QUESTION-ANSWERING SYSTEMS AS EFFICIENT SOURCES OF TERMINOLOGICAL INFORMATION: AN EVALUATION

Abstract

Question-answering systems (or QA Systems) stand as a new alternative for Information Retrieval Systems. Most users often need to retrieve specific information about a factual question in order to obtain a whole document. We conducted a study to evaluate the efficiency of QA systems as terminological sources for physicians, specialized translators, and users in general. To this end we analyzed the performance of one open-domain QA system, START, and one restricted-domain QA system, MedQA. The research entailed a collection of two hundred definitional questions (*What is…?),* either general or specialized, from WebMed. We studied the sources that QA systems used to retrieve the answers, and later applied different evaluation measures to mark the quality of answers. Both QA systems were determined to be appropriate for the retrieval of terminology, proving reliable as sources and supplying correct answers.

Keywords

Question-answering systems, Evaluation of QA systems, definitional questions, sources of health information, MedQA, Start.

Introduction

Question-answering systems (heretofore QA Systems) can be viewed as a new alternative to the more familiar Information Retrieval Systems. These systems try to offer detailed, understandable answers to factual questions, in order 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. The QA systems try to make retrieval easier through the short-answer question models.[3-4] Accordingly, users do not have to read the full text of documents such as a web page, an article of scientific journal, etc., in order to arrive at the information they need because the QA system shows the correct answer by means of a number, a noun, a short phrase or a concise extract of text.

The 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, to be comprehended in a natural way for humans.[5] QA systems follow three main steps: first of all, the systems retrieve the documents to obtain relevant sentences about the search term; they retrieve and select the sentences; and finally, they choose non-redundant definition sentences from the overall results of sentence retrieval, to

delimit the response.[6-7] The objective pursued is for the systems to retrieve only the correct information to answer the users´ questions.[8] Evaluation is one of the most important dimensions in QA systems, as the process of assessing, comparing and ranking is key to monitoring progress in the field.[9] The main component of these systems consists of measuring modules, which analyze the tagged sentences in the documents selected, and compare them with the question in order to find the most similar sentence.[10-11] Generally speaking, QA systems feature very simple and user-friendly interfaces, and rely on methods of linguistic analysis and natural language processing in the different phases of operation. For example, when dealing with the questions posed by the users, they identify their component parts and then determine the kind of answer anticipated.[12] The ones that allow users to query in different languages are known as multilingual QA systems.

Although these systems have enriched the possibilities of information retrieval tools, QA systems are also hampered by certain restrictions. To date, most of them are not open-domain systems or general systems,[13] but rather restricted-domain systems specializing in a particular field. Moreover, all of these QA systems are based on prototypes; that is, they are available as demos, and only in a few cases have been marketed. A further problem can be found in the design of these systems: what is needed is a more interactive QA procedure that allows for real feedback between questions and answers, and user communication with the system on a conversational level.

While not many QA systems are available on the Internet, we do have some open-domain QA systems such as START[14], developed at the Massachusetts Institute of Technology, it is a very atypical which includes calls to OMNIBASE, a system that integrates heterogeneous data sources using an *object-property-value model* (Katz et al., 2002); NSIR[15], developed by the University of Michigan; or Qualim[16], financed by Microsoft; in addition to some restricted-domain QA systems including MedQA[17], developed by Columbia University. In the case of NSIR and Qualim, answers are constructed on the basis of information provided by Google[18] and Wikipedia[19], respectively. Although START also retrieves information from Wikipedia, it uses other specialized sources such as directories, databases, dictionaries, or encyclopaedias. Meanwhile, MedQA retrieves information from the medical database Medline, specialized dictionaries, Wikipedia and certain search engines like Google.

In the Web setting, the overload of information may be perceived more acutely than in other contexts. When users pose a given question by means of search engine tools (including directories or metasearchers), the systems tend to retrieve an excessive number of web pages, many of which are not relevant or useful in light of the users' needs. Professionals in different areas claim that QA systems constitute a good method to obtain specialized information in quick and efficient manner. [20-22]

In a study by Ely,[23] participating physicians spent on the average less than two minutes looking for information to resolve clinical queries, although many of their

questions remained unanswered. Regarding this point, some researchers have shown that the physicians trust QA systems as search methods for specialized information retrieval.[21,24] The general public increasingly consults knowledge resources like the Web as well: before or after seeing a doctor, for themselves or for relatives, to obtain information about the nature of a disease, the indications and contraindications of a treatment, etc.[12]

While researchers have looked into various aspects of QA systems in recent years, one facet that is widely overlooked is the formal evaluation of this tool and the results it supplies. Indeed, no study to date has focused specifically on the information sources from which responses are derived. This is the main aim of our line of research. Ideally, QA systems should create coherent definitions in a dynamic way, and ones that contain and summarize the most descriptive information contained in a document collection, in view of the specific term or focus of the user query.[12,25]

Our objective led us to use definition-type questions in order to evaluate two QA systems and determine the different sources behind the retrieval of medical information. In the sections below we describe the questions used, the QA systems analyzed and the measures of evaluation applied. Finally, we show our results and briefly expound some conclusions.

