Evaluación de los sistemas QA de dominio abierto frente a los de dominio especializado en el ámbito biomédico

María-Dolores Olvera-Lobo[1,2], Juncal Gutiérrez-Artacho[3]

[1] CSIC, Unidad Asociada Grupo SCImago, Madrid, España

[2] Departamento de Biblioteconomía y Documentación,
Universidad de Granada, Granada, España

[3] Departamento de Traducción e Interpretación,
Universidad de Granada, Granada, España

{molvera, juncalgutierrez}@ugr.es

Resumen. Los sistemas QA se presentan como una alternativa a los sistemas tradicionales de recuperación de información tratando de ofrecer respuestas precisas a preguntas factuales. Hemos realizado un estudio para evaluar la eficiencia de estos sistemas como fuentes terminológicas para los especialistas y para usuarios en general. Con este fin, se ha evaluado el funcionamiento de cuatro sistemas QA, dos especializados en el dominio biomédico (MedQA y HONqa) y dos de dominio general (START y QuALiM). El estudio ha utilizado una colección de 150 preguntas biomédicas definicionales (*What is…?)*, obtenidas del sitio web médico WebMD. Para determinar el funcionamiento, se han evaluado las respuestas ofrecidas utilizando una serie de medidas específicas (precisión, MRR, TRR, FHS).
El estudio permite confirmar que los cuatro sistemas son útiles para la recuperación de información definicional en este ámbito, ya que han proporcionado respuestas coherentes y precisas con un grado de aceptabilidad adecuado.

Palabras clave: Sistemas QA de dominio abierto, Sistemas QA de dominio especializado, Evaluación del funcionamiento, Información biomédica

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. Por ello, los profesionales de diversos ámbitos comienzan a reconocer la utilidad de otros tipos de sistemas, como los sistemas de búsqueda de respuestas (en inglés *question-answering systems*, en adelante sistemas QA), como método para la obtención de información especializada de forma rápida y efectiva [1-3].

La Recuperación de Información (en adelante, RI) se ha entendido como el proceso, totalmente automático, en el que dada una consulta (que, supuestamente, expresa la necesidad de información del usuario) y una colección de documentos, el sistema devuelve una lista ordenada de documentos potencialmente relevantes para esa consulta [4]. Un sistema de RI con funcionamiento óptimo recuperaría todos los documentos relevantes (lo que implica una exhaustividad completa) y sólo aquellos documentos que son relevantes (precisión perfecta). El modelo tradicional de RI lleva consigo muchas restricciones

implícitas tales como: a) la suposición de que los usuarios del sistema buscan documentos (textos completos), no respuestas, y que son los documentos, como tales, los que responden y satisfacen una consulta; b) que el proceso debe ser directo y unidireccional en lugar de interactivo; y c) que la consulta y el documento están escritos en la misma lengua.

Un paso en la evolución hacia la mejora de la RI son los sistemas QA. Se presentan como una alternativa a los tradicionales sistemas de RI tratando de ofrecer respuestas precisas y comprensibles a preguntas factuales, en lugar de presentar al usuario una lista de documentos relacionados con la búsqueda [5], de modo que el usuario no ha de leer documentos completos para obtener la información requerida. El desarrollo de los sistemas QA toma un importante impulso en el seno de la conferencia sobre recuperación de información TREC (*Text REtrieval Conference)* −principalmente a partir de TREC-8 [6]− la cual, desde 1992, constituye un foro internacional para aunar e incentivar la investigación en diferentes ámbitos de la recuperación de información.

Según un estudio de Ely y otros [7], los especialistas médicos tardan más de dos minutos de promedio en buscar información relativa a las preguntas que les surgen y, a pesar del tiempo empleado, muchas de ellas no consiguen obtener la respuesta adecuada. En este sentido, varios trabajos han demostrado la confianza de los especialistas médicos en el uso los sistemas QA como método de búsqueda y recuperación de información especializada [2,8], así como que los pacientes también han aumentado sus consultas en estos sistemas antes y después de ver al médico para obtener información sobre la naturaleza de la enfermedad, las indicaciones y contraindicaciones de los tratamientos, entre otros [9].

