Evaluation of Open- vs. Restricted-Domain Question Answering Systems in the Biomedical Field

María-Dolores Olvera-Lobo
CSIC, Unidad Asociada Grupo SCImago, Madrid, Spain & 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 (hereinafter 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 one hundred fifty 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. 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 QA systems are valid and useful for obtaining definitional medical information in that they offer coherent and precise answers.

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.

1 Correspondence to: Juncal Gutiérrez-Artacho, Facultad de Traducción e Interpretación, C/ Buensuceso 11 s/n, Universidad de Granada, juncal@ugr.es
Keywords: Question answering systems; evaluation; biomedical information; Open-Domain Question answering systems, Restricted-Domain Question answering systems.

1. Introduction

Information Retrieval is a discipline focused in 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: a) the assumption that users want full-text documents, rather than answers, and that the query will satisfied with these documents; b) that the process is direct and unidirectional rather than interactive; c) 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 (hereinafter 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].

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 one hundred fifty 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 exact and correct answer -in the form of a number, a noun, a short phrase, or a brief piece of text [7] – for user's questions. Analysis of question, search and choosing answer are three important issues in a QA system so that it includes at least three following process: 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 word or keywords. The system then locates and extracts one or several answers from different sources of information, depending on the specialized area of the question [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 a 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. The ones that allow users to query in different languages are known as multilingual QA systems.

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 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; a restricted-domain QA systems EAGLI (Engine for question-Answering in Genomics Literature); among others.

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

START\textsuperscript{4}, 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 Internet. 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\textsuperscript{5}. This system is a question answering demo, financed by Microsoft, retrieves textual information, by means of the encyclopedia 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 send to a web search engine (either Google or Yahoo) [24-25].

In a study by Ely and colleagues [26], 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 [9-10]. The general public increasingly consults knowledge resources like the

---

\textsuperscript{2} TextMap Question Answer, http://brahms.isi.edu:8080/textmap/
\textsuperscript{3} EAGLI, http://eagl.unige.ch/EAGLi/
\textsuperscript{4} START, http://start.csail.mit.edu/
\textsuperscript{5} QuALiM, http://demos.inf.ed.ac.uk:8080/qualim/
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 [21].

HONqa\(^6\) is a domain-restricted QA system operated by the Health On the Net Foundation, which is a Swiss nonprofit 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 analyzed in our study, as it provides information in English, Italian and French, although we only have focused in 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 in a certain illness.

Finally, MedQA, Medical Question Answering System (nowadays this QA system is not available in their Web page) 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 [28]. It also has a user-friendly interface. It retrieves information from a wide array of sources, including Wikipedia, Medline or Medline Plus. Nowadays, MedQA is not available. To identify definitional sentences, MedQA implemented a set of lexicosyntactic patterns generated automatically from a large set of precompiled definitional sentences that were collected using the Google’s service. 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 [29].

Although in last years, researchers have studied different aspects of QA systems, there are not yet studies enough to evaluate how these tools work. In our study, we have focused on how these four QA systems these systems work to generate phrases with dynamic, coherent definitions and with the most interesting information [16, 30].

3. Methodology

We took a sample of one hundred fifty definitional questions about different medical issues as the basis of this study (see figure 1). The questions were obtained from the webpage WebMD\(^7\), 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 set of questions used passed the internal consistency test with a Cronbach’s alpha coefficient of 0.997.

---

\(^6\) HONqa, http://services.hon.ch/cgi-bin/QA10/qa.pl/

\(^7\) WebMD, http://www.webmd.com/
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 health care [31] - and, as restricted-domain QA systems MedQA and HONqa.

After asking questions in the four systems, several medical professionals and students of Medical Studies 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 [32], 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 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.

\[
MRR = \frac{1}{q} \sum_{i=1}^{q} \frac{1}{F_i}
\]

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

\[
\text{Precision} = \frac{\text{relevant answers} \cap \text{retrieved answers}}{\text{retrieved answers}}
\]

Recall is the fraction of the documents, or answers, that are relevant to the queries that are successfully retrieved. Recall is not been calculated automatically because this measure depends on all the answers retrieved and the data “lost” or not retrieved [33]. The method of polling (combining) of all the outputs for the same request is used to establish relative recall.

\[
\text{Recall} = \frac{\text{relevant answers} \cap \text{retrieved answers}}{\text{relevant answers}}
\]

We have considered as the total number of relevant answers that would be analyzed in this study is 750 answers per QA system, because we have evaluated a maximum of 5 answers for each of the 150 questions. So an exhaustive and precise QA system will retrieve a total of 750 correct answers.

4. Result and discussion

After asking the 150 questions in the four QA systems, the first five answers from each system were analyzed, 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 more answers for some questions, while others did not offer the five answers. The average of answers retrieved by each QA system was very significant, because it is really higher in HONqa (Table 1).