Methodology

We took a sample of two hundred definitional questions about different medical issues as the basis of this study. The questions were obtained from the webpage WebMD[26], a US health portal providing valuable health information and support, tools for managing health problems, and specialized background on a number of illnesses. It was created by health specialists who aspire to explain, briefly yet credibly, in-depth medical information, reference material, and online community programs.

The collection of two hundred questions was created 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. We chose around 250 factual questions about different health issues, specified in Table 1. 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 what sources they retrieve data from. This led us to finally choose 200 questions to be answered by both systems.

Authors Ely and colleagues[27] suggest a classification of five hierarchical categories to categorize medical questions. Firstly they distinguish between clinical and non-clinical questions: the clinical questions are further divided into general and specific; general questions are divided into evidence and no-evidence; and in turn, the evidence questions are divided into intervention and no-intervention. While not all

these categories were appropriate for our purposes, they served as the foundation for our classification (Table 1).

Table 1. Categories of reference of definitional questions.

START, a QA system allowing users to pose questions about various health issues, can respond to even very specialized questions within the area of health care.[28] It has a dynamic yet easy interface, and responds quickly. 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[17] is a specialized QA system that analyses thousands of documents to arrive at a coherent response. Because it works specifically in the area of health care, its sources are more specialized.[24] It also has a user-friendly interface, but it is slower than START. It retrieves information from a wide array of sources, including *Wikipedia, Medline* or *Medline Plus*.

After presenting the questions to both QA systems, we analyzed and evaluated the answers obtained, and identified the source or sources used by the system. Answers were marked as: incorrect (0 points), inexact (1 point) or correct (2 points), according to the guidelines of CLEF (Cross Language Evaluation Forum)[29]. To be judged as correct, the answer had to respond accurately to the question asked, not use more than 100 words in its response, and not contain irrelevant information. All the questions that were answered correctly yet did not fulfil these criteria were considered inexact. Likewise, we recorded the response time and the partial or total

repetitions of information by the systems. The mark obtained by each question was the baseline for application of further evaluation measures, explained below.[30]

Mean Reciprocal Rank (MRR) is a statistical tool for 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. In these cases, 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 questions in the list of results.

We used the measurement of "precision" in the evaluation of information retrieval. It is understood as the capacity of system 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).

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

Results

After posing 200 questions in our QA systems, we identified the sources used by them to obtain the answers. START provided answers to the medical questions from six sources, appearing in this order: *Wikipedia*, *American Medical Association* (the only specialized source used by START), *The Internet Movie Data Bases*, *Webopedia.com*, *Yahoo* and *Merriam Webster Dictionary*.

Very briefly, *Wikipedia* is a widely used online encyclopaedia able to offer information about different issues in several languages. The website *American Medical Association*[31] offers useful information about health for patients and physicians. *The Internet Movie Database* (IMBD)[32] is an American movie site, available in some languages, with data about movies, series and actors from all over the world. *Yahoo*[33] is a directory that categorizes web pages under different subjects. *Webopedia.com*[34] is an online computer dictionary and internet search engine for internet terms and technical support. And finally, the *Merriam-Webster Dictionary*[35] is a free dictionary and thesaurus more strictly speaking, with definitions, etymology, pronunciation, etc. for each entry.

The source that offered more answers was *Wikipedia*, with a total of 182. Second was *Merriam Webster Dictionary* with 84 answers –although 31 of these answers repeated exactly the same information, for which reason we rejected them. *American Medical Association*, the only specialized source, gave 36 answers. The sources providing the fewest answers were *The Internet Movie Data Base* (IMDB), *Yahoo* and *Webopedia.com*, with 5, 2 and 1 answers obtained, respectively.

Table 2. Sources used by START

In evaluating the quality of the results by the START sources (Table 3), *Wikipedia* was found to be the source giving more 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 general dictionary *Merriam-Webster Dictionary* offered 45 correct answers, 7 inexact ones and only one incorrect answer. The *American Medical Association* supplied just one correct answer and 35 inexact answers. The only response obtained through *Webopedia.com* was considered correct, whereas all the answers of *IMDB* and *Yahoo* were incorrect.

Table 3. Answers provided by START

The number of answers retrieved by MedQA was higher than for START, and most sources were of a specialized nature. *Medline*[36] answered all the questions. This bibliographic database created by the *U.S. National Library of Medicine* includes citations and specialized articles from approximately 5000 selected journals, from 1966 to the present.