El funcionamiento de los sistemas QA se basa en los modelos de respuestas cortas [10] puesto que ofrece la respuesta potencialmente correcta en forma de un número, un sustantivo, una frase corta o un fragmento breve de texto. Existen diferentes patrones a la hora de plantear las preguntas en los sistemas QA, la mayoría se caracterizan por aceptar preguntas expresadas a través de partículas interrogativas (qué, cómo, quién, por qué, cuándo, dónde), o de forma imperativa. Los sistemas QA proceden a la construcción de respuestas coherentes expresadas en lenguaje natural [11]. Planteada la pregunta en el motor de búsqueda del sistema, éste procede a analizar la pregunta separando la palabra o palabras claves y determinando el tipo de respuesta esperada, luego se localiza y extrae una respuesta a partir de diferentes fuentes –dependiendo de la cobertura temática del sistema se utilizarán unas u otras fuentes de información [12]–, y finalmente, se evalúa y elimina aquella información redundante o que no responde correctamente a la pregunta planteada para, posteriormente, elaborar y presentar una o varias respuestas concretas que supuestamente satisfacen la necesidad del usuario [13,14]. Una de las dimensiones más importantes de estos sistemas QA es el proceso de la evaluación de las respuestas, ya que además de evaluarlas, las compara y ordena [15]. Algunos de estos sistemas utilizan algún módulo destinado a la comparación de las preguntas planteadas por el usuario con el pasaje del documento o documentos seleccionados que, potencialmente, recogen la respuesta adecuada [16,17].

Los sistemas QA suelen tener una sencilla interfaz con un motor de búsqueda en el que los usuarios plantean su pregunta, algunos de ellos facilitan la lista de las últimas cuestiones introducidas para facilitar al usuario la comprensión acerca del funcionamiento del mismo. Para el tratamiento y gestión de las preguntas, los sistemas QA aplican algoritmos y métodos de análisis lingüístico y de procesamiento del lenguaje natural con el fin de identificar sus componentes y determinar el tipo de respuesta esperada [9]. Este análisis consiste normalmente en utilizar una variedad de tipos de preguntas estándar en los que se reemplazan ciertas palabras por las etiquetas aceptadas por el sistema [18].

Los sistemas QA pueden ser de dominio general –si puede atender consultas de temas muy diversos, como START [19] o QuALiM [20] – o de dominio específico, si se centran en un ámbito determinado, como MedQA [21] o HONqa [22]. Otro de los aspectos clave de estos sistemas es que el establecer una relación sistema-usuario no unidireccional y una interacción en el proceso de búsqueda ayudaría al sistema a encontrar mejores respuestas, y al usuario a encontrar la respuesta más rápidamente. No obstante, todavía es necesario profundizar en el diseño de estos sistemas interactivos que hagan posible la existencia de un verdadero *feedback* entre preguntas y respuestas, y que el usuario se comunique a nivel conversacional con el sistema.

A pesar del avance que supone el poder contar con herramientas de búsqueda de información de este tipo, los sistemas QA presentan algunas restricciones como que muchos de los sistemas han sido desarrollados únicamente como prototipos, o *demos*, y sólo en casos muy poco frecuentes se han comercializado.

Si bien en los últimos años se han analizado diversos aspectos de los sistemas QA, aún son escasos los estudios que evalúan el funcionamiento de estas herramientas. Estos sistemas deben generar frases con definiciones dinámicas y coherentes que contengan y resuman la información más descriptiva que posean en su colección de documentos sobre el término o foco de la pregunta del usuario [23, 13].