<table>
<thead>
<tr>
<th></th>
<th>Total number of answer retrieved</th>
<th>Average answers retrieved</th>
<th>Total number of answer assessed</th>
<th>Total correct answers</th>
<th>% correct answers</th>
<th>Total incorrect answers</th>
<th>% incorrect answers</th>
<th>Total inexact answers</th>
<th>% inexact answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>HONqa</td>
<td>6635</td>
<td>44.23</td>
<td>726</td>
<td>343</td>
<td>47.24</td>
<td>325</td>
<td>44.76</td>
<td>58</td>
<td>7.99</td>
</tr>
<tr>
<td>QuALIM</td>
<td>441</td>
<td>3</td>
<td>406</td>
<td>166</td>
<td>40.88</td>
<td>180</td>
<td>44.33</td>
<td>60</td>
<td>14.78</td>
</tr>
<tr>
<td>MedQA</td>
<td>802</td>
<td>5.34</td>
<td>705</td>
<td>329</td>
<td>46.66</td>
<td>245</td>
<td>34.75</td>
<td>131</td>
<td>18.58</td>
</tr>
<tr>
<td>START</td>
<td>236</td>
<td>1.6</td>
<td>234</td>
<td>164</td>
<td>70.08</td>
<td>27</td>
<td>11.54</td>
<td>43</td>
<td>18.38</td>
</tr>
</tbody>
</table>

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

Table 1. Measures for evaluating the quality of answers

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 increase substantially, especially in the case of HONqa (with 44.23 answers), while MedQA is slightly better than the open-domain systems (with an average of 5.34 answers for each question).

The correct answers are present to the greatest degree in START (70.08%). In the two restricted-domain QA systems, this average decreases – MedQA (46.66%) and HONqa (47.24%) – and QuALIM is the most deficient, with 40.88% of answers correct. Figure 3 shows the number of correct, inexact and incorrect answers from the four QA systems analyzed.
In relation to the inexact answers, the QA system with the lowest average is 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 is 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.

<table>
<thead>
<tr>
<th>System</th>
<th>MRR</th>
<th>TRR</th>
<th>FHS</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>HONqa</td>
<td>0.75</td>
<td>1.15</td>
<td>0.55</td>
<td>47.24%</td>
<td>45.73%</td>
</tr>
<tr>
<td>QuALiM</td>
<td>0.65</td>
<td>0.77</td>
<td>0.59</td>
<td>40.88%</td>
<td>22.13%</td>
</tr>
<tr>
<td>MedQA</td>
<td>0.87</td>
<td>1.29</td>
<td>0.76</td>
<td>46.66%</td>
<td>43.87%</td>
</tr>
<tr>
<td>START</td>
<td>0.67</td>
<td>0.81</td>
<td>0.64</td>
<td>70.08%</td>
<td>21.87%</td>
</tr>
</tbody>
</table>

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

Table 2. Evaluation measures for QA systems

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 behavior 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 1 to 3.

However, the TRR metric is greater is 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 a greater number of results is offered.

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 recall measure, and an average of 3 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 Methodology. 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 have considered the inexact answers as correct to calculate the precision, the values would increase, in some cases, considerably (Table 3).

<table>
<thead>
<tr>
<th></th>
<th>HONqa</th>
<th>QuALiM</th>
<th>MedQA</th>
<th>START</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>52.75%</td>
<td>42.66%</td>
<td>65.24%</td>
<td>88.46%</td>
</tr>
</tbody>
</table>

Table 3. Precision considering correct and inexact answers

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

<table>
<thead>
<tr>
<th>MRR</th>
<th>MRR</th>
<th>TRR</th>
<th>FHS</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>1.00</td>
<td>0.959*</td>
<td>0.700</td>
<td>-0.265</td>
<td>0.837</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>0.041</td>
<td>0.300</td>
<td>0.735</td>
<td>0.163</td>
<td></td>
</tr>
<tr>
<td>TRR</td>
<td>Pearson</td>
<td>0.959*</td>
<td>1.000</td>
<td>0.472</td>
<td>-0.316</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>0.041</td>
<td>0.528</td>
<td>0.684</td>
<td>0.043</td>
<td></td>
</tr>
<tr>
<td>FHS</td>
<td>Pearson</td>
<td>0.700</td>
<td>0.472</td>
<td>1.000</td>
<td>0.093</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>0.300</td>
<td>0.528</td>
<td>0.907</td>
<td>0.803</td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>Pearson</td>
<td>-0.265</td>
<td>-0.316</td>
<td>0.093</td>
<td>1.000</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>0.735</td>
<td>0.684</td>
<td>0.907</td>
<td>0.613</td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>Pearson</td>
<td>0.837</td>
<td>0.957*</td>
<td>0.197</td>
<td>-0.387</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>0.163</td>
<td>0.043</td>
<td>0.803</td>
<td>0.613</td>
<td></td>
</tr>
</tbody>
</table>

* The correlation is significant at 0.05

Table 4. Correlation of measures

5. Conclusions

Today, users are more demanding of the retrieved information, as much in regards to the quality as in regards to the quantity, just as in response time. Because of all this, the QA systems could be one of the future systems of information retrieval on the Internet since they intend to meet the needs and demands of the current users. None the less, 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 analyzing and evaluating such systems and their characteristics, in order to promote their development and commercialization.
Through analyzing 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 QA 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 works [15], we compared the sources used by the restricted-domain QA systems with those used by the general-domain systems we could see significant differences in the typology and specialization, as would be expected. A comparison of the sources used by START and QuALiM confirms the absence of significant differences, as both systems use Wikipedia as the 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, encyclopedias and databases specializing in the biomedical field, HONqa opts for Web sites 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 [34] have explored various possibilities for improvements, such as the use of ontologies, which contributes to increase the quality of the answers obtained by formalizing the relevant information from the domain in question. In addition, these approaches together others are slowly attracting more researchers who are experienced in handling the results they produce. This data suggests 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 [35]. 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 remains 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 showed that the QA systems are a useful tool to retrieve information quickly and accurately.

6. References


