Health Information and Libraries Journal

Health Libraries Group | cilip Chartered Institute of Library and Information Professionals

DOI: 10.1111/j.1471-1842.2010.00896.x

Question-answering systems as efficient sources of terminological information: an evaluation

María-Dolores Olvera-Lobo* & Juncal Gutiérrez-Artacho†

*Grupo Scimago, Unidad Asociada CSIC, Madrid, Spain & University of Granada, and †University of Granada

Abstract

Background: Question-answering systems (or QA Systems) stand as a new alternative for Information Retrieval Systems. Most users frequently need to retrieve specific information about a factual question to obtain a whole document.

Objectives: The study evaluates the efficiency of QA systems as terminological sources for physicians, specialised translators and users in general. It assesses the performance of one open-domain QA system, START, and one restricted-domain QA system, MedQA.

Method: The study collected two hundred definitional questions (*What is*…?), either general or specialised, from the health website WebMD. Sources used by the open-domain QA system, START, and the restricted-domain QA system, MedQA, were studied to retrieve answers, and later a range of evaluation measures (precision, Mean Reciprocal Rank, Total Reciprocal Rank, First Hit Success) were applied to mark the quality of answers.

Results: It was established that both systems are useful in the retrieval of valid definitional healthcare information, with an acceptable degree of coherent and precise responses from both. The answers supplied by MedQA were more reliable that those of START in the sense that they came from specialised clinical or academic sources, most of them showing links to further research articles.

Conclusions: Results obtained show the potential of this type of tool in the more general realm of information access, and the retrieval of health information. They may be considered a good, reliable and reasonably precise alternative in alleviating the information overload. Both QA systems can help professionals and users can obtain healthcare information.

Keywords: decision support techniques, evaluation studies as topic, information storage and retrieval, natural language processing, MedQA, START

Key Messages

Implications for Practice

- Question-answering systems (QA systems) are a useful tool for retrieving data and terminological information.
- The evaluative method can be replicated for other QA systems and other areas of knowledge.
- Question-answering systems help in identifying users information needs.

Implications for Policy

- Question-answering systems are set to become one of the key tools available to retrieve and organise health information.

Introduction

Question-answering systems (QA Systems) can be viewed as a new alternative to the more familiar

Correspondence: Juncal Gutiérrez-Artacho, University of Granada.
E-mail: juncalgutierrez@ugr.es

Information Retrieval Systems. These systems try to offer detailed, understandable answers to factual questions, to retrieve a collection of documents related to a particular search.[1] In recent years, the development of QA systems has been encouraged and furthered through the TREC meetings (*Text REtrieval Conference*)[2] – mainly since TREC-8.[3] This Conference has proven to be an important international forum, putting together and improving research efforts behind the different aspects of information retrieval.

Question-answering systems endeavour to make retrieval easier through the short-answer question models.[4–6] Accordingly, users do not have to read the full text of documents either from a scientific article or a web page, to obtain the required information because the QA system shows the correct answer by means of a number, a noun, a short phrase or a concise extract of text.

Questions used in QA systems can be expressed using interrogative adverbs (*who*, *what*, *which*, *how*, *when*, *where*), or in imperative form (tell me, show, list…). Once the question is provided, the QA systems extract natural language answers.[7] QA systems follow these main steps:

- Systems retrieve documents to obtain relevant sentences about the search term, using questions posed by the users;
- they identify their components parts;
- determine the kind of answer anticipated;[8]
- they retrieve and select the sentences;
- they choose non-redundant definition sentences from the overall results of sentence retrieval, to delimit the response.[9,10]

The objective of the systems is to retrieve only correct information to answer the users' questions.[11] Evaluation is one of the most important dimensions in QA systems, as the process of assessing, comparing and ranking is key to monitor progress in the field.[12,13] The main component of these systems consists of measuring modules, which analyse tagged sentences in selected documents, and compare them with the question to find the most similar sentence.[14,15] Generally speaking, QA systems feature very simple and user-friendly interfaces, and rely on methods of linguistic analysis and natural language. The ones that allow users to query in different languages are known as multi-lingual QA systems.

All these QA systems are based on prototypes; that is, they are available as demos, like askEd,[16] only a few have they been marketed like Wolfram-Alpha.[17] Demos are not regularly upgraded and the design is not satisfactory therefore they present more problems than the marketed versions. A more interactive QA procedure that allows for real feedback between questions and answers, and user communication with the system on a conversational level is needed.

