Procesamiento del Lenguaje Natural, nº39 (2007), pp. 45-52 recibido 18-05-2007; aceptado 22-06-2007

Evaluación de Sistemas de Búsqueda de Respuestas con restricción de tiempo

Fernando Llopis[1], **Elisa Noguera**[1], **Antonio Ferrández**[1] y **Alberto Escapa**[2]
[1]Grupo de Investigación en Procesamiento del Lenguaje Natural y Sistemas de Información
Departamento de Sistemas y Lenguajes Informáticos
[2]Departamento de Matemática Aplicada
Universidad de Alicante
{elisa,llopis,antonio}@dlsi.ua.es // alberto.escapa@ua.es

Resumen: Las investigaciones sobre la evaluación de los sistemas de Búsqueda de Respuestas (BR) sólo se han centrado en la evaluación de la precisión de los mismos. En este trabajo se desarrolla un procedimiento matemático para explorar nuevas medidas de evaluación en sistemas de BR considerando el tiempo de respuesta. Además, hemos llevado a cabo un ejercicio para la evaluación de sistemas de BR en la campaña CLEF-2006 usando las medidas propuestas. La principal conclusión es que la evaluación del tiempo de respuesta puede ser un nuevo escenario para la evaluación de los sistemas de BR.
Palabras clave: Evaluación, Búsqueda de Respuestas

Abstract: Previous works on evaluating the performance of Question Answering (QA) systems are focused in the evaluation of the precision. Nevertheless, the importance of the answer time never has been evaluated. In this paper, we developed a mathematic procedure in order to explore new evaluation measures in QA systems considering the answer time. Also, we carried out an exercise for the evaluation of QA systems within a time constraint in the CLEF-2006 campaign, using the proposed measures. The main conclusion is that the evaluation of QA systems in realtime can be a new scenario for the evaluation of QA systems.
Keywords: Evaluation, Question Answering

1. Introducción

El objetivo de los sistemas de Búsqueda de Respuestas (BR) es localizar, en colecciones de texto, respuestas concretas a preguntas. Estos sistemas son muy útiles para los usuarios porque no necesitan leer todo el documento o fragmento de texto para obtener la información requerida. Preguntas como: *¿Qué edad tiene Nelson Mandela?, o ¿Quién es el presidente de los Estados Unidos?, ¿Cuando ocurrió la Segunda Guerra Mundial?* podrían ser contestadas por estos sistemas. Los sistemas de BR contrastan con los sistemas de Recuperación de Información (RI), ya que estos últimos tratan de recuperar los documentos relevantes respecto a la pregunta, donde la pregunta puede ser un simple conjunto de palabras clave (ej. edad Nelson Mandela, presidente Estados Unidos, Segunda Guerra Mundial,...).

La conferencia anual *Text REtrieval Conference (TREC*[1]*)*, organizada por el *National Institute of Standards and Technology (NIST)*, tiene como objetivo avanzar en el estudio de la RI y proveer de la infraestructura necesaria para una evaluación robusta de las metodologías de la recuperación textual. Este modelo ha sido usado por el *Cross-Language Evaluation Forum (CLEF*[2]*)* en Europa y por el *National Institute of Informatics Test Collection for IR Systems (NTCIR*[3]*)* en Asia, los cuales investigan el problema de la recuperación multilingüe. Desde 1999, TREC tiene una tarea específica para la evaluación de sistemas de BR (Voorhees y Dang, 2005). En las competiciones CLEF (Magnini et al., 2006) y NTCIR (F. et al., 2002) se han introducido también la evaluación de los sistemas de BR. Esta evaluación consiste en localizar las respuestas a un conjunto de preguntas en una colección de documentos, analizando los documentos de forma automática.

En estas evaluaciones, los sistemas tienen hasta una semana para responder al conjunto de preguntas. Esto es un problema en la evaluación de sistemas de BR porque nor-

[1]http://trec.nist.gov

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

[3]http://research.nii.ac.jp/ntcir

ISSN: 1135-5948

malmente son muy precisos, pero a la vez muy lentos, y esto hace muy difícil la comparación entre sistemas. Por esta razón, el objetivo de este trabajo es aportar un nuevo escenario para la evaluación de sistemas de BR con restricción de tiempo.

