Revista Española de Documentación Científica
36(2), abril-junio 2013, e009
ISSN-L:0210-0614. doi: http://dx.doi.org/10.3989/redc.2013.2.921

ESTUDIOS / *RESEARCH STUDIES*

Evaluación del rendimiento de los sistemas de búsqueda de respuestas de dominio general

María-Dolores Olvera-Lobo*, Juncal Gutiérrez-Artacho**

* CSIC, Unidad Asociada Grupo SCImago, Madrid; Departamento de Información y Documentación, Universidad de Granada
** Departamento de Traducción e Interpretación, Universidad de Granada.
Correo-e: molvera@ugr.es; juncalgutierrez@ugr.es

Recibido: 06-12-2011; 2ª versión: 31-01-2012; Aceptado: 07-02-2012

Cómo citar este artículo/ Citation: Olvera-Lobo, M. D.; Gutiérrez-Artacho, J. (2013). Evaluación del rendimiento de los sistemas de búsqueda de respuestas de dominio general. *Revista Española de Documentación Científica*, 36(2):e009. doi: http://dx.doi.org/10.3989/redc.2013.2.921

Resumen: Los sistemas de búsqueda de respuestas (SBR) son una alternativa a los tradicionales sistemas de recuperación de información tratando de ofrecer respuestas precisas y comprensibles a preguntas factuales, en lugar de presentar al usuario una lista de documentos relacionados con su búsqueda. Se ha evaluado la eficacia de cuatro SBR disponibles en la Web —*QuALiM, SEMOTE, START, y TrueKnowledge*—, mediante una amplia muestra de preguntas de definición, factuales y de lista, pertenecientes a distintos dominios temáticos. Se utilizó una colección de 500 preguntas cuyas respuestas fueron valoradas por los usuarios y, posteriormente, se aplicaron varias medidas para su evaluación (MRR, TRR, FHS, MAP y precisión). Se observa que *START* y *TrueKnowledge* presentan un nivel aceptable de respuestas correctas, precisas y en una secuencia bien ordenada. Los resultados obtenidos revelan el potencial de esta clase de herramientas en el ámbito del acceso y la recuperación de información de dominio general.

Palabras clave: Recuperación de información; sistemas de búsqueda de respuestas; evaluación; medidas de evaluación; World Wide Web.

Performance Analysis in Web-based Question Answering Systems

Abstract: Information overload is felt more strongly on the Web than elsewhere. Question-answering systems (QA systems) are considered as an alternative to traditional information retrieval systems, because they give correct and understandable answers rather than just offering a list of documents. Four answer search systems available online have been analyzed: START, QuALiM, SEMOTE, and TrueKnowledge. They were analyzed through a wide range of questions that prompted responses of definitions, facts, and closed lists pertaining to different thematic areas. The answers were analyzed using several specific measurements (MRR, TRR, FHS, MAP and precision). The results are encouraging and they show that these systems, although each one different, are potentially valid for precise information retrieval of diverse types and thematic areas.

Keywords: Question-Answering Systems; performance analysis; definitional questions; factoid question; list questions; evaluation.

1. INTRODUCCIÓN

En el entorno de la Web la sobrecarga de información se deja sentir aún más que en otros contextos. De esta forma, en demasiadas ocasiones, al plantear una determinada consulta en las herramientas de búsqueda de información web (buscadores, directorios o metabuscadores) el número de páginas web recuperadas resulta excesivo y no todas ellas son relevantes ni útiles para los objetivos del usuario. Los sistemas de búsqueda de respuestas o SBR (en inglés, *question-answering systems*) se presentan como una alternativa a los tradicionales sistemas de recuperación de información tratando de ofrecer respuestas precisas y comprensibles a preguntas factuales, en lugar de mostrar al usuario una lista de documentos relacionados con su búsqueda (Jackson y Schilder, 2005). El funcionamiento de los sistemas de BR se basa en los modelos de respuestas cortas (Blair-Goldensohn y otros, 2004), y la ventaja principal que ofrece al usuario es que éste no ha de consultar documentos completos para obtener la información requerida puesto que el sistema ofrece la respuesta correcta en forma de un número, un sustantivo, una frase corta o un fragmento breve de texto (Pérez-Coutiño y otros, 2004).

Puesto que la búsqueda de respuestas se presenta como un avance destacado en la mejora de la recuperación de información (Kolomityets y Moens, 2011), se hace necesario determinar su eficacia para el usuario final. Con este objetivo se ha realizado un estudio donde se evalúa el rendimiento y la calidad de las respuestas de los principales SBR de dominio general disponibles en la Web (*QuALiM*[1], *SEMOTE*[2], *START*[3] y *TrueKnowledge*[4]) ante preguntas de diversos tipos (de definición, factuales y de lista) y temas (Arte y Literatura, Biología, Personajes, Historia, Economía o Deportes, entre otros), para lo que se aplican diferentes medidas de evaluación. A continuación se detalla el análisis realizado. Los objetivos del trabajo son comparar y evaluar las respuestas ofrecidas por cada sistema de BR de dominio general ante 500 preguntas factuales, de definición y de lista, de modo que podamos confirmar su relevancia y eficacia.

