What about trust in a question answering system?

Sarah Cruchet[a], Arnaud Gaudinat[a], Célia Boyer[a], Thomas Rindflesh[b]

[a] *Health On the Net Foundation, Geneva, Switzerland*
[b] *National Library of Medicine, LHC, Washington, USA*

Abstract and Objective

A question answering system (QA) aims to answer a question asked in a natural form. This paper presents a solution using an existing medical QA developed by HON and investigates whether the relevance and reliability of the answers extracted conforms to the standards of the quality and trustworthy health web pages analyzed. The evaluation focuses on the comparison of the results between the QA searches through trustworthy health documents and the Google database (which include certified and non-certified websites). Results are merged and classified by a medical expert. We use Trec-eval measures for the evaluation. As a result, for a set of 100 questions, we obtain a MAP of 59% and a MRR of 76% for QAHON_honcode. According to our results, the trustworthiness of the database used influence the relevance and accuracy of the answers retrieved by the HON QA.

Keywords:

Trust, Reliability, Information retrieval.

Methods

Classical search engines submerge the user with thousand of links to documents which perhaps contain the answer to their question. Furthermore, the user is unsure of the quality of results. A question answering system (QA) aims to provide directly an answer to a question asked in a natural form. The HON foundation has dedicated itself to the promotion and improvement of the quality of online medical information through its ethical and quality code of conduct, the HONcode[1]. That's why HON became involved in the development of a QA applied to health and where answers are extracted only from HONcode certified Websites. Our study is based on the comparison of the results of the QA developed by HON when applied to two different databases (DB) of online health documents: those of the HONcode certified websites (QAHON_honcode) and Google (QAHON_google). It is important to notice that for both DBs, 32 documents are retrieved and then analyzed by the QA. As the QA is common to the 2 approaches, we compare the quality of results retrieved using one DB of web pages respecting HONcode standards and the Google one with no filter except Google's "Page Rank". For this evaluation we used the method described by Sneiderman C.A. & AL [1]. The first 10 responses were considered and have been "anonymised". Then the answers obtained from the two approaches were graded by a health professional by using the following scale: A + (very relevant and reliable), A (relevant and reliable), 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. We used Trec-eval to evaluate our results. Six measures were taken into account to evaluate the system (MAP=Mean Average Precision; BPREF=non relevant document printed before relevant, compute precision after R answers extracted; MRR=reciprocal rank of the first relevant document, P@x=rate of relevant document in the top x answers). We consider that an answer is relevant if it gets an A or a B.

Results

Results are globally better for QAHON_honcode than for QAHON_google: 59% of answers coming from the HONcode database are relevant against 36% for Google. In addition, out of the first 5 answers found by QAHON_honcode, more than half answer exactly the question with appropriate answer. Another measure shows that the first answer using the Google DB has a better average rank than the one from the HONcode DB. The first results are the most important because the internet users will not read all answers given by the system. Nevertheless, it should be noted that the Google results include non-HONcode certified websites as well as HONcode certified websites. In relation to the impact of taking into account the trustworthiness of documents, our research shows a significant difference in results when using the whole Web (Google DB) as opposed to using a selection of trustworthy sites (HONcode DB) with the same QA. The quality of responses is 23% better with QAHON_honcode. Even if Google is a general search engine and HONcode is a health one, we believe that the differences in relevance are not related to this fact because 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: the response can syntactically answering the question and being false in term of quality.

References

[1] Sneiderman C.A. & AL Knowledge-Based Methods to Help Clinicians Find Answers in MEDLINE, JAMIA 2007.

[1] Ethical code of Conduct (www.hon.ch/HONcode/Conduct.html)