La investigación sobre Recuperación de la Información en español

Carlos G. Figuerola, Facultad de Documentación, Universidad de Salamanca
Paseo de S. Vicente s/n, 37007 Salamanca
tf.: +34 (9)23 294580 fax +34 (9)23 294582
e-mail: figue@gugu.usal.es

Resumen:
La Recuperación de la Información es un área de investigación fuertemente relacionado con las Ciencias de la Documentación y con la Informática. Los principales modelos propuestos dependen en mayor o menor medida del idioma, tanto de los documentos como de las consultas que se formulan. Pese a ser el español una de las lenguas más habladas, el trabajo de investigación y experimentación en Recuperación de la Información sobre colecciones documentales en español es poco abundante. Se pasa revista a los principales trabajos publicados en esta línea, se comentan los principales problemas encontrados, así como los resultados obtenidos.

1. Introducción

La Recuperación de la Información (*Information Retrieval* o *IR* en inglés) es un campo fuertemente relacionado con las Ciencias de la Documentación y con la Informática que, independientemente del contenido informal que queramos encontrar en la expresión, hace referencia a un área temática perfectamente definida y con una historia o tradición de investigación de más de treinta años. El concepto y contenidos de la IR aparecen perfectamente delimitados en diversos trabajos, entre los cuales se puede destacar la conocida obra de Rijsbergen (Rijsbergen, 1979), o algunos de los de Salton (Salton y McGill, 1983) y (Salton, 1989).

Algunos de los objetivos de la investigación en IR pueden resumirse en:

a) la recuperación de registros poco o nada estructurados, como pueden ser documentos a texto completo, o al menos resúmenes de los mismos. Se entiende que se trata de recuperación en función de los temas o contenidos de éstos, y no de otros atributos como autor(es), fechas, editoriales, revistas, etc.., fácilmente tratables como campos de cualquier sistema estándar de gestión de bases de datos.

b) consultas formulables en lenguaje natural, o al menos de un modo próximo a éste

c) discriminación o evaluación del grado de ajuste del documento recuperado a la consulta formulada o a las necesidades informativas del usuario; estas dos cosas no son necesariamente equivalentes.

Como se ha mencionado antes, la investigación en IR data ya de hace bastantes años. Un buen resumen de su evolución, hasta mediados de la

década de los 80 puede encontrarse en el trabajo de Belkin y Corft (Belkin, N. y Croft, W.B. , 1987) . Hasta esta época, aproximadamente, se producen los modelos teóricos y se plantean diversas alternativas técnicas que son la base de muchos sistemas de recuperación operativos actualmente.

No es preciso insistir en la importancia que ha adquirido la IR; cada vez hay más cantidad de documentos legibles por ordenador. Buena parte de lo que hoy se imprime en papel se escribe y diseña mediante ordenador, quedando también grabado electrónicamente. Además, existe una cantidad creciente de documentos que ni siquiera se imprimen, puesto que su uso y distribución se realiza a través de sistemas informáticos. El Web es un buen ejemplo de esto.

La mayor parte de las técnicas propuestas por los investigadores en *IR* utilizan, en mayor o menor medida, algún tipo de conocimiento lingüístico. Resulta paradójico, en este sentido, que siendo el español una de las lenguas más extendidas, la investigación en recuperación de la información en español resulte tan exigua. Esto es cierto, desde luego, para lo que se refiere a la formulación de modelos generales teóricos, pero también para lo que es investigación experimental, en el sentido de la producción y ajuste de técnicas que implementen lo propuesto en los modelos teóricos.

Así, no contamos en España con demasiados trabajos relacionados con la Recuperación de la Información, entendida tal como hemos expresado más arriba. Entre los existentes, cabe citar el de Valle Bracero y Fernández García (Valle Bracero, A. y Fernández García, J.A., 1983) , así como el de Simón Granda y Lema Garzón (Simón Granda, J. y Lema Garzón, E., 1983) . En la misma línea de interés hacia la indización automática están los de Gil Leiva y

Rodríguez Muñoz (Gil Leiva, I. y Rodríguez Muñoz, J.V., 1996 y 1997). Igualmente, aunque desde un punto de vista más descriptivo, cabe destacar el artículo de Codina (Codina, Ll., 1995).

2. Las conferencias TREC

Probablemente, la investigación más importante realizada sobre recuperación de información en documentos en español ha sido la llevada a cabo en algunas de las conferencias TREC (*Text REtrieval Conference*). En realidad, puede afirmarse que las TREC constituyen uno de los esfuerzos más significativos de investigación experimental en IR en general realizado hasta la fecha. Auspiciadas por el *National Institute of Standards and Technology* (NIST) y por la *Defense Advanced Research Projects Agency* (DARPA) norteamericanas, dichas conferencias comenzaron en 1992 (TREC-1) y vienen celebrándose con periodicidad anual hasta la fecha, de manera que en 1998 se ha celebrado TREC-7. Básicamente, el objetivo de las conferencias es el siguiente: ofrecer a los investigadores participantes una amplia colección de documentos (varios Gigabytes), una colección de consultas, unos sistemas normalizados de presentación y evaluación de resultados, y la propuesta de varias tareas o 'tracks' que los investigadores deben llevar a cabo: consultas ad-hoc, filtrado o encaminamiento de documentos, etc.. La idea es poder establecer comparaciones fiables entre los distintos sistemas empleados por los investigadores participantes, dado que todos operan con las mismas colecciones y las mismas consultas, y presentan sus resultados en la misma

forma; obviamente, utilizan sistemas y técnicas diferentes, que es lo que se trata de comparar (Harman, D.K., 1994).