En este trabajo hemos llevado a cabo un estudio para evaluar la eficiencia de los sistemas de búsqueda de respuestas como fuentes terminológicas para los especialistas y para usuarios en general. Con este fin, se ha analizado y evaluado el funcionamiento de cuatro sistemas de búsqueda de respuestas, dos especializados en el dominio biomédico (MedQA y HONqa) y dos de dominio general (START y QuALiM). Para ello se ha utilizado una colección de 150 preguntas biomédicas definicionales y se han evaluado las respuestas ofrecidas utilizando medidas específicas aplicables a este tipo de sistemas considerando además las fuentes utilizadas por éstos para extraer tales respuestas. A continuación se describe con detalle la metodología utilizada y los principales resultados obtenidos.

2 Metodología

En nuestro análisis se utilizaron 150 preguntas de definición sobre diversos temas médicos. La colección de preguntas utilizadas se ha obtenido tras plantear la expresión "*What is*" en el motor interno de búsqueda del sitio web WebMD [24], un portal estadounidense creado por especialistas médicos para dar respuesta a las incertidumbres de los pacientes y en el que se ofrece información sobre una amplia lista de temas médicos de diferente grado de especialización. El sitio web proporcionó más de 6000 preguntas. Finalmente, seleccionamos para nuestro estudio las 150 preguntas que obtuvieron respuesta en los cuatro sistemas analizados, los cuales se eligieron debido a que sus bases de datos presentan una extensa cobertura y están actualizadas. El conjunto de preguntas utilizadas superó el test de validez interna con un alfa de Cronbach de 0,997.

Los sistemas QA utilizados fueron: START, QuALiM, MedQA y HONqa. START, desarrollado por el *Massachusets Institute of Technology* es un sistema que permite a los usuarios plantear preguntas sobre temas diversos por lo que también debe ser capaz de responder preguntas especializadas de dominio médico [25]. Las fuentes de información de las que extraen las respuestas son muy variadas, entre las que se encuentran página o sitios web generales como *Wikipedia,* diccionarios de uso general, *Internet Public Library*, *WorldBook*, *The World Factbook 2008*, entre otros, así como otras páginas o sitios especializados en una determinada área como diccionarios y enciclopedias especializados, etc. Por su parte, el segundo sistema QA de dominio abierto analizado fue QuALiM. Este sistema, financiado por *Microsoft*, recupera tanto información textual −para lo que utiliza únicamente la enciclopedia *Wikipedia*− como gráfica −extraída del buscador de imágenes de Google− [26].

Por su parte, MedQA, un sistema QA de dominio especializado desarrollado en la Universidad de Columbia se trata de un sistema especializado que analiza miles de documentos para generar un párrafo que responda correctamente a preguntas especializadas en temas médicos, por lo que sus fuentes de información difieren parcialmente de los anteriores [8]. Utiliza una amplia gama de fuentes de información, entre las que se encuentran *Wikipedia*, *Medline*, *Medline Plus*, un par de ellas más. Por último, HONqa está desarrollado por la *Health On the Net Foundation*, una organización suiza sin ánimo de lucro cuyo objetivo es promover la creación de información médica de calidad y veraz. Es el único sistema multilingüe ya que se puede recuperar información en inglés, italiano y francés [27]. Las fuentes de información que utiliza son muy variadas y suelen ser portales médicos generales o especializados en una determinada enfermedad.

Tras plantear las preguntas en los cuatro sistemas, varios profesionales médicos valoraron las respuestas como incorrectas, inexactas o correctas. Se consideró una respuesta correcta aquélla que respondía adecuadamente a la pregunta planteada, no utilizaba más de 100 palabras para generar la respuesta y no contenía información que no fuera relevante a la pregunta. Todas las respuestas que, a pesar de contestar correctamente a la pregunta no se ajustaban al resto de criterios, fueron valoradas como inexactas. La valoración realizada para cada respuesta sirvió de base para la aplicación de las medidas de evaluación del funcionamiento de los sistemas [28], las cuales se describen a continuación.

