What about trust in the Question Answering world?

Journal:	*AMIA 2009 Annual Symposium*
Manuscript ID:	AMIA-0578-A2009
Manuscript Type:	Paper
Date Submitted by the Author:	13-Mar-2009
Complete List of Authors:	Cruchet, Sarah; Health On the Net foundation gaudinat, arnaud; Health On the Net, SIM; Health On the Net foundation Rindflesch, Thomas; Nationa Library of Medicine, LHC Boyer, Celia; Health On the Net foundation

What about trust in the Question Answering world?

Sarah Cruchet[1], Arnaud Gaudinat[1], Tom Rindflesch[2,] Célia Boyer[1],
[1]Health On the Net Foundation, Geneva, [2] Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda

Abstract

Objectives: Current search engines are based on keyword queries, which are not natural for the user Further, the user often has to dig within each page to find the answer. It is also difficult for the common user to judge the quality of a document retrieved on the Internet? This paper presents a solution using an existing question answering system and investigates whether the quality of the answers extracted depends on the quality of the health web pages analysed.

Design: This paper presents two separate evaluations of the QA system developed by HON. The first evaluation is systematic and validates the effectiveness of our system. The second focuses on the use of QA for site selection of quality health care, (those with HONcode) or any site on the Web.

Measurements: Results are merged and classify by a medical expert (in A =relevant, B = lead to an answer, C = not relevant). We use Trec eval for the evaluation of our system.

Results: Systematic evaluation: For a set of 100 questions, 70 are well answered.

Qualitative evaluation: For a set of 100 questions, we obtain a MAP of 59% and a MRR of 76% for QAHON_honcode.

Conclusion: According to our results the trustworthiness of the database used influence the quality and accuracy of the answers retrieved by the HON question answering system.

Introduction

The number of documents for a given request is constantly increasing but the search for information for the lay person has not changed at all! The user is always challenged in terms of relevance and quality of results and he/she must read the entire document hoping to reach a satisfactory answer. In addition, asking a question using keywords is not natural [1, 2] and can produce inaccurate results. While some rely on the advent of the Semantic Web to generate new intelligent search engines, others hope to facilitate the interaction with search engines by providing question answering (QA) systems. These are based mainly on heuristics and natural language processing (NLP) algorithms and the use of redundancy of information. The goal of such systems is to provide relevant answers to a question in natural language. Thus the user directly accesses the information required and does not have to go through all the documents proposed by the search engine. In this age of mobility, short and precise answers with, for example, a voice interface would be beneficial for this type of system.

Furthermore, campaigns such as TREC evaluation [3] have demonstrated the feasibility and interest of QA. The best of systems evaluation reached nearly 71% of accuracy in 2005 [4]. Regarding the medical field, interest is growing as shown by the number of papers in this field [5-8]. A QA system applied to the medical field is now available online. This is MedQA [9], a system which addresses issues of type definition and is intended for biologists, to facilitate access to the biomedical literature.

Another issues is that the creation of a website is within the reach of everyone, hence the variability of information found online. In the case of health information, there is a direct impact on the well being of individuals [10]. Thus, initiatives such as the HONcode [11-13] (code of good practice for health websites with 8 ethical principles) and vertical search tools try to offer complementary services and alternatives to popular search engines. The HONcode has been developed with the consensus of site managers by the Health On the Net Foundation (HON), a non-profit organization aimed at promoting quality and trustworthy health information. Current studies conducted on QA are interested only in the relevance of answers, not on its quality. 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.

One of our hypothesis is that quality of an answer is strongly related to its relevancy. This paper presents two separate evaluations of *QA*, as system developed by the HON. The first is a systematic evaluation to validate the effectiveness of the system presented. The second focuses on the use of *QA*; on the one hand, it uses a selection of quality health websites (those of HONcode) and on the other, any other website on the Internet. For each assessment, materials and methods will be presented.

We will then present the results and discuss a few ways to better integrate the qualitative dimension in the *QA* systems. Finally, we conclude and present future prospects.

