Procesamiento del Lenguaje Natural, núm. 36 (2006), pp. 39-46 recibido 01-03-2006; aceptado 01-06-2006

Tratamiento de la Variación Sintáctica mediante un Modelo de Recuperación Basado en Localidad*

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

Resumen: La aplicación de información sintáctica en el modelo de recuperación basado en documentos imperante en la actualidad ha sido probada sin excesivo éxito en numerosas ocasiones, debido mayormente a los problemas que supone la integración de este tipo de información en el modelo. En este artículo proponemos el empleo de un modelo basado en localidad aplicado a la reordenación de resultados, el cual aborda el problema de la variación lingüística sintáctica mediante medidas de similaridad basadas en la distancia entre palabras. Se estudian dos aproximaciones cuya efectividad ha sido evaluada sobre el corpus CLEF de documentos en español.
Palabras clave: Recuperación de Información, variación lingüística sintáctica, modelo basado en localidad, fusión de datos.

Abstract: To date, attempts for applying syntactic information in the document-based retrieval model dominant have led to little practical improvement, mainly due to the problems associated with the integration of this kind of information into the model. In this article we propose the use of a locality-based retrieval model for reranking, which deals with syntactic linguistic variation through similarity measures based on the distance between words. We study two approaches whose effectiveness has been evaluated on the CLEF corpus of Spanish documents.
Keywords: Information Retrieval, syntactic linguistic variation, locality-based model, data fusion.

1. Introducción

El procesamiento sintáctico ha sido empleado repetidamente en el ámbito de la Recuperación de Información (RI) para hacer frente a la *variación lingüística sintáctica* presente en los textos (Perez-Carballo y Strzalkowski, 2000; Hull et al., 1997), si bien su empleo en el caso del español ha sido poco estudiado hasta ahora (Alonso, Vilares, y Darriba, 2002; Vilares y Alonso, 2003). Estas técnicas precisan de algún tipo de analizador sintáctico, para lo cual es necesario contar con una gramática apropiada, por sencilla que sea. Sin embargo, aún cuando dicha información sintáctica pueda ser convenientemente extraída del texto, persiste todavía el problema de cómo incorporar dicha información al sistema. La aproximación más común, consistente en una combinación ponderada de términos simples y términos multipalabra —formados por términos simples relacionados sintácticamente—, no logra siempre resolver adecuadamente los problemas derivados de la sobrevaloración que el sistema tiende a dar a los términos complejos en detrimento de los términos simples (Mitra et al., 1997).

En este contexto, el empleo de técnicas pseudo-sintácticas basadas en distancias entre términos se presenta como una alternativa práctica que evita dichos problemas, al no ser necesaria gramática o analizador alguno, y al integrar de modo consistente la información obtenida, tanto a nivel de la aparición de los términos en sí, como de su proximidad, frecuentemente ligada a la existencia de una relación sintáctica entre los mismos.

En este artículo proponemos la utilización de un *modelo basado en localidad*, sustentado sobre similaridades basadas en distancias, como complemento a las técnicas clásicas de Recuperación de Información basadas en la indexación de términos simples, con el fin de

* Parcialmente financiado por el Ministerio de Educación y Ciencia y FEDER (TIN2004-07246-C03-02), y por la Xunta de Galicia (PGIDIT05PXIC30501PN, PGIDIT05PXIC10501PN, PGIDIT05SIN044E).

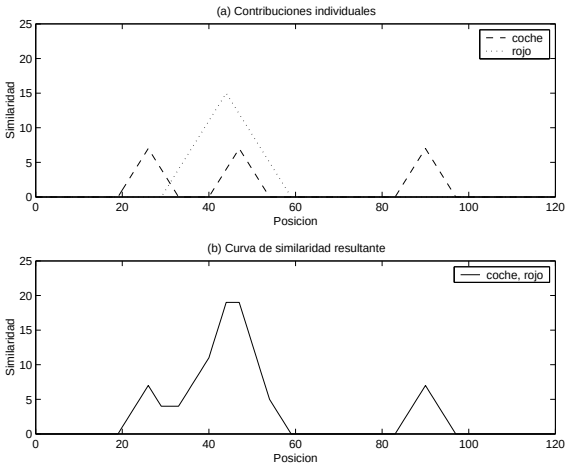

Figura 1: Modelo basado en localidad: (a) posiciones del texto con aparición de términos de la consulta y sus áreas de influencia; y (b) curva de similaridad resultante

incrementar la precisión de los documentos devueltos por el sistema.