Este artículo está organizado de la siguiente forma: la sección 2 describe la evaluación de los sistemas de BR en el CLEF-2006. La sección 3 presenta una nueva propuesta de medidas de evaluación para sistemas de BR. La sección 4 describe el experimento llevado a cabo en el CLEF-2006 dentro del contexto de la BR. Finalmente, la sección 5 aporta las conclusiones y el trabajo futuro.

2. Evaluación de sistemas de BR en CLEF-2006

El objetivo en la tarea de BR en el CLEF es promover el desarrollo de los sistemas de BR dotando de una infraestructura para la evaluación de estos sistemas. Esta tarea tiene un creciente interés para la comunidad científica. En esta sección nos hemos centrado en describir los principales elementos de la tarea principal de BR en el CLEF-2006. Para más información consultar (Magnini et al., 2006).

2.1. Colección de preguntas

El conjunto de preguntas estaba formado por 200 preguntas, de las cuales 148 eran preguntas de tipo *factoid*, 42 de tipo *definition* y 10 de tipo *list*.

- Una pregunta *factoid* realiza la consulta sobre hechos o eventos. Por ejemplo, *¿Cuál es la capital de Italia?*. Se consideraron 6 tipos de respuesta esperada para estas preguntas: PERSONA, TEMPORAL, LOCALIZACIÓN, ORGANIZACIÓN, MEDIDA y OTRAS.
- Las preguntas de tipo *definition* requieren infomación sobre definiciones de gente, cosas u organizaciones. Un ejemplo de pregunta de este tipo podría ser: ¿Quién es el presidente de España?. Los tres tipos de respuesta para preguntas de tipo definición están divididos en: PERSONA, ORGANIZACIÓN, OBJECTO y OTROS.
- Una pregunta de tipo *list* requiere información de diferentes instancias de gente, objetos o datos, como *Lista los paises de Europa de Este*.

Fueron introducidas 40 preguntas con restricción temporal para los diferentes tipos de preguntas (*factoid*, *definition* y *list*). Concretamente, fueron introducidas tres tipos de restricciones temporales: FECHA, PERÍODO y EVENTO. *¿Quién ganó el Premio Nobel de la Paz en 1992?* es un ejemplo de pregunta con restricción de FECHA.

Además, hubieron varias preguntas que no tenían respuesta dentro de la colección. Estas respuestas son llamadas *NIL*. La importancia de éstas es porque los sistemas deben detectar si hay respuesta dentro de la colección y sino devolver la respuesta de tipo *NIL*.

Los participantes tuvieron una semana para enviar los resultados. Esto significa que los sistemas pueden ser muy lentos, lo cual no es una característica deseable para los sistemas de BR.

2.2. Evaluación de las respuestas

Las respuestas devueltas por cada participante fueron manualmente juzgadas por asesores nativos. En particular, cada idioma se coordinó por un grupo de asesores. Cada respuesta fue juzgada como: R (correcta) si la respuesta era correcta y estaba soportada por los fragmentos de texto devueltos, W (incorrecta) si la respuesta no era correcta, X (inexacta) si la respuesta contenía menos o más información de la requerida por la pregunta y U (no soportada) si los fragmentos de texto no contenían la respuesta, no fueron incluidos en el fichero de respuestas o no provenían del documento correcto.

2.3. Medidas de evaluación

Las respuestas fueron evaluadas principalmente usando la medida de evaluación: *accuracy*. También, se consideraron otras medidas: *Mean Reciprocal Rank (MRR)*, *K1* y *Confident Weighted Score (CWS)*.

$$accuracy = \frac{r}{n} \tag{1}$$

La medida *accuracy* se define como la proporción de respuestas correctas sobre el total de preguntas. Solamente se permite una respuesta por pregunta. Esto se obtiene con la fórmula (1), donde r es el número de respuestas correctas devueltas por el sistema y n es el número total de preguntas. Esta medida ha sido usada desde el CLEF-2004. La principal razón del uso de esta medida es porque normalmente sólo se evaluá una respuesta por pregunta.

$$MRR = \frac{1}{q}\sum_{i=1}^{q}\frac{1}{far_i} \quad (2)$$

En la conferencia QA@CLEF-2003, se usó la medida *MRR*, ya que en esa ocasión se permitieron 3 respuestas por pregunta. En cambio, este año se ha usado como medida adicional únicamente para evaluar los sistemas que devuelven más de una respuesta por pregunta. Esta medida asigna el valor inverso de la posición en la que la respuesta correcta fue encontrada, o cero si la respuesta no fue encontrada. El valor final es la media de los valores obtenidos para cada pregunta. MRR asigna un valor alto a las respuestas que están en las posiciones más altas de la clasificación. Esta medida está definida con la fórmula (2), donde q es el número de preguntas y far_i es la primera posición en la cual una respuesta correcta ha sido devuelta.