While not many QA systems are available on the Internet, there are some open-domain QA systems such as START. START is atypical, it includes calls to OMNIBASE, a system that integrates heterogeneous data sources using an *object-property-value model*;[18] NSIR,[19] developed by the University of Michigan; or Qualim,[20] financed by Microsoft; there are also some restricted-domain QA systems including MedQA. In the case of NSIR and Qualim, answers are constructed on the basis of information provided by Google[21] and Wikipedia,[22] respectively. Although START also retrieves information from Wikipedia, it uses other specialised sources such as directories, databases, dictionaries, or encyclopaedias. Meanwhile, MedQA retrieves information from the medical database MEDLINE, specialised dictionaries, Wikipedia and certain search engines like Google.

Information overload is more acute on the Web than in other contexts. When users pose a given question by means of search engine tools (including directories or metasearchers), they may retrieve an excessive number of web pages, many of which are not relevant or useful. Professionals in different areas claim that QA systems constitute a good method of obtaining specialised information quickly and efficiently.[23–25]

In a study by Ely *et al.*[26] participating physicians spent on average <2 min looking for information to resolve clinical queries, although many of their questions remained unanswered. Some studies have shown that physicians trust QA systems as search methods for specialised information retrieval.[25,27] The general public increasingly explore Web resources to obtain information about the disease before or after consulting a doctor.[8]

While researchers have looked into various aspects of QA systems in recent years, one facet that is widely overlooked is the formal evaluation

Table 1 Categories of reference of definitional questions

Question	Pain	Inflammation	Disease	Syndrome	Infection	Treatment	Others
Number	8	16	97	11	10	38	15

of this tool and the results it supplies. No study to date has focused specifically on information sources from which responses are derived. This was the main aim of our study.

Ideally, QA systems should create coherent definitions which contain and summarise the most descriptive information contained in a document collection, in view of the specific term or focus of the user query.[8,28]

Our study aimed to evaluate the quality and efficiency of two QA Systems, an open-domain QA system, START,[4] developed at the Massachusetts Institute of Technology, and a restricted-domain QA systems, MedQA,[5] which is specialised in the biomedical domain and developed by Columbia University. These were chosen because they have been used in several studies and have always given good results in the retrieval of general or specific information.[29]

Although QA systems offer different kinds of information depending on the factual question posed, our study focused on health questions. It was not our intention to evaluate the coverage of the databases sources of QA systems START and MedQA, but merely to appraise how they work and from what sources they retrieve data. We studied all the sources used by the both QA systems.

Methodology

A sample of two hundred questions about different medical issues were used as the basis of this study. The questions were obtained from the web page WebMD,[30] a US health portal created by health specialists providing valuable health information with on a number of illnesses.

The two hundred questions were obtained using the expression '*What is*…?' (i.e. what is irritable bowel syndrome?) in the internal search engine of the website; and in turn, WebMD provided a list of some 6000 responses in their characteristic question-answer format.

The questions were about different health issues (Table 1), were to be answered by both systems. Although other authors, like Ely *et al*[31] have proposed a classification of more generic questions, we have decided to create one based on the most generic questions of this taxonomy (Table 1).

START, which has a dynamic but easy interface, is a QA system allowing users to pose questions about various health issues, answering very specialised questions within the area of healthcare.[31] Information is retrieved from a very wide list of sources, such as *World Book*, *The World Factbook 2008*, *START KB*, *Internet Public Library*, and many others.

Meanwhile, MedQA, which has a user-friendly interface and uses more specialised, sources, analyses thousands of documents to arrive at coherent answers specifically within the area of healthcare.[32] It retrieves information from a wide array of sources, including *Wikipedia*, *Medline* or *Medline Plus*.

After presenting the questions to both QA systems, the answers were analysed and evaluated and the source or sources used by the system were identified. Answers were marked as: incorrect (0 points), inexact (1 point) or correct (2 points), according to one of the methods of evaluation proposed in the guidelines of Cross Language Evaluation Forum.[33] A student, a physician and a general user formed a group to judge the two hundred questions as correct, inexact or incorrect. To be judged as correct, the answer had to respond accurately to the question, the response could not use more than 100 words, with no irrelevant information. All the questions answered correctly but not fulfilling these criteria were considered inexact.[34] The response time and the partial or total repetitions of information by the systems were recorded.

