mecanismos para obtener el peso asociado a un término, aunque en general se aplica el clásico mecanismo de pesado TF-IDF (Salton, 1973). En este modelo se conjuga la capacidad que tiene un término para representar el contenido del documento de acuerdo a su frecuencia de aparición en dicho documento con el poder de dicho término para discriminar un documento de otro. Así, si un término se repite mucho en un documento se considera muy representativo de su contenido. Pero se realiza un ajuste en función de cuán frecuente es ese término en la colección documental. Un término que aparece en casi todos los documentos es muy poco efectivo para distinguir un documento de otro.

2.2. Módulo de indización

Nuestro motor de recuperación incluye un módulo de indización que realiza el procesado automático del texto de los documentos y consultas hasta obtener los términos índice que los representan. El módulo de indización realiza las siguientes acciones: en primer lugar quita los acentos a las vocales acentuadas y elimina las palabras casi vacías de contenido semántico (preposiciones, artículos, conjunciones, etc.) o aquellas que se repiten mucho en la colección documental; todas ellas forman la lista de palabras vacías. Seguidamente aplica un proceso de detección de nombres propios, que será descrito en el siguiente apartado. Después aplica un proceso de lematización consistente en eliminar las terminaciones «-as», «-es», «-os», «-a», «-e» y «-o» de todas las palabras, salvo las marcadas como nombres propios. La lematización consiste en reducir una palabra a una forma simple para remitir a ella a las de su misma familia por razones de economía. Aunque su intención es eliminar variantes de género y número, en realidad este proceso de lematización no sigue ninguna regla lingüística, pues se aplica sobre cualquier tipo de palabra sin considerar su categoría gramatical. La utilizamos porque hemos comprobado que mejora notablemente la recuperación (Figuerola et al., 2004).

3. Detección de nombres propios

En español los nombres propios empiezan con mayúscula, por eso el proceso automático que detecta nombres propios supone que el texto de entrada está en minúscula, salvo inicio de oración o presencia de nombres propios. También pueden encontrarse otro tipo de palabras, como siglas o abreviaturas que pueden estar en mayúsculas.

La detección de nombres propios no es un problema sencillo de resolver, pues a menudo encontramos palabras que empiezan en mayúsculas por otros motivos, como por ejemplo, para realizar énfasis, o en estructuras que no forman oración por sí mismas (detrás de dos puntos o en un texto entrecomillado). Este es uno de los problemas con los que nos hemos encontrado en nuestros experimentos.

En general, dado que la colección documental proviene de textos periodísticos, no suelen emplearse mayúsculas para resaltar contenidos. Sin embargo, nos encontramos frecuentemente con palabras que comienzan en mayúsculas después de dos puntos y en texto entrecomillado. Para solucionar este problema hemos decidido ampliar el número de caracteres que pueden ser considerarlos separadores de oraciones.

Además del problema que acabamos de indicar, tenemos dos tipos de nombres propios: simples y compuestos. Los nombres propios simples son los formados por una sola palabra. Los compuestos pueden estar formados por una

sucesión de nombres propios simples (José María López), o pueden formarse con la inclusión de los nexos «de/del», «de la/las/los», e «y» entre nombres propios simples (Juan de Austria, Mar del Río, Cereceda de la Sierra, Palacio de los Castro, Juan de las Indias, Construcciones y Contratas). La inclusión del nexo «y» suele llevar a innumerables errores, por eso en nuestros experimentos no hemos detectado nombres propios unidos por este nexo.

Para marcar un nombre propio hemos utilizado llaves en los extremos, y si éste es compuesto, se ha intercalado un guión de subrayado «_» como carácter de unión. Por ejemplo, {Luis}, {José_María_López}, {Cereceda_de_la_Sierra}.

El criterio general para decidir si una palabra es nombre propio es que empiece por mayúscula. Esto es cierto si la palabra no está al principio de una oración, que es donde se puede dar la ambigüedad. Así pues, el primer paso es dividir el texto de entrada en oraciones. El separador de oraciones por excelencia es el punto. En menor medida también lo son el cierre de admiración e interrogación cuando cierran un enunciado. En gran cantidad de casos de texto entrecomillado, como es lo típico en textos periodísticos, es también habitual encontrarse con otros signos que actúan de separadores, como los dos puntos, las comillas dobles y simples, el apóstrofo, o incluso otros. En nuestros experimentos hemos supuesto que todos los signos que acabamos de mencionar son separadores de oraciones.

