Article

Journal of Information Science
37(2) 152–162
© The Author(s) 2011
Reprints and permission: sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0165551511398575
jis.sagepub.com

Open- vs. Restricted-Domain QA Systems in the Biomedical Field

María-Dolores Olvera-Lobo
CSIC, Unidad Asociada Grupo SCImago, Madrid, Spain and Universidad de Granada, Department of Library and Information Science, Granada, Spain

Juncal Gutiérrez-Artacho
Universidad de Granada, Department of Translation and Interpretation, Granada, Spain

Abstract

Question answering systems (QA systems) stand as a new alternative for information retrieval systems. 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 analysed the performance of two open-domain and two restricted-domain QA systems. The research entailed a collection of 150 definitional questions from WebMed. We studied the sources that QA systems used to retrieve the answers, and later applied a range of evaluation measures to mark the quality of answers. Through analysing the results obtained by asking the 150 questions in the QA systems MedQA, START, QuALiM and HONqa, it was possible to evaluate the systems' operation through applying specific metrics (MRR, FHS, TRR, Precision, Recall). Despite the limitations demonstrated by these systems, it has been confirmed that these four QA systems are valid and useful for obtaining definitional medical information in that they offer coherent and precise answers.

Keywords

biomedical information; evaluation; open-domain question answering systems; question answering systems; restricted-domain question answering systems

1. Introduction

Information retrieval (IR) is a discipline focused on the problems of information items' selection from a storage system in order to facilitate retrieval for the users' needs [1–5]. Traditionally, IR is understood as a fully automatic process that responds to a user query by examining a collection of documents and returning a sorted document list that should be relevant to the user requirements as expressed in the query [1]. Simply stated, it could be said that retrieval implies finding certain requested information in a storage system or database of information [6]. An optimal IR system recovers *all* the relevant documents (implying an exhaustive search, i.e. a high recall) and *only* the relevant documents (implying perfect accuracy, that is to say, a high precision). This traditional model involves many implied restrictions: the assumption that users want full-text documents, rather than answers, and that the query will be satisfied with these documents; that the process is direct and unidirectional rather than interactive; and finally, that the query and document share the same language.

Information overload is felt more strongly on the web than elsewhere. All too often a query made with a web search tool (search engine or meta-search engine) results in the retrieval of too many pages – many of which are useless or irrelevant to the user. Question answering systems (QA systems) are an evolutionary improvement in IR systems. As alternative traditional IR systems, they give correct and understandable answers to factual questions [7] – rather than just offering a list of documents related to the search. The benefit is that users do not have to read whole documents to find the desired information. Therefore, professionals from various areas are beginning to recognize the usefulness of these systems, for quickly and effectively finding specialized information [8–10].

Corresponding author:
Juncal Gutiérrez-Artacho, Facultad de Traducción e Interpretación, C/ Buensuceso 11 s/n, Universidad de Granada, Granada, Spain.
Email: juncalgutierrez@ugr.es

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 analysed the performance of two open-domain QA systems, START and QuALiM, and two restricted-domain QA systems, MedQA and HONqa. The research entailed a collection of 150 definitional questions (What is…?), either general or specialized, from WebMed. We studied and analysed the performance of each QA system through to the answers retrieved by them in order to assess their efficacy in the biomedical field, using multiple evaluation measures.

2. Open- vs restricted-domain QA systems on the web

QA systems have attracted major attention since the TREC-8 (Text REtrieval) conference on information retrieval [11]. They are based on short-answer models [12] and their aim is to find an exact and correct answer – in the form of a number, a noun, a short phrase or a brief piece of text [7] – for users' questions. Analysis of question, search and choosing an answer are three important issues in a QA system so it includes at least the three following processes: question processing, document processing and answer processing [13].