Table 4. Sources used by MedQA

The *Dictionary of Cancer Terms*[37] (created by the *U.S. National Institute of Cancer*) and *Wikipedia* offered 192 and 191 answers, respectively. *Google* is appraised by previous authors as one of the best sources for answering definitional questions;[24] this search engine offered 174 answers in our experience, though 34 were rejected as repetitions. *Dorland's Illustrated Medical Dictionary,*[38] another non-free dictionary for health issues, gave 143 answers. *Medline Plus,*[39] as a multilingual medical portal with information about medication, disease and other health issues, features a medical encyclopaedia, tutorials and videos for patients; it gave us 105 answers. The multilingual glossary of *Technical and Popular Medical Terms,*[40] set up by *The European Commission* and executed by *Heymans Institute of Pharmacology* and *Mercator School,* provided 29 results. The *National Immunization Program Glossary*[41] of the *U.S. Department of Health & Human Services* supplied just 3 answers.

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 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 harvest 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 "penalized" the QA systems by not considering these answers as valid.

Table 5. Answers shown by MedQA

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 remitted irrelevant information about *Dorland's* itself (copyright, edition and other non-pertinent information) in most responses. *Medline* and the *Dictionary of Cancer Terms* gave more incorrect answers, and this dictionary 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 was specifically designed and evaluated on definitional question-answering.

Calculation of the time of response (time elapsing before appearance of results on the screen) 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 to 4 seconds, while MedQA was considerably slower –with a minimum of 10 seconds and a maximum of 135 seconds. Overall, nearly 50% of the queries were solved in a period between 26 and 35 seconds (Figure 1). 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 1. 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 comes to confirm that the more specialized system offers a more adequate coverage by subject for the sort of query collection used here; and aside 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).

Table 6. Measures for evaluating the quality of answers

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 us, in this case, 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 a somewhat more realistic or convenient measure, 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. Since MedQA provides a greater amount of 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 inexact answers were also included as relevant (84% with START and 67% for MedQA) Therefore, we may affirm that the more specialized system produces a greater degree of documental noise –that is, that the correct responses are accompanied by numerous incorrect and/or inexact one.

Discussion

The results obtained by presenting 200 questions to the two separate systems analysed here, START and MedQA, allowed us to subsequently 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 health-care information, with responses from both proving coherent and precise to an acceptable degree. However, as one might expect, the answers supplied by MedQA were more reliable that those of START in the sense that they came from specialized clinical or academic sources, most of them showing links to research articles addressing the matter at 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 source of the information. The systems give priority in the display of results to those sources that consistently provide answers (like *Wikipedia* or *Google),* regardless of the criteria of 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 of this type of tool in the more general realm of information access, as they may be a good, reliable and reasonably precise alternative on occasions, alleviating informational overload. They are able to provide concrete results quickly and easily. Recent studies[9,42] 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 formalizing all relevant information from the thematic domain of reference.

References

1 Jackson, P. & Schilder, F. Natural Language Processing: Overview". In: Brown, K. (eds). *Encyclopedia of Language & Linguistics*, 2nd. Ed. Amsterdam: Elsevier Press., 2005: 503–518.

2 Access to Text REtrieval Conference (TREC). Available from: 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 8th Text REtrieval Conference*, vol. 500-246 in NIST Special Publication, NIST, Gaithersburg, Md, 1999: 107–130.

4 Blair-Goldensohn, S.B. & Schlaikjer, A.H. Answering Definitional Questions: A Hybrid Approach. *New Directions In Question Answering*, 2004, 4: 47-58.

5 Costa, L.F. & Santos, D. Question Answering Systems: a partial answer. *SINTEF*, Oslo, 2007.

6 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.

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

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

9 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*, 2005: 9.

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

11 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, California, 2003: 463–468.

12 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.

13 Frank, A., Kirefer, H.U., Xu, F., Uszkoreit, H., Crysmann, B., Jörg, B. & Shäfer, U. Question answering from structured knowledge sources. *Journal of*

Applied Logic, Special Issue on Questions and Answers: Theoretical and Applied Perspectives, 2006: 5, 20–48

14 Access to START (Natural Language Question Answering System). Available from: http://start.csail.mit.edu/

15 Access to NSIR (Question Answering System). Available from: http://tangra.si.umich.edu/clair/NSIR/html/nsir.cgi/

16 Access to QuaLiM (Question Answering Demo). Available from: http://demos.inf.ed.ac.uk:8080/qualim/

17 Access to MedQA. Available from: http://monkey.ims.uwm.edu:8080/MedQA/

18 Access to Google. Available from: http:// www.google.com/

19 Access to Wikipedia. Available from: http://www.wikipedia.org/

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

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

22 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.