2. *Recuperación de Información Basada en Localidad*

2.1. Modelo de Recuperación

En el modelo de recuperación imperante en RI, denominado *basado en documentos*, el usuario solicita del sistema los documentos relevantes a su consulta o necesidad de información. Por otra parte, el modelo *basado en localidad* propuesto por de Kretser y Moffat (de Kretser y Moffat, 1999b; de Kretser y Moffat, 1999a) va un paso más allá y busca las *posiciones* concretas del texto que pueden resultar relevantes al usuario.

La *Recuperación de Pasajes* (Kaszkiel y Zobel, 2001) es una aproximación intermedia que persigue identificar aquellas secciones del documento —*pasajes*— relevantes para la consulta. En este modelo, una vez que el documento original ha sido dividido en pasajes, éstos son procesados y ordenados mediante técnicas tradicionales. Sin embargo restan por resolver problemas acerca de cómo definir el concepto de pasaje, su tamaño, grado de superposición, etc. (Llopis, 2003).

Por el contrario, el modelo basado en localidad considera la colección a indexar no como un conjunto de documentos, sino como una secuencia de palabras donde cada aparición de un término de la consulta ejerce una influencia sobre los términos circundantes. Dichas influencias son aditivas, de forma que la contribución de diferentes apariciones

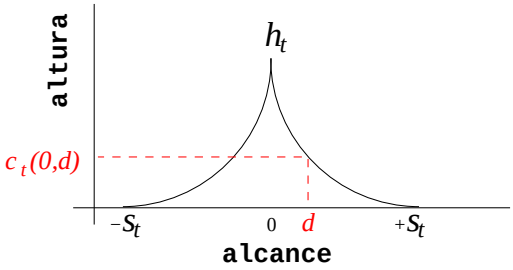

Figura 2: Función de contribución de similaridad c_t de forma circular.

de términos de la consulta pueden sumarse, dando lugar a una medida de similaridad, tal y como muestra la figura 1. Aquellas áreas del texto con una mayor densidad de términos de la consulta, o con términos de mayor peso, darán lugar a picos en la curva de influencia resultante, señalando posiciones del texto potencialmente relevantes. Todo ello sin necesidad de particionar artificialmente el documento como en el caso de la Recuperación de Pasajes.

2.2. Cálculo de Similaridades

A continuación, describiremos el modelo basado en localidad propuesto originalmente por de Kretser y Moffat (de Kretser y Moffat, 1999b; de Kretser y Moffat, 1999a). En este modelo la medida de similaridad o relevancia es calculada únicamente sobre aquellas posiciones donde aparecen términos de la consulta, reduciendo de este modo el coste computacional asociado.

La contribución a dicha similaridad por parte de un término de la consulta viene dada por una *función de contribución de similaridad* c_t definida en base a los siguientes parámetros (de Kretser y Moffat, 1999a):

- La *forma* de la función, siendo la misma para todos los términos.
- La *altura máxima* h_t de la función, que se da en la posición del término que ejerce la influencia.
- El *alcance* s_t de la función, es decir, su radio de influencia.
- La distancia en palabras entre los dos términos considerados, $d = |x - l|$, donde l es la posición del término de la consulta que ejerce la influencia y x la posición sobre la que se desea calcular la medida de similaridad.

Si bien en (de Kretser y Moffat, 1999a) se describen diversas formas de función, experimentos previos mostraron un mejor comportamiento de la función circular (cir) en el caso del español. Dicha función, cuya representación gráfica se muestra en la figura 2, se define mediante la ecuación:

$$c_t(x,l) = h_t \cdot \sqrt{1-(d/s_t)^2} \qquad (1)$$

con $c_t(x,l) = 0$ para $|x-l| > s_t$, y equivalente a los cuadrantes de dos círculos normalizados con centros en $(h_t, -s_t)$ y (h_t, s_t).

Por su parte, la altura máxima h_t asociada a un término t se calcula como función inversa de su frecuencia en la colección:

$$h_t = f_{Q,t} \cdot log_e\Big(\frac{N}{f_t}\Big) \qquad (2)$$

donde N es el número total de términos en la colección, f_t el número de apariciones del término t en la colección y $f_{Q,t}$ la frecuencia del término t en la consulta Q.

En lo que respecta al alcance s_t de la influencia de un término t, ésta viene dada también por el inverso de su frecuencia en la colección, pero normalizada en base a la frecuencia media:

$$s_t = \frac{n}{N} \cdot \frac{N}{f_t} = \frac{n}{f_t} \qquad (3)$$

siendo n el número de términos diferentes en la colección, es decir, el tamaño del vocabulario.