Mean Reciprocal Rank (MRR) asigna el valor inverso de la posición en la que la respuesta correcta fue encontrada (1 si es la primera, ½ si es la segunda, ¼ si es la cuarta, y así sucesivamente), o cero si la respuesta correcta no fue encontrada. Esta medida únicamente considera la primera respuesta correcta encontrada en 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 a las respuestas que están en las posiciones más altas de la clasificación.

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

Total Reciprocal Rank (TRR) es útil para evaluar la existencia de varias respuestas correctas ofrecidas por un sistema ante una misma pregunta. En estos casos no es suficiente considerar únicamente la primera respuesta correcta en las evaluaciones, por lo que TRR las tiene todas en cuenta y asigna un peso a cada respuesta de acuerdo con su posición en la lista de resultados recuperados. Así, si la segunda y la cuarta respuesta de una lista de resultados son correctas para una pregunta el valor de TRR será ½ + ¼ = ¾.

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).

Además, se utilizó una medida clásica en la evaluación de la recuperación de información, la precisión. Ésta se define como la capacidad del sistema para recuperar documentos (o respuestas, en el caso de los sistemas QA) que sean relevantes a las consultas (o preguntas) planteadas y que estén bien ordenados (en el caso de que los sistemas establezcan un ranking de resultados).

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

3 Resultados y Discusión

Tras plantear las 150 preguntas en los cuatro sistemas QA se analizaron las cinco primeras respuestas de cada uno de estos sistemas, al ser el promedio de respuestas recuperadas por la mayoría de ellos y debido a que los usuarios, pretendidamente, utilizarían este tipo de sistemas para la recuperación rápida de información, centrando su atención en las primeras respuestas. Para ciertas preguntas sin embargo, algunos sistemas QA ofrecieron un número superior y otros no llegaron a ofrecer las cinco respuestas.

El volumen de respuestas recuperadas por los sistemas de dominio abierto es inferior a los de dominio restringido, siendo el menor el de START (con 1,6 respuestas como media) seguido de cerca de QuALiM (con 3 respuestas). En los sistemas QA de dominio especializado los resultados aumentan sustancialmente, sobre todo en el caso de HONqa (con 44,23 respuestas), mientras que en MedQA es ligeramente superior a la de los sistemas QA de dominio abierto (5,34 respuestas a cada pregunta de promedio).

Las respuestas correctas están presentes en mayor medida en START (70,08%), en los dos sistemas QA de dominio especializado este promedio desciende,−MedQA (46,67%) y HONqa (47,25%)− y QuALiM se presenta como el más deficiente con el 40,89% de respuestas correctas. La figura 1 muestra el número de respuestas correctas, inexactas e incorrectas de los cuatro sistemas QA analizados.