Los sistemas de BR devuelven las respuestas sin un orden establecido (simplemente se usa el mismo orden que en el conjunto de preguntas), aunque es opcional, algunos pueden asignar a cada respuesta un valor de confianza (entre 0 y 1). Este valor se utiliza para calcular dos medidas adicionales: *CWS* y *K1*. Estas medidas tienen en cuenta la precision y la confianza. De cualquier forma, la confianza es un valor opcional que sólo algunos sistemas de BR asignan, y solamente estos sistemas podrían ser evaluados con estas medidas. Para más información consultar (Magnini et al., 2006).

2.4. Limitaciones de las actuales evaluaciones en BR

En la actualidad, hay varios aspectos en las evaluaciones de los sistemas de BR que podrían ser mejorados: (1) los participantes tienen varios días para responder a las preguntas, (2) el tiempo de respuesta no se evaluá, esto causa que los sistemas tengan un buen rendimiento, pero que sean sistemas demasiado lentos, y (3) la comparación entre sistemas de BR puede ser difícil si tienen diferente tiempo de respuesta. En consecuencia, el análisis del rendimiento involucra la evaluación de la eficiencia y de la eficacia de los sistemas de BR.

La motivación de este trabajo es estudiar la evaluación de los sistemas de BR con restricción de tiempo. Concretamente, hemos propuesto nuevas medidas de evaluación que combinan la precisión y el tiempo de respuesta de los sistemas. Para evaluar el tiempo de respuesta de los sistemas, hemos llevado a cabo un experimento en el CLEF-2006 aportando un nuevo escenario para comparar sistemas de BR. Observando los resultados obtenidos por los sistemas, podemos argumentar que este es un prometedor paso para cambiar la dirección en la evaluación de los sistemas de BR.

3. *Nuevas aproximaciones sobre la evaluación de los sistemas de BR*

El problema mencionado anteriormente puede ser reformulado de forma matemática. Consideramos que la respuesta de cada sistema S_i puede ser caracterizada en este problema como un conjunto de pares de números reales ordenados (x_i, t_i). El primer elemento de cada par representa la precisión del sistema y el segundo la eficiencia. De este modo, la tarea de BR puede ser representada geométricamente como un conjunto de puntos localizados en un subconjunto $D \subseteq \mathbb{R}^2$. Nuestro problema puede ser solventado aportando un método que permita ordenar los sistemas S_i de acuerdo a un criterio prefijado que valore tanto la precisión como la eficiencia. Este problema es de la misma naturaleza que otros problemas tratados en la Teoría de Decisión.

Una solución a este problema puede ser obtenido introduciendo un preorden total, a veces referido como quasiorden, en D. Una relación binaria $\preceq$ en un conjunto D es un preorden total si es reflexivo, transitivo y si dos elementos (cualesquiera) de D son comparables entre si. En concreto, podemos definir un quasiorden en D con la ayuda de una función con dos variables de tipo real $f : D \subseteq \mathbb{R}^2 \rightarrow I \subseteq \mathbb{R}$, de modo que: $(a, b) \preceq (c, d) \Leftrightarrow f(a, b) \leq f(c, b), \forall\,(a, b), (c, d) \in D$.

Nos referiremos a esta función como función de clasificación. Una de las ventajas de este procedimiento es que la función de clasificación contiene toda la información relativa al criterio elegido para clasificar los distintos sistemas S_i.

Matemáticamente, todos los elementos que están situados en la misma posición en la clasificación pertenecen a una misma curva de nivel en la función de clasificación. Específicamente, las curvas de *iso-ranking* están caracterizadas por todos los elementos de D que completan la ecuación $f(x, t) = L$, siendo

L un número real en la inversa de f, I.