De este modo, la medida de similaridad $C_Q(x)$ asignada a la posición x del documento en la cual aparece un término de la consulta Q se calcula como:

$$C_Q(x) = \sum_{t \in Q} \sum_{\substack{l \in I_t \\ |l-x| \leq s_t \\ term(x) \neq term(l)}} c_t(x,l) \qquad (4)$$

donde I_t es el conjunto de posiciones donde ocurre un término t de la consulta Q, y donde $term(w)$ denota el termino asociado a la posición w. En otras palabras, la medida de similaridad o relevancia asociada a una posición es la suma de las influencias ejercidas por los demás términos de la consulta presentes en el documento y dentro de cuyo alcance se encuentra, exceptuando otras apariciones del término existente en la posición considerada (de Kretser y Moffat, 1999b).

Finalmente, la medida de relevancia $sim(D,Q)$ asignada a un documento D respecto a una consulta Q vendrá dada en función de las similaridades asignadas a las apariciones de términos de la consulta que dicho documento contenga. Este punto se comenta en mayor detalle en el siguiente apartado.

2.3. Adaptaciones del Modelo.

Dado que el modelo basado en localidad permite trabajar a un nivel de detalle mayor que las técnicas clásicas de RI, al identificar no sólo los documentos relevantes sino también concretar las posiciones de interés dentro de los mismos, hemos optado por emplear este modelo en nuestros experimentos. Para ello ha sido necesario realizar ciertas adaptaciones de acuerdo con nuestras necesidades, las cuales nos diferencian del planteamiento original del modelo.

El planteamiento elegido a la hora de integrar la similaridad basada en distancias dentro de nuestro sistema de RI, ha sido el del postprocesado de los documentos previamente obtenidos mediante un sistema de recuperación clásico basado en documentos, con intención de incrementar la precisión de los primeros documentos devueltos. Este primer conjunto de documentos devuelto por el sistema es a continuación procesado empleando el modelo basado en localidad, tomando la ordenación final obtenida en base a distancias como aquélla a devolver al usuario.

Otra de las principales diferencias respecto al modelo original es el del empleo de la lematización (Graña, Chappelier, y Vilares, 2001) en lugar del *stemming* a la hora de la normalización de consultas y documentos, dado su mejor comportamiento en el caso del español (Vilares et al., 2002).

Por otra parte, debemos señalar que los parámetros de altura máxima h_t y alcance s_t utilizados durante la reordenación se calculan en base a los parámetros globales de la colección, y no en base a los parámetros locales al subconjunto de documentos devueltos, para así evitar los problemas derivados de la correlación que esto conllevaría.[1]

Finalmente, a la hora de calcular la relevancia $sim(D,Q)$ de un documento D respecto a una consulta Q, en lugar del algoritmo iterativo del modelo original (de Kretser y Moffat, 1999a), nuestra solución calcula dicha medida de relevancia como la suma de

[1] Por ejemplo, el parámetro f_t de número de apariciones de un término t es el número de apariciones de t en toda la colección, no el número de apariciones de t en el conjunto de documentos a reordenar.

las medidas de similaridad individuales de las apariciones de términos de la consulta en dicho documento:

$$sim(D,Q) = \sum_{\substack{x \in D \\ term(x) \in Q}} C_Q(x) \qquad (5)$$

3. Resultados Experimentales con Distancias

Nuestra aproximación ha sido probada sobre el corpus monolingüe para español del CLEF[2], conformado por las siguientes colecciones de documentos y *topics* asociados a partir de los cuales generar las consultas:

CLEF 2001-02·A: colección de entrenamiento y estimación de parámetros, formada por 215.738 teletipos de la agencia española de noticias EFE[3] correspondientes al año 1994, siendo sus *topics* asociados aquéllos de número impar empleados en las ediciones 2001 y 2002 del CLEF.

CLEF 2001-02·B: colección de evaluación similar a la anterior, si bien emplea los *topics* de número par.

CLEF 2003: colección de evaluación, formada por 454.045 teletipos de EFE correspondientes a los años 1994 y 1995, y que usa los *topics* del CLEF 2003.

Dichos *topics* están formados por tres campos: *título*, un breve título como su nombre indica; *descripción*, una somera frase de descripción; y *narrativa*, un pequeño texto especificando los criterios de relevancia.

Debemos precisar que aquellos *topics* con menos de 6 documentos relevantes fueron eliminados, ya que en dichos casos la modificación en la posición de uno o dos documentos devueltos puede acarrear cambios muy marcados en los resultados obtenidos para dicha consulta, distorsionando así los resultados globales (Hull et al., 1997).