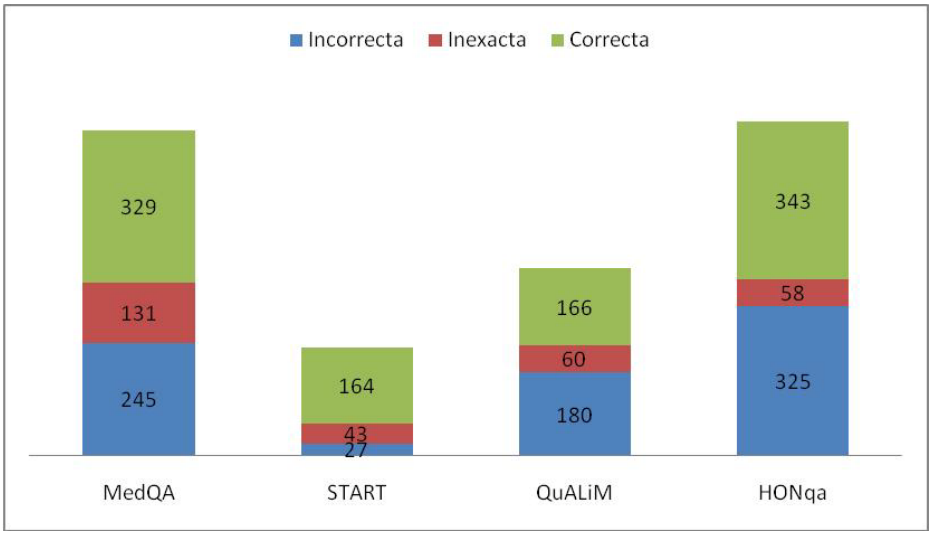

Figura 1. Respuestas incorrectas, inexactas y correctas

En relación a las respuestas inexactas, sistema QA con el promedio inferior es HONqa (7,97%), ofreciendo el resto de los sistemas QA un promedio similar (MedQA, 18,58%; START, 18,38%; QuALiM, 14,78%).

Las respuestas incorrectas en HONqa (44,78%) y QuALiM (44,33%) alcanzan una cifra similar. En MedQA es algo inferior (34,75%) aunque continúa siendo elevada siendo START (11,54%) el sistema que presenta menos respuestas inexactas.

El valor obtenido al aplicar la medida de evaluación MRR y FHS indica que MedQA ordena mejor las respuestas, de manera que la primera respuesta correcta aparece en los primeros lugares de la lista de resultados. Esto es significativo puesto que no emplea ningún algoritmo de ranking para llevar a cabo este proceso sino que siempre recurre a la misma ordenación de las respuestas, según de la fuente de la que provengan. FHS resulta una medida muy relevante puesto que los usuarios, en muchas ocasiones, tienden a centrarse en la primera respuesta recuperada obviando el resto.

Tabla 1. Medidas de evaluación de los sistemas QA

	MRR	**TRR**	**FHS**	**Precisión**
MedQA	0,87	1,29	0,76	0,54
START	0,67	0,81	0,64	0,76
QuALiM	0,65	0,77	0,59	0,56
HONqa	0,75	1,15	0,55	0,46

En la situación contraria nos encontramos con el otro sistema QA de dominio especializado, HONqa, ya que ofrece el FHS con valor más bajo, aún cuando se trata del segundo sistema con un promedio superior en MRR. El comportamiento de los dos sistemas QA de dominio abierto es muy similar y no hay una gran diferencia entre ambas medidas, lo que se explica si se tiene en cuenta que, para cada pregunta, las respuestas recuperadas suelen oscilar entre 1 y 3.

No obstante, la medida TRR es superior en los sistemas de dominio especializado puesto que pondera el valor de cada respuesta correcta en función del lugar que ocupa en la lista de resultados. Necesariamente, este valor aumente, al ofrecer un mayor número de resultados.

El valor de la precisión en los sistemas QA de dominio abierto −principalmente en START− ha sido superior al de los sistemas QA de dominio especializado ya que éstos últimos presentan una alta tasa de ruido documental.

Como se puede observar, ninguna de las medidas aplicadas presenta valores muy altos. Esta circunstancia se ha visto claramente influida por el hecho de que las condiciones requeridas para evaluar una respuesta como correcta tenían un alto nivel de exigencia. En muchas ocasiones, por ejemplo en MedQA y en START se encontraron respuestas correctas con más de 100 caracteres, lo que provocó que fueran consideradas como inexactas, como se indicó en el apartado Metodología. En Honqa, por ejemplo, se encontraron respuestas correctas que, al no responder a la pregunta de forma totalmente concreta y precisa, fueron consideradas también inexactas. La siguiente tabla muestra la correlación existente entre las métricas utilizadas en este estudio.