El procedimiento de clasificación propuesta para evaluar los sistemas en la tarea de BR es de tipo ordinal. Esto significa que no se debe hacer una conclusión sobre la diferencia numérica absoluta sobre la diferencia de los valores numéricos para dos sistemas en la función de clasificación. La única información relevante es la posición relativa en la clasificación de los sistemas en la tarea de evaluación de BR. De hecho, si consideramos una nueva función de clasificación construida componiendo la función de clasificación inicial con un estricto incremento de la función, el valor numérico asignado a cada sistema cambiará, pero la clasificación obtenida será la misma que inicialmente.

En la aproximación desarrollada en este artículo, la precisión x_i del sistema S_i es calculada con la medida de evaluación *Mean Reciprocal Rank (MRR)*, de modo que $x_i \in [0, 1]$. La eficiencia se mide considerando el tiempo de respuesta de cada sistema, de modo que, tener un tiempo de respuesta pequeño significa tener una buena eficiencia.

Para definir una función de clasificación realista, es necesario establecer algunos requirimientos adicionales. Estas propiedades están basadas en el comportamiento intuitivo que debe cumplir la función. Por ejemplo, como aproximación inicial, vamos a establecer las siguientes condiciones:

1. La función f debe ser continua en D.
2. El límite superior de I se obtiene con $\lim_{t\to 0} f(1,t)$. En el caso que I no tenga límite superior, tendremos $\lim_{t\to 0} f(1,t) = +\infty$.
3. El límite inferior de I se obtiene con $f(0,1)$.

La primera condición se ha impuesto por conveniencia matemática, aunque se podría interpretar en términos de simplificación de argumentos. Cabe destacar que este requerimiento excluye la posibilidad que, si suponemos que dos sistemas están en distintas posiciones en la clasificación, una pequeña variación en la precisión o la eficiencia, pueda alterar los valores de la clasificación. La segunda condición está relacionada con el hecho que, si suponemos un sistema definido por el par $(1, 0)$ siempre debería estar en la primera posición en la clasificación. Finalmente, la última condición implica que el par $(1, 0)$ debería estar en la última posición.

3.1. Función de clasificación independiente del tiempo (MRR_2)

Como primer ejemplo de función de clasificación, consideramos $MRR_2(x, t) = x$. El preorden inducido por esta función es semejante al orden lexicográfico, a veces llamado orden alfabético. Para esta función de clasificación tenemos que:

1. La función inversa de MRR_2 está en el intervalo $[0, 1]$.
2. La función MRR_2 es continua en D.
3. $\lim_{t\to 0} MRR_2(1, t) = 1$.
4. $MRR_2(0, 1) = 0$.

De modo que, la función cumple las condiciones establecidas previamente. Por otro lado, las curvas de *iso-ranking* de la función son de la forma $x = L$, $L \in [0, 1]$ cuya representación es una familia de segmentos verticales con una unidad de longitud (veáse la figura 1). El preorden construido por esta función de clasificación sólo valora la precisión de los sistemas.

3.2. Función de clasificación con dependencia temporal inversa ($MRRT$)

Como el primer ejemplo de función de clasificación no valora la eficiencia de los sistemas, vamos a considerar la función $MRRT(x, t)$. Suponemos que en este caso la función de clasificación es inversamente proporcional a la eficiencia (tiempo de respuesta) y directamente proporcional a la precisión. En particular, esta función verifica las siguientes propiedades:

1. La función inversa de $MRRT$ está en el intervalo $[0, +\infty)$.
2. La función $MRRT$ es continua en D.
3. $\lim_{t\to 0} MRRT(1, t) = +\infty$.
4. $MRRT(0, 1) = 0$.

Las curvas de *iso-ranking* asociadas a la función son de la forma $x = L$, $L \in [0, 1]$. Geométricamente, estas curvas son una familia de segmentos que pasan por el punto

$(0, 0)$ y con una pendiente de $1/L$ (veáse la figura 2). De este modo, los sistemas con mejor eficiencia, es decir, un tiempo de respuesta pequeño, obtendrán un mejor valor de x y una posición alta en la clasificación. Así mismo, aunque la función de clasificación es de naturaleza ordinal, es deseable que la función inversa este acotada entre 0 y 1, ya que esto facilita su intuitiva representación, condición que no se cumple por esta función.

