Trustworthiness and relevance in Web-based clinical question answering

Sarah Cruchet, Célia Boyer, Lonneke van der Plas
Health on the Net Foundation
Geneva, Switzerland

Abstract. Question answering systems try to give precise answers to a user's question posed in natural language. It is of utmost importance that the answers returned are relevant to the user's question. For clinical QA, the trustworthiness of answers is another important issue. Limiting the document collection to certified websites helps to improve the trustworthiness of answers. On the other hand, limited document collections are known to harm the relevancy of answers. We show, however, in a comparative evaluation, that promoting trustworthiness has no negative effect on the relevance of the retrieved answers in our clinical QA system. On the contrary, the answers found are in general more relevant.

Keywords. Web-based clinical question answering (cQA), HONcode, Search engine

Introduction

With the constant growth of information on the Internet we are all confronted with the problem of how to find answers to the questions we have. Question answering (QA) is an application of natural language processing (NLP) that tries to fulfill that need and has been receiving a lot of attention since the late 90s with evaluation campaigns such as TREC [1].

Still, the most widely used medium to search for information on the Web to date are search engines, such as Google. Question answering systems differ from search engines in two ways: Firstly, question answering systems are designed to deal with full-fledged questions in natural language and not just keywords. Secondly, they return a precise answer to the user’s question and not a list of relevant documents. We can easily see that these two characteristics make QA systems stand out in terms of user-friendliness, compared to search engines. There is no need for the user to translate an information need into keywords. They can just pose the question in their natural language. The fact that users receive precise answers and not long lists of documents is particularly useful with the recent growth in use of mobile phones to access the Internet, when, due to small monitors, scrolling down a long list of documents is particularly inconvenient.

The recent extraordinary performance of Watson, the QA system developed by IBM, in the TV quiz Jeopardy! put QA into the spotlights. The QA system was able to beat the biggest all-time money winner on Jeopardy!. This success has also had an
impact on the awareness of the use of question answering for the medical domain. In AMIA 2011 several papers discussed clinical QA (cQA), [2] [3] among others, and there was a panel discussion.

The relevance and trustworthiness of the answers returned is of utmost importance in QA systems, and the latter especially for clinical QA. As mentioned above, QA systems return answers and not documents. This gives the user more efficient information seeking abilities, but it also removes some context. Context, that could be used to check the source of the answer, its relevance and its trustworthiness. In other words, if we want to help the user with precise answers to their questions, we better make sure the answers are relevant and trustworthy.

Performance of QA systems is generally evaluated in terms of the relevance of returned answers. In this paper, we would like to introduce an aspect that is particularly important for the medical domain: the trustworthiness of the returned answers.

The World Wide Web is a great source of information. Whereas most QA systems were previously constrained to use closed document collections to find answers to the user's questions, current QA systems are given access to the Web. [4] have compared the performance of question answering systems when varying the size of text collections and found that QA systems benefit strongly from having access to larger document collections. Even more so because larger document collections contain redundancy and this redundancy can be brought to the benefit of QA systems [4].

However, when looking for health information, it is questionable whether giving the user access to unlimited sources of information on the Internet is desirable. The Internet is a valuable source of information but also open to abuse. Moreover, in the medical domain, the retrieved information has been found to have a direct impact on the well-being of individuals [5].

To guide the user in finding trustworthy health information, initiatives such as the HONcode [6-8] offer certification of health web sites with a code of good practice that includes 8 ethical principles. Health websites certified by the HON Foundation are deemed to be transparent with respect to financial sources, give clear references to the sources of the information they provide and adjust the content to the audience targeted, among others. The certification process functions as a strong separator for health websites. In a comparative evaluation, [9] showed that of all health websites that are not in the process of certification by HON only 0.6 % are in compliance with the HON principles. At the time we ran the QA experiments the 6800 sites were HON certified.

HON has developed the first clinical QA system that aims at providing users with trustworthy answers by obtaining answers exclusively from the websites certified by the foundation. Although promoting trustworthiness is important for cQA, it is clear from the discussion above that limiting the document collection might hurt the system's performance in terms of relevancy. The 6800 certified websites are a very small subset of the websites on the Web and QA systems are known to benefit from the redundancy found in large document collections [4]. We might reason that the HON certified websites are of better quality and are therefore more relevant. Although the HON certification process, that aims to promote trustworthiness of websites, might improve the quality of websites, this is not a direct result of the appliance of the ethical principles. And even if so, smaller amounts of high-quality information are not a guarantee to find relevant answers to a broad range of questions.

In this paper, we try to answer the following question: Does limiting the document collection based on trustworthiness hurt the performance of a clinical QA system in terms of relevancy?
To answer this question, we ran a comparative evaluation where the same QA system is given access to either the whole World Wide Web or the HONcode-certified websites only. To preview results, we show that promoting trustworthiness has no negative effect on the relevance of the retrieved answers. On the contrary, the answers found are in general more relevant.