Asimismo se emplearon dos tipos de consultas, las denominadas consultas *cortas*, generadas a partir de los campos *título* y *descripción*, y las denominadas consultas *largas*, que emplean la totalidad de los campos del *topic*. En el caso de las consultas largas, se ha doblado la relevancia asignada al campo *título*, al concentrar éste la semántica básica de la consulta.

En lo que respecta a la indexación inicial de términos lematizados (*lem*), se empleó el conocido motor de indexación vectorial SMART (Buckley, 1985), empleando un esquema de pesos `atn·ntc` (Salton y Buckley, 1988). Por otra parte, con objeto de mejorar en lo posible el rendimiento final del sistema resultante, se partirá de un conjunto inicial de documentos tan bueno como sea posible. Para ello aplicaremos realimentación mediante expansión de consultas con el algoritmo de Rocchio (Rocchio, 1971):

$$Q_1 = \alpha Q_0 + \beta \sum_{k=1}^{n_1} \frac{R_k}{n_1} - \gamma \sum_{k=1}^{n_2} \frac{S_k}{n_2} \qquad (6)$$

donde Q_1 es vector de la consulta final, Q_0 es el vector de la consulta inicial, R_k es el vector del k-ésimo documento relevante, S_k es el vector del k-ésimo documento no relevante, n_1 es el número de documentos relevantes examinados, n_2 es el número de documentos no relevantes examinados, y α, β y γ son, respectivamente, los parámetros que controlan las contribuciones relativas de la consulta original, los documentos relevantes, y los documentos no relevantes. En concreto, expandiremos automáticamente la consulta inicial con los t=10 mejores términos de los cinco primeros documentos devueltos (n_1=5), con unas contribución relativas α=0.8, β=0.1, γ=0 para consultas cortas y α=1.2, β=0.1, γ=0 para consultas largas.

Los resultados obtenidos se muestran en la tabla 1. El rendimiento del sistema se ha medido en base a los parámetros recogidos en cada fila: número de consultas empleadas, número de documentos devueltos, número de documentos relevantes esperados, número de documentos relevantes devueltos, precisión media no interpolada para todos los documentos relevantes, precisión-R, precisión en los primeros niveles estándar de cobertura, y precisión a los n documentos devueltos. La primera columna de cada grupo recoge los resultados de la línea base, la indexación de lemas con realimentación (*lem*), mientras que la segunda columna muestra los resultados obtenidos tras la ordenación de *lem* mediante distancias (*cir*). Para cada parámetro se han marcado en negrita los valores para los que se ha obtenido una mejora respecto a la línea base.

[2] `http://www.clef-campaign.org`

[3] `http://www.efe.es`

corpus	*CLEF 2001-02·A*				*CLEF 2001-02·B*				*CLEF 2003*			
consulta	*cortas*		*largas*		*cortas*		*largas*		*cortas*		*largas*	
técnica	*lem*	*cir*	*lem*	*cir*	*lem*	*cir*	*lem*	*cir*	*lem*	*cir*	*lem*	*cir*
#consultas	46	=	46	=	45	=	45	=	47	=	47	=
#docs. dev.	46k	=	46k	=	45k	=	45k	=	47k	=	47k	=
#rlvs. esp.	3007	=	3007	=	2513	=	2513	=	2335	=	2335	=
#rlvs. dev.	2767	=	2779	=	2376	=	2406	=	2240	=	2223	=
Pr. no int.	.5220	.4668	.5604	.4714	.4773	.4278	.5392	.4831	.5024	.3924	.5207	.4005
Pr.-R	.4990	.4651	.5366	.4652	.4599	.4205	.5104	.4592	.4912	.3921	.4871	.3911
Pr. a 0 %	.8221	**.8835**	.8895	**.8979**	.8210	**.8233**	.8710	.8678	.8145	**.8230**	.8301	**.8415**
Pr. a 10 %	.7490	**.7870**	.8028	**.8143**	.6861	**.7197**	.7619	**.8084**	.7369	.6626	.7421	.6518
Pr. a 20 %	.6866	**.6883**	.7352	.7017	.6319	**.6378**	.6929	.6808	.6632	.5663	.6758	.5737
Pr. a 30 %	.6573	.6148	.6996	.6066	.5688	.5464	.6497	.6000	.6019	.5098	.6304	.5030
Pr. a 40 %	.5997	.5267	.6541	.5372	.5289	.4827	.6202	.5438	.5638	.4439	.5975	.4400
Pr. a 50 %	.5456	.4656	.6005	.4728	.5017	.4322	.5733	.4987	.5410	.4077	.5479	.3956
Pr. a 5 docs.	.6609	**.6913**	.6957	**.7261**	.5956	**.6000**	.6844	.6533	.5872	.5532	.6213	.5574
Pr. a 10 docs.	.6457	.6391	.6848	.6522	.5600	.5444	.6178	.6089	.5596	.5064	.5872	.4979
Pr. a 15 docs.	.5884	**.5899**	.6435	.5971	.5274	.5111	.5822	.5556	.5305	.4624	.5504	.4652
Pr. a 20 docs.	.5630	.5446	.6043	.5674	.5011	.4822	.5533	.5189	.4883	.4181	.5266	.4277
Pr. a 30 docs.	.5225	.4848	.5580	.4971	.4444	.4215	.5081	.4733	.4433	.3702	.4667	.3780
Pr. a 100 docs.	.3507	.3052	.3598	.3048	.2940	.2780	.3191	.3022	.2770	.2400	.2853	.2404