Una vez dividido el texto en oraciones es necesario reconocer los nombres propios; pero existen nombres propios que contienen caracteres especiales como el guión (rally París-Dakar) o el apostrofo (generalmente en nombres propios en otros idiomas: *d'Orsay*, *N'sang*, *D'Aubuisson*, *Bar'am*, *Aujourd'hui*). Para los primeros se ha considerado que el guión une nombres propios simples; para los segundos, dado que dependen de reglas de formación propias de cada idioma, simplemente se ha sustituido el apóstrofo por un espacio en blanco (*d Orsay*).

Los nombres propios empiezan en mayúscula y su detección en mitad de la oración no plantea problema, una vez hecha la salvedad de las situaciones indicadas, que podemos resolver utilizando más separadores de oraciones. La ambigüedad se produce al inicio de la oración, pues la palabra puede ser nombre propio. La detección se complica porque se deben recoger también nombres compuestos.

Para resolver la ambigüedad se necesitan dos diccionarios, uno de nombres propios (simples y compuestos) y otro de palabras de inicio de oración que no pueden ser nombres propios. A este diccionario lo hemos denominado *diccionario de no propios*. El proceso que sigue el sistema cuando detecta un posible nombre propio es chequear primero el diccionario de nombres propios y después el de no propios. Veamos algunos ejemplos.

En la oración «Los hombres de […]», el algoritmo detectará «Los» como candidato a nombre propio. En primer lugar, se chequea el diccionario de nombres propios. En este caso «Los» no está en ese diccionario; pero puede ser un nombre propio no recogido en el mismo, y no podemos asegurar nada. Así pues, se chequea el diccionario de nombres no propios: se pasa a minúscula el término y se comprueba si está en el diccionario. En este caso «los» está, lo cual indica que no puede ser nombre propio, de manera que no se marca como tal.

Veamos otro ejemplo. Cuando el algoritmo reciba la oración «La Coruña tiene [...]», detectará como candidato a nombre propio «La_Coruña». Se pasa entonces a comprobar si está en el diccionario de nombres propios, en el que efectivamente se encuentra, de modo que «La Coruña» se marca como nombre propio: «{La_Coruña} tiene [...]».

En el ejemplo «La Universidad de Jaén posee [...]», el algoritmo detectará como posible nombre propio «La_Universidad_de_Jaén». Se comprueba primero si está en el diccionario de nombres propios; pero no aparece (seguramente sí estará «Universidad_de_Jaén», pero no con el artículo), entonces se comprueba que la palabra inicial «la», que es la ambigua en la detección al comienzo de la oración, está en el diccionario de no propios, como efectivamente es, de modo que «La» se quita del nombre y se marca el resto, «La {Universidad_de_Jaén} posee [...]».

Frecuentemente, se dará el caso en que la primera palabra de la oración no esté en ninguno de los dos diccionarios. Tenemos dos alternativas. La primera es considerar al candidato propuesto como nombre propio. Es la opción más conservadora. Pero obliga a tener un diccionario de nombres no propios muy grande para evitar tener mucho ruido (muchos nombres propios marcados sin serlo). La segunda es no considerar al candidato propuesto como nombre propio. Es una opción muy restringida, y ello obliga a tener un diccionario de nombres propios muy extenso, para evitar tener mucho silencio (muchos nombres propios no marcados).

En nuestro caso, dado que los documentos son noticias de prensa, hemos optado por la primera opción, pues el número potencial de nombres propios es muy elevado.

Vemos que es muy importante la construcción de los dos diccionarios. En particular, en el diccionario de palabras que no pueden ser nombres propios deben estar incluidas las palabras vacías de la colección. En la práctica también es importante un diccionario de nombres propios que empiecen por palabra vacía (La_Coruña, El_Mundo, Los_Ángeles, etc.).

Hay que decir que el trabajo para crear los diccionarios de nombres propios y no propios, necesarios para este proceso, ha supuesto un enorme esfuerzo de varios días de revisión manual. Aunque parte del proceso se realizó automáticamente (obteniendo nombres propios en situaciones no comprometidas, es decir, cuando no están al comienzo de las frases), las tareas de revisión de ambos diccionarios han constituido una labor larga y pesada.

3.1. Términos índice de nombres propios

Finalmente, sólo queda determinar, una vez que se han marcado los nombres propios, cómo se almacenan éstos al ser tratados como términos índice. Dado que nuestro motor de recuperación acepta consultas en lenguaje natural, nuestro criterio ha sido convertir todos los términos índice en minúscula, salvo aquéllos expresamente marcados como nombres propios, que se almacenan en mayúsculas.