Method

We use the English version of HON's multilingual QA system presented in [10]. The overall performance of this system has been discussed in an independent evaluation [11]. It outperformed all QA systems under evaluation that have a reasonable coverage, in terms of precision at the first 5 answers. In terms of MRR, the HON QA system came second out of the four systems evaluated. Although the overall architecture is rather conventional, the components are heavily adapted to the medical domain, as we will discuss below:

1) The question analysis module (described in detail in [10]) is designed to identify the medical topic of the question and the expected answer type. It uses semantic resources such as the UMLS to extract medical terms and semantic types. The classifier is trained using machine learning.

2) The query generation module turns the user’s question, formulated in natural language, into a query that can be processed by the search engine. In addition, it expands the query with synonyms for medical terms that it gathers from the UMLS lexical resources. For example, if a user asks “What is diabetes?”, the query generation module will create a query that contains “diabetes” and a second query that contains the synonym “diabetes mellitus”.

3) The document retrieval module collects the documents that the search engine found for the several queries. Before passing these documents on to the next module, the documents are analyzed semantically in order to extract medical terms and semantic types, that will be used in the answer extraction module later on.

4) The passage extraction module finds the most suitable passages according to the medical topic of the question and the expected answer type, as determined by the first module.

5) The answer extraction module deals with selecting the best answers in the passages according to the question, the expected answer type and the redundancy among the candidate answers.

The evaluation framework is based on a study conducted at the National Library of Medicine (NLM), Bethesda, USA on information retrieval systems [13]. We compare two systems. The first, Certified_only, is the system whose document collection is limited to the database of 6800 HONcode certified sites. The second, Unrestricted, is the same QA system, but it now has access to the World Wide Web.

1 We would like to stress that we opted for the HONcode-certified websites and not the scientific articles provided in MEDLINE, for example, for a reason. Such a MEDLINE QA system would have the advantage of providing high quality medical information. However, the content is hard to read for the average user. Between the results from an average site on the World Wide Web and the scientific articles from MEDLINE, websites that are HONcode-certified seem a fair alternative.
For each system only the first 10 answers for each question are taken into account. They are given to a medical expert with the task to grade the answer with respect to the level of relevancy\(^2\): A+ (very relevant), A (relevant), A- (not the whole answer), B+ (leading to answer), B (may lead to the answer), B- (unclear), C (not relevant).

We used Trec-eval [12] to evaluate our results. Six standard metrics 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 returned before the relevant ones. It determines the rank of relevant documents.
3. R precision (R-pre): it computes precision after \(R\) answers extracted, where \(R\) is the number of relevant answers for the topic.
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 in [13].

**Results**

Table 1. Results for the QA system having access to certified websites only compared to the system having access to the World Wide Web for a large number of standard metrics.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Certified only</th>
<th>Unrestricted</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.59</td>
<td>0.36</td>
</tr>
<tr>
<td>Bpref</td>
<td>0.50</td>
<td>0.34</td>
</tr>
<tr>
<td>R-pre</td>
<td>0.59</td>
<td>0.38</td>
</tr>
<tr>
<td>MRR</td>
<td>0.76</td>
<td>0.86</td>
</tr>
<tr>
<td>P@5</td>
<td>0.54</td>
<td>0.36</td>
</tr>
<tr>
<td>P@10</td>
<td>0.32</td>
<td>0.22</td>
</tr>
</tbody>
</table>

In Table 1, we see the performance of the two systems with respect to the six standard metrics. The mean average precision is much higher for the restricted system, and, as we see in row 6, out of the first 5 answers found by the restricted system, more than half are relevant against only 36\% for the unrestricted system. The only measure where the unrestricted system performs better is MRR: the first relevant answer is ranked higher on average. However, this metric has been often criticised, because it is agnostic to the number of correct answers found overall. It takes only the rank of the first relevant answer into account.

---

\(^2\) The answers are scrambled, so that the expert cannot tell what system produced the answers.
Discussion

Trustworthiness of answers returned to the user is an important issue in clinical QA. Limiting the document collection to HONcode-certified health websites, thereby discarding the remainder of the World Wide Web, provides a way of controlling the trustworthiness of answers. However, QA systems are known to fare well with large document collections and benefit from the redundancy found in the data.

In this paper, we showed that limiting the document collection to the certified websites does not hurt the performance of our clinical QA system in terms of relevancy. On the contrary, the focus on trustworthiness leads to improvement of the system overall.

The QA system has been released as a beta version on the HON website. The users have the possibility to rank the results given by the QA system. These human judgments will allow us to evaluate the relevance of the system’s proposed ranking with respect to its use by non-clinicians. In future work we would like to use these human rankings to improve the QA system to be able to better serve the citizens with trustworthy and relevant answers to their medical questions.

The research published in this article has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 257528 (KHRESMOI).

References