Although there are various templates for making queries in QA systems, most of these systems understand questions expressed with interrogative particles (*who*, *what*, *where*, *why*, *when and how*), while some understand the imperative form (*tell me*). Then QA systems proceed to do coherent questions according to natural language [14]. When a query is entered into the interface, the system proceeds to analyse the question by separating the words or keywords. The system then locates and extracts one or several answers from different sources of information, depending on the question's specialized area [15]. Subsequently, the system evaluates and eliminates redundant information, or information that does not respond correctly to the question, and submits one or more prepared responses to the user [16–17]. 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 [18]. The main component of these systems consists of measuring modules, which analyse the tagged sentences in the documents selected, and compare them with the question in order to find the most similar sentence [19–20]. 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. Multilingual QA systems allow users to query in different languages.

These systems usually have a simple interface where users can enter their queries, while some offer a list of recent queries to help users understand how the system works. QA systems handle these queries by applying algorithms and methods of linguistic analysis, as well as using natural language processing to identify the components and determine the expected response [21]. This analysis usually uses a variety of standard questions in which certain words are replaced by labels accepted by the system [7].

While the development of QA systems represents progress, the systems nevertheless suffer restrictions. Many were only developed as prototypes, or demonstration versions, and few were marketed. Some researchers have designed and created systems that were presented and discussed at various forums and conferences. However, because the usefulness of the systems was limited to very specific contexts, or because of problems of implementation, only a few of these systems were later developed for end users.

QA systems may be general domain or domain-specific, also known as open- and restricted-domain QA systems. General domain systems answer questions from diverse fields and domain-specific systems, which focus on a specialized area, use specific linguistic resources that enable more precise answers to be given. It can find several examples of both types of QA systems available on the web, i.e. open-domain QA system *TextMap Question Answer* (http://brahms.isi.edu:8080/textmap/) and a restricted-domain QA systems *EAGLI* (*Engine for question-Answering in Genomics Literature*; http://eagl.unige.ch/EAGLi/) among others.

We analysed: START, QuALiM, MedQA and HONqa. We chose these four QA systems because they have been used in previous research works about this field, they have been available for few years, and they have been developed by significant research groups.

START (http://start.csail.mit.edu/) Natural Language Question Answering System is a publicly accessible information access system that has been available for use on the internet since 1993. START answers natural language questions by presenting components of text and multi-media information drawn from a set of information resources that are hosted locally or accessed remotely through the web. These resources contain structured, semi-structured and unstructured information [22–23]. It has a dynamic yet easy interface. 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.

Another open-domain QA system is QuALiM (http://demos.inf.ed.ac.uk:8080/qualim/). This is a question answering demo, financed by Microsoft, which retrieves textual information by means of Wikipedia and graphic information, through the Google image search engine. QuALiM is a pattern-based QA system that searches the web for answers. Each of its patterns contains a syntactic description that matches a subclass of questions, a set of syntactic descriptions of potential answer sentences, and semantic information concerning the appropriate answer type for the question class. When asked a question, QuALiM will search all of the patterns' question descriptions and retain those that match the question. The matching pattern's information about potential answer sentence formulations is used to create rather specific quoted search queries that are sent to a web search engine (either Google or Yahoo) [24–25].

In a study by Ely and colleagues [26], participating physicians spent on 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 physicians trust QA systems as search methods for specialized IR [9–10]. The general public also increasingly consults knowledge resources like the web: 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, and so on [21].

HONqa (http://services.hon.ch/cgi-bin/QA10/qa.pl/) is a domain-restricted QA system operated by the Health On the Net Foundation, which is a Swiss non-profit organization that aims to promote the development of quality, reliable medical information. The HON Foundation has dedicated itself to the maintenance and improvement of the quality of online medical information. With this view, HON has developed a QA applied to medical research where responses obtained are from all the sites certified by the foundation [27]. HONqa is the only multilingual system analysed in our study, as it provides information in English, Italian and French, although we have only focused on the information retrieved in English. The sources of information used by the system are varied, and its search engines are usually general medical or specialized towards a certain illness.