3.3. Función de clasificación exponencial inversa con dependencia del tiempo $MRRT_e$

Debido a las desventajas presentadas en las funciones anteriores, hemos propuesto una nueva función que también depende de la precisión y de la eficiencia del sistema, aunque la eficiencia tiene un menor peso que la precisión en esta función. A continuación, vamos a introducirla:

$$MRRT_e(x, t) = \frac{2x}{1 + e^t}, \qquad (3)$$

siendo e^t la función exponencial de la eficiencia. Esta función cumple las siguientes condiciones:

1. La inversa de $MRRT_e$ está en el intervalo $[0, 1)$.
2. La función $MRRT_e$ es continua en D.
3. $\lim_{t \to 0} MRRT_e(1, t) = 1$.
4. $MRRT_e(0, 1) = 0$.

Las curvas de *iso-ranking* son de la forma $2x/(1 + e^t) = L$, $L \in [0, 1)$, estando representadas en la figura 3. Si suponemos un sistema ideal, es decir, que responde instantáneamente ($t = 0$), entonces el valor de esta función coincidiría con el valor de la función de precisión. En cambio, la dependencia funcional del tiempo modula el valor de la función, de modo que, cuando el tiempo incrementa, la función decrece. De cualquier forma, esta dependencia es más suave que en la función anterior. Además, si consideramos un sistema S, únicamente obtendremos la misma clasificación que él si consideramos sistemas cuya precisión y eficiencia varian en un rango particular, no sólo para un valor pequeño de la precisión.

4. Evaluación en el CLEF-2006

Como se ha descrito anteriormente, nosotros consideramos el tiempo como parte fundamental en la evaluación de los sistemas de BR. En acuerdo con la organización del CLEF, llevamos a cabo una tarea experimental en el CLEF-2006, cuyo objetivo era evaluar los sistemas de BR con una restricción de tiempo. Éste fue un experimento innovador para la evaluación de los sistemas de BR y fue una iniciativa para aportar un nuevo escenario en la evaluación de los sistemas de BR. El experimento sigue las mismas directrices que la tarea principal, descrita en la sección 2, pero considerando el tiempo de respuesta.

4.1. Participantes

En total, 5 grupos participaron en este ejercicio experimental. Los grupos participantes fueron: *daedalus* (España) (de Pablo-Sánchez et al., 2006), *tokyo* (Japón) (Whittaker et al., 2006), *priberam* (Portugal) (Cassan et al., 2006), *alicante* (España) (Ferrández et al., 2006) y *inaoe* (Mexico) (Juárez-Gonzalez et al., 2006). Todos estos sistemas participaron también en la tarea principal del CLEF-2006 y tienen experiencia en investigación en sistemas de BR.

4.2. Evaluación

En esta sección se presentan los resultados de la evaluación de los 5 sistemas que participaron en el experimento. Por un lado, se presenta la precisión y la eficiencia obtenida por estos sistemas. Por otro lado, se presentan las puntuaciones obtenidas por cada uno de ellos con las diferentes medidas, las cuales combinan la precisión y la eficiencia (presentada en la sección 2.3).

La tabla 1 muestra el resumen de los resultados obtenidos con las diferentes medidas de evaluación (MRR, t, MRRT, $MRRT_e$). Se muestran todos los resultados en una sola tabla para hacer más fácil la comparación entre las diferentes medidas. También se muestra la posición (pos) obtenida por cada sistema con respecto a cada medida.

4.2.1. Evaluación de la precisión y del tiempo de respuesta

La precisión de los sistemas de BR fue evaluada en el experimento con la medida MRR (ver la sección 2.3). Nosotros usamos esta medida porque los sistemas enviaron tres respuestas por pregunta. La evaluación de los sistemas con esta medida se presenta en la

Participante	MRR	pos	t	pos	MRRT	pos	$MRRT_e$	pos
daedalus1	**0.41**	**1º**	0.10	4º	3.83	4º	**0.38**	**1º**
tokyo	0.38	2º	1.00	6º	0.38	6º	0.20	6º
priberam	0.35	3º	**0.01**	**1º**	**32.13**	**1º**	0.34	**2º**
daedalus2	0.33	4º	0.03	3º	8.56	3º	0.32	3º
inaoe	0.3	5º	0.38	5º	0.78	5º	0.24	**4º**
alicante	0.24	6º	0.02	2º	16.23	2º	0.23	**5º**

Cuadro 1: Evaluación de los resultados obtenidos con las diferentes medidas de evaluación

tabla 1. Por otra parte, los tiempos de respuesta se midieron en segundos ($tsec$), aunque en la tabla se presenta el tiempo de respuesta (t) normalizado para cada sistema con respecto a $tmax$, o tiempo de respuesta del sistema menos rápido. Es decir, t es igual a $tsec/tmax$.