Traditional information retrieval systems use recall and precision to measure performance. In our study, we have proved that it is only necessary to evaluate precision. In addition, the mark obtained by each question was the baseline for the

application of further evaluation measures, drawn from a 2001 study by Raved *et al.*[35] All these measures were chosen because they showed different aspects of the QA systems. All the measures used are described:

Mean Reciprocal Rank (MRR) a statistical tool evaluating any process that produces a list of possible answers to a query. The reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct answer (for example, if a question gets the correct answer in the 1st place, it will receive a score of 1, it would be ½ if it is in the 2nd place, 1/3 in the 3rd place...). If the answer is not found, a score of 0 is assigned. MRR can be used with several correct answers, but it only takes into account the first correct answer found.

Total Reciprocal Rank (TRR) is useful, when there is more than one correct answer to a question. It is not sufficient to consider the first correct answer in evaluations; instead, TRR takes into consideration all the correct answers and assigns a weight to each according to its ranking in the list provided by the system. For example, if the QA system provides two correct answers (the first and the third ones), the TRR will be 1/1 + 1/3.

First Hit Success (FHS) assigns 1 if the first answer returned by the system is correct and 0 if it is not. This measure, then, only accepts the first answer in the list of results. For a user who relies only on the QA system for retrieving answers, most probably the user only accepts the first answer returned by the system. If we solely consider the first answer retrieved to each question and assume that the QA systems' databases can provide answers to all the questions. Then the average of FHS represents the recall ratio of a QA system.

The measurement of 'precision' was used in the evaluation of information retrieval. The system should be able to retrieve documents or answers (in the case of QA systems) relevant to the query and well ranked (in the case of systems ranking the results).

$$\text{precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of documents retrieved}}$$

Results

After posing 200 questions in our QA systems, we identified the sources used by them to obtain answers. START provided answers to the medical questions from six sources (as shown in Table 2).

Sources used by START were: *Wikipedia*[22] a widely used online encyclopaedia offering information about different issues in several languages.

American Medical Association[36] a website which is the only specialised source used by START, it offers useful information about health for patients and physicians.

The Internet Movie Database (IMBb)[37] an American movie site, available in English, Spanish and Portuguese, with data about movies, series and actors from all over the world.

Yahoo[38] a directory that categorises web pages under different subjects.

Webopedia.com[39] an online computer dictionary and internet search engine for internet terms and technical support.

Merriam-Webster Dictionary[40] a free dictionary and thesaurus more strictly speaking, with definitions, etymology, pronunciation, etc. for each entry.

Wikipedia was the source offering the most answers with a total of 182. Second was *Merriam-Webster Dictionary* with 84 answers – although 31 of these were repetitions, these were rejected. Other answers provided by START are given in Table 3.

Table 2 Sources used by START

Sources	Answers obtained
Wikipedia	182
Merriam-Webster Dictionary	84 (31 repetitions)
American Medical Association	36
IMDB	5
Yahoo	2
Webopedia.com	1
Total	310

Table 3 Answers provided by START

Source	Correct	Inexact	Incorrect
Wikipedia	104	42	36
Merriam-Webster Dictionary	45	7	1
American Medical Association	1	35	0
Webopedia.com	1	0	0
Yahoo	0	0	2
IMDB	0	0	5
Total	151	84	44