Finally, MedQA, Medical Question Answering System (nowadays this QA system is not available on the webpage) is a specialized QA system that analyses thousands of documents to arrive at a coherent response. Because it works specifically in the area of healthcare, its sources are more specialized [28]. It also has a user-friendly interface. It retrieves information from a wide array of sources, including *Wikipedia*, *Medline* or *Medline Plus*. To identify definitional sentences, MedQA implemented a set of lexico-syntactic patterns generated automatically from a large set of precompiled definitional sentences that were collected using Google. The lexicon-syntactic patterns were ranked based on the ratio of their occurrences in the definitional and in the non-definitional sentences. This system suffered from incorrect noun phrases being extracted by its shallow syntactic chunked from the questions [25].

Although researchers have recently studied different aspects of QA systems, there are not yet enough studies to evaluate how these tools work. In our study, we have focused on how these four QA systems work to generate phrases with dynamic, coherent definitions and with the most interesting information [16, 29].

3. Methodology

In total, a sample of 150 definitional questions about different medical issues formed the basis of this study (Figure 1). We decided to use definitional questions because they do not demand specific data (such as a date, name or place) as the QA

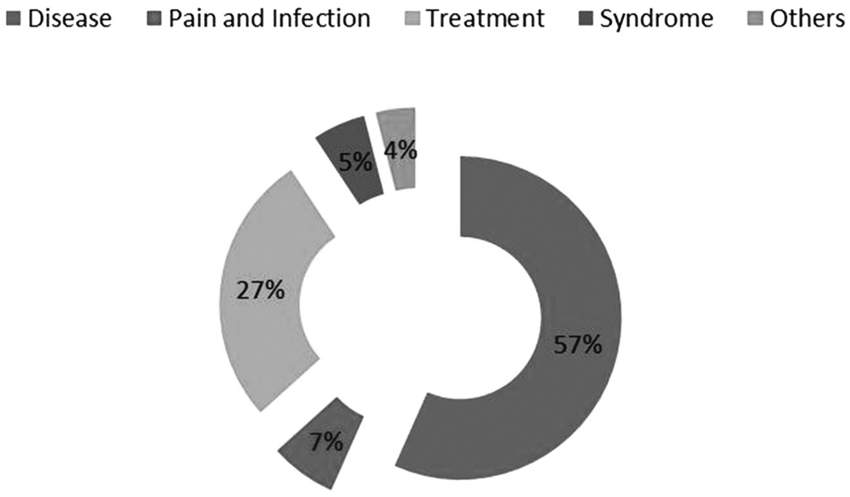

Figure 1. Category of question reference.

system builds an answer. So, these systems have to identify and obtain relevant information in their sources used, and have to summarize and develop the best answer possible as a definition. The questions were obtained from the webpage WebMD (http://www.webmd.com/), 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 programmes. The set of questions used passed the internal consistency test with a Cronbach's alpha coefficient of 0.997.

We used START and QuALiM as open-domain QA systems – although allowing users to pose questions about various health issues, they can respond to even very specialized questions within the area of healthcare [30] – and, as restricted-domain QA systems, MedQA and HONqa.

After asking questions in the four systems, a group of two medical professionals and three medical students evaluated the answers as incorrect, inexact or correct. Correct answers were those that answered the question adequately, were expressed in fewer than 100 words and did not contain information that was irrelevant to the question. Responses that answered the question correctly but that did not meet the rest of the criteria were evaluated as inexact. The evaluations performed on each answer created a base for the application of the evaluation measures on the system's operation [31], measures which are described next.

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 first place, it will receive a score of 1, it will be 1/2 if it is in the second place, 1/3 in the third place, and so on). 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.

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

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 have also used the classic measures of evaluation in IR: precision and recall [4]. *Precision* is understood as the capacity of a system to retrieve documents, or answers (in the case of QA systems), relevant to the query.

$$\text{Precision} = \frac{\left|\{\text{relevant answers}\}\cap\{\text{retrieved answers}\}\right|}{\{\text{retrieved answers}\}}$$

Recall is the fraction of the documents, or answers, that are relevant to the queries that are successfully retrieved. Recall is not calculated automatically because this measure depends on all the answers retrieved and the data 'lost' or not retrieved [32]. The method of polling (combining) of all the outputs for the same request is used to establish relative recall:

$$\text{Recall} = \frac{\left|\{\text{relevant answers}\}\cap\{\text{retrieved answers}\}\right|}{\{\text{relevant answers}\}}$$

The total number of relevant answers that would be analysed in this study is 750 answers per QA system, because we evaluated a maximum of five answers for each of the 150 questions. So an exhaustive and precise QA system will retrieve a total of 750 correct answers.