Lo interesante, desde el punto de vista del presente trabajo, de las conferencias TREC, es que en algunas de ellas se introdujo una colección de documentos en español, con su correspondiente colección de consultas, también en español; planteándose como una de las tareas de la conferencia la recuperación en español. La primera conferencia en la que se introdujo la *spanish track* fue TREC-3, continuando durante TREC-4 y TREC-5. Las actas de TREC-6 no han sido publicadas en el momento de redactar este trabajo, y, aunque TREC-7 se ha celebrado recientemente (NIST, 1998), parece que no ha habido en ella ninguna tarea específica para documentos en español.

En lo que sigue del presente artículo, intentaremos señalar las principales dificultades encontradas por los investigadores al aplicar sus técnicas de recuperación a una colección de documentos en español. En este sentido, es importante precisar que, salvo pequeñas excepciones, casi todos emplearon con la colección en español las mismas técnicas utilizadas con el inglés, con las modificaciones, claro está, derivadas de la propia lengua. Una descripción de las técnicas utilizadas en *IR* está, desde luego, fuera del alcance de este artículo; el lector puede recurrir a obras referenciadas más arriba, especialmente (Rijsbergen, K. Van, 1979), (Salton, G. y McGill, M., 1983), (Belkin, N. y Croft, B.W., 1987) e incluso (Codina, Ll., 1995). Sin embargo, y debido a la importancia que revisten algunos extremos de estas técnicas a la hora de apreciar las diferencias lingüísticas, permítasenos detenernos un poco en algunos aspectos de dichas técnicas.

Todas, en mayor o menor medida, toman en cuenta la distribución de

frecuencias de términos (individuales o complejos) en la colección de documentos y en cada documento en particular, así como en cada consulta. En función de diversos factores (propios de las distintas técnicas aplicadas), cada término (individual o complejo) adquiere una importancia determinada en cada documento en que aparece; esta importancia suele expresarse mediante un valor numérico, conocido como peso de ese término en ese documento. Lo mismo sucede con los términos que conforman cada pregunta o consulta que se hace a la colección de documentos. La resolución de tales consultas es un proceso en el cual se calcula mediante diversos sistemas un coeficiente o tasa de similaridad entre dicha consulta y cada uno de los documentos de la colección; y esta similaridad se computa a partir, precisamente, de los pesos de los términos comunes a consulta y documento. Naturalmente, los documentos de mayor coeficiente de similaridad son los que el sistema de recuperación considera más relevantes.

3. Palabras Vacías

Uno de los problemas mayores encontrados es el de las palabras vacías. En efecto, parece que la construcción de listas de palabras vacías en español representa un escollo importante; una razón para ello es, desde luego, el desconocimiento de la lengua por parte de los investigadores, todos ellos angloparlantes. Pero otro es la carencia (o su desconocimiento, al menos) de corpus y estudios estadísticos para el idioma español.

En contra de lo que pudiera pensarse, la eliminación o no de palabras

vacías no es simplemente una cuestión de tamaños de índices y ficheros invertidos (y de tiempo de procesamiento); más aún, con el precio que la memoria tiene actualmente, éste sería un problema nimio. El problema radica en que, con sistemas que atribuyen pesos a términos, y que operan con éstos, las palabras vacías introducen un factor de ruido considerable. Veamos un ejemplo: un sistema 'clásico' de cálculo de pesos de términos, utilizado ampliamente en *IR* (Harman, D., 1992) , aplica la idea de que la capacidad discriminatoria de un término es inversamente proporcional a su frecuencia. Probablemente el coeficiente más utilizado para expresar esto es el conocido como *Inverse Document Frequency* (IDF), una de cuyas fórmulas más habituales es (Sparck Jones, K., 1972):

$$IDF_t = \log_2 \frac{N}{n_t} + 1$$

donde

t es el término en cuestión

N es el número de documentos en la colección

n_t es el número de documentos en que aparece el término *t*

Otras variantes para el cálculo de la IDF puede encontrarse en el trabajo de Harman ya citado (Harman, D., 1992). Dado, sin embargo, que conviene tener también en cuenta la frecuencia de un término determinado en un documento cualquiera, se suele buscar algún tipo de relación entre la frecuencia general de un término en toda la colección y su frecuencia en un documento determinado. Es decir, que se busca asignar un peso o importancia a cada uno de los términos, pero dentro de cada documento. Así, un término

podría ser poco significativo acerca del contenido de un documento determinado, pero el mismo término, sin embargo, podría resultar más representativo en otro documento diferente.