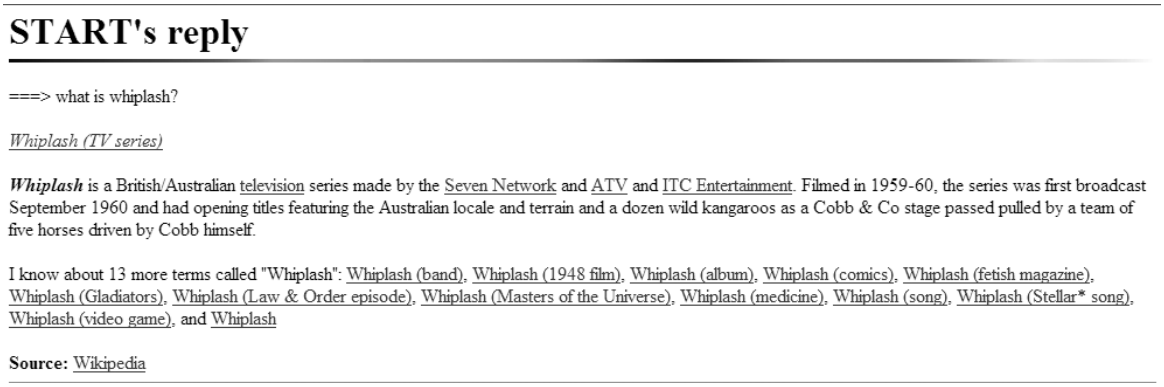

Figure 1 The intermediating 'window' of an inexact answer

In evaluating the quality of results by the START sources, *Wikipedia* was found to be the source giving most correct answers (104), with 42 answers that were inexact and 36 others that were incorrect. Some of the inexact answers pointed to an intermediating 'window' of sorts with several options related with the query, the question was not answered as our study expected (Fig. 1).

The number of answers retrieved by MedQA was higher than for START, and most sources were of a specialised nature. See Table 4.

Sources used by MedQA were: *Medline*,[41] a bibliographical database created by the *U.S. National Library of Medicine* includes citations and specialised articles from approximately 5000 selected journals, from 1966 to the present.

Dictionary of Cancer Terms[42] created by the *U.S. National Institute of Cancer.*

Google a search engine.

Dorland's Illustrated Medical Dictionary[43] another non-free dictionary for health issues.

Medline Plus,[44] a multi-lingual medical portal with information about medication, disease and other health issues, features a medical encyclopaedia, tutorials and videos for patients.

Technical and Popular Medical Terms,[45] a multi-lingual glossary set up by *The European Commission* and executed by *Heymans Institute of Pharmacology* and *Mercator School.*

National Immunization Program Glossary[46] of the *U.S. Department of Health & Human Services*.

Answers provided by MedQA are given in Table 5. Although *Google* is believed by previous authors as one of the best sources for answering definitional questions;[28] 34 answers were rejected as repetitions.

The two QA systems evaluated here gave similar figures for repeated answers (31 repetitions in START and 34 in MedQA). In START, all the repetitions were exactly identical, and came from the

Table 4 Sources used by MedQA

Sources	Answer obtained
Medline	200
Dictionary of Cancer Terms	192
Wikipedia	191
Google	174 (34 repetitions)
Dorland's Illustrated Medical Dictionary	143
Medline Plus	105
Technical and Popular Medical Terms	29
National Immunization Program Glossary	3
Total	1037

Table 5 Answers shown by MedQA

Source	Correct	Inexact	Incorrect
Google	122	26	26
Wikipedia	117	31	43
Medline Plus	95	1	9
Dictionary of Cancer Terms	51	0	140
Technical and Popular Medical Terms	21	3	5
Dorland's Illustrated Medical Dictionary	14	94	35
Medline	12	61	127
National Immunization Program Glossary	2	0	1
Total	434	216	386

same sources (*Merriam-Webster Dictionary*). In MedQA, the repetitions offered more or less the same answer, but their sources were different (*Wikipedia* and *Google*). Although a question may give different yet equally valid answers at a given time, when the same answer is repeated, users tend to feel confused, and the list of results increases unnecessarily. This is why we 'penalised' the QA systems by not considering these answers as valid.

As we see in Table 5, there were five sources providing more correct answers than inexact or incorrect ones: these were: *Medline Plus*, *Wikipedia*, *Google*, *Technical and Popular Medical Terms* and *National Immunization Program Glossary*. The only source supplying a majority of inexact answers was *Dorland's Illustrated Medical Dictionary*, which gave irrelevant information about *Dorland's* itself (copyright, edition and other non-pertinent information). *Medline* and the *Dictionary of Cancer Terms* gave more incorrect answers, and the latter sometimes offered irrelevant or incorrect information. *Medline* is a bibliographical database, and it rarely showed definitions about specific terms, but instead supplied extracts from studies (or abstracts) by health specialists or other researchers. Thus, we may infer that the questions were not expressed in the best possible terms. This is due to MedQA which was specifically designed and evaluated on definitional question answering.

Calculation of the response time for each question led us to some interesting findings. The values obtained were quite different for the two systems: the average response time for START was 2–4 s, while MedQA was considerably slower – with a minimum of 10 s and a maximum of 135 s. Overall, nearly 50% of the queries were solved in a period between 26 and 35 s (Fig. 2). During the wait, MedQA tells users that operations are underway at that moment – first of all, the system looks over *Google*, then in *Medline*, and finally, it removes all the redundant answers to generate the coherent ones.