Tabla 1: Resultados obtenidos mediante reordenación por distancias (*cir*) de la lematización con realimentación (*lem*)

Como muestran los resultados, la reordenación por distancias ha producido una disminución general del rendimiento del sistema, salvo para los primeros niveles de cobertura y primeros documentos devueltos, donde en algunos casos los resultados son similares o incluso mejores. Podemos concluir, pues, que esta primera aproximación no ha demostrado ser de demasiado interés práctico.

4. Fusión de Datos mediante Intersección

4.1. Justificación

Dado que el número de documentos relevantes devueltos es el mismo, la caída en el rendimiento del sistema en ésta primera aproximación sólo puede deberse a una mala ordenación de los resultados por el modelo basado en distancias. Por esta razón decidimos estudiar la variación en la distribución de documentos relevantes y no relevantes en los K primeros documentos devueltos. Comentaremos únicamente los resultados obtenidos empleando consultas cortas para el corpus CLEF 2001-02·A, mostrados en la tabla 2, ya que los resultados obtenidos en dicho estudio son muy similares para los demás corpus y tipos de consultas.

Cada fila muestra los resultados obtenidos al comparar los K primeros documentos devueltos por el sistema mediante indexación de lemas con realimentación (*lem*) —conjunto de resultados L— con aquéllos devueltos tras su reordenación mediante distancias (*cir*) —conjunto de resultados D. Las columnas muestran los resultados obtenidos para cada uno de los parámetros considerados: número medio de nuevos relevantes obtenidos mediante distancias ($D \setminus L$), número medio de relevantes perdidos con distancias ($L \setminus D$), número medio de relevantes que se mantienen ($L \cap D$), coeficiente de superposición de relevantes (R_{sup}), precisión de *lem* a los K primeros documentos ($Pr(L)$), precisión a los K documentos tras la reordenación por distancias ($Pr(D)$), y precisión en los documentos comunes a ambas aproximaciones dentro de sus K primeros documentos ($Pr(L \cap D)$). En la parte derecha de la tabla se muestran sus equivalentes para el caso de los documentos no relevantes: número medio de no relevantes añadidos, perdidos y comunes, y grado de superposición de no relevantes.

A partir de estos resultados se pueden extraer diversas conclusiones de importancia. En primer lugar, observamos que el número de documentos relevantes obtenidos por ambas aproximaciones dentro de sus K primeros documentos es muy similar —si bien algo menor para las distancias—, tal como se puede apreciar en las cifras absolutas de documentos relevantes entrantes y salientes y en

	Docs. relevantes							Docs. no relevantes			
K	$D \setminus L$	$L \setminus D$	$L \cap D$	R_{sup}	$Pr(L)$	$Pr(D)$	$Pr(L \cap D)$	$D \setminus L$	$L \setminus D$	$L \cap D$	N_{sup}
5	1.93	1.78	1.52	0.45	**0.66**	**0.69**	***0.80***	1.15	1.30	0.39	0.24
10	3.24	3.30	3.15	0.49	**0.65**	**0.64**	***0.76***	2.61	2.54	1.00	0.28
15	4.17	4.15	4.67	0.53	**0.59**	**0.59**	***0.72***	4.35	4.37	1.80	0.29
20	4.59	4.96	6.30	0.57	**0.56**	**0.54**	***0.72***	6.65	6.28	2.46	0.28
30	5.61	6.74	8.93	0.59	**0.52**	**0.48**	***0.68***	11.22	10.09	4.24	0.28
100	7.00	11.48	23.52	0.72	**0.35**	**0.31**	***0.49***	45.43	40.96	24.04	0.36
200	5.54	9.63	37.35	0.83	**0.23**	**0.21**	***0.36***	90.50	86.41	66.61	0.43
500	2.35	3.39	52.67	0.95	**0.11**	**0.11**	***0.16***	167.43	166.39	277.54	0.62

Tabla 2: Distribución de documentos relevantes y no relevantes tras la reordenación mediante distancias. Corpus CLEF 2001-02·A, consultas cortas

las precisiones a los K documentos de ambas aproximaciones. Esto nos permite confirmar que se trata de un problema de mala ordenación de los documentos.