Solamente queda resolver la duda de cómo almacenar los nombres compuestos. Por ejemplo, dado un nombre propio como «Villaseco de los Gamitos», ¿debería tomarse solamente el nombre compuesto como término índice, «VILLASECO_DE_LOS_GAMITOS», o también los nombres simples que componen el compuesto, esto es, «VILLASECO» y «GAMITOS»? Démonos cuenta de que esta situación se agrava en lenguaje periodístico, pues es habitual referirse a las personas por sus apellidos («Zapatero» por «José Luis Rodríguez Zapatero»). Varios autores que han trabajado con detección de grupos nominales (Fajan, 1989; Krovetz, 1997) indican que deben considerarse términos índice los nombres propios compuestos y también los simples que forman un compuesto, aunque debe determinarse el peso asociado a cada uno de ellos.

3.2. Detección de siglas y abreviaturas

Un apunte más, la detección de nombres propios debe incorporar también la detección de siglas y abreviaturas en mayúsculas. En nuestros experimentos hemos supuesto que se trata de casos particulares de nombres propios que poseen todos sus caracteres en mayúscula. No obstante, su proceso de detección debe aplicarse por separado, dado que se puede incurrir en errores, sobre todo por los nexos, como, por ejemplo, en el texto «[...] leí el ABC de Juan que estaba [...]», que diera {ABC_de_Juan} como nombre propio.

En nuestro estudio, hemos considerado que las siglas están formadas por las iniciales en mayúsculas de las palabras de las que deriva. Si entre las mismas se han colocado puntos, se trata de abreviaturas. No obstante, la distinción no es tan clara en la colección documental, pues las reglas de estilo no eran las mismas que ahora (recordemos que se trata de documentos escritos en el año 1994). En la colección documental hemos encontrado situaciones de siglas con y sin punto (CSIC y C.S.I.C.), y también abreviaturas con y sin puntos (S.A.R. y SAR, por Su Alteza Real), por eso hemos decidido marcar ambas situaciones de la misma manera.

Hemos usado el mismo criterio que para el marcado de nombres propios, esto es, utilizar llaves. Así, por ejemplo, la sigla «AB» y la abreviatura «A.B.» quedan marcadas como {AB} en el proceso. Cuando se trata de un solo carácter en mayúsculas, solo se considera abreviatura si lleva punto, «A.», marcándose como tal «{A}.».

Aunque hoy día las abreviaturas de términos en mayúscula plural se escriben con punto y espacios en blanco (FF. AA., por Fuerzas Armadas), en la colección documental nos encontramos con abreviaturas en plural sin puntos, como en «CC OO» o «EE UU». El proceso de detección de siglas y abreviaturas se ha configurado para que den como resultado {CCOO} y {EEUU}.

Cuando la sigla no contenga puntos, pero acabe en punto, se tratará de fin de oración, de modo que se conservará éste («[…] ABC.» quedará marcado como «[…] {ABC}.»).

Hemos detectado varias situaciones en las que las siglas están construidas incorrectamente. Un caso muy frecuente es el de las siglas en plural con incorporación de una *s* minúscula, derivado de la costumbre inglesa, seguido o no de apóstrofo, como por ejemplo, «ONGs» y «PC's». Estas siglas deben marcarse como {ONG} y {PC}. También se da el caso de abreviaturas sin punto final (A.B, A.B.C, AA.BB), que deben quedar marcadas como {AB}, {ABC} y {AABB}.

4. Experimentos

En nuestras pruebas hemos utilizado solamente el campo TEXT de los documentos (véase la Figura 1, anteriormente), pues es el que permite la detección de nombres propios. Hay que decir que, para obtener conclusiones más robustas,

hemos ejecutado los experimentos tanto con el campo *título* de las consultas, como con el campo *descripción* (véase la Figura 2).

Se han realizado las pruebas que se describen a continuación. La intención es analizar no solamente la forma de considerar los nombres simples y compuestos como términos índice, sino también la ponderación asociada.