2. SISTEMAS DE BÚSQUEDA DE RESPUESTAS

Desde el punto de vista de la recuperación de información, el uso del lenguaje natural favorece el acceso a los contenidos al permitirle al usuario recurrir a su forma habitual de expresión. Los SBR normalmente presentan una sencilla interfaz con un motor de búsqueda mediante el cual los usuarios pueden formular su pregunta, e incluso algunos proporcionan un listado de las últimas cuestiones introducidas para ayudarles a entender cómo han de plantearlas. Ciertamente, estos sistemas intentan emular el comportamiento del lenguaje humano por lo que tratan de entender la pregunta formulada en lenguaje natural y proporcionar respuestas adecuadas. En otras palabras, la interpretación del lenguaje natural por el sistema es un proceso esencial en el desarrollo de los SBR (Belkin y Vickery, 1985; Sultan, 2006). Tanto es así que el análisis de la pregunta, así como la búsqueda y la extracción de las respuestas son tres importantes tareas llevadas a cabo por los SBR, las cuáles implican, al menos, el procesamiento de las preguntas, el procesamiento de los documentos y el procesamiento de las respuestas (Kangavari y otros, 2008).

Los primeros sistemas de búsqueda de respuestas surgieron en los años 60 y utilizaban bases de datos de dominio restringido con información estructurada. Ejemplos clásicos son Baseball (Green y otros, 1961), una base de datos de partidos de béisbol -*How many games did the Yankees play in July?*, Lunar (Woods y otros,1972), una base de datos de análisis químicos de las misiones lunares de Apollo -*What is the average concentration of aluminium in high alkali rocks?*- o Chat-80 (Warren, 1981), una base de datos geográficos -*Which is the largest African country?*- con una versión moderna que convierte la pregunta en SQL. Otro tipo de sistemas de BR son los sistemas de diálogo como el clásico Eliza (Weizenbaum, 1966). Este sistema simulaba un psicoanalista y puede considerarse precursor de los actuales *chatterbot* –software diseñado para simular una conversación inteligente con uno o más humanos por medio de texto y/o audio-. Por último, los antecesores más inmediatos de los sistemas web de búsqueda de respuestas, en los que aquí nos centramos, son los sistemas de búsqueda en documentos de texto, los cuáles tomaron un importante impulso a partir de la conferencia TREC-8 (Text REtrieval Conference) (Voorhees, 1999).

En el tratamiento y la gestión de las preguntas, los SBR aplican algoritmos y métodos de análisis lingüístico y de procesamiento del lenguaje natural para identificar sus componentes y determinar la clase de respuesta esperada (Zweigenbaum, 2005). El tipo de preguntas que suelen permitir son las denominadas preguntas factuales, de definición y de lista. Las preguntas factuales son las relacionadas con datos o hechos concretos, nombres propios, etc., se expresan mediante partículas interrogativas (*who*, *what*, *where*, *when, how*) y persiguen una respuesta concreta y rápida (un nombre, una fecha, un lugar, una cantidad). Este tipo de preguntas constituye la mayoría de las consultas (*"who won the Nobel Prize for Literature in 1994?", what actress starred in "The Lion in Winter"?, "when was the telegraph invented?", "how did Jimi Hendrix die?" "where are the Rocky Mountains"*?). Las preguntas de definición, como su nombre indica, persiguen obtener la definición de un término, organización, etc., y están formuladas como *"what is…?"* (*"what is angiotensin*?"). En estos casos, las respuestas más relevantes serán las que ofrezcan información de manera eficiente, con el menor número de palabras, pero de construcción similar a las entradas de una enciclopedia

(Greenwood y Saggion, 2004; Olvera-Lobo y Gutiérrez-Artacho, 2011a). Finalmente, las preguntas de lista son aquellas que solicitan un cierto número de respuestas de un mismo tipo y suelen plantearse de forma imperativa (*"tell me…", "name all of London's airports"* o *"List 5 pharmaceutical companies that manufacture antibiotics"*).