Materials and Methods

We use the multilingual *QA* system of HON presented in [14]. In this section, we summarize the architecture of the system. It is based on a detection module of the question type (with a learning approach) and using semantic resources such as those present in UMLS to guide the system, particularly in the choice of answers. *QA* has a relatively conventional architecture for a QA system but with specific characteristics for each module: 1 / QuestionAnalyzer is designed to identify the question of the user to better anticipate the search for information and selection of the response. 2 / The QueryGenerator can generate one or more applications based on various search engines and the issue discussed earlier. 3 / DocumentRetrieval module handles different query search engines. 4 / PassageExtractor The module identifies the best passages in the light of the question and the answer expected. 5 / AnswerExtractor module deals with selecting the best answers in the passages according to the question, the expected response and redundancy of some answers. 6 / A display module allows the user to highlight the different responses found in a discrete manner, with the opportunity to draw the best information in context. Note that the original module was enhanced recently with the aim of achieving a comprehensive analysis of research results (i.e. inserted between the DocumentRetriever and PassageRetriever) to further assist in the selection of passages and answers. This module is based on an Ngram analysis of words in all the documents returned, filtered by the semantic types expected.

Systematic evaluation

For the first evaluation of our system, we have a corpus of 377 questions in English as well as 12 sites from which they originate. These questions were collected manually on Internet discussion forums and Frequently Asked Questions (FAQ) dealing with the medical field. This collection was originally made for a task of learning in the context of identifying the type of question in the QuestionAnalyzer module of our *QA* system [14]. As the medical field is very large, we have focused on issues related to disease. Indeed, according to [15], 64% of the health topics searched online are related to disease. In addition, we have chosen a particular disease, to avoid learning a classification system based on disease rather than the type of questions. The disease chosen is diabetes as it

was used in the PIPS European project of the six framework program (IST-2002-2.3.1.11) [16], which motivated the initial development of our QA system. However, in this study it was not only the question which interested us but also the response. Thus, a subset of 100 questions with their responses were manually selected for these 12 sites, to create a corpus of reference. The 12 sites were downloaded in 2008, corresponding to 633 different pages, and were indexed locally to overcome variations due to the inherent dynamism of the Web. Here are 3 sample questions and their answers:
Q1: Who is at greatest risk of developing type 2 diabetes?
Url1: http://www.nutritionaustralia.org/Food_Facts/FAQ/diabetes_detailfaq.html
A1: Until recently, type 2 diabetes was considered as the 'adult onset' form - that is, it was rarely seen other than in middle-aged and older people. However, it is now affecting younger people as well.
Q2: What is metabolic syndrome?
Url2: http://www.idf.org/home/index0689.html
A2: The metabolic syndrome is a cluster of the most dangerous heart attack risk factors: Prediabetes and diabetes, abdominal obesity, changes in cholesterol and high blood pressure.
Q3: What are the different types of diabetes?
Url3: http://diabeticgourmet.com/faq/entry/22/34/index.html
A3: The three main types of diabetes are: Type 1 Diabetes, Type 2 Diabetes, Gestational Diabetes.

Qualitative evaluation

Our study is based on the methodology of a study conducted at the U.S. National Library of Medicine (NLM) on information retrieval systems [5]. It is based on the comparison of 2 versions of QA developed by HON. The first system, QAHON_honcode is based on research using only the database of 6800 HONcode certified sites. The second QAHON_google, is the use of Google results through our QA system. The QA system is similar to the 2 systems compared; however, the research, information, and resources are different. As part of a QA system, it is much more interesting to assess the responses returned by the system rather than the documents.

For the second evaluation we used the method described in [5] and used a part of our body of the first issue of evaluation. So for each system only the first 10 responses were returned. These responses are mixed and "anonymous", the information obtained

from the database is submitted to the judgment of an expert with the of grading the answers using letters: A + (very relevant) , A (relevant), A-(not the whole answer), B + (leading to response), B (may lead to the answer), B-(unclear), C (not relevant). The rating reflects the adequacy of the response to the question and the relevance of the document from which the answer was extracted.

We used Trec-eval to evaluate our results. Six measures were taken into account to evaluate the system:

1. Mean Average Precision (MAP) : it computes the average precision after each relevant answer extracted
2. Binary Preference (Bpref): it considers non relevant answers printed before the relevant ones. It determines the rank of relevant documents.
3. R precision (R-pre): it computes precision after R answers extracted.
4. Mean Reciprocal Rank (MRR): it is the reciprocal of the rank at which the first relevant document was found.
5. Precision at five documents (P@5): rate of relevant documents in the top five answers.
6. Precision at ten documents (P@10): rate of relevant documents in the top ten answers.

We consider that an answer is relevant if it gets an A or a B, as do Sneiderman et al. [5].