4. Results and discussion

After asking the 150 questions in the four QA systems, the first five answers from each system were analysed, as five was the average number of answers retrieved by most of the systems, and because users would, presumably, use these types of systems to obtain information quickly, paying most attention to the first answers. Nonetheless, some QA systems offered

Table 1. Measures for evaluating the quality of answers

	Total number of answer retrieved	Average answers retrieved	Total number of answer assessed*	Total correct answers	Percentage correct answers	Total incorrect answers	Percentage incorrect answers	Total inexact answers	Percentage inexact answers
HONqa	6635	44.23	726	343	47.24	325	44.76	58	7.99
QuALiM	441	3	406	166	40.88	180	44.33	60	14.78
MedQA	802	5.34	705	329	46.66	245	34.75	131	18.58
START	236	1.6	234	164	70.08	27	11.54	43	18.38

* Only the first five answers retrieved by QA systems in each question were assessed.

more answers for some questions, while others did not offer the five answers. The average number of answers retrieved by each QA system was very significant, because it is higher in HONqa (Table 1).

The open-domain systems retrieved fewer answers than the restricted-domain systems, with START in last place (with an average of 1.6 answers), followed closely by QuALim (with 3 answers). In the restricted-domain QA systems, the results increased substantially, especially in the case of HONqa (with 44.23 answers), while MedQA was slightly better than the open-domain systems (with an average of 5.34 answers for each question).

The correct answers were present to the greatest degree in START (70.08%). In the two restricted-domain QA systems, this average decreased – MedQA (46.66%) and HONqa (47.24%) – and QuALim was the most deficient, with 40.88% of answers being correct. Figure 2 shows the number of correct, inexact and incorrect answers from the four QA systems analysed.

In relation to the inexact answers, the QA system with the lowest average was HONqa (7.99%). The remaining QA systems offered similar averages (MedQA, 18.58%, START, 18.38% and QuALiM, 14.78%).

The incorrect answers in HONqa (44.76%) and QuALiM (44.33%) were similar figures. In MedQA, the value was somewhat lower (34.75%) although still high, and START (11.54%) was the system that presented the fewest inexact answers.

The value obtained by applying the MRR and FHS evaluation measures indicates that MedQA best ranks answers, as the first correct answer appears at the top of the list of results (Table 2). This is significant in that the system does not use any sort of rank algorithm to carry out this process but rather it always recurs to the same order of answers based on the knowledge source. FHS turns out to be a very relevant metric, as users often tend to focus on the first answer obtained, disregarding the rest.

The other restricted-domain QA system, HONqa, is in the opposite situation, as it offers the lowest FHS value but has the second highest average in MRR. The behaviour of the two open-domain QA systems is quite similar, and there is not a great deal of difference between the two metrics, which is explained by the fact that the answers retrieved for each question usually fluctuate from one to three.

However, the TRR metric is greater in the restricted-domain systems as it considers the value of each correct answer based on its position in the list of results. Naturally, this value increases when more results are offered.