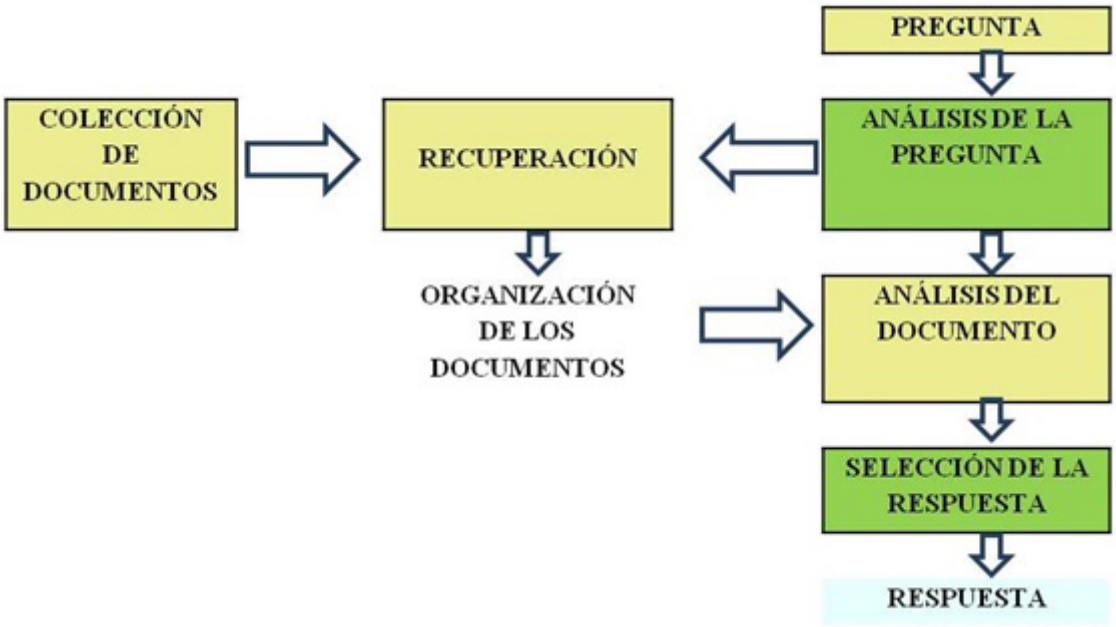
Como parte de la arquitectura de los SBR, el módulo de procesamiento de documentos se encarga de realizar una primera selección de los documentos o párrafos que se pueden considerar como relevantes para la pregunta planteada (véase figura 1). Las fuentes de información que utilizan los sistemas para seleccionar estos documentos son de lo más variadas y van desde la omnipresente *Wikipedia* hasta enciclopedias, diccionarios o bases de datos especializadas de gran prestigio como *Medline* (Olvera-Lobo y Gutiérrez-Artacho, 2011b). La elección de las fuentes de información es una decisión habitualmente condicionada por el hecho de que se trate de un SBR de dominio general –y, por tanto, capaz de atender consultas de temas muy diversos, como *START Natural Language Question Answering* o *NSIR Question Answering System*[5]– o de dominio específico –si se centran en un ámbito temático determinado, como *HONQA Health On the Net Foundation*[6] (Crouch y otros, 2005; Olvera-Lobo y Gutiérrez-Artacho, 2011c) or *EAGLi Engine for question-Answering in Genomics Literature* [7] (Abdou y otros, 2006) –.

Por último, el módulo de procesamiento de la respuesta lo forman dos importantes componentes, el destinado a la extracción de la respuesta y el dedicado a la validación de la misma. Las respuestas candidatas se extraen de los documentos que son recuperados por el motor de búsqueda del SBR. Tras ello, mediante el filtrado y la ordenación de las respuestas candidatas, se validan las respuestas que serán las que finalmente se muestren al usuario (Kangavari y otros, 2008). El objetivo de esta etapa es eliminar cualquier pasaje incorrecto o redundante que se encuentre en la lista recuperada por el SBR. No siempre la respuesta será única y el sistema puede proveer varias respuestas correctas que satisfagan la necesidad del usuario (Cui y otros, 2004; Rodrigo y otros, 2010; Tsur, 2003).

3. MÉTODO Y MATERIALES

La colección de preguntas que se ha utilizado en este estudio incluye preguntas factuales, de definición y de lista en lengua inglesa, y se creó a partir de las colecciones de preguntas de evaluación propuestas por dos de las principales conferencias sobre recuperación de información a nivel internacional, TREC (*Text Retrieval Conference*) y CLEF (*Cross-Language Evaluation Forum*). Las colecciones de evaluación generadas en estos foros son utilizadas por los participantes para llevar a cabo la evaluación de sus sistemas, de manera que los resultados obtenidos puedan compararse con los de los demás. Partiendo de las colecciones de preguntas de los años 2000 a 2004 se obtuvo una serie de casi 2000 preguntas de definición, factuales y de lista –para las que existen métodos de evaluación claramente definidos (Voorhees, 2002)– que versaban sobre diferentes temas y especialidades (véanse tablas I a III).

Figura 1. Arquitectura general de un SBR



Finalmente, se utilizaron para la evaluación las 500 preguntas (véase tabla 4), tanto de dominio general como específico, que obtuvieron respuesta por parte de los cuatro sistemas analizados, a saber, *QuaLiM, SEMOTE* −sistemas que recientemente han dejado de estar operativos− *START, y TrueKnowledge*. Se trata de SBR gratuitos, monolingües, de dominio general, disponibles en la Web, y que ofrecen una amplia cobertura temática ante diferentes tipos de preguntas.

QuALiM era un sistema financiado por *Microsoft* y desarrollado por el investigador Michael Kaisser de la Universidad de Edimburgo. Aunque se definía como una *demo*, se trataba de un sistema que contaba con un funcionamiento aceptable y recuperaba tanto información textual −para lo que utilizaba únicamente la enciclopedia *Wikipedia*− como gráfica −extraída del buscador de imágenes de Google− (Kaisser, 2008). Se caracterizaba por presentar una interfaz muy sencilla y breves explicaciones con ejemplos. Por su parte, *SEMOTE* era un sistema que permitía a los usuarios plantear preguntas sobre diferentes dominios temáticos. Se caracterizaba por utilizar una amplia variedad de recursos −desde páginas web dedicadas a temas específicos hasta portales web de más amplia cobertura− para extraer las repuestas a las preguntas planteadas. Además, los resultados ofrecidos solían ser bastante exhaustivos. Como se ha indicado, estos dos últimos sistemas no se encuentran actualmente operativos si bien en el momento de realizar nuestra evaluación sí estaban en funcionamiento y extraían las respuestas de fuentes de información actualizadas.