La forma más estándar de calcular dicho peso consiste en multiplicar la frecuencia del término en el documento en cuestión por su IDF (Salton, G., 1987):

$$peso_{td} = frec_{td} \times IDF_{t}$$

El Cuadro 1 muestra algunos de los términos pertenecientes a un registro de una colección de documentos experimental; los documentos están en español, y consisten en resúmenes de artículos sobre diversos temas relacionados con las Ciencias de la Documentación (Gómez Díaz, R. y López de San Román, E. 1998). Junto con los términos, aparecen sus frecuencias en ese documento y su IDF (correspondiente, naturalmente, a toda la colección). Para cada término aparece también su peso, calculado en la forma descrita más arriba. Los términos están ordenados decrecientemente en función de su peso.

Pues bien, a pesar de utilizar una función inversa como el IDF, que hace disminuir el peso de los términos más frecuentes, que vienen a coincidir con lo que entendemos como palabras vacías, la elevada frecuencia en el propio documento de estas palabras hacen que acaben obteniendo un peso elevado. De una forma apriorística, cabría haber esperado lo contrario: que la aplicación del IDF hubiera dejado con muy poco peso esos términos; sin embargo no es así, y claramente se ve que las palabras vacías introducen un factor

considerable de ruido.

En este punto, la mayor parte de las listas de palabras vacías utilizadas fueron, al menos en los primeros experimentos, bastante pobres. Probablemente una de las listas más elaboradas ha sido la construída por Buckley y sus colegas, los cuales elaboraron una lista de 342 palabras (Buckley, C. y otros, 1994). Dicha lista ha sido ampliamente difundida y utilizada por otros investigadores, distribuyéndose actualmente junto con el célebre programa SMART y las colecciones de prueba que vienen con éste (Cornell, 1998). De otro lado, investigadores interesados en el empleo de expresiones compuestas (al menos por dos palabras), además del uso convencional de términos simples, necesitaron eliminar también expresiones o frases vacías. Este ha sido el caso de Allan y sus colegas (Allan, J. y otros 1995), los cuales trabajaron con la siguiente lista de 'frases vacías':

hay
indicaciones de
cuáles son
cómo van
tendrá
información sobre

Sólo éstas. Nótese, además que estos autores catalogan como *phrases* palabras simples, debido a que en inglés algunas de éstas se traducen por más de una palabra (por ejemplo hay --> 'there is').

4. Lematización

Debido a que, de una forma u otra, la mayor parte de los sistemas de recuperación utilizan las frecuencias de los términos, la existencia de palabras derivadas de otras puede alterar seriamente los resultados de los cálculos. Es deseable, en consecuencia, reducir dichos derivados a un solo término, sea éste una palabra real o una raíz. De esta forma, palabras como bibliotecario, bibliotecología, biblioteconomía, etc.. serían tenidas como si fuesen el mismo término. La influencia de este hecho sobre cualquier tipo de cálculo basado en frecuencias de términos parece evidente.

En idiomas morfológicamente más complejos que el inglés, esta cuestión puede resultar de gran importancia (Paice, C.D., 1996). La mayor parte de los sistemas de recuperación experimentados en inglés utilizan sistemas de lematización basados en el conocido algoritmo de Porter (Porter, M.F., 1980) . En realidad, la expresión ‘lematización’ no resulta muy correcta, puesto que lo que hacen los sistemas basados en Porter es lo que los anglosajones denominan *stripping*, es decir, la eliminación de la palabra de una terminación que coincide con alguna de las recogidas en una lista de sufijos. El sistema de Porter establece una serie de garantías: se elimina la terminación coincidente más larga, siempre y cuando la palabra mantenga una longitud superior a un límite, que debe establecerse experimentalmente. Porter elimina también prefijos, aunque parece que la incidencia de éstos es menor que la de los sufijos.

Para el idioma inglés, este sistema funciona razonablemente bien, y se aplica fácilmente. Existen desde hace mucho listas sencillas de sufijos ingleses

que producen buenos resultados; Rijsbergen, en su obra ya citada (Rijsbergen, K. van, 1979), reproduce una de 250 elementos, que ha sido ampliamente utilizada.

En otras lenguas, como el español, el mero *stripping* puede no resultar suficiente, ya que los sufijos no se pegan sin más a una determinada raíz, sino que, en el proceso de 'pegado', dicha raíz puede sufrir modificaciones importantes. De ahí que el tema de la lematización haya sido objeto de investigación específica para diversas lenguas además del inglés; así, hay trabajos referentes al francés (Savoy, J., 1993), al esloveno (Willet, P. y Popovic, M., 1992), o incluso al malayo (Ahmad, F. y otros, 1996). No conocemos ningún trabajo similar para el español.