- *Prueba 0*. Se trata de la prueba base en la que no hay detección de nombres propios, que nos sirve de referencia para determinar la mejoría, de existir, del resto de pruebas.
- *Prueba A*. Se trata de considerar como términos índice los nombres propios simples y compuestos tal como se marcan en los documentos y consultas. El valor de ponderación para estos términos será igual que para el resto de términos, es decir, no hay ponderación diferente.
- *Prueba B0*. Se trata de considerar como términos índice los nombres propios simples, incluso cuando forman parte de nombres compuestos. No se consideran nombres compuestos. El valor de ponderación para estos términos será igual que para el resto de términos, es decir, no hay ponderación diferente.
- *Prueba B1*. Es igual que la Prueba B0, pero se aplica una ponderación 1,5 veces superior a los nombres propios respecto del resto de términos índice.
- *Prueba B2*. Es igual que la Prueba B0, pero se aplica ponderación doble a los nombres propios.
- *Prueba C0*. Se trata de considerar como términos índice los nombres propios simples y compuestos, y también los simples que forman los compuestos, y con igual peso que éstos. No hay distinción en la ponderación de los nombres propios.
- *Prueba C1*. Es igual que la Prueba C0, pero se aplica una ponderación 1,5 veces superior a los nombres propios.
- *Prueba C2*. Es igual que la Prueba C0, pero se aplica una ponderación doble a los nombres propios.
- *Prueba D0*. Se trata de considerar como términos índice los nombres propios simples y compuestos, y también los simples que forman los compuestos, pero con mitad de ponderación que éstos.
- *Prueba D1*. Es igual que la Prueba D0, pero se aplica una ponderación a los nombres propios superior en 1,5 veces al resto de términos. Los nombres propios simples que forman parte de compuestos mantienen su relación de ponderación con el compuesto.
- Prueba D2. Es igual que la Prueba D0, pero se aplica una ponderación doble a los nombres propios.

En la Tabla 1 podemos ver las características de estas pruebas.

Prueba	**0**	**A**	**B**	**C y D**
Términos índice en los documentos (campo TEXT)				
Total	279.088	621.651	325.893	694.944
Media	114,1	109,8	116,7	124,6
Desviación	56,0	54,56	57,5	61,7
Máximo	658	651	675	764
Mínimo	0	0	0	0
Fichero índice	193MB	193MB	197MB	210MB
Términos índice en las consultas (campo *título*)				
Total	122	116	124	133
Media	2,6	2,5	2,7	2,8
Desviación	0,8	0,8	0,8	1,0
Máximo	5	5	5	5
Mínimo	1	1	1	1
Términos índice en las consultas (campo *descripción*)				
Total	291	274	297	318
Media	8,4	7,9	8,4	8,8
Desviación	2,6	2,7	2,6	2,8
Máximo	15	15	15	15
Mínimo	4	4	4	4

Tabla I. *Características de la colección en las diferentes pruebas.*

Podemos ver que el número de términos aumenta considerablemente cuando se consideran nombres propios tal como aparecen, simples y compuestos (Prueba A). Al incluir exclusivamente nombres simples, incluso cuando forman parte de compuestos (Pruebas B), el número de términos aumenta en más del 15%. Cuando se incluyen además los nombres compuestos y los simples que los forman (Pruebas C y Pruebas D), el número de términos índice de la colección se duplica. Asimismo, también es mayor el número medio de términos índice por documento.

4.1. Evaluación

Para evaluar los resultados de los experimentos hemos utilizado las medidas habituales en los sistemas de recuperación de información: precisión media (media ponderada respecto de la precisión de cada pregunta), y los diagramas precisión-exhaustividad. También hemos utiliza-

do la precisión a 10 documentos vistos (normalmente se denota como P@10). Esta medida tiene mucha importancia en los motores de recuperación que prestan sus servicios en entornos reales. Varios estudios en este tipo de sistemas muestran que la mayoría de usuarios no se interesan más que por la primera pantalla de resultados, es decir, aproximadamente los primeros diez documentos, y de éstos, solamente ven el contenido de unos ocho (Jansen et al, 2000; Jones et al., 2000). Por ello, es de especial interés la medida de precisión P@10.

5. Resultados

En la Tabla 2 están los resultados experimentales para los dos campos de las consultas. Podemos ver que la detección de nombres propios no sólo no mejora los resultados respecto de la *Prueba 0*, sino que en general los empeora. Los resultados menos malos se consiguen considerando términos índice únicamente los nombres propios simples, incluso cuando forman parte de nombres compuestos (*Pruebas B*). Además tampoco parece que la modificación en la ponderación de los nombres propios proporcione una mejora de la recuperación.