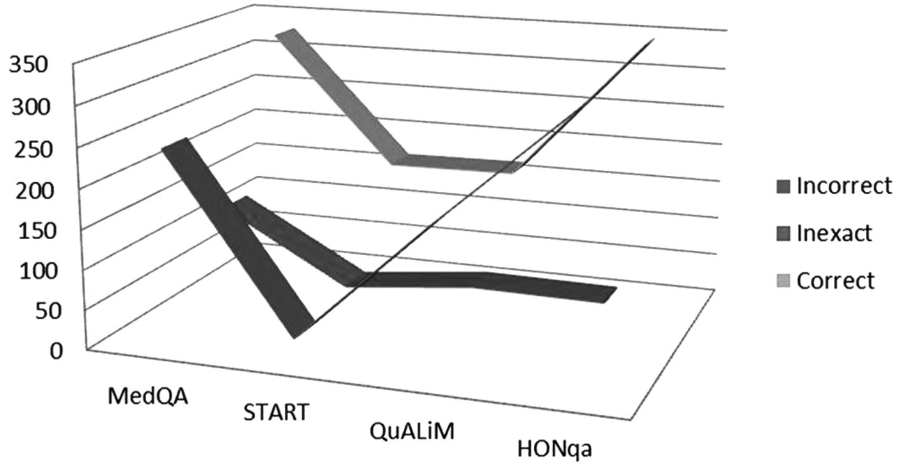

Figure 2. Incorrect, inexact and correct answer.

Table 2. Evaluation measures for QA systems

	MRR	TRR	FHS	Precision (%)	Recall (%)
HONqa	0.75	1.15	0.55	47.24	45.73
QuALiM	0.65	0.77	0.59	40.88	22.13
MedQA	0.87	1.29	0.76	46.66	43.87
START	0.67	0.81	0.64	70.08	21.87

FHS: First Hit Success; MRR: Mean Reciprocal Rank; TRR: Total Reciprocal Rank.

The precision value in the open-domain QA system START was higher than in the restricted-domain QA systems because the latter ones demonstrate a high rate of noise. However, the recall shows us that the two open-domain QA systems are less exhaustive than the restricted-domain ones. So START presents a recall value of 21.87% and an average of retrieved answers of 1.6, and the values of QuALiM are 22.13%, referring to the recall measure, and an average of three retrieved answers.

As can be observed, none of the applied metrics present very high values, a situation which has clearly been influenced by the high standards set for an answer to be evaluated as correct. In many cases, for example in MedQA and START, there were correct answers with more than 100 characters, which caused them to be considered inexact, as indicated in Section 3. Likewise, in HONqa, there were correct answers that were considered inexact because they did not respond to the question in a totally specific and precise way. So if we considered the inexact answers as correct to calculate the precision, the values would increase, in some cases considerably (Table 3).

Finally, the correlation between the metrics used in this study (Table 4) shows a significant correlation between TRR and MRR because they focus on similar characteristics of the answers retrieved by QA systems.

5. Conclusions

Today, users are more demanding of the retrieved information, as much with regard to the quality as with regard to the quantity, just as in response time. QA systems could therefore be one of the future systems of IR on the internet as they intend to meet the needs and demands of the current users. Nonetheless, despite an increase in QA system research, the open-domain systems are scarce and not all of them function adequately. In as much, it is necessary to continue analysing and evaluating such systems and their characteristics, in order to promote their development and commercialization.

Table 3. Precision considering correct and inexact answers

	HONqa	QuALiM	MedQA	START
Precision (%)	52.75	42.66	65.24	88.46

Table 4. Correlation of measures

		MRR	TRR	FHS	Precision	Recall
MRR	Pearson	1.000	0.959*	0.700	–0.265	0.837
	Significance (2-tailed)		0.041	0.300	0.735	0.163
TRR	Pearson	0.959*	1.000	0.472	–0.316	0.957*
	Significance (2-tailed)	0.041		0.528	0.684	0.043
FHS	Pearson	0.700	0.472	1.000	0.093	0.197
	Significance (2-tailed)	0.300	0.528		0.907	0.803
Precision	Pearson	–0.265	–0.316	0.093	1.000	–0.387
	Significance (2-tailed)	0.735	0.684	0.907		0.613
Recall	Pearson	0.837	0.957*	0.197	–0.387	1.000
	Significance (2-tailed)	0.163	0.043	0.803	0.613	

* The correlation is significant at 0.05.

Through analysing the results obtained by asking the 150 questions in the QA systems MedQA, START, QuALiM and HONqa, it was possible to evaluate the systems' operation through applying specific metrics. Despite the limitations demonstrated by these systems, as they are not accessible to everyone and they are not always completely developed, it has been confirmed that these four systems are valid and useful for obtaining definitional medical information in that they offer coherent and precise answers.