4.2.2. Evaluación de los resultados con MRR_2

La evaluación global de los sistemas de BR, combinando precision y tiempo de respuesta con la medida $MRRT_2$ (ver sección 3) es la misma que usando sólo la medida MRR (ver sección 1), porque esta medida valora primero la precisión, y después valora el tiempo en el caso que la precisión sea la misma entre varios sistemas. En este caso, como la precisión es distinta, los sistemas quedarían ordenados por su MRR.

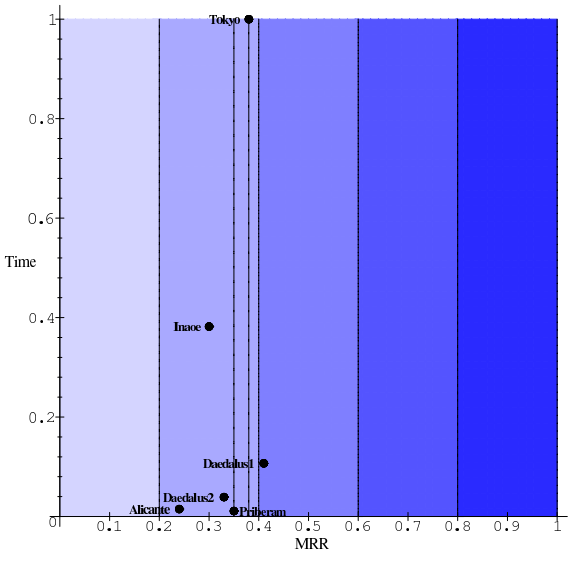

Figura 1: Comparativa de los resultados obtenidos para cada sistema con la medida de evaluación MRR_2 (Preorden lexicográfico).

Gráficamente, una curva de iso-ranking contiene a todos los sistemas con el mismo valor de MRR y cualquier valor de tiempo de respuesta. Es decir, el criterio para establecer la clasificación es el mismo que la precisión obtenida para evaluar los sistemas de BR. Las limitaciones de este procedimiento, las cuales han sido argumentadas en este trabajo, son claras si consideramos por ejemplo los sistemas *priberam* y *tokyo* en la figura 1. Podemos observar como *tokyo* está en segunda posición en el ranking y el sistema *priberam* está el tercero. En cambio, la diferencia en la precisión de los dos sistemas es muy pequeña, 0.38 vs. 0.35, mientras que la eficiencia del sistema *priberam* es mucho mejor que la eficiencia del sistema *tokyo*. En consencuencia, sería razonable que el sistema *priberam* precediera al sistema *tokyo*. Esto es imposible con esta clase de medidas que son independientes del tiempo.

4.2.3. Evaluación de los resultados con MRRT

La evaluación de los sistemas con la medida MRRT (ver la sección 3) se presenta en la tabla 1. También, para cada sistema se muestra la posición en la lista que ha obtenido con esta medida.

Como podemos observar en la tabla, *priberam* obtuvo el mejor valor de MRRT (32.13) con un t de 0.01 y un MRR de 0.35. Además, también se puede observar que la primera prueba enviada por *daedalus* (daedalus1) obtuvo el mejor MRR con 0.41, en cambio esta prueba no fue la más rápida (0.10). En consecuencia, esta prueba obtuvo un bajo MRRT (0.08). La segunda prueba enviada por *daedalus* (daedalus2) obtuvo un MRR más bajo que el anterior (0.33), en cambio obtuvo un mejor t (0.03), por esta razón esta segunda prueba obtuvo un mejor MRRT que la primera prueba.