Tabla 2. Correlación entre las medidas

		MRR	**TRR**	**FHS**
MRR	Pearson Correlation	1,000	,959*	,700
	Sig. (2-tailed)		,041	,300
	N	4,000	4	4
TRR	Pearson Correlation	,959*	1,000	,472
	Sig. (2-tailed)	,041		,528
	N	4	4,000	4
FHS	Pearson Correlation	,700	,472	1,000
	Sig. (2-tailed)	,300	,528	
	N	4	4	4,000
Precisión	Pearson Correlation	-,450	-,606	,201
	Sig. (2-tailed)	,550	,394	,799
	N	4	4	4

* La correlación es significativa al nivel 0,05.

La tabla muestra que las medidas que tienen alta correlación son MRR y FHS y por otro lado TRR y precisión, de manera inversa, pero tan sólo MRR y TRR tienen una correlación significativa ($p<0,05$).

4 Conclusiones

El análisis de los resultados obtenidos al plantear las 150 preguntas en los sistemas QA, MedQA, START, QuALiM y HONqa, ha permitido evaluar su funcionamiento aplicando métricas específicas. A pesar de las restricciones que muestran estos sistemas, ya que no son accesibles para todos y no se encuentran siempre desarrollados completamente, se ha comprobado que los cuatro sistemas QA son válidos y útiles para la recuperación de información definicional médica, puesto que ofrecieron respuestas coherentes y precisas.

Otro aspecto interesante se refiere a las fuentes de información utilizadas por cada uno de estos sistemas QA. Si comparamos las fuentes utilizadas por los sistemas QA de dominio especializado con aquellas utilizadas por los sistemas QA de dominio general, vemos grandes diferencias en la tipología y especialización, como era de esperar. Al comparar las fuentes utilizadas por START y QuALiM comprobamos que no existe gran diferencia, ya que ambos utilizan Wikipedia como fuente principal, aunque START se nutre de más fuentes para su recuperación. Sin embargo, en las fuentes de los dos sistemas QA de dominio especializado vemos grandes diferencias en las fuentes. Mientras que MedQA utiliza diccionarios, enciclopedias y bases de datos especializados en el dominio biomédico, HONqa se decanta por sitios web especializados en este dominio.

Los resultados son esperanzadores al mostrar este tipo de herramientas como una nueva posibilidad en el ámbito de la recuperación de información precisa, fiable y concreta en un periodo breve de tiempo. En este sentido, algunos autores [29,22] han explorado varias posibilidades de mejora, tales como el uso de ontologías, las cuales contribuyen a elevar el nivel de calidad de las respuestas obtenidas puesto que formaliza la información relevante del dominio en cuestión.

5 Bibliografía

1. Crouch, D., Saurí, R., Fowler., A.: AQUAINT Pilot Knowledge-Based Evaluation: Annotation Guidelines. Tech. rep., Palo Alto Research Center (2005)

2. Lee, M., Cimino, J.; Zhu; H.R., Sable; C., Shanker; V., Ely, J., Yu., H.: Beyond Information Retrieval – Medical Question Answering. En: AMIA (2006)

3. Yu, H., Lee, M., Kaufman, D., Ely;, J., Osheroff, J.A., Hripcsak, G., Cimino, J.: Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians. Journal of Biomedicine Informatics. 4, pp. 236–251 (2007)

4. Baeza-Yates, R.; Ribeiro-Nieto, B.: Modern Information Retrieval. Nueva York: ACM Press; Addison-Wesley (1999)

5. Pérez-Coutiño, M., Solorio, T., Montes-y-Gómez, M., López-López, A., Villaseñor-Pineda, L.: Toward a Document Model for Question Answering Systems. En Proceedings of the Second International Atlantic Web Intelligence Conference. AWIC 2004. Cancun, México (2004)

6. Voorhees, E.M.: The TREC 8 Question Answering Track Report. En Voorhees, E.M. y D.K. Harman (eds.), Proceedings of the 8th Text REtrieval Conference, vol. 500–246, pp. 107–130. NIST, Gaithersburg, Maryland (1999)

7. Ely, J.W., Osheroff, J., Gorman, P.N., Ebell, M.H., Chambliss, M.L., Pifer, E.A., Stavri., P.Z.: A taxonomy of generic clinical questions: classification study. BMJ, vol. 321, pp. 429–432 (2000)