En segundo lugar debemos referirnos a los coeficientes de superposición de documentos relevantes (R_{sup}) y no relevantes (N_{sup}). Estos coeficientes, definidos en (Lee, 1997), indican el grado de superposición entre el conjunto de documentos relevantes o no relevantes de dos conjuntos de documentos devueltos. Para dos ejecuciones run_1 y run_2, dichos coeficientes se definen como:

$$R_{sup} = \frac{2\ |Rel(run_1) \cap Rel(run_2)|}{|Rel(run_1)| + |Rel(run_2)|} \quad (7)$$

$$N_{sup} = \frac{2\ |Nonrel(run_1) \cap Nonrel(run_2)|}{|Nonrel(run_1)| + |Nonrel(run_2)|} \quad (8)$$

donde $Rel(X)$ y $Nonrel(X)$ representan, respectivamente, el conjunto de documentos relevantes y no relevantes devueltos en la ejecución X.

Como podemos apreciar en la tabla 2, los factores de superposición de los documentos relevantes son considerablemente mayores que los de los no relevantes. De esta forma ambas aproximaciones devuelven un conjunto similar de documentos relevantes, pero un conjunto diferente de documentos irrelevantes. Se cumple, pues, la denominada *propiedad de la superposición desigual* (Lee, 1997), que dice que diferentes ejecuciones deben devolver conjuntos similares de documentos relevantes a la vez que devolver conjuntos disimilares de no relevantes como primer indicador de la efectividad que tendría la fusión de datos de ambas.

En tercer lugar, y en relación al punto anterior, puede verse que la precisión en los documentos comunes a ambas aproximaciones dentro de sus K primeros documentos ($Pr(L \cap D)$) es mayor que las precisiones alcanzadas tanto por lemas ($Pr(L)$) como por distancias ($Pr(D)$); o lo que es lo mismo, la probabilidad de que un documento sea relevante es mayor cuando es devuelto por ambas aproximaciones.

Conforme a estas observaciones, se planteó una nueva aproximación para la reordenación, esta vez basada en la fusión de datos.

4.2. Descripción del Algoritmo

La *fusión de datos* es una técnica de combinación de evidencias consistente en la combinación de resultados devueltos empleando diferentes representaciones de consultas o documentos, o mediante múltiples técnicas de recuperación (Fox y Shaw, 1994; Lee, 1997).

En nuestro caso hemos optado por una aproximación basada no en la combinación de puntuaciones en base a similaridades (Fox y Shaw, 1994; Lee, 1997) o rango (Lee, 1997), sino en un criterio booleano para el cual, una vez fijado un valor K, los documentos son devueltos en el siguiente orden:

1. En primer lugar, los documentos pertenecientes a la intersección de los K primeros documentos de ambas aproximaciones: $L_K \cap D_K$. El objetivo perseguido es el de incrementar la precisión en los primeros documentos devueltos.
2. A continuación, los documentos pertenecientes a los K primeros documentos de ambas aproximaciones que no estén en la intersección: $(L_K \cup D_K) \setminus (L_K \cap D_K)$. El objetivo es añadir a los primeros documentos devueltos aquellos documentos relevantes devueltos únicamente mediante la aproximación basada en distancias, sin perjudicar la ordenación de aquéllos devueltos únicamente por la indexación de lemas.

3. Finalmente, los restantes documentos devueltos por los lemas: $L \setminus (L_K \cup D_K)$.

donde L es el conjunto de resultados devuelto por *lem*, L_K el conjunto de K primeros resultados devuelto con *lem*, y D_K el conjunto de K primeros resultados devuelto mediante la reordenación por distancias.

Con respecto a la ordenación interna de los resultados, se tomará como referencia, por sus mejores resultados, la ordenación obtenida mediante la indexación de lemas (*lem*). De esta forma cuando se devuelva un subconjunto S de resultados, los documentos que lo conforman se devolverán en el mismo orden relativo que existía entre ellos cuando eran devueltos por *lem*.[4]

5. Resultados Experimentales con Fusión de Datos

Tras experimentos previos de puesta a punto, se optó finalmente por emplear un valor $K = 30$ en el caso de consultas cortas y $K = 50$ en el caso de las largas.