Results

Systematic evaluation

We searched the system's ability to find the answers collected manually in the 12 sites from which were produced the questions of the test base, or a corpus of   633   pages. By a simple text pattern matching, we considered that the system provided a correct answer when a system's response was contained in the manual selection of the responses. The system gets a good response rate of 70%. This result is positive, although unfortunately, it is not an indicator for precision, because the system responses that are not part of the body of reference are not recognized when they may be right. In addition, the QA system found on average, 3.2 responses per question and a maximum of only 2 responses were collected manually during the formation of the collection of reference.

Qualitative evaluation

System	QAHON_honcode	QAHON_google
MAP	**0.59**	0.36
Bpref	**0.50**	0.34
R-pre	**0.59**	0.38
MRR	0.76	**0.86**
P@5	**0.54**	0.36
P@10	**0.32**	0.22

Table 1. Results obtained with the soft evaluation

The first observation that can be made is that the results are better for QAHON_honcode than for QAHON_google. Indeed, QAHON_honcode obtained a MAP of 0.59 against 0.36 for QAHON_google. This means that of all answers given by QAHON_honcode to a question, 59% are relevant against 36% for QAHON_google. In addition, out of the first 5 answers found by QAHON_honcode, more than half correspond exactly to the question. The only measure where QAHON_google performs better is MRR where the first reply has a better average rank than for QAHON_honcode
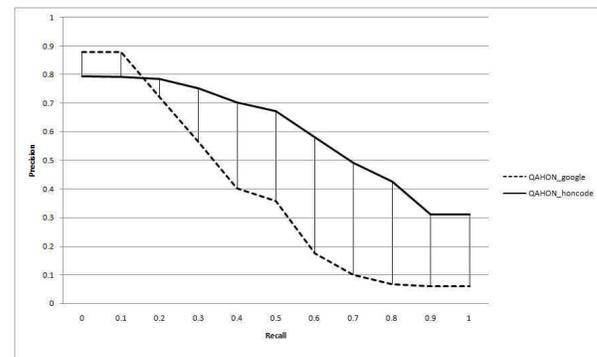

Figure 1. Precision/Recall at 11 points (dotted line=QAHON_goole)

Figure 1 shows the precision / recall for the 2 systems. We note that QAHON_google is better in the first 2 readings, following which QAHON_honcode is consistently, the better system. This means that for all the replies given by the system, QAHON_honcode is often more relevant than GoogleQA.

Discussion

Although the first evaluation is limited in its approach, it allowed us to validate our QA system

and it will be very useful for us to continue with fine tuning. Indeed, there is much room for improvement and interpretation of the missing results will improve the base system.

In relation to the impact of taking into account quality, our results show a significant difference in results when using the Web in its entirety as opposed to a selection of quality sites and those with a similar system of QA. There is a better response of on average 23% (difference of MAP between the two systems for a QA system based on high quality sites). Even if Google is a general search engine and that HONcodehunt is a health one, we consider that the differences of relevancy are not related to this fact. Indeed, this study uses question instead of query (usually simple keyword(s)), and so the questions are specific enough to be clearly related to the health domain.

Of course, in this study only the subjective relevance of the evaluator is measured, while it is difficult to separate the semantic relevance of the quality of the response.

Since we use a method of assessment [5], we can compare some results keeping in mind that the evaluation focuses on the documents found while ours focuses on the answers. Thus it is interesting to note that our system outperforms them for the first three types of measures (MAP, Bpref and R-prec) following which, it is less for the other 3 measures (MPR, P @ 5 and P @ 10). In our opinion, the first 3 steps are more representative of a system that gives a good overview of the question posed. In our case, where users are more often lay people, this may be more relevant.

Use of reliable resources

In this article we presented the use of resources of different quality in the context of a QA system. The assumption is that documents selected according to editorial standards or quality favor the quality of the QA system responses (like a system of classical information retrieval on the Web). In addition, following the method of selection, editorial policy, third-party certification, selection by professionals or collaborative rating poses the question of weighting the different resources and its combination with the semantic relevance score.

Exceptional Use, for example the forums

Using the Web as a resource is more subject to noise than the use of MEDLINE, for example. The existence of forums or collaborative QA systems illustrates this problem, often because it can reveal areas where the drafting of responses is unknown. As part of the Web, it is essential to recognize the forums as such to enable filtering.

Using the date as an index of quality of information

The date is an important criterion for the quality of information to be included in a system of question responses. A QA system of quality should take into account this dimension in its strategy of filtering and ranking or at least give the date of creation for each response submitted to the end user. Indeed, a statement may be true at a given time though may not be so in a few months or years later.