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

24 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.

25 Blair-Goldensohn, S.B., McKeow, K.R. & Schlaikjer, A.H. A hybrid Approach for QA Track Definitional Questions. *Proceedings of TREC*, 2003: 336-343.

26 Access to WebMD. Available from: http://www.webmd.com/

27 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*, 2000: 321, 429–432.

28 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: *Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems (NLDB 2002)*. 2002: 230–234.

29 Access to Cross Language Evaluation Forum (CLEF). Available from: http://www.clef-campaign.org/

30 Raved, D.R., Qi, H., Wu, H. & Fan, W. *Evaluating Web-based Question Answering Systems*. Technical Report, University of Michigan, 2001.

31 Access to American Medical Association (AMA). Available from: http://www.ama-assn.org/

32 Access to Internet Movie Database (IMDb). Available from: http://www.imdb.com/

33 Access to Yahoo. Available from: http://www.yahoo.com/

34 Access to Webopedia. Available from: http://www.webopedia.com/

35 Access to Merriam-Webster. Available from: http://www.merriam-webster.com/

36 Access to Medline. Available from: http://www.ncbi.nlm.nih.gov/pubmed/

37 Access to Dictionary of Cancer Terms. Available from: http://www.cancer.gov/dictionary/

38 Access to Dorland's Illustrated Medical Dictionary. Available from: http://www.dorlands.com/wsearch.jsp/

39 Access to MedlinePlus. Available from: http://medlineplus.gov/

40 Access to Glossary of Technical and Popular Medical Terms. Available from: http://users.ugent.be/~rvdstich/eugloss/welcome.html/

41 Access to National Immunization Program Glossary. Available from: http://www.cdc.gov/vaccines/about/terms.htm/

42 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*, 2008: 66, 759 – 788.

Table 1

Question	PAIN	INFLAMMATION	DISEASE	SYNDROME	INFECTION	TREATMENT	OTHERS
Number	8	16	97	11	10	38	15

Table 1. Categories of reference of definitional questions.

Table 2

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 2. Sources used by START

Table 3

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**

Table 3. Answers provided by START

Table 4

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 4. Sources used by MedQA

Table 5

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**

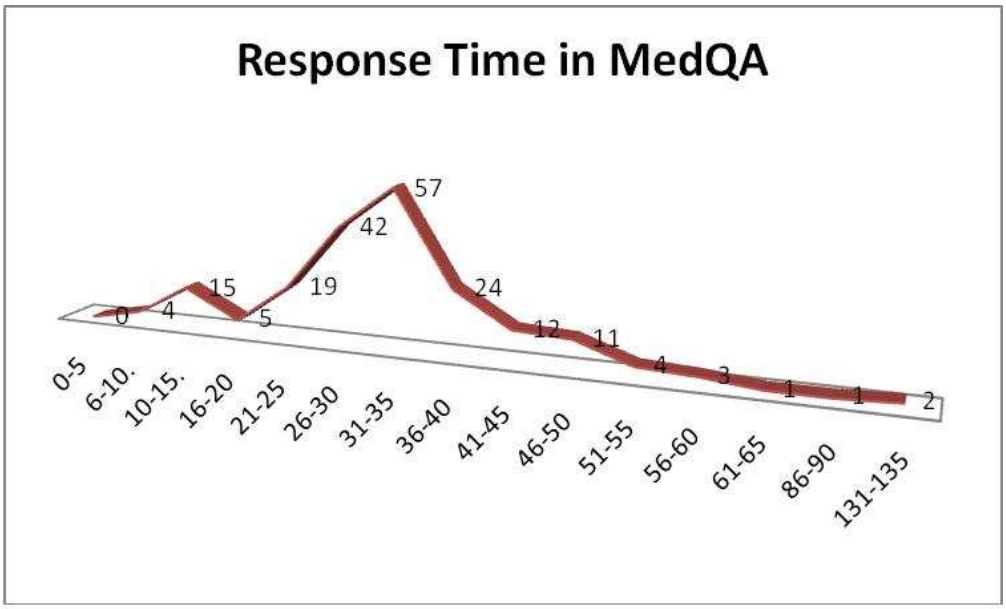
Table 5. Answers shown by MedQA

Table 6

	Average answers retrieved per question	Average correct answers per question	Average incorrect answers per question	Average inexact answers per question	MRR	FHR	TRR	Precision (1)	Precision (2)
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%

(1)Taking only correct responses into account
(2)Taking both correct and inexact responses into account

Table 6. Measures for evaluating the quality of answers



Analysis of frequencies according to the response time in MedQA
199x120mm (96 x 96 DPI)