Por lo que se refiere a los experimentos TREC con documentos en español, la mayor parte de ellos aplicaron técnicas de *stripping* basadas en Porter, como ya se ha dicho. Su principal problema consistió, pues, en contar con listas de sufijos lo suficientemente exhaustivas. Sorprende, en este sentido, la parvedad de bastantes de ellas. Por ejemplo, la manejada en el trabajo de Buckley y sus compañeros (Buckley, C. y otros, 1994), ya citado, constaba tan sólo de los sufijos 'as', 'es', 'os', 'a', y 'e', añadiendo un procedimiento para cambiar las 'z' finales (resultantes después de haber eliminado alguno de esos sufijos) por 'c'. El experimento realizado por el equipo de Wilkinson (Wilkinson, R. y otros, 1995) aplicó una lista de 30 sufijos o terminaciones, la mayor parte de ellas correspondientes a las distintas flexiones de verbos regulares.

Por cierto, que el tema de los verbos irregulares parece haber sido una auténtica obsesión para muchos de los participantes en las conferencias TREC. No es para menos; Gey y sus colegas (Gey, F.C. y otros, 1995)

consiguieron hasta 2.750 verbos irregulares flexionados en TREC-4, lista que ampliaron hasta 3.375 en TREC-5 (Gey, F.C. y otros, 1996). Esta parece ser, con mucho, la lista más exhaustiva en lo que a formas flexionadas de verbos se refiere. Sin embargo, simples manuales abreviados para estudiantes de español como lengua extranjera recogen más de 10.000 verbos irregulares (*Manual práctico* ..., 1992).

En cualquier caso, no está clara la importancia de las formas verbales como términos de alta significatividad o peso dentro de los documentos, y, probablemente, mucho menos en las consultas. Los estudios efectuados por Gil Leiva y Rodríguez Muñoz, citados anteriormente (Gil Leiva, I. y Rodríguez Muñoz, J.V., 1996 y 1997), aunque circunscritos a las expresiones clave utilizadas (manualmente) para indizar documentos en español, muestran una escasa presencia de las formas verbales en dichas expresiones clave. Aunque desde luego esto no es directamente extrapolable a la totalidad del documento, sí que puede hacer pensar que la importancia de los verbos pudiera ser relativa. Esta relatividad tiene visos de ser aún mayor en lo que se refiere a las consultas; en efecto, cabe al menos sospechar una cierta relación entre la forma de plantear una necesidad informativa (consulta) y la manera de expresar el contenido temático de un documento (expresiones clave). En cualquier caso, la aplicación de las listas de algunos miles de verbos irregulares mencionadas antes tampoco se tradujo en una mejora sustancial de los resultados, aunque bien es cierto que en dichos resultados intervienen otros factores.

5. Análisis morfológico y sintáctico

Algunas técnicas empleadas en inglés (e intentadas también en español) para mejorar la efectividad en la recuperación aplican, aunque de forma limitada, análisis morfológico y sintáctico de los documentos. Se pretende con ello varias cosas: indizar no sólo por términos simples, sino también por expresiones compuestas por varias palabras, así como pesar de forma diferente los términos en función de su categoría gramatical.

La aplicación de este tipo de técnicas requiere lo que en inglés se conoce como *part_of_speech tagger* (POS *tagger*), un programa que, de forma automática, es capaz de reconocer si una palabra es un sustantivo o un verbo, etc.. El asunto no es trivial, y no se resuelve simplemente con un diccionario, dado que existen abundantes ambigüedades en las funciones gramaticales de las palabras que deben ser resueltas. Los investigadores que trabajan en esta línea se quejan de carecer de un POS *tagger* para español, e intentan abordar la cuestión con reconocedores de sustantivos extremadamente simples. Así, Allan y otros, en un trabajo ya mencionado (Allan, J. y otros, 1995), emplean ciertas reglas para reconocer sustantivos: palabras, por ejemplo, que comienza por mayúscula, suelen ser nombres, al igual que aquéllas que tienen determinadas terminaciones. En este sentido, identifican once de tales terminaciones; la primera de todas ellas es -dor, de la cual ponen como ejemplo una palabra española y su correspondiente traducción al inglés: matador (*bull fighter*). Anecdótico, tal vez, pero significativo.

Otros investigadores consiguieron POS *taggers* para español (Smeaton, A.F. y otros, 1995), pero incluso esos mismos investigadores ponen en duda la

solvencia de tales programas.

6. N-gramas

El uso de n-gramas constituye un enfoque radicalmente divergente de los vistos hasta ahora. Básicamente, un n-grama es una especie de ventana de n caracteres de tamaño que se va desplazando a través de todo el texto de documentos y consultas. Así, por ejemplo, un texto consistente en la palabra ‘recuperación’, descompuesto en n-gramas de n=3, produciría la siguiente lista de trigramas: ‘_re’, ‘rec’, ‘ecu’, ‘cup’, ‘upe’, ‘per’, ‘era’, ‘rac’, ‘aci’, ‘cio’, ‘ion’, ‘on_’ (el símbolo _ representa el espacio en blanco que se considera separa unas palabras de otras). Los n-gramas así obtenidos, resultantes de la descomposición de un documento, pueden tratarse de igual forma que los términos en el modelo vectorial clásico, considerando cada n-grama como un término; y lo mismo con las consultas. De manera, que, en función de sus frecuencias y del número de documentos de la colección, pueden calcularse sus IDFs, pesos, similaridades con vectores de consultas, etc..