Prueba	**0**	**A**	**B0**	**B1**	**B2**	**C0**	**C1**	**C2**	**D0**	**D1**	**D2**
Campo *título* de las consultas											
Prec. media	*0,4040*	-16,7	-6,8	-7,1	-7,2	-9,1	-10,4	-10,5	-35,4	-37,4	-9,9
P@10	*0,5300*	-8,3	-0,8	-2,3	-2,6	-3,8	-4,2	-5,7	-25,7	-28,3	-6,4
Campo *descripción* de las consultas											
Prec. media	*0,4129*	-16,0	-1,0	-0,8	0,0	-5,4	-6,3	-5,0	-60,2	-59,4	-4,4
P@10	*0,5160*	-6,6	1,2	2,7	0,8	-3,5	-3,1	-4,3	-51,6	-51,6	-3,5

Tabla II. *Resultados de las pruebas de detección de nombres propios. Se han indicado los porcentajes de variación respecto de la Prueba 0.*

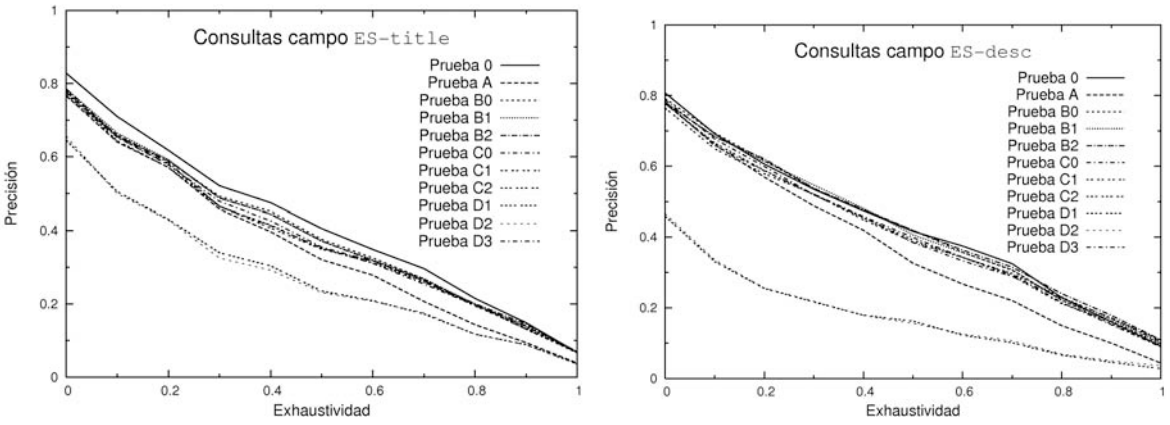

Figura 3. *Comparación de pruebas de detección de nombres propios.*

En la Figura 3 se muestran las curvas precisión-exhaustividad para los experimentos. Vemos que para el campo *título* (ES-Title) de las consultas ninguna prueba de detección de nombres propios supera a la prueba sin detectarlos. Para el campo *descripción* (ES-desc) alguna curva es superior en algunos momentos, pero no tanto como para recomendar su uso, sobre todo si tenemos en cuenta el coste computacional que ello supone: proceso de detección de nombres propios, aumento de términos índice, aumento del tamaño del fichero inverso, y consecuentemente también del tiempo de respuesta del sistema. En resumen, podemos decir que la detección de nombres propios no mejora la recuperación, al revés, la empeora.

Ello ratifica que a menudo la aplicación de técnicas lingüísticas en Recuperación de Información no siempre ofrece los resultados que se esperan de ellas. Después de todo el largo tiempo y esfuerzo invertido en la construcción del sistema de detección de nombres propios, siglas y abreviaturas, y en la creación de los diccionarios necesarios, vemos que no hay mejoría en la recuperación.

6. Conclusiones

En este trabajo hemos realizado una leve incursión en el campo del procesamiento del lenguaje natural (PLN), con la intención de mejorar el proceso de indización automática de documentos y consultas. Se ha tratado de identificar expresiones multipalabra. En particular, hemos descrito un mecanismo para la detección de nombres propios en español, y hemos analizado, utilizando una colección documental de gran tamaño, su repercusión en Recuperación de Información. Para ello, hemos analizado diferentes formas de contabilizar nombres propios simples y compuestos, y diferentes maneras de ponderarlos. En todos los experimentos hemos comprobado que la detección de nombres propios no mejora los resultados de recuperación, a pesar de que la indización sí se mejora. Los resultados son peores cuando se detectan nombres propios compuestos, debido fundamentalmente a que se introducen variaciones en los nombres de las mismas entidades; si bien, el efecto se reduce al considerar también como términos índice los simples que los forman.