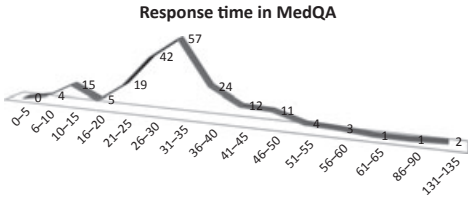

Figure 2 Analysis of frequencies according to the response time in MedQA

In identifying the sources used by the two systems, we applied specific measures for the evaluation of information retrieval.

Table 6 indicates that the average number of answers retrieved for each question is considerably higher with MedQA (5.2) than with START (1.41). Moreover, MedQA gave, on the average, more correct responses per question, 2.17, as compared with the 0.94 of START. This finding confirms that the more specialised system offers a more adequate subject coverage for the sort of query used here. Apart from the greater yield of responses provided by MedQA, the average offerings of incorrect and inexact responses are also greater under this system (1.93 and 1.08, respectively) than with the general-domain system START (0.22 incorrect and 0.25 inexact ones).

As we explained in the section on Methods, MRR calculates the inverse value of the first correct answer, whereas FHS simply evaluates if the first answer was correct or not. The two measures show that MedQA ranks their results more adequately, because the first correct answer tends to appear in the first place of the list (more frequently than with START). This proves very important, as no algorithm is involved in the ranking process. These systems, then, maintain the ranking of answers as determined by the source they came from. In terms of user-friendliness, FHS might be better, because users usually focus on the first answer retrieved.

The measure TRR is lower in MedQA, however. This figure takes into account not just the first one, but all the correct responses supplied by the system, and weights the value of the correct response in light of its placement within the list of results. As MedQA provides more results, the correct responses in the lower positions of the ranking receive less weight, and the TRR drops with respect to that of the START, which consistently yielded fewer responses.

Finally, we assessed the precision of the two systems. The value obtained for START precision was higher (67% relevant responses) than for MedQA (42%). The percentages increased if the

Table 6 Measures for evaluating the quality of answers

	Average answers retrieved per question	Average correct answers per question	Average incorrect answers per question	Average inexact answers per question	MRR	FHS	TRR	Precision*	Precision†
MedQA	5.18	2.17	1.93	1.08	0.86	0.75	0.40	42%	63%
START	1.41	0.94	0.22	0.25	0.60	0.61	0.59	67%	84%

MRR, Mean Reciprocal Rank; TRR, Total Reciprocal Rank; FHS, First Hit Success.
*Taking only correct responses into account.
†Taking both correct and inexact responses into account.

inexact answers were also included as relevant (84% with START and 67% for MedQA) Therefore, we may affirm that the more specialised system produces a greater degree of documental noise – that is, that the correct responses are accompanied by numerous incorrect and/or inexact one.

Discussion

Results obtained for the two systems analysed, START and MedQA, allowed us to evaluate their effectiveness and their use of different information sources. Despite certain limitations on the part of both systems (a lack of accessibility for the general public, and insufficient development in some specific areas), we were able to confirm that both are very useful in the retrieval of valid definitional healthcare information, with responses from both proving coherent and precise to an acceptable degree. They also help in understanding the information collected and are set to become one of the key tools available to index and organise health information.

As one might expect, the answers supplied by MedQA were more reliable that those of START in the sense that they came from specialised clinical or academic sources, and gave links to research articles on the subject in hand.

Another interesting finding is that the responses do not appear under a truly representative ranking of relevance, but rather, with both systems, results are shown in a pre-established order according to the source. The systems give priority in the display of results to sources that consistently provide answers (like *Wikipedia* or *Google*), regardless of the reliability and credibility that should be demanded of scientific information. Notwithstanding, we did observe that MedQA always makes use of *Medline* in responding to queries, which can be interpreted as a sign of reliability, yet not necessarily of precision.

Results are encouraging in that they point to the potential for this type of tool in the more general realm of information access. They are a good, reliable and reasonably precise alternative to help with information overload. They provide concrete results quickly and easily, enabling users to spend less time in the retrieval of information. Recent studies[12,47,48] have explored various possible means of enhancing the performance of such QA systems, for instance through the incorporation of ontology, which would heighten the quality of the answers obtained by structuring, inter-relating and formalising all relevant information from the thematic domain. In addition, other approaches such as computational grammars are slowly attracting experienced researchers in handling the results they produce. This data suggests that we may see unexpected changes in the future. This area deserves to be studied and evaluated in future research.