La ventaja (presunta) de los n-gramas es que permiten obviar problemas como los errores tipográficos, frecuentes, por ejemplo, en documentos grabados con OCRs. Del mismo modo, los n-gramas deberían permitir abordar con éxito la cuestión de las palabras con la misma raíz, pero con distintos sufijos, sin necesidad de hacer lematización. Pensemos, en efecto, en las palabras ‘bibliotecas’, ‘bibliotecarios’, ‘biblioteconomía’. Aunque todas ellas producen listas distintas de n-gramas, parte de esos n-gramas serán comunes,

puesto que la primera parte o raíz de esas palabras es la misma. Esos n-gramas comunes serían, en consecuencia, de mayor peso y aumentarían la similaridad con consultas que contuvieran palabras con raíz ‘bibliot’, y que producirían también n-gramas del mismo tipo.

En este punto, hay que indicar que el tamaño de n es crucial, y que puede incidir directamente en la efectividad de los n-gramas producidos. Este tamaño se suele fijar de modo experimental, pero los valores más frecuentes son de n=3 (Smeaton, A.F. y otros, 1994), n=4 (Cavnar, W.B., 1994), o n=5 (Grossman, D.A. y otros, 1995).

N-gramas de texto en español fueron obtenidos hace ya varios años (Gutiérrez Muñoz, F. y otros, 1989). Aunque ligeramente diferentes debido a la metodología de obtención (normalizando las palabras previamente mediante la eliminación de acentos), parece haber algunas discrepancias con los obtenidos, por ejemplo, por Smeaton y sus compañeros (Smeaton, A.F. y otros, 1994). Este cita los 20 trigramas más frecuentes, que no siempre coinciden con los de Gutiérrez Muñoz. Otra diferencia notable consiste en que, en este mismo trabajo de Smeaton, se hace mención del hecho de que las frecuencias de los trigramas no siguen una distribución zipfiana, lo cual llama la atención, y sugiere que una estrategia basada en pesos de trigramas no producirá buenos resultados, como, en líneas generales, parece que ha ocurrido. La excepción parece ser el trabajo citado de Cavnar (Cavnar, W.B., 1994), un ferviente defensor, por otra parte, de esta técnica.

7. Resultados

Hemos intentado efectuar una comparación entre los resultados obtenidos para los experimentos de recuperación en español y los obtenidos con documentos en inglés. Es necesario advertir que no es posible comparar con un mínimo de rigor resultados de recuperaciones en colecciones de documentos distintas, y con cuestionarios o consultas diferentes. Sin embargo, parece desde luego de interés saber si las técnicas utilizadas producen unos rendimientos homologables a los que se obtienen con documentos y consultas en inglés.

La efectividad de los experimentos se mide en los términos habituales de precisión y exhaustividad (‘recall’) y ello permite efectuar algún tipo de valoración, aún cuando ésta tenga un valor meramente aproximativo. Así, hemos calculado la curva de precisión interpolada con los valores medios facilitados para los experimentos de recuperación de documentos en español de TREC-4 (Harman, D., 1995). Y hemos hecho la misma operación con los resultados de las recuperaciones en inglés, tanto para TREC-4 como para TREC-3. En realidad, para los resultados sobre documentos en inglés sólo hemos utilizado los resultados de los experimentos centrados en la ‘adhoc task’, puesto que el resto consisten en operaciones de filtrado de textos, encaminamiento de documentos, recuperación de textos defectuosos, etc.., que se alejan un poco de las características de los experimentos en español.

Hemos considerado conveniente construir las curvas para TREC-3 y TREC-4 (inglés), debido a que hay una diferencia notoria entre ambas: el tamaño de las consultas. Así, las de TREC-4 (inglés) fueron mucho más cortas

que las de la edición anterior, produciendo, como era de esperar, peores resultados. Las consultas en español son, por lo que se refiere a tamaño, más similares a las de TREC-3 (inglés).

Las curvas resultantes se muestran en la figura1 y, a pesar de lo problemático de la comparación, tal como se ha comentado antes, parece que indican claramente que la recuperación en español produjo de forma generalizada peores resultados que la recuperación en inglés. Si, además, tenemos en cuenta la cuestión del tamaño de las consultas, y comparamos con los resultados de TREC-3, la situación es todavía más desventajosa.

No es difícil aventurar la razón, a tenor de lo que se ha expuesto en las páginas precedentes. La carencia del conocimiento y las herramientas lingüísticas adecuadas plantean serios problemas en la aplicación de técnicas y modelos de recuperación. Merece la pena, sin embargo, detenerse en algunos resultados particulares. Parece que la lematización produce mejoras en los resultados; algunos experimentos (Gey, F.C. y otros, 1995) informan de una mejora de varias centésimas en la precisión media debido exclusivamente a la lematización, y la misma conlusión se extrae (9.40 %) de otros trabajos en la misma línea (Broglio, J. y otros, 1994). Hearst y sus colegas (Hearst, M. y otros, 1995) afirman, por otra parte, aplicar un sistema de lematización basado en análisis de ‘morfología inflexional’, consiguiendo mejoras del 7.5 %; no explican, sin embargo, el contenido concreto de dicha morfología inflexional.