Using the redundancy of information

Redundancy of information is one of the main strategies of current QA systems to identify relevant answers. From the viewpoint of the quality of information, the assumption may be the same, since the same information repeated on several separate sites is certainly more likely to be credible

Crossing responses in the event of non-consensus

In addition to the previous section, it must be emphasized that the plurality of responses is equally important (as long as the redundancy condition is met). To have several different answers is valid in both cases, when the matter concerns a list of items (e.g. what are the risk factors of lung cancer), or when there is no scientific consensus on the issue.

Contradiction or Justification with professional resources

The use of databases such as MEDLINE can be complementary to check some assertions found on the Web and increase the quality of the QA system.

The QA modules to be modified, to take into account, the quality of information

The document retriever certainly is the module most relevant for taking into account the quality of information. This is where resources are indicated and can be weighted. Filters on forums can also be added during indexing to avoid getting responses with poor authority. The management of dates is, on the other , more difficult to implement. Indeed, if databases such as MEDLINE can be searched with a date, the Web may remain insensitive to the search with a time dimension. The result of the search with Google by specifying time intervals is the most blatant example in recent years. This is particularly due to the lack of rigor and diversity statements to update pages by the Webmaster. Principle 4 of the HONcode guarantees that these dates are accurate on these websites, even if their expression is left up to the editor. There was a similar trend in the use of CMS, which allows an automatic update of date with

the possibility of adding this information as meta-information.

Semantic Web and the quality of information

In the introduction, we mentioned the promise of the Semantic Web to offer enormous possibilities as well as accurate information retrieval (as in the case QA). But again one must be careful, because those who produce or collect that information will have goals that may be very different from pure philanthropy. Indeed the information from the Semantic Web will certainly be much less readable in terms of traceability, and only third party safeguards can guarantee a certain neutrality or at least the ability to obtain independent details on this information.

Health Literacy

The use of a QA system based solely on MEDLINE always has the advantage of providing high quality scientific methodology. However one can doubt the value of this single system in so far as it might be difficult for the average user to interpret these responses. So, considering results from general sites and results which are too scientific, websites which are HONcode certified appear as a fair alternative with a pre-calculated indicator of health literacy levels.

Conclusion

The solutions proposed in this paper to take into account the variability of the quality of information on the medical web are valid in both the classical information retrieval in the context of use of question and question answering systems. In the case of safe use of resources such as MEDLINE in a QA, using methods discussed in this paper is less critical. However, the use of meta-information such as taking into account the date is crucial.

Redundancy of information, although a crude indicator, can be a very good indicator of relevance if one ensures that the documents used are reliable and independent. On the other hand a good question answering system must also highlight the diversity of responses.

In the interface, the user should at all times have access to the answer in its context for a better traceability of the response.

Acknowledgements

References

1. C. Kwok, & AL, Scaling question answering to the Web, In Proceedings of WWW'10, *2001*
2. D. ROUSSINOV, How Question Answering Technology Helps to Locate Malevolent Online Content , Intelligence and Security Informatics volume 3495/2005, 2005.
3. http://trec.nist.gov/
4. E. M. Voorhees and Hoa T. Dang. 2006. Overview of the TREC 2005 question answering track. In Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005).
5. C.A. SNEIDERMAN & AL Knowledge-Based Methods to Help Clinicians Find Answers in MEDLINE, JAMIA 2007
6. D. DEMMER-FUSMAN & AL, Answering Clinical Questions with Knowledge-Based and Statistical Techniques, Computational Linguistics 2007
7. D. DEMMER-FUSMAN & AL, Combining resources to find answers to biomedical questions, TREC 2007
8. J. Lin & AL, Semantic Clustering of Answers to Clinical Questions, AMIA 2007
9. Y. HONG & AL, Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians, Journal of Biomedical Informatics, 2007
10. E. HORVITZ, Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search, Microsoft study, 2008
11. M. SELBY & AL, Health On the Net Foundation Code of Conduct for Medical and Health Websites. MedNet 96 - European Congress on the Internet in Medicine, Brighton, U.K., Oct. 14 to 17, 1996
12. C. BOYER & AL, Health On the Net foundation: assessing the quality of health web pages all over the world, MedInfo, 2007
13. http://www.hon.ch/HONcode/Conduct.html
14. S. CRUCHET & Supervised approach to recognize question type in a QA system for health, MIE, 2008
15. S.FOX, http://www.pewinternet.org/pdfs/PIP_Online_Health_2006.pdf, 2006
16. http://www.pips.eu.org/