Tabla I. Procedencia de las preguntas de la muestra

	CLEF	**TREC**	**Total**
Nº Preguntas	597	1383	1980

Tabla II. Preguntas por año

	Año					
	2000	**2001**	**2002**	**2003**	**2004**	**Total**
Nº Preguntas	730	475	100	475	200	1980

Tabla III. Temas a los que se refieren las preguntas de la muestra

Temas de las preguntas de evaluación									
Arte y Literatura	269	**Ciencia**	251	**Deportes**	91	**Economía**	156	**General**	324
Geografía	178	**Historia**	255	**Medicina**	86	**Personajes**	219	**Política**	151
Total									1980

Tabla IV. Preguntas según el tipo de respuesta esperada

	P. definición	**P. factuales**	**P. de lista**
Nº Preguntas	127	348	25

Rev. Esp. Doc. Cient., 36(2), abril-junio 2013, e009. ISSN-L: 0210-0614. doi: http://dx.doi.org/10.3989/redc.2013.2.921

START es un sistema desarrollado por el *Massachusets Institute of Technology* que permite a los usuarios plantear preguntas sobre temas muy diversos, ya sean especializados o no (Katz y otros, 2007). Cuenta con una sencilla interfaz y sus tiempos de respuesta son considerablemente plausibles (Olvera-Lobo y Gutiérrez-Artacho, 2010). Las fuentes de información de las que extraen las respuestas son muy variadas, entre las que se encuentran sitios web de cobertura amplia como *Wikipedia,* diccionarios de uso general, *Internet Public Library, WorldBook, The World Factbook 2008,* entre otros, así como sitios web dedicados a un determinado ámbito temático como diccionarios y enciclopedias especializados, etc.

Por último, *TrueKnowledge*, desarrollado por una empresa londinense, se caracteriza por extraer las respuestas de numerosos recursos utilizando tanto su base de datos, como diversos sitios web, y las propias respuestas ofrecidas por los usuarios. Al igual que *QuALiM*, en ocasiones, también recupera información visual.

Figura 2. Interfaz de QuALiM: Página de resultados

Figura 3. Interfaz de SEMOTE: Página principal

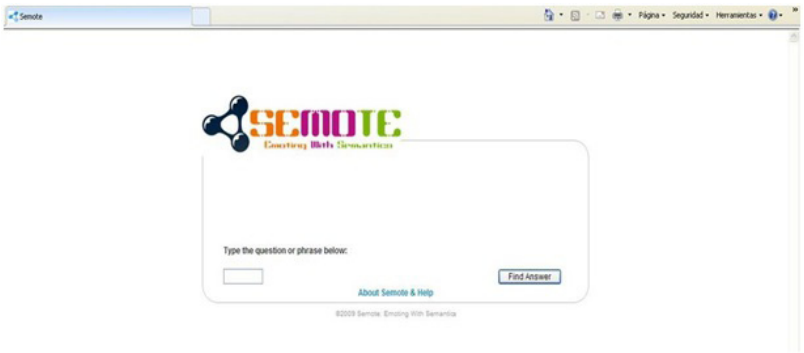
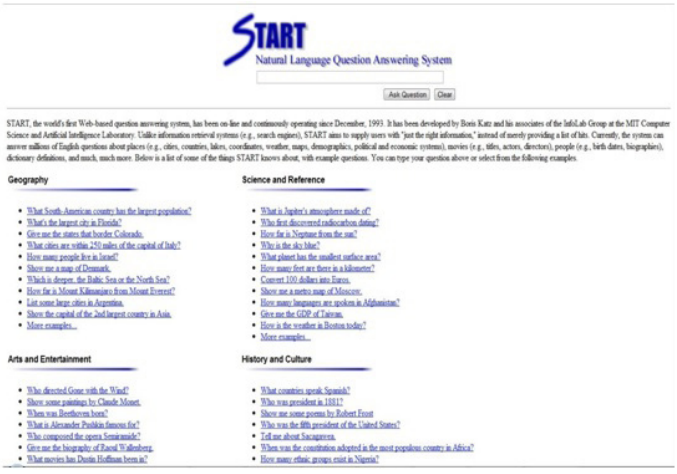


Figura 4. Interfaz de START: Página principal



Las respuestas ofrecidas por cada sistema fueron juzgadas por un grupo de estudiantes de tres diferentes grados de la Universidad de Granada y de edades variadas. El grupo de estudiantes se dividió en grupos de tres personas (uno de cada especialidad) y analizaron una parte de la muestra conjuntamente como incorrectas, inexactas o correctas siguiendo la metodología de evaluación propuesta en CLEF (Peters, 2009). Se consideraron correctas aquellas preguntas que respondían a la consulta de forma adecuada y no añadían información irrelevante. Todas las respuestas que satisfacían la consulta pero incorporaban información irrelevante fueron consideradas inexactas. Finalmente, se calificaron como incorrectas las respuestas cuyo contenido era irrelevante a la pregunta formulada al sistema. A partir de la valoración de las respuestas obtenidas, se aplicaron diferentes medidas para la evaluación del funcionamiento de los SBR.