Aunque en otros campos de la Recuperación de Información la detección automática de nombres propios obtiene muy buenos resultados, como por ejemplo, en la extracción de información, en nuestros experimentos de recuperación clásica no esperábamos estos resultados, aunque sí los temíamos, pues son muchos los artículos que han demostrado la nula o poco positiva influencia de las técnicas de PLN aplicadas a tareas clásicas de RI; véanse, por ejemplo, los artículos de Voorhess (1994) o Lewis y Voorhees (1996). A menudo, ello se ha achacado al alto coste computacional que supone llevar los resultados de laboratorio a grandes sistemas reales (véanse las conclusiones de Strzalkowski et al. (1999) y Gonzalo et al. [2002]), o por la falta de perfeccionamiento de las técnicas lingüísticas. Sin embargo, el algoritmo de detección de nombres propios es lo suficientemente robusto y fácil de llevar a la práctica, como para haber esperado mejores resultados, máxime tratándose de un proceso de desambiguación terminológica.

Hay que decir, y esto es un aspecto en el que ya estamos trabajando, que en nuestro algoritmo falta un proceso de normalización de términos, el de unificación. La unificación terminológica debe darse en una doble vertiente; en primer lugar, son muy frecuentes los errores de tecleado en nombres propios, máxime si éstos son extranjeros. En segundo lugar, es muy común, sobre todo en lenguaje periodístico, omitir alguna parte del nombre propio cuando la persona o personaje es muy conocido. No obstante, creemos que el perfeccionamiento de la subrutina de detección de nombres propios no va a provocar un cambio sustancial en la recuperación, a la vista de los resultados que hemos obtenido en este trabajo.

Referencias

Fagan, Joel (1989). The effectiveness of a non-syntactic approach to automatic phrase indexing for document retrieval. // Journal of the American Society for Information Science. 40:2 (1989) 115-132.

Figuerola, Carlos G.; Zazo, Ángel F.; Rodríguez, Emilio; Alonso Berrocal, José Luis (2004). La recuperación de información en español y la normalización de términos. // Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial. 8:22 (2004) 135-45.

Gonzalo, Julio; Peñas, Anselmo; Verdejo, Felisa (2002). La indexación con técnicas lingüísticas en el modelo clásico de recuperación de información. // Sanchís, E.; Moreno, I; Gil, I. (ed.). Primeras Jornadas de Tratamiento y Recuperación de Información JOTRI-2002. Valencia: Universidad Politécnica de Valencia (2002) 97-106.

Jansen, Bernard J.; Spink; Amanda; Saracevic, Tefko (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. // Information Processing & Management. 36:2 (2000) 207-227.

Jones, Steve; Cunningham, Sally Jo; McNab, Rodger J.; Boddie, Stefan (2000). A transaction log analysis of a digital library. // International Journal on Digital Libraries. 3:2 (2000) 152-169.

Krovetz, Robert (1997). Homonymy and polysemy in information retrieval. // 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. Canadá: ACL / Morgan Kaufmann, (1997) 72-79.

Lewis, David; Voorhees, Ellen (1996). Natural language processing for information retrieval. // Communications of the ACM, 39:1 (1996) 92-101.

Peters, Carol (2007). Cross Language Evaluation Forum (CLEF) [en línea], http://www.clef-campaign.org/ (20-03-2007).

Salton, Gerard; Yang, C.S. (1973). On the specification of term values in automatic indexing. // Journal of Documentation. 29:4 (1973) 351-372.

Strzalkowski, Tomek; Perez-Carballo, Jose; Karlgren, Jussi; Hulth, Anette; Tapanainen, Pasi; Latineen, Timo (1999). Natural language information retrieval: TREC-8 report. // Voorhees; E.; Harman, D. (ed.). The Eighth Text REtrieval Conference (TREC 8). NIST Special Publication. 500-246. (1999).

Voorhees, Ellen (1994). Query expansion using lexical-semantic relations. // Croft, Bruce; van Rijsbergen, C.J. (ed.). Proceedings of the 17th Annual International ACM-SIGIR. Dublín (Ireland): ACM/Springer, (1994) 61-69.

Voorhees, Ellen (2007). Message Understanding Conference (MUC) [en línea]. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/ (20-03-2007).