Another interesting aspect concerns the sources of information used by each of these QA systems. In previous work [15], we compared the sources used by the restricted-domain QA systems with those used by the general-domain systems to ascertain if we could see significant differences in the typology and specialization. A comparison of the sources used by START and QuALiM confirms the absence of significant differences, as both systems use Wikipedia as their main source, although START recurs to more sources for its retrieval. Nonetheless, we see big differences in the sources used by the two restricted-domain QA systems. While MedQA uses dictionaries, encyclopaedias and databases specializing in the biomedical field, HONqa opts for websites specializing in this area.

The results are encouraging because they present this type of tool as a new possibility for gathering precise, reliable and specific information in a short period of time. In this sense, some authors [33] have explored various possibilities for improvements, such as the use of ontologies, which will increase the quality of the answers obtained by formalizing the relevant information from the domain in question. In addition, these approaches together with others are slowly attracting more researchers who are experienced in handling the results they produce. These data suggest that we may see unexpected changes in the future and this area deserves to be studied and evaluated in future research.

QA systems have been extended in recent years to explore critical new scientific and practical dimensions [34]. Additional aspects, such as interactivity (often required for clarification of questions or answers), answer reuse, and knowledge representation and reasoning to support question answering, have been explored. Future research may explore what kinds of questions can be asked and answered about social media, including sentiment analysis. Another key aspect of these systems is that the system–user relationship is two-way. Establishing an interaction helps QA systems find better answers and, in turn, the QA system helps users find answers more quickly. However, it is still necessary to deepen the interactive design of these systems and enable true feedback between questions and answers so that users communicate with the system in a conversational manner. Finally, retrieval has shown that the QA systems are a useful tool to retrieve information quickly and accurately.

References

[1] R. Baeza-Yates and B. Ribeiro-Nieto, *Modern Information Retrieval* (ACM Press, New York, 1999).

[2] R. R. Korfhage, *Information Storage and Retrieval* (Wiley Computer Publishing, New York, 1997).

[3] G. Salton, Automatic processing of foreign language documents, *Journal of American Society for Information Sciences* 21 (1970) 187–194.

[4] G. Salton and M.J. Mc Gill, *Introduction to Modern Information Retrieval* (McGraw-Hill, New York, 1983).

[5] C.J. Rijsbergen, *Information Retrieval*, 2nd Edition (Butterworths, London, 1979).

[6] C.T. Meadow, *Text Information Retrieval Systems* (Academic Press, San Diego, 1993).

[7] M. Pérez-Coutiño, T. Solorio, M. Montes-y-Gómez, A. López-López and L. Villaseñor-Pineda, Toward a document model for question answering systems, *Proceedings of the Second International Atlantic Web Intelligence Conference* (2004)

[8] D. Crouch, R. Saurí and A. Fowler, *AQUAINT Pilot Knowledge-Based Evaluation: Annotation Guidelines* (Palo Alto Research Center, Palo Alto, 2005). Available at: www2.parc.com/isl/groups/nltt/papers/aquaint_kb_pilot_evaluation_guide.pdf (accessed 25 October 2010).

[9] M. Lee, J. Cimino, H.R. Zhu, C. Sable, V. Shanker, J. Ely and H. Yu, Beyond information retrieval – medical question answering, *AMIA 2006 Annual Symposium Proceedings* (2006) 469–473.

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

[11] E.M. Voorhees, The TREC 8 question answering track report, *Proceedings of the 8th Text REtrieval Conference* (1999) 107–130.

[12] S.B. Blair-Goldensohn, K.R. McKeow and A.H Schlaikjer, A hybrid approach for QA track definitional questions, *Proceedings of TREC 2003* (Gaithersburg, MD, 2003) 336–343.

[13] M.R. Kangavari, S. Ghandchi and M. Golpour, A new model for question answering systems, *World Academy of Science, Engineering and Technology* 42 (2008).

[14] L.F. Costa and D. Santos, *Question answering systems: a partial answer* (SINTEF, Oslo, 2007).