Una de las medidas utilizadas para evaluar los sistemas de RI en general, que también se aplica a los SBR en particular (Fukumoto y otros, 2004; Voorhees y Tice, 1999), es *Mean Reciprocal Rank* (MRR). Esta medida asigna el valor inverso de la posición en la que la respuesta correcta fue encontrada (1 si es la primera, 1/2 si es la segunda, 1/3 si es la tercera, y así sucesivamente), o cero si la respuesta correcta no fue encontrada. Según esta medida, solamente hay una respuesta correcta dentro de la lista de resultados ofrecidos por el sistema, y el valor final es el promedio de los valores obtenidos para cada pregunta. MRR asigna un valor alto si las respuestas correctas se encuentran en las posiciones más altas del *ranking* de resultados.

$$MRR = \frac{1}{q}\sum_{i=1}^{q}\frac{1}{rank_i}$$

En este análisis se han aplicado además otras medidas específicamente desarrolladas para la evaluación de SBR en el entorno de la Web como son FHS y TRR (Radev y otros, 2001). Efectivamente, en los SBR la importancia de recuperar los resultados más relevantes en primer lugar es tan esencial como la recuperación en sí. Y es que estos sistemas suelen ser utilizados por usuarios que persiguen recuperar información rápida y eficaz en un breve espacio de tiempo y, por ello, no suelen examinar un número muy alto de respuestas. La medida *First Hit Success* (FHS) asigna valor 1 si la primera respuesta ofrecida es correcta, y valor 0 si no lo es (por lo que sólo considera la respuesta que aparece en primer lugar en la lista de resultados).

Por su parte, *Total Reciprocal Rank* (TRR) resulta una medida bastante útil para evaluar la existencia de varias respuestas correctas ofrecidas por un sistema ante una misma pregunta y asigna un valor a cada respuesta de acuerdo con su posición en la lista de resultados recuperados. Si en una lista de respuestas aparece varias veces repetida la correc-

Figura 5. Interfaz de TrueKnowledge: Página principal

ta, el usuario puede considerarla como más fiable. En estos casos no es suficiente tener en cuenta únicamente la primera respuesta correcta en las evaluaciones, por lo que TRR las tiene en cuenta a todas. Así, si la primera y la tercera respuesta de una lista de resultados son correctas para una pregunta el valor de TRR será 1/1+ 1/3.

La precisión, basada en la relevancia, es una de las medidas tradicionales de la RI (Harman, 1998) que más viene utilizándose para la evaluación del funcionamiento de los sistemas de RI desde los años 50 (Cleverdon, 1997). Impulsada por Salton y McGill (1983), sigue contando en la actualidad con gran aceptación y consenso en la comunidad investigadora, tal como lo demuestra el hecho de que reputados foros como *Text REtrieval Conference* la incorporen a su modelo de evaluación. La precisión refleja la capacidad del sistema para recuperar documentos (respuestas, en el caso de los SBR) que sean relevantes a la consulta (o pregunta) planteada.

Precisión de la recuperación de información

$$Precisión = \frac{\{Número\ de\ documentos\ relevantes\}}{\{Número\ de\ documentos\ recuperados\}}$$

Precisión de la búsqueda de respuestas

$$Precisión = \frac{\{Número\ de\ respuestas\ relevantes\}}{\{Número\ de\ respuestas\ recuperadas\}}$$

Las medidas tradicionales de evaluación se han ido enriqueciendo con otras que las completan y complementan. Una de las medidas más comunes en la "comunidad TREC" es la *Mean Average Precision* (MAP), la cual genera un único valor que resume el rendimiento de un sistema a distintos niveles de cobertura. Efectivamente, para los sistemas que devuelven una secuencia ordenada de documentos o respuestas, es necesario también considerar el orden en el que se presentan los documentos recuperados. MAP mide el promedio medio de precisión para un conjunto de preguntas formuladas cuyas respuestas son ordenadas por el sistema siguiendo un ranking de relevancia. Cuando se realiza la evaluación utilizando MAP, para cada consulta se calcula la media de los valores de precisión obtenidos cada vez que se encuentra un documento relevante. El valor final para el conjunto de consultas permite determinar qué sistema demuestra una mejor eficacia en la recuperación (Buckley y Voorhees, 2000). En MAP, Q es el conjunto total de preguntas.

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}$$

Por su parte, la precisión promedio (*Average precision* o AveP), incluida en MAP, permite medir la eficacia en la ordenación de las respuestas relevantes en la lista de resultados recuperados puesto que calcula el valor promedio de precisión para cada lugar del ranking en el que aparecen las respuestas relevantes. Un sistema con buen funcionamiento situará las respuestas relevantes en las primeras posiciones de la lista. P(r) es la precisión del sistema en la posición r de la lista de resultados y rel(r) es una función binaria que indica si el documento recuperado r es relevante a la consulta (valor 1) o no (valor 0).

$$AveP = \frac{\sum_{r=1}^{n}\left(P\ (r) \times rel\ (r)\right)}{número\ de\ respuestas\ relevantes}$$

4. RESULTADOS Y DISCUSIÓN

Tras plantear las 500 preguntas en los cuatro SBR evaluados se analizaron las respuestas ofrecidas por cada uno de ellos. Los resultados indican que el número total de respuestas recuperadas en *SEMOTE* −5000 y un promedio de 10 respuestas para cada pregunta− fue bastante superior al resto, seguido , aunque de lejos, por *QuaLim* con algo más de un tercio de respuestas recuperadas −1871 y 3,7 de promedio−. En los otros dos sistemas evaluados el total de respuestas recuperadas fue similar (744 respuestas para *START* y 766 en *TrueKnowledge*, es decir, un promedio de 1,5 y 1,5 respectivamente). Los resultados obtenidos en trabajos previos (Olvera-Lobo y Gutiérrez-Artacho, 2011a) ofrecían un número aproximado de respuestas recuperadas por *QuALiM* y *START* (3 y 1,6 respectivamente).