Gráficamente, podemos ver los diferentes valores obtenidos en la figura 2. Por ejemplo, el sistema *alicante*, cuya presición es 0.24 y t es 0.02, está en la misma posición en la clasificación que *priberam*, siendo su precisión

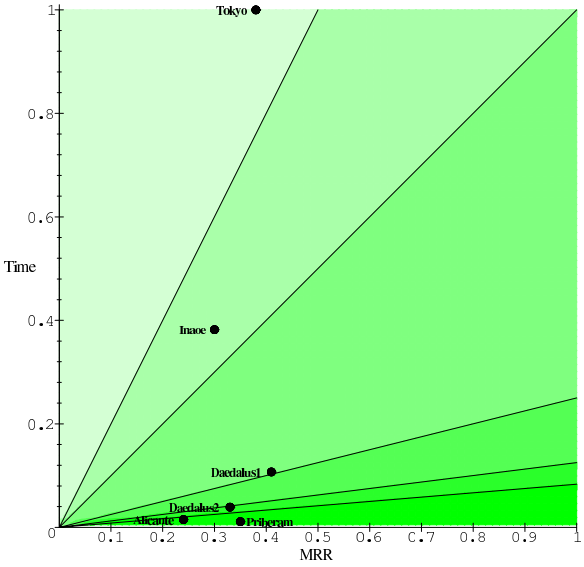

Figura 2: Comparativa de los resultados obtenidos por cada sistema con la medida de evaluación $MRRT$ en sus curvas de iso-ranking.

mejor (0.35). La posición de cualquier sistema en la clasificación, puede ser igualada por un sistema de menor precisión pero con una mayor eficiencia, y en particular esto puede ocurrir aún teniendo un valor pequeño en la precisión. Esto es una desventaja porque se valora mucho la eficiencia de los sistemas y, en nuestra opinión, el factor principal debe de ser la precisión, aunque la eficiencia también sea valorada.

4.2.4. Evaluación de los resultados con $MRRT_e$

La medida de evaluación MRRT, presentada en la sección anterior, fue usada en la tarea de BR con restricción de tiempo dentro del CLEF-2006. Consideramos que esta medida valora demasiado el tiempo, por lo tanto, hemos propuesto una medida alternativa más adecuada para la evaluación de sistemas de BR con restricción de tiempo. La nueva medida, descrita en la sección 3, ha sido diseñada para penalizar aquellos sistemas que tienen un elevado tiempo de respuesta.

Como muestra la tabla 1, *daedalus1* y *priberam* obtienen los mejores resultados con la medida $MRRT_e$ (0.38 y 0.34 respectivamente). La disminución de resultados de *priberam* (de 0.35 a 0.34), en términos de MRR, no es significativa porque tiene un tiempo de respuesta muy pequeño (0.01), al igual que *alicante* (de 0.24 a 0.23). En cambio, el valor de $MRRT_e$ de daedalus1 reduce su valor de MRR en mayor grado (de 0.41 a 0.38), porque tiene un t más elevado (0.10) que los anteriores. Finalmente, *inaoe* y *tokyo* han sido penalizados significativamente por tener unos tiempos de respuesta muy elevados.

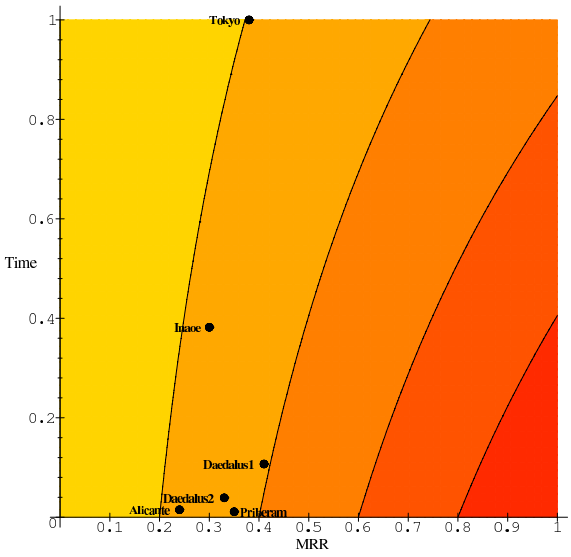

Figura 3: Comparativa de los resultados obtenidos por cada sistema con la medida de evaluación $MRRT_e$ en sus curvas de iso-ranking.

Gráficamente, podemos comparar los distintos valores de $MRRT_e$ en la figura 3. También se puede observar en la figura que para obtener la misma posición en el ranking que, p.ej. un sistema con una precisión de 0.4 y un t de 0.2, su precisión oscilará entre (0.36, 0.76) y su t variará entre 0 y 1 dependiendo de su precisión. Estas características hacen la medida de evaluación $MRRT_e$ adecuada para la evaluación de sistemas de BR con restricción de tiempo.