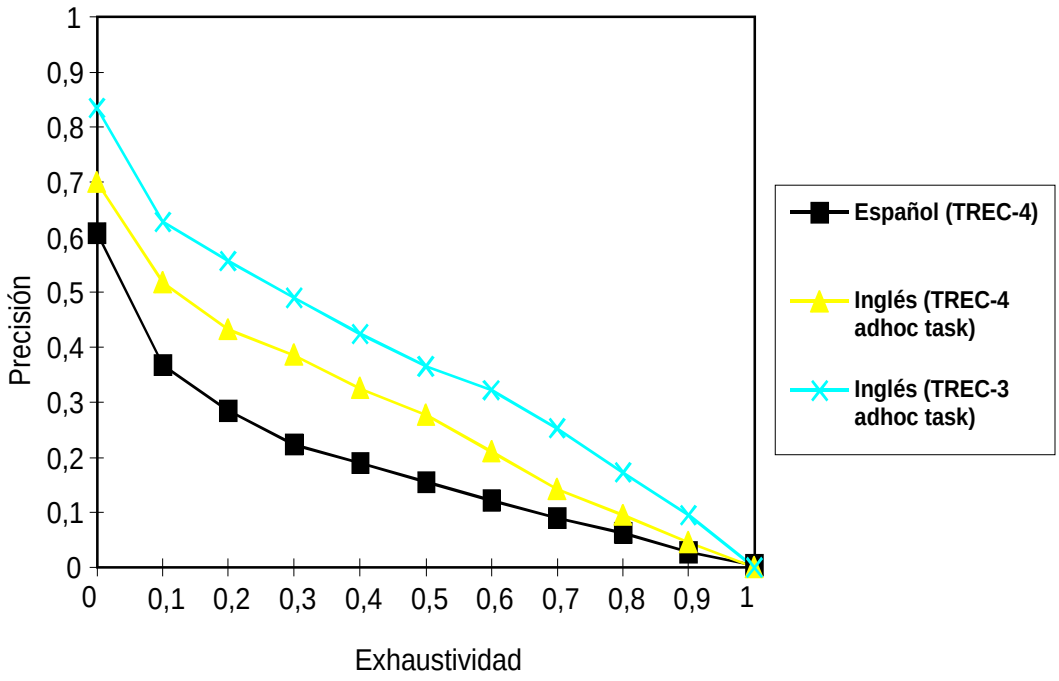
En sentido contrario, Wilkinson y colegas obtienen los mismos resultados con lematización y sin ella (Wilkinson, R. y otros, 1995).

Las técnicas basadas en la detección de las categorías gramaticales de los términos, y en la indización de términos complejos, por lo general parejas

de nombres, parecen producir mejoras claramente. Así, el trabajo citado de Hearst (Hearst, M. y otros, 1995) informa de mejoras en los resultados mediante la asignación de peso doble a los sintagmas nominales. Hull y otros investigadores (Hull, D.A. y otros, 1997) también consiguieron mejoras mediante tratamiento especial de los sintagmas nominales, pero descubrieron que esto funcionaba mejor cuando las consultas eran especialmente cortas.

8. Conclusiones

La investigación en IR sobre documentos en español no es precisamente abundante, y buena parte de la que se ha hecho proviene del mundo anglosajón. En general, parece que los modelos propuestos son aplicables al español, pero para ello es preciso implementar un cierto conocimiento lingüístico. Este conocimiento es tanto más importante cuanto que, según se ha visto, las técnicas que utilizan reconocimiento de categorías gramaticales y análisis morfológico consiguen mejoras en la efectividad en la recuperación. Este es un campo totalmente abierto para nuestros investigadores.



Figuera 1. Resultados aproximados TREC

termino	frecuencia	idf	peso
INFORMATICA	3	4,21	12,64
HERRAMIENTA	2	5,01	10,03
LA	7	1,18	8,28
INFORMACION	3	2,65	7,97
INVESTIGAR	1	7,58	7,58
AHI	1	7,17	7,17
DE	7	1,01	7,11
EXPLOSION	1	6,88	6,88
GESTIONAR	1	6,32	6,32
CAPACES	1	6,19	6,19
INFORMATIVA	1	6,19	6,19
NACE	1	5,97	5,97
RAPIDA	1	5,71	5,71
FACIL	1	5,63	5,63
SEAN	1	5,50	5,50
SOPORTES	1	5,50	5,50
BUSQUEDA	1	5,18	5,18
PROCESOS	1	5,05	5,05
BIBLIOTECAS	2	2,49	4,98
HERRAMIENTAS	1	4,53	4,53
COMO	2	2,26	4,53
FORMA	1	4,46	4,46
NUEVOS	1	4,40	4,40
DOCUMENTALES	1	4,38	4,38

Tabla 1. Términos de un documento, frecuencias en el documento, IDFs y pesos.