Si se tiene en cuenta la ratio de respuestas correctas respecto al número total de respuestas recuperadas por cada sistema se observa que es *START* el de funcionamiento más eficaz, con el 84,3% de respuestas correctas, mientras que *SEMOTE*, el sistema con más respuestas totales recuperadas, ha sido el que presenta un porcentaje inferior (31,8%). Es decir, los SBR que menos respuestas promedio recuperaron (*START, TrueKnowledge y QuaLiM*) fueron sin embargo más eficaces, lo que constata que una larga lista de respuestas no garantiza que éstas sean mejores ni más precisas.

En lo que a respuestas incorrectas se refiere *SEMOTE* −con un 57,5%− es el que mayor índice presenta seguido de *QuaLiM* con una proporción también bastante considerable −36,2%−. Frente a éstos, tanto *START* como *TrueKnowledge* −8,3% y 13,2%, respectivamente− revelan una ratio de respuestas incorrectas manifiestamente inferior. Por último, y en relación al tipo de respuestas que aquí se ha considerado como inexactas, la presen-

cia de las mismas en general no ha sido demasiado elevada, si bien en algún sistema ha superado el 20%. Estos datos mejoran sutilmente los resultados obtenidos en evaluaciones previas de los SBR (Olvera-Lobo y Gutiérrez-Artacho, 2010, 2011a), en donde *QuALiM* y *START* fueron evaluados y comparados con SBR de dominio especializado. Aunque las preguntas en las evaluaciones previas se restringían exclusivamente a las de definición y en un ámbito de especialización, se comprueba que los sistemas siguen siendo eficaces en la recuperación de información.

Los datos que arrojan las medidas de evaluación utilizadas ilustran el comportamiento de estos sistemas considerando además la eficacia en la ordenación de las respuestas. Los resultados, en general, pueden considerarse bastante satisfactorios y, por tanto, evidencian que estos sistemas son potencialmente útiles para recuperar información concisa de distinto tipo y dominio temático (véase tabla VI). El valor de MRR, medida que considera únicamente el lugar donde aparece la primera respuesta correcta en la lista de resultados, es bastante elevado en todos los sistemas analizados excepto en *SEMOTE* (0,4). En este sentido, FHS –aún más exigente con la ordenación de las respuestas correctas– es una medida muy destacada puesto que los usuarios, en muchas ocasiones, tienden a centrarse en la primera respuesta recuperada obviando el resto. Se observa que más del cincuenta por ciento de las primeras respuestas ofrecidas en todos los sistemas han sido correctas. El sistema que presenta un valor superior en FHS ha sido *START* (0,9) frente a *SEMOTE*, sistema que obtuvo el valor inferior (0,6). En TRR los valores mejoran para todos los sistemas puesto que se tiene en cuenta el lugar que ocupan en el ranking todas las respuestas correctas y no sólo las que aparecen en primer lugar.

Al analizar la precisión se observa que, como para las otras medidas, son *START* y *TrueKnowledge* los SBR que ofrecen mejores resultados. Además, si se flexibiliza el nivel de exigencia y se incluye en el cálculo de la precisión, no sólo las respuestas valoradas como correctas sino también las denominadas inexactas –que igualmente incluyen la información requerida pero con cierto ruido– los valores se incrementan, en algunos casos considerablemente. Por su parte, MAP, una medida ampliamente usada que ofrece una idea global del funcionamiento del sistema, muestra el mismo patrón de comportamiento que las medidas anteriores. Efectivamente, se observa que, excepto para el caso de TRR, hay una alta correlación entre las medidas usadas en este estudio (véase tabla VII).

Tabla V. Respuestas recuperadas en los cuatro SBR

Sistemas de búsqueda de respuestas	Total de respuestas	Promedio respuestas	Respuestas correctas	Respuestas inexactas	Respuestas incorrectas
QuaLiM	1871	3,7	47,7% (892)	16,1% (302)	36,2% (677)
SEMOTE	5000	10	31,8% (1588)	10,8% (538)	57,5% (2874)
START	744	1,5	84,3% (627)	7,4% (55)	8,3% (62)
TrueKnowledge	766	1,5	67,4% (516)	19,5% (149)	13,2% (101)

Tabla VI. Medidas de evaluación

Sistemas de búsqueda de respuestas	MRR	FHS	TRR	P	P*	MAP
QuaLiM	0,72	0,68	1,05	0,48	0.64	0,59
SEMOTE	0,38	0,55	1,01	0,32	0.42	0,35
START	0,91	0,89	1,05	0,84	0.92	0,93
TrueKnowledge	0,82	0,83	0,89	0,67	0.87	0,78

MRR: Mean Reciprocal Rank; FHS: First Hit Success; TRR: Total Reciprocal Rank; P: precisión; P*: precisión incluyendo también las respuestas inexactas; MAP: Mean Average Precision.