[15] M.D. Olvera-Lobo and J. Gutiérrez-Artacho, Question-answering systems as efficient sources of terminological information: evaluation, *Health Information and Library Journal* 27(4) (2010) 268–276.

[16] H. Cui, M.Y. Kan, T.S. Chua and J. Xiao, A comparative study on sentence retrieval for definitional question answering, *SIGIR Workshop on Information retrieval for Question Answering* (Sheffield, 2004).

[17] O. Tsur, *Definitional question-answering using trainable text classifiers* (Institute of Logic Language and Computation (ILLC), University of Amsterdam, 2003).

[18] G.O. Sing, C. Ardil, W. Wong and S. Sahib, Response quality evaluation in heterogeneous question answering system: a black-box approach, *Proceedings of World Academy of Science, Engineering and Technology*, 9 (Lisbon, 2005).

[19] E. Alfonseca, M. De Boni, J.L. Jara, S. Manandhar, A prototype question answering system using syntactic and semantic information for answer retrieval, *Proceedings of the 10th Text Retrieval Conference (TREC-10)* (Gaithersburg, MD, 2002)

[20] P. Jacquemart and P. Zweigenbaum, Towards a medical question-answering system: a feasibility study, *Proceedings of Medical Informatics Europe (MIE '03), Volume 95 of Studies in Health Technology and Informatics* (IOS Press, San Palo, CA, 2003) 463–468.

[21] P. Zweigenbaum, Question answering in biomedicine, *Proceedings Workshop on Natural Language Processing for Question Answering, EACL 2003* (2005) 1–4.

[22] B. Katz, Using English for indexing and retrieving. In: P.H. Winston and S.A. Shellard (eds), *Artificial intelligence at MIT: expanding frontiers*, volume 1 (MIT Press, Cambridge, MA, 1990).

[23] M. Kaisser, *Acquiring syntactic and semantic transformations in question answering,* (PhD thesis, University of Edinburgh, Edinburgh, 2009).

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

[25] I. Fahmi. *Automatic term and relation extraction for medical question answering system* (PhD thesis, University of Gronigen, Gronigen, Netherlands, 2009).

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

[27] S. Cruchet, A. Gaudinat, T. Rindflesch and C. Boyer, What about trust in the question answering world? *AMIA 2009 Annual Symposium* (San Francisco, CA, 2009).

[28] S.B. Blair-Goldensohn, K.R. McKeow and A.H Schlaikjer, A hybrid approach for QA track definitional questions, *Proceedings of TREC 2003* (Gaithersburg, MD, 2003) 336–343.

[29] H. Yu and D. Kaufman, A cognitive evaluation of four online search engines for answering definitional questions posed by physicians, *Pacific Symposium on Biocomputing* 12 (2007) 328–339.

[30] B. Katz, S. Felshin, D. Yuret, A. Ibrahim, J. Lin, G. Martion, A.J. McFarland and B. Temelkuran, Omnibase: uniform access to heterogeneous data for question answering, *Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems* (Estocolmo, NLDB, 2002) 230−234.

[31] J.R. Taylor, *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*, (University Science Books, 1999), 128–129.

[32] T. Saracevic, Evaluation of evaluation in information retrieval, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Special issue of SIGIR Forum* (1995) 138–146.

[33] P. Buitelaar, P. Cimiano, P. Frank, M. Hartung and S. Racioppa, Ontology-based information extraction and integration from heterogeneous data sources, *International Journal of Human–Computer Studies* 66 (2008) 759−788.

[34] M.T. Maybury, *New Directions in Question Answering* (AAAI/MIT Press, Cambridge, MA, 2004).

Appendix

1. What is abortion?
2. What is acne?
3. What is acupuncture?
4. What is ADHD?
5. What is allergic asthma?
6. What is alopecia aerate?
7. What is altitude sickness?
8. What is amebiasis?
9. What is anesthesia?
10. What is ankle sprain?
11. What is anthrax?
12. What is aortic aneurysm?
13. What is aortic valve regurgitation?
14. What is aromatherapy?
15. What is asthma?
16. What is atherosclerosis?
17. What is ayurveda?