5. *Conclusiones y trabajos futuros*

Principalmente, la evaluación de sistemas de BR ha sido estudiado en profundidad en tres foros de investigación: TREC, CLEF y NTCIR. Aunque, en estos foros sólo se han centrado en evaluar la precisión de los sistemas, y no se ha valorado su eficiencia (consideramos el tiempo de respuesta como medida de eficiencia) en ninguna ocasión. En la mayor parte de los casos, los sistemas suelen ser muy eficaces pero muy poco eficientes. Por esta razón, hemos estudiado en este trabajo la evaluación de sistemas de BR valorando también su tiempo de respuesta.

Para la evaluación de los sistemas de BR, hemos propuesto tres medidas (MRR_2,

$MRRT$, $MRRT_e$) para evaluar los sistemas con restricción de tiempo. Estas medidas están basadas en la medida *Mean Reciprocal Rank (MRR)* y el tiempo de respuesta. Como resultados preliminares, hemos visto que $MRRT_2$ sólo valora la precisión y $MRRT$ valora demasiado el tiempo. Hemos solventado este inconveniente proponiendo una nueva medida llamada $MRRT_e$. Esta medida combina el MRR y el tiempo de respuesta, penalizando a los sistemas que tienen un tiempo de respuesta elevado. Cabe mencionar, que está basada en una función exponencial. En conclusión, la nueva medida MRR_e permite clasificar los sistemas considerando su precisión y su tiempo de respuesta.

Además, hemos llevado a cabo una tarea en el CLEF-2006 para evaluar sistemas de BR con restricción de tiempo (siendo la primera vez que se organiza una evaluación de estas características). Este experimento nos ha permitido establecer los criterios para la evaluación de sistemas de BR en un nuevo escenario. Afortunadamente, este experimento fue recibido con una gran expectación tanto por los participantes, como por los organizadores.

Finalmente, las futuras direcciones que vamos a seguir son: valorar otras variables como el *hardware* de los sistemas, e insertar nuevos parámetros de control para poder dar más importancia a la precisión o a la eficiencia.

Bibliografía

Cassan, A., H. Figueira, A. Martins, A. Mendes, P. Mendes, C. Pinto, y D. Vidal. 2006. Priberam's Question Answering System in a Cross-Language Environment. En *WORKING NOTES CLEF 2006 Workshop*.

de Pablo-Sánchez, C., A. González-Ledesma, A. Moreno, J. Martínez-Fernández, y P. Martínez. 2006. MIRACLE at the Spanish CLEF@QA 2006 Track. En *WORKING NOTES CLEF 2006 Workshop*.

F., Junichi, Tsuneaki K., , y Fumito M. 2002. An Evaluation of Question Answering Task. En *Third NTCIR Workshop on Research in Information Retrieval, Question Answering and Summarization*, October.

Ferrández, S., P. López-Moreno, S. Roger, A. Ferrández, J. Peral, X. Alvarado, E.Noguera, y F. Llopis. 2006. AliQAn and BRILI QA Systems at CLEF 2006. En *WORKING NOTES CLEF 2006 Workshop*.

Juárez-Gonzalez, A., A. Téllez-Valero, C. Denicia-Carral, M. Montes y Gómez, y L. Villase nor Pineda. 2006. INAOE at CLEF 2006: Experiments in Spanish Question Answering. En *WORKING NOTES CLEF 2006 Workshop*.

Magnini, B., D. Giampiccolo, P. Forner, C. Ayache, P. Osenova, A. Pe nas, V. Jijkoun, B. Sacaleanu, P. Rocha, y R. Sutcliffe. 2006. Overview of the CLEF 2006 Multilingual Question Answering Track. En *WORKING NOTES CLEF 2006 Workshop*.

Voorhees, E. y H. Trang Dang. 2005. Overview of the TREC 2005 Question Answering Track. En *TREC*.

Whittaker, E. W. D., J. R. Novak, P. Chatain, P. R. Dixon, M. H. Heie, y S. Furui. 2006. CLEF2006 Question Answering Experiments at Tokyo Institute of Technology. En *WORKING NOTES CLEF 2006 Workshop*.