Tabla VII. Correlación entre medidas

Sistemas de búsqueda de respuestas	MRR	TRR	FHS	Precisión	MAP
MRR	**1**	**-0,09**	**0,96***	**0,96***	**0,97***
TRR		**1**	**-0,21**	**-0,14**	**-0,11**
FHS			**1**	**0,99****	**0,99****
Precisión				**1**	**0,99****
MAP					**1**

MRR: Mean Reciprocal Rank; FHS: First Hit Success; TRR: Total Reciprocal Rank; MAP: Mean Average Precision.
* La correlación es significativa a 0,05 ($p<0,05$).
** La correlación es significativa a 0,01 ($p<0,01$).

5. CONCLUSIONES

El usuario actual confía en recuperar información específica y de calidad que responda a sus necesidades. Los SBR presentan una interesante alternativa a la recuperación de información en Internet intentando satisfacer sus exigencias y demandas. Sin embargo, a pesar del aumento de esta clase de sistemas y del avance que supone el poder contar con herramientas de búsqueda de información de este tipo, los SBR disponibles en la Web son escasos y no todos proporcionan una cobertura adecuada. De hecho, las investigaciones que se vienen realizando y que culminan en interesantes propuestas plasmadas en diferentes publicaciones, foros y congresos, salvo contadas excepciones –y ya sea porque su utilidad se limita a contextos muy concretos, o bien por sus dificultades de implementación–, no se desarrollan para el usuario final.

En este análisis se han evaluado cuatro SBR de dominio general accesibles desde la Web mediante una colección de 500 preguntas cuyas respuestas, conforme a la metodología TREC, fueron juzgadas como correctas, incorrectas o inexactas por estudiantes y especialistas en diferentes campos temáticos. En base a estas valoraciones se aplicaron diferentes medidas de evaluación mediante las que se ilustra claramente la eficacia del funcionamiento de los sistemas analizados. Los dos sistemas que obtuvieron peores resultados (*QuALiM y SEMOTE*), recientemente han desaparecido y ya no se encuentran operativos en la Web para el usuario final. Uno de los principales problemas que presentan estos sistemas es que las bases de datos internas del sistema no se actualizan con regularidad, presentando en ocasiones resultados obsoletos. Sin embargo, la mayor fuente de información de los SBR son portales, páginas web y bases de datos especializadas de reconocido prestigio, por lo que las respuestas son en su mayoría satisfactorias para el usuario.

El estudio realizado revela resultados alentadores debido a que presentan este tipo de herramienta como una nueva posibilidad para obtener información precisa y fiable en un corto período de tiempo.

6. NOTAS

[1] http://demos.inf.ed.ac.uk:8080/qualim/ (Disponible hasta noviembre de 2011)

[2] http://www.semote.com (Disponible hasta agosto de 2011)

[3] http://start.csail.mit.edu/

[4] http://www.trueknowledge.com

[5] http://tangra.si.umich.edu/clair/NSIR/html/nsir.cgi

[6] services.hon.ch/cgi-bin/QA10/qa.pl

[7] http://eagl.unige.ch/EAGLi/

7. BIBLIOGRAFÍA

Abdou, S.; Savoy, J.; Ruch P. (2006) Dépister efficacement de l'information dans une banque documentaire: L'exemple de MEDLINE. En *Actes du XXIVème Congrès INFORSID*, 129-143.

Belkin, N.J.; Vickery, A. (1985). *Interaction in Information Systems: A Review of Research from Document Retrieval to Knowledge-based systems* (LIR Report No 35). Londres: The British Library.

Blair-Goldensohn, S.; McKeown, K.; Schlaikjer, A. H. (2004). Answering Definitional Questions: A Hybrid Approach. En: Maybury, M.T. (ed.). New Directions in Question Answering, Palo Alto: AAAI Press, 47-58.

Buckley, C.; Voorhees, E. M. (2000). Evaluating evaluation measure stability. *SIGIR '00 Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 33-40.

Cleverdon, C. (1997). The Cranfield tests on index languages devices. En Sparck Jones, K. y Willett, P. (eds.), *Readings in information retrieval.* San Francisco: Morgan Kaufmann, 47-59.

Cui, H.; Kan, M. Y.; Cua, T. S.; Xiao, J. (2004). A Comparative Study on Sentence Retrieval for Definitional Question Answering. *SIGIR Workshop on Information retrieval for Question Answering* (IR4QA), Sheffield.

Crouch, D.; Saurí, R.; Fowler, A. (2005). AQUAINT Pilot Knowledge-Based Evaluation: Annotation Guidelines. *Palo Alto Research Center*. Disponible en: http://www2.parc.com/isl/groups/nltt/papers/aquaint_kb_pilot_evaluation_guide.pdf

Fukumoto, J.; Kato, T.; Masui, F. (2003). Question Answering Challenge (QAC-1) an evaluation of question answering tasks at the NTCIRWorkshop 3. *Proceedings of AAAI Spring Symposium on New Directions in Question Answering*, 122-133.

Green, B. F.; Wolf, A. K.; Chomsky, C.; Laughery, K. (1961). Baseball: An Automatic Question Answerer. En: *Proceedings of the Western Joint Computer Conference*, v.19, pp. 219–224.

Greenwood, M. A.; Saggion, H. (2004). A Pattern Based Approach to Answering Factoid, List and Definition Questions. *Proceedings of the 7th RIAO Conference (RIAO 2004)*, 232–243.

Harman, D. K. (1998). Text retrieval conferences (TRECs): providing a test-bed for information retrieval systems. *Bulletin of the American Society for Information Science,* vol. 24 (4), 11-13.