Conclusions

Health information and libraries need current terminological information to organise and index information. Different studies in Information Retrieval have shown that QA systems are a useful tool for retrieving information quickly and accurately. In this study, we have investigated the effectiveness of these systems in the retrieval of health information, and the main differences between an open-domain QA system, like START, and a restricted-domain QA system, like MedQA.

Question-answering systems may provide a different way for physicians and users in general to seek biomedical information and identify tools to limit human work.

References

1 Jackson, P. & Schilder, F. Natural language processing: overview. In: Brown, K. (ed). *Encyclopedia of Language & Linguistics*, 2nd edn. Amsterdam: Elsevier Press, 2005: 503–518.

2 Access to Text REtrieval Conference (TREC). Accessible at: http://trec.nist.gov/.

3 Voorhees, E. M. The TREC 8 Question Answering Track Report. In: Voorhees, E. M. & Harman, D. K. (eds). *Proceedings of the Eighth Text REtrieval Conference*, vol. 500-246 in NIST Special Publication, Gaithersburg, MD: NIST, 1999: 107–130.

4 Access to START (Natural Language Question Answering System). Accessible at: http://start.csail.mit.edu/.

5 Access to MedQA. Accessible at: http://monkey.ims.uwm.edu:8080/MedQA/.

6 Blair-Goldensohn, S. B. & Schlaikjer, A. H. Answering definitional questions: a hybrid approach. *New Directions In Question Answering* 2004, **4**, 47–58.

7 Costa, L. F. & Santos, D. *Question Answering Systems: A Partial Answer*. Oslo: SINTEF, 2007.

8 Zweigenbaum, P. Question answering in biomedicine. In: De Rijke, M. & Webber, B. (eds). *Proceedings Workshop on Natural Language Processing for Question Answering*. Budapest: ACL, EACL, 2003: 1–4.

9 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.

10 Mollá, D. & Vicedo, J. L. *Question-Answering in Restricted Domains*. Menlo Park, CA: AAAI Press, 2005.

11 Tsur, O. *Definitional Question-Answering Using Trainable Text Classifiers*. PhD Thesis. Amsterdam: Institute of Logic Language and Computation (ILLC), University of Amsterdam, 2003.

12 Sing, G. O., Ardil, C., Wong, W. & Sahib, S. Response quality evaluation in heterogeneous question answering system: a black-box approach. *International Journal of Information Technology*, **2**, 4, 2006.

13 Fahmi, I. Automatic term and relation extraction for medical question answering system. Groningen Dissertations of Linguistics, 2009, 72.

14 Alfonseca, E., De Boni, M., Jara, J. L. & Manandhar, S. A prototype question answering system using syntactic and semantic information for answer retrieval. In: E. M., Voorhees and D. K., Harman. (eds). *Proceedings of the 10th Text Retrieval Conference (TREC-10)*. Gaithersburg, NIST, Gaithersburg, MD, 2002.

15 Jacquemart, P. & Zweigenbaum, P. Towards a medical question-answering system: a feasibility study. In: Beux, P. L. & Baud, R. (eds). *Proceedings of Medical Informatics Europe (MIE '03)*, vol. 95 of Studies in Health Technology and Informatics, San Palo, CA, 2003: 463–468.

16 Access to asked (Automatic Multilingual Question Answering System). Accessible at: http://asked.jp/edw/pc/.

17 Access to WolframAlpha computational knowledge engine. Accessible at: http://www.wolframalpha.com/.

18 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. In: Johannesson, P. (ed). *Proceedings of the Seventh International Workshop on Applications of Natural Language to Information Systems (NLDB 2002)*, Stockholm, Sweden, Lecture Notes in Computer Sciences, Springer Verlag, 2002: 230–234.

19 Access to NSIR (Question Answering System). Accessible at: http://tangra.si.umich.edu/clair/NSIR/html/nsir.cgi/.

20 Access to QuaLiM (Question Answering Demo). Accessible at: http://demos.inf.ed.ac.uk:8080/qualim/.

21 Access to Google. Accessible at: http://www.google.com/.

22 Access to Wikipedia. Accessible at: http://www.wikipedia.org/.

23 Crouch, D., Saurí, R. & Fowler, A. AQUAINT pilot knowledge-based evaluation: annotation guidelines. Tech. rep., Palo Alto Research Center, 2005.