8. Yu, H., Kaufman., D.: A cognitive evaluation of four online search engines for answering definitional questions posed by physicians. Pacific Symposium on Biocomputing, vol. 12, pp. 328–339 (2007)

9. Zweigenbaum, P.: Question answering in biomedicine. En De Rijke, M. y B. Webber (eds.), Proceedings Workshop on Natural Language Processing for Question Answering, EACL 2003, pp. 1–4. ACL, Budapest (2005)

10. Blair-Goldensohn, S.B., McKeow, K.R., Schlaikjer., A.H .: A hybrid Approach for QA Track Definitional Questions. Proc. of TREC 2003, pp. 336–343, Gaithersburg, Maryland (2003)

11. Costa, L.F., Santos, D.: Question Answering Systems: a partial answer. SINTEF. Oslo (2007)

12. Olvera-Lobo, M. D., Gutiérrez-Artacho, J.: Question-Answering Systems as Efficient Sources of Terminological Information: Evaluation. Health Information and Library Journal (2010).

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

14. Tsur, O.: Definitional Question-Answering Using Trainable Text Classifiers. Tesis doctoral. Institute of Logic Language and Computation (ILLC), University of Amsterdam (2003)

15. Sing, G.O., Ardil, C., Wong, W., Sahib., S.: Response Quality Evaluation in Heterogeneous Question Answering System: A Black-box Approach. Proceedings of World Academy of Science, Engineering and Technology, vol. 9, Lisbon (2005)

16. Alfonseca, E., De Boni, M., Jara, J.L., Manandhar., S.: A prototype Question Answering system using syntactic and semantic information for answer retrieval. En Proceedings of the 10th Text Retrieval Conference (TREC-10). Gaithersburg, Maryland (2002)

17. Jacquemart, P., Zweigenbaum, P. : Towards a Medical Question-Answering System: a Feasibility Study. En: Proceedings of Medical Informatics Europe (MIE '03), vol. 95 of Studies in Health Technology and Informatics, pp. 463–468. IOS Press, San Palo, Calif, USA (2003)

18. Pérez-Coutiño, M., Solorio, T., Montes y Gómez, M., López López, A., Villaseñor Pineda, L: The Use of Lexical Context in Question Answering for Spanish. Workshop of the Cross-Language Evaluation Forum, Trondheim (2004)

19. START, http://start.csail.mit.edu/

20. QuALiM, http://demos.inf.ed.ac.uk:8080/qualim/

21. MedQA, http://monkey.ims.uwm.edu:8080/MedQA/

22. HONqa, http://services.hon.ch/cgi-bin/QA10/qa.pl/

23. Blair-Goldensohn, S.B., Schlaikjer., A.H.:. Answering Definitional Questions: A Hybrid Approach. New Directions In Question Answering, vol. 4, pp. 47−58 (2004)

24. WebMD, http://www.webmd.com/

25. Katz, B., Felshin, S., Yuret, D., Ibrahim, A., Lin, J., Martion, G., McFarland, A.J., Temelkuran, B.: Omnibase: Uniform Access to Heterogeneous Data for Question Answering. En: Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems, Estocolmo (NLDB 2002), pp. 230−234 (2002)

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

27. Cruchet, S., Gaudinat, A., Rindflesch, T., Boyer, C.: What about trust in the Question Answering world? En: AMIA 2009 Annual Symposium, San Francisco (2009)

28. Raved, D.R., Qi, H., Wu, H., Fan., W.: Evaluating Web-based Question Answering Systems. Technical Report, University of Michigan (2001)

29. Buitelaar, P., Cimiano, P., Frank, P., Hartung, M., Racioppa., S.: Ontology-based information extraction and integration from heterogeneous data sources. Int. J. Human-Computer Studies, vol. 66, pp. 759−788 (2008)