Jackson, P.; Schilder, F. (2005). Natural Language Processing: Overview. En: Brown (ed.), *Encyclopedia of Language & Linguistics,* 2. Amsterdam, Elsevier Press, 503-518.

Kaisser, M. (2008). The QuALiM question answering demo: supplementing answers with paragraphs drawn from Wikipedia. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session.* Stroudsburg: Association for Computational Linguistics, 32–35.

Kangavari, M.R.; Ghandchi, S.; Golpour, M. (2008). A New Model for Question Answering Systems. *World Academy of Science, Engineering and Technology,* vol. 42, 506-513.

Katz, B.; Borchardt, G.; Felshin, S.; Shen, Y.; Zaccak, G. (2007). Answering English questions using foreign-language, semistructured sources. *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007)*. Irvine: IEEE Computer Society, 439-445.

Kolomityets, O.; Moens, M. F. (2011). A Survey on Question Answering Technology from an Information Retrieval Perspective. *Information Sciences* (en prensa), 2011.

Olvera-Lobo, M. D.; Gutiérrez-Artacho, J. (2010). Question-Answering Systems as Efficient Sources of Terminological Information: Evaluation. *Health Information and Library Journal,* vol. 27 (4), 268-276.

Olvera-Lobo, M. D.; Gutiérrez-Artacho, J. (2011a). Evaluation of Open -vs. Restricted- Domain Question Answering Systems in the Biomedical Field. *Journal of Information Science,* vol. 37 (2), 152-162.

Olvera-Lobo, M. D.; Gutiérrez-Artacho, J. (2011b). Language resources used in multi-lingual Question Answering Systems. *Online Information Review,* vol. 35 (4), 543-557.

Olvera-Lobo, M. D.; Gutiérrez-Artacho, J. (2011c). Multilingual Question-Answering System in Biomedical Domain on the Web: An Evaluation. *Multilingual and Multimodal Information Access Evaluation, Lecture Notes in Computer Science,* vol. 6941, 83-88.

Pérez-Coutiño, M.; Solorio, T.; Montes y Gómez, M.; López López, A.; Villaseñor Pineda, L. (2004). The Use of Lexical Context in Question Answering for Spanish. *Workshop of the Cross-Language Evaluation Forum* (CLEF 2004), 377-384 http://www.clef-campaign.org/2004/working_notes/CLEF2004WN-Contents.html [11 octubre 2011].

Peters, C. (2009). What Happened in CLEF 2009: Introduction to the Working Notes*. Working Notes for the CLEF 2009 Workshop*. http://www.clef-campaign.org/2009/working_notes/ [11 septiembre 2011].

Radev, D. R.; Qi, H.; Wu, H.; Fan, W. (2001). *Evaluating Web-based Question Answering Systems*. Informe técnico, University of Michigan.

Rodrigo, A.; Pérez-Iglesias, J.; Peñas, A.; Garrido, G.; Araujo, L. (2010). A Question Answering System based on Information Retrieval and Validation, *Notebook Papers/LABs/Workshops* (*CLEF 2010)* http://clef2010.org/index.php?page=pages/proceedings.php [5 septiembre 2011].

Salton, G.; McGill, J. (1983). *Introduction to modern information retrieval.* New York: McGraw-Hill.

Sultan, M. (2006). *Multiple Choice Question Answering*. Tesis doctoral. Sheffield: University of Sheffield.

Tsur, O. (2003). *Definitional Question-Answering Using Trainable Text Classifiers.* Tesis doctoral. Ámsterdam: University of Amsterdam.

Voorhees, E. M. (1999). The TREC 8 Question Answering Track Report. *Proceedings of the 8th Text Retrieval Conference*. *http://trec.nist.gov/pubs/trec8/papers/qa_report.pdf*

Voorhees, E. M. (2002). Overview of the TREC 2002 Question Answering Track. *Proceedings of the Eleventh Text Retrieval Conference*. http://comminfo.rutgers.edu/~muresan/IR/TREC/Proceedings/t11_proceedings/t11_proceedings.html [5 septiembre 2011].

Voorhees, E. M.; Tice, D. (1999). The TREC-8 question answering track evaluation. En: Voorhees, E. y Harman, D., *Proceedings of the Eighth Text Retrieval Conference*. Gaithersburg, MD: NIST Publicación Especial. http://comminfo.rutgers.edu/~muresan/IR/TREC/Proceedings/t8_proceedings/t8_proceedings.html [5 septiembre 2011].

Warren, D. (1981). Efficient Processing of Interactive Relational Database Queries Expressed in Logic. *Proceedings Seventh International Conference on Very Large Data Bases*, v.7. Cannes, VLDB Endowment, v.7., pp. 272-283.

Weizenbaum, J. (1966). Eliza: A computer program for the study of natural language communication between man and machine. *Communications of the ACM*. v.9, n.1, pp.36-45.

Woods, W. A.; Kaplan, R. M.; Nash-Webber, B. (1972). The Lunar Sciences Natural Language Information System. En: *BBN Final Report 2378*. Cambridge: Bolt, Beranek and Newman.

Zweigenbaum, P. (2005). Question answering in biomedicine. *Proceedings Workshop on Natural Language Processing for answering.* Budapest: ACL, EACL 2003, 1-4.