24 Lee, M., Cimino, J., Zhu, H. R., Sable, C., Shanker, V., Ely, J. & Yu, H. *Beyond Information Retrieval – Medical Question Answering*. Washington, DC: AMIA, 2006.

25 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* 2007, **4**, 236–251.

26 Ely, J. W., Osheroff, P. N., Ebell, M., Bergus, G., Barcey, L., Chambliss, M. & Evans, E. Analysis of questions asked by family doctors regarding patient care. *British Medical Journal* 1999, **319**, 358–361.

27 Yu, H. & Kaufman, D. A cognitive evaluation of four online search engines for answering definitional questions posed by physicians. *Pacific Symposium on Biocomputing* 2007, **12**, 328–339.

28 Blair-Goldensohn, S. B., McKeow, K. R. & Schlaikjer, A. H. A hybrid approach for QA track definitional questions. *In Proceedings of the 12th Text Retrieval Conference (TREC 2003), Gaithersburg, Maryland,* 2003, 336–343.

29 Katz, B., Felshin, S., Yuret, D., Ibrahim, A., Lin, J., Marton, G., Jerome McFarland, A. & Temelkuran, B. *Proceedings of the Seventh International Workshop on Applications of Natural Language to Information Systems (NLDB 2002)*. Omnibase: Uniform Access to Heterogeneous Data for Question Answering, 2002.

30 Access to WebMD. Accessible at: http://www.webmd.com/.

31 Ely, J. W., Osheroff, J. A., Gorman, P. N., Ebell, M. H., Chambliss, M. L., Pifer, E. A. & Stavri, P. Z. A taxonomy of generic clinical questions: classification study. *British Medical Journal* 2000, **321**, 429–432.

32 Katz, B., Felshin, S., Yuret, D., Ibrahim, A., Lin, J., Martion, G., McFarland, A. J. & Temelkuran, B. Uniform access to heterogeneous data for question answering. In: Johannesson, P. (ed). *Proceedings of the Seventh International Workshop on Applications of Natural Language to Information Systems (NLDB 2002, Stockholm, Sweden, Lecture Notes in Computer Sciences, Springer Verlag)*. Omnibase, 2002: 230–234.
33 Access to Cross Language Evaluation Forum (CLEF). Accessible at: http://www.clef-campaign.org/
34 Cao, Y. G., Ely, J., Antieau, L. & Yu, H. Evaluation of the clinical answering presentation. *Proceedings of the Workshop on BioNLP, Boulder, Colorado,* 2009, 171–178.
35 Raved, D. R., Qi, H., Wu, H. & Fan, W. *Evaluating Web-Based Question Answering Systems*. Technical Report, Michigan: University of Michigan, 2001.
36 Access to American Medical Association (AMA). Accessible at: http://www.ama-assn.org/.
37 Access to Internet Movie Database (IMDb). Accessible at: http://www.imdb.com/.
38 Access to Yahoo. Accessible at: http://www.yahoo.com/.
39 Access to Webopedia. Accessible at: http://www.webopedia.com/.
40 Access to Merriam-Webster. Accessible at: http://www.merriam-webster.com/.
41 Access to Medline. Accessible at: http://www.ncbi.nlm.nih.gov/pubmed/.
42 Access to Dictionary of Cancer Terms. Accessible at: http://www.cancer.gov/dictionary/.
43 Access to Dorland's Illustrated Medical Dictionary. Accessible at: http://www.dorlands.com/wsearch.jsp/.
44 Access to MedlinePlus. Accessible at: http://medlineplus.gov/.
45 Access to Glossary of Technical and Popular Medical Terms. Accessible at: http://users.ugent.be/~rvdstich/eugloss/welcome.html/.
46 Access to National Immunization Program Glossary. Accessible at: http://www.cdc.gov/vaccines/about/terms.htm/.
47 Buitelaar, P., Cimiano, P., Frank, P., Hartung, M. & Racioppa, S. Ontology-based information extraction and integration from heterogeneous data sources. *International Journal of Human-Computer Studies* 2008, **66**, 759–788.
48 Cruchet, S., Gaudinat, A., Rindflesch, T. & Boyer, C. What about trust in the Question Answering world? AMIA 2009 Annual Symposium, 2009.

Received *2 June 2009;* Accepted *10 May 2010*