REFERENCIAS

PORTER, M.F. (1980): “An algorithm for suffix stripping”, *Program*, 14, 130-137

AHMAD, F., YUSOFF, M. y SEMBOK, M.T. (1996): “Experiments with a Stemming Algorithm for Malay Words”, JASIS, 47(12), 909-918

ALLAN, J., BALLESTEROS, L., CALLAN, J.P., CROFT, B.W. y LU, Z. (1995): “Recents Experiments with INQUERY”, en HARMAN, D.K. de.: *The Fourth Text Retrieval Conference*, NIST SP 500-236, Gaithersburg, Maryland, 1995, <*http://trec.nist.gov/pubs/trec4/papers/umass.ps*> (consulta el 01-02-1998)

BELKIN, N. y CROFT, W.B. (1987): "Retrieval Techniques", *Annual Review of Information Science & Technologies*, 22, 109-145

BROGLIO, J., CALLAN, J.P., CROFT, W.B. & NACHBAR, D.W. (1994): “Document Retrieval and Routing Using the INQUERY System”, en HARMAN, D.K. editor: *Overview of the Third Text Retrieval Conference, NIST SP 500-226*, Gaithersburg, Maryland, 1994, <*http://trec.nist.gov/pubs/trec3/papers/umass.revised.ps*> (consulta el 01-02-1998)

BUCKLEY, C., SALTON, G. ALLAN, J. y STINGHAL, A. (1994): “Automatic Query Expansion Using SMART: TREC-3”, en HARMAN, D.K. editor: *Overview of the Third Text Retrieval Conference, NIST SP 500-226*, Gaithersburg, Maryland, 1994, <*http://trec.nist.gov/pubs/trec3/papers/cornell.new.ps*> (consulta el 01-02-1998)

CAVNAR, W.B. (1994): “Using An N-Gram Based Document Representation With A Vector Processing Retrieval Model”, en HARMAN, D.K. editor: *Overview of the Third Text Retrieval Conference, NIST SP 500-226*, <*http://trec.nist.gov/pubs/trec3/papers/cavnar_ngram_94.ps*> (consulta el 01-02-1998)

CODINA, LL. (1995): “Teoría de recuperación de información: modelos fundamentales y aplicaciones a la gestión documental”, *Information World en español*, 38 (octubre 1995), 18-22

GEY, F.C., CHEN, A., HE, J. & MEGGS, J. (1995): “Logistic Regression at TREC-4: Probabilistic Retrieval from Full Text Document Collections”, en HARMAN, D.K. de.: *The Fourth Text Retrieval Conference*, NIST SP 500-236, Gaithersburg, Maryland, 1995, <*http://trec.nist.gov/pubs/trec4/papers/berkeley.ps*> (consulta el 01-02-1998)

GEY, F.C., CHEN, A., HE, J., XU, L. & MEGGS, J. (1996): “Term Importance, Boolean Conjunct Training, Negative Terms, and Foreign Language Retrieval: Probabilistics Algorithms at TREC-5”, en HARMAN, D.K. y VOORHEES, E.M. eds.: I*nformation Technology: The Fifth Text Retrieval Conference (TREC-5)*, NIST SP 500-238, Gaithersburg, Maryland, 1996, <*http://trec.nist.gov/pubs/trec5/papers/brkly.trec5.main.ps*> (consulta el 01-02-1998)

GIL LEIVA, I. y RODRÍGUEZ MUÑOZ, J.V. (1997): “Análisis de los descriptores de diferentes áreas del conocimiento indizadas en bases de datos del CSIC. Aplicación a la indización automática”, *Revista Española de Documentación Científica*, 20(2), 150-160

GIL LEIVA, I. y RODRÍGUEZ MUÑOZ, J.V. (1996): “Tendencias en los sistemas de indización automática. Estudio evolutivo”, *Revista Española de Documentación Científica*, 19(3), 273-291

GROSSMAN, D.A., HOLMES, D.O., FRIEDER, O., NGUYEN, M.D. & KINGSBURY, C.E. (1995): “Improving Accuracy and Run-Time Performance for TREC-4”, en HARMAN, D.K. ed..: *The Fourth Text Retrieval Conference*, NIST SP 500-236, Gaithersburg, Maryland, 1995, <*http://trec.nist.gov/pubs/trec4/papers/gmu.ps*> (consulta el 01-02-1998)

GUTIERREZ MUÑOZ, F., REY GUTIERREZ, G. del y REY GUERRERO, A. del (1989): “Recuento estadístico de palabras, letras, digramas y trigramas en títulos de artículos”, *Revista Española de Documentación Científica*, 12(2), 160-167

HARMAN, D. (1992): “Ranking Algorithms”, en Frakes, W.B. y Baeza-Yates, R. eds.: *Information Retrieval. Data Structures & Algorithms*, New Jersey: Prentice-Hall, 1992, 363-392