La tabla 3 recoge los resultados obtenidos para la nueva aproximación. Al igual que antes, la primera columna de cada grupo muestra los resultados para la línea base, la indexación de lemas con realimentación (*lem*), mientras que la segunda columna muestra los resultados obtenidos tras su ordenación mediante fusión por intersección (*cir*).

Podemos apreciar que las mejoras obtenidas con la reordenación mediante fusión son consistentes, especialmente en el caso de la precisión a los n documentos devueltos —tal como se pretendía—, si bien dichas mejoras también se extienden al resto de parámetros estudiados, siendo algo menores en el caso del corpus CLEF 2003.

6. Conclusiones y Trabajo Futuro

A lo largo de este artículo se ha planteado la utilización de un modelo de recuperación basado en distancias entre palabras, o basado en localidad, para tratar el problema de la variación lingüística de carácter sintáctico presente en los textos.

Se han considerado dos aproximaciones, ambas enfocadas a la reordenación de resultados, en este caso obtenidos mediante indexación de lemas de palabras con contenido. La primera aproximación, que asumía el orden obtenido mediante la aplicación del modelo basado en localidad como el orden final a devolver, no obtuvo buenos resultados. Tras analizar el comportamiento del sistema se optó por emplear una aproximación basada en la fusión de datos, y que emplea la intersección de los conjuntos de documentos devueltos por ambos sistemas como guía para la reordenación. Esta segunda aproximación resultó fructífera, obteniendo una mejora consistente y general.

En lo que respecta al trabajo futuro, pretendemos dar capacidad al sistema para el tratamiento de variantes morfosintácticas de una expresión (Jacquemin y Tzoukermann, 1999), así como de relaciones de sinonimia ponderada (Fernández-Lanza, Graña, y Sobrino, 2003).

[4] Es decir, si la secuencia original en *lem* era *d2-d3-d1* y se toma un subconjunto {*d1,d2*} a devolver, los documentos se obtendrían en el mismo orden relativo original: *d2-d1*.

Bibliografía

Alonso, M. A., J. Vilares, y V. M. Darriba. 2002. On the usefulness of extracting syntactic dependencies for text indexing. En vol. 2464 de *Lecture Notes in Artificial Intelligence*. Springer-Verlag, pág. 3–11.

Buckley, C. 1985. Implementation of the SMART information retrieval system. Technical Report TR85-686, Department of Computer Science, Cornell University.

de Kretser, O. y A. Moffat. 1999a. Effective document presentation with a locality-based similarity heuristic. En *Proc. of SIGIR '99, Berkeley, USA*, pág. 113–120.

de Kretser, O. y A. Moffat. 1999b. Locality-based information retrieval. En *Proc. of 10th Australasian Database Conference (ADC '99), Auckland, New Zealand*, pág. 177–188.

Fernández-Lanza, S., J. Graña, y A. Sobrino. 2003. Introducing FDSA (Fuzzy Dictionary of Synonyms and Antonyms): Applications on Information Retrieval and Stand-Alone Use. *Mathware & Soft Computing*, 10(2-3):57–70.

Fox, E. A. y J. A. Shaw. 1994. Combination of multiple searches. En *The 2nd Text REtrieval Conference (TREC-2), Gaithersburg, USA*, pág. 243–252.

Graña, J., J.-C. Chappelier, y M. Vilares. 2001. Integrating external dictionaries into stochastic part-of-speech taggers. En