18. What is back pain?
19. What is binge eating disorder?
20. What is bipolar disorder?
21. What is borderline personality disorder?
22. What is bradycardia?
23. What is bulimia nervosa?
24. What is campylobacteriosis?
25. What is capsaicin?
26. What is carbon monoxide poisoning?
27. What is carpal tunnel syndrome?
28. What is cataract?
29. What is central venous catheter?
30. What is chelation therapy?
31. What is chemotherapy?
32. What is chickenpox?
33. What is Chinese medicine?
34. What is chiropractic?
35. What is chlamydia?
36. What is chronic pain?
37. What is circumcision?
38. What is cochlear implant?
39. What is colorectal cancer?
40. What is COPD?
41. What is Crohn’s disease?
42. What is cystic fibrosis?
43. What is deep vein thrombosis?
44. What is degenerative disc disease?
45. What is depression?
46. What is diabetic ketoacidosis?
47. What is diabetic retinopathy?
48. What is disseminated intravascular coagulation?
49. What is diverticulosis?
50. What is dry mouth?
51. What is dyslexia?
52. What is endometriosis?
53. What is ephedra?
54. What is epididymitis?
55. What is epilepsy?
56. What is esophageal spasm?
57. What is fibromyalgia?
58. What is flatfoot?
59. What is flu?
60. What is flu mist?
61. What is gastric bypass surgery?
62. What is gastroenteritis?
63. What is gastro esophageal reflux?
64. What is gastro paresis?
65. What is genital herpes?
66. What is gestational diabetes?
67. What is gingivitis?
68. What is gonorrhea?
69. What is Guillain-Barre syndrome?
70. What is gynecomastia?
71. What is healing touch?
72. What is heart attack?

73. What is heart failure?
74. What is hemophilia?
75. What is hepatitis C?
76. What is herniated disc?
77. What is high blood pressure?
78. What is homeopathy?
79. What is Huntington's disease?
80. What is hydrocele?
81. What is hydrotherapy?
82. What is hypospadias?
83. What is hypothermia?
84. What is hypothyroidism?
85. What is inflammatory bowel disease?
86. What is insomnia?
87. What is irritable bowel syndrome?
88. What is kava?
89. What is kernicterus?
90. What is lactose intolerance?
91. What is laryngitis?
92. What is latex allergy?
93. What is laxative?
94. What is lipoma?
95. What is listeriosis?
96. What is lung cancer?
97. What is melanoma?
98. What is metabolic syndrome?
99. What is naturopathic medicine?
100. What is oophorectomy?
101. What is orthodontics?
102. What is osteoporosis?
103. What is overactive bladder?
104. What is patellofemoral pain syndrome?
105. What is perimenopause?
106. What is pheochromocytoma?
107. What is pityriasis rosea?
108. What is pleurisy?
109. What is pneumothorax?
110. What is prediabetes?
111. What is presbyopia?
112. What is primary biliary cirrhosis?
113. What is prostate cancer?
114. What is radiation therapy?
115. What is retinal detachment?
116. What is rheumatoid arthritis?
117. What is root canal treatment?
118. What is rosacea?
119. What is roseola?
120. What is rotavirus?
121. What is salmonellosis?
122. What is SARS?
123. What is saw palmetto?
124. What is scabies?
125. What is scarlet fever?
126. What is schizophrenia?
127. What is sciatica?

128. What is scoliosis?
129. What is shigellosis?
130. What is sickle cell disease?
131. What is sinusitis?
132. What is sleep apnea?
133. What is smallpox?
134. What is spermatocyte?
135. What is spondylolisthesis?
136. What is stem cell transplant?
137. What is tartar?
138. What is TENS?
139. What is testicular cancer?
140. What is TMJ?
141. What is tonsillitis?
142. What is Tourette syndrome?
143. What is transient ischemic attack?
144. What is trichomoniasis?
145. What is tuberculosis?
146. What is vaginitis?
147. What is ventricular tachycardia?
148. What is vesicoureteral reflux?
149. What is virtual colonoscopy?
150. What is whiplash?