HARMAN, D. (1995): “Overview of The Fourth Text Retrieval Conference (TREC-4)”, en HARMAN, D.K. ed.: *The Fourth Text Retrieval Conference*, NIST SP 500-236, Gaithersburg, Maryland, 1995, <*http://trec.nist.gov/pubs/trec4/overview.ps*> (consulta el 01-02-1998)

HARMAN, D. K. (1994): “Overview of the Third Text Retrieval Conference”, en HARMAN, D.K. editor: *Overview of the Third Text Retrieval Conference, NIST SP 500-226*, Gaithersburg, Maryland, 1994 <*http://trec.nist.gov/pubs/trec3/overview.ps*> (consulta el 01-02-1998)

HEARST, M., PEDERSEN, J., PIROLLI, P. & SCHUTZE, H. (1995): “Xerox Site Report: Four TREC-4 Tracks”, en HARMAN, D.K. ed..: *The Fourth Text Retrieval Conference*, NIST SP 500-236, Gaithersburg, Maryland, 1995, <*http://trec.nist.gov/pubs/trec4/xerox.ps*> (consulta el 01-02-1998)

HULL, D.A., GREFENSTETE, G., SCHULZE, B.M., GAUSSIER, E., SCHUTZE, H. & PEDERSEN, J.O. (1997): “Xerox TREC-5 Site Report: Routing, Filtering, NLP, and Spanish Tracks”, en HARMAN, D.K. y VOORHEES, E.M. eds.: I*nformation Technology: The Fifth Text Retrieval Conference (TREC-5)*, NIST SP 500-238, Gaithersburg, Maryland, 1996, <*http://trec.nist.gov/pubs/trec5/papers/real-xerox.ps*> (consulta el 01-02-1998)

La colección forma parte de un trabajo de experimentación en IR llevado a cabo por Raquel Gómez Díaz y Eva López de San Román, aún sin finalizar, y,consiguientemente, no publicado.

Manual práctico de conjugación, Barcelona: Larousse-Planeta, 1992

SALTON, G.(1987): "On the relationship between theoretical retrieval models", *Informetrics 87/88*, Diepenbeeck (Bélgica), 1987, pp. 263-270.

PAICE, C.D.(1996): "Method for Evaluation of Stemming Algorithms Based on Error Counting", *JASIS*, 47(8), 632-649

RIJSBERGEN, K. van (1979): *Information retrieval*, London, 1979

SALTON, G. (1989): *Automatic Text Processing*, Reading: Adisson-Wesley, 1989

SALTON, G. y MCGILL, M. (1983): *Introduction to Modern Information Retrieval*, New York: McGraw-Hill, 1983

SALTON, G.(1987): "On the relationship between theoretical retrieval

SAVOY, J. (1993): “Stemming of french words based on gramatical categories”, *JASIS*, 44(1), 1-9

SIMON GRANDA, J. y LEMA GARZON, E. de (1990): “Primeras experiencias sobre el análisis de textos en castellano aplicado a la indexación automática de información”, *Terceras Jornadas Españolas de Documentación Científica Automatizada*, 1990, 1255-1270.

SMEATON, A.F., KELLEDY, F. & O’DONNEL, R. (1994): “Indexing Structures Derived from Syntax in TREC-3: System Description”, en HARMAN, D.K. editor: *Overview of the Third Text Retrieval Conference, NIST SP 500-226*, <*http://trec.nist.gov/pubs/trec3/papers/dublin.ps*> (consulta el 01-02-1998)

SMEATON, A.F., KELLEDY, F. & O’DONNEL, R. (1995): “TREC-4 Experiments at Dublin City University: Thresholding Posting Lists, Query Expansion with WordNet and POS Tagging of Spanish”, en HARMAN, D.K. de.: *The Fourth Text Retrieval Conference*, NIST SP 500-236, Gaithersburg, Maryland, 1995, <*http://trec.nist.gov/pubs/trec4/papers/dublin.ps*> (consulta el 01-02-1998)

SPARCK JONES, K. (1972): “A Statistical Interpretation of Term Specifity and Its Application in Retrieval”, *Journal of Documentation*, 28(1), 11-20

VALLE BRACERO, A. y FERNANDEZ GARCIA, J.A. (1983): “Automatización de la indización y coordinación de descriptores”, *Revista Española de Documentación Científica*, 6(1), 9-16

WILKINSON, R., ZOBEL, J. & SACKS-DAVIS, R. (1995): “Similarity Measures for Short Queries”, en HARMAN, D.K. de.: *The Fourth Text Retrieval Conference*, NIST SP 500-236, Gaithersburg, Maryland, 1995, <*http://trec.nist.gov/pubs/trec4/papers/citri.ps*> (consulta el 01-02-1998)

WILLET, P. y POPOVIC, M. (1992): “The effectiveness of Stemming for Natural Language Access to Slovene Textual Data”, *JASIS*, 43(5), 384-390

(11)<*http://trec.nist.gov*> (consulta el 01-02-1998)

(17) <*ftp://ftp.cs.cornell.edu/pub/smart*> (consulta el 01-02-1998)