corpus	*CLEF 2001-02·A*				*CLEF 2001-02·B*				*CLEF 2003*			
consulta	*cortas*		*largas*		*cortas*		*largas*		*cortas*		*largas*	
técnica	*lem*	*cir*	*lem*	*cir*	*lem*	*cir*	*lem*	*cir*	*lem*	*cir*	*lem*	*cir*
#consultas	46	=	46	=	45	=	45	=	47	=	47	=
#docs. dev.	46k	=	46k	=	45k	=	45k	=	47k	=	47k	=
#rlvs. esp.	3007	=	3007	=	2513	=	2513	=	2335	=	2335	=
#rlvs. dev.	2767	=	2779	=	2376	=	2406	=	2240	=	2223	=
Pr. no int.	.5220	**.5327**	.5604	.5589	.4773	.4768	.5392	**.5497**	.5024	.4977	.5207	.5167
Pr.-R	.4990	**.5126**	.5366	**.5433**	.4599	.4551	.5104	**.5188**	.4912	.4737	.4871	.4865
Pr. a 0 %	.8221	**.8386**	.8895	**.9091**	.8210	**.8248**	.8710	**.8751**	.8145	**.8163**	.8301	.8257
Pr. a 10 %	.7490	**.7758**	.8028	**.8256**	.6861	**.7191**	.7619	**.7740**	.7369	.7283	.7421	**.7540**
Pr. a 20 %	.6866	**.7193**	.7352	**.7528**	.6319	**.6426**	.6929	**.7188**	.6632	**.6737**	.6758	**.6834**
Pr. a 30 %	.6573	**.6844**	.6996	.6922	.5688	**.5818**	.6497	**.6784**	.6019	.6015	.6304	**.6391**
Pr. a 40 %	.5997	**.6164**	.6541	**.6610**	.5289	**.5470**	.6202	**.6460**	.5638	**.5672**	.5975	.5876
Pr. a 50 %	.5456	**.5644**	.6005	**.6026**	.5017	.4909	.5733	**.5996**	.5410	.5359	.5479	.5327
Pr. a 5 docs.	.6609	**.6739**	.6957	**.7217**	.5956	**.6178**	.6844	**.6933**	.5872	**.6298**	.6213	**.6553**
Pr. a 10 docs.	.6457	**.6761**	.6848	**.7065**	.5600	**.5756**	.6178	**.6400**	.5596	**.5745**	.5872	**.5979**
Pr. a 15 docs.	.5884	**.6188**	.6435	**.6449**	.5274	**.5393**	.5822	**.6000**	.5305	**.5390**	.5504	**.5560**
Pr. a 20 docs.	.5630	**.5826**	.6043	**.6185**	.5011	**.5089**	.5533	**.5722**	.4883	**.5074**	.5266	.5170
Pr. a 30 docs.	.5225	.5225	.5580	**.5652**	.4444	.4444	.5081	**.5148**	.4433	.4433	.4667	**.4716**
Pr. a 100 docs.	.3507	.3502	.3598	.3539	.2940	**.3011**	.3191	**.3304**	.2770	**.2789**	.2853	.2809

Tabla 3: Resultados obtenidos mediante reordenación por fusión con intersección (*cir*) de la lematización con realimentación (*lem*)

Proc. of RANLP 2001, Tzigov Chark, Bulgaria, pág. 122–128

Hull, D. A., G. Grefenstette, B. M. Schulze, E. Gaussier, H. Schütze, y J. O. Pedersen. 1997. Xerox TREC-5 site report: Routing, filtering, NLP, and Spanish tracks. En *The 5th Text REtrieval Conference (TREC-5), Gaithersburg, USA*, pág. 167–180.

Jacquemin, C. y E. Tzoukermann. 1999. NLP for term variant extraction: synergy between morphology, lexicon and syntax. En *Natural Language Information Retrieval*, vol. 7 de *Text, Speech and Language Technology*. Kluwer Academic Publishers, pág. 25–74.

Kaszkiel, M. y J. Zobel. 2001. Effective ranking with arbitrary passages. *Journal of the American Society of Information Science*, 52(4):344–364.

Lee, J. H. 1997. Analyses of multiple evidence combination. En *Proc. of SIGIR '97, Philadelphia, USA*, pág. 267–276. ACM Press.

Llopis, F. 2003. *IR-n: Un sistema de Recuperación de Información basado en Pasajes*. PhD. Thesis, Universidad de Alicante.

Mitra, M., C. Buckley, A. Singhal, y C. Cardie. 1997. An analysis of statistical and syntactic phrases. En *Proc. of RIAO-97, 5th International Conference "Recherche d'Information Assistee par Ordinateur", Montreal, Canada*, pág. 200–214.

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

Rocchio, J.J., 1971. *The SMART Retrieval System–Experiments in Automatic Document Processing*, capítulo Relevance feedback in information retrieval, pág. 313–323. Prentice-Hall.

Salton, G. y C. Buckley. 1988. Term-weighting approaches in automatic retrieval. *Information Processing & Management*, 24(5):513–523.

Vilares, J., M. A. Alonso, F. J. Ribadas, y M. Vilares. 2002. COLE experiments at CLEF 2002 Spanish monolingual track. En *Working Notes for the CLEF 2002 Workshop, Rome, Italy*, pág. 153–160.

Vilares, J. y M. A. Alonso. 2003. A grammatical approach to the extraction of index terms. En *Proc. of RANLP 2003, Borovest, Bulgaria*, pág. 500–504.