A Question Answering System over Chronic Diseases and Epigenetics Knowledge

Luciana Almansa, Gabriel Rubio, Alessandra Alaniz Macedo[1]

[1]DCM - FFCLRP - University of São Paulo (USP)
Av Bandeirantes 3900 – Ribeirão Preto – SP – Brazil

Abstract. *Medical records describe patients' health conditions and help experts to decide on treatments. The scientific biomedical knowledge can improve the prevention and treatment of diseases and promote innovation and discovery in health. However, healthcare professionals may have difficulty in searching for relevant scientific information due to lack time and constant literature update. The present work proposes a Question Answering (Q&A) architecture to support a more focused search for information about chronic diseases. A user question in natural language initiates the search for answering and promoting knowledge such as a learning healthcare system. To evaluate the system, we employ a reference collection on epigenetics and chronic disease and calculate performance measures like precision, recall and F-measure.*

1. Introduction

Healthcare professional and experts use information obtained from medical reports on patient's treatment [Bhat et al. 2016]. Shortliffe & Cimino [Shortliffe and Cimino 2013] believe that "in the near future, medical systems known as *"Learning Healthcare Systems"* (LHS) will work with information from medical combined with other kinds of information such as scientific papers to produce knowledge". Healthcare professionals can then employ this knowledge to manage health institution or even patients' health. The creation of LHS is a challenge for information systems.

Scientific papers are important sources of information. Most of these papers are stored in on-line repositories such as Pubmed[1]. Unfortunately, the search for interesting papers has become increasingly complex for various reasons: a lot of documents are included in the on-line repositories every day; professionals lack the time to read papers; papers are published in various languages; each specialty requires a different language; and professionals use many different templates to present information [Cohen and Hunter 2008, Rzhetsky et al. 2009, Shortliffe and Cimino 2013]. In this context, Information Retrieval (IR) and Question-Answering (Q&A) frameworks can build systems that can manipulate searches for specialized information.

According to Baeza Yates & Ribeiro Neto [Baeza-Yates and Ribeiro-Neto 1999], IR systems help users to search for important information and facilitate the organization of data. A shortcoming of IR systems is the huge number of papers they retrieve [Kolomiyets and Moens 2011]. In contrast, Q&A systems provide short and direct answers to user search [Athenikos and Han 2010]. For LHS, Q&A system can focus and start the process of learning healthcare information.

This work presents the *Question-Answering Surveillance* architecture (QASF) and evaluates its modules separately. The QASF receives a question about chronic diseases as well as epigenetics[2] information and looks for answers in a collection of scientific papers. The QASF exploits linguistic and knowledge resources to support the processing of LHS. Architectures and frameworks support the development of systems. QASF-supported systems can accelerate the health expert's search for answers.

2. Architecture of a Question-Answering Surveillance Framework

Question-Answering (Q&A) systems basically return short and direct answers to users. This kind of system usually relies on Information Extraction, Text Mining and Information Retrieval techni-

[1]Pubmed is an on-line repository of information with more than 25 million of journals, citations and biomedical books (http://www.ncbi.nlm.nih.gov/pubmed).

[2]Research works have suggested that people exposed to risk factors, such as food shortage, at the beginning of life can alter gene expression. This alteration can result in increased risk of chronic disease in the adult life. Epigenetics is the field that studies the alteration in gene expression [Barker 2001]

ques [Allam and Haggag 2012], and its essential architecture comprises the three following modules: (i) Question Processing, (ii) Answer Processing and (iii) Document Processing. The architecture of QASF has the three traditional Q&A modules, but the last module comprises abstract classes (program-code-template) reused from two software frameworks [Macedo et al. 2016b, Macedo et al. 2016c] created by our team. An information system [Almansa and Macedo 2016] motivated the development of the QASF. Figure 1 depicts the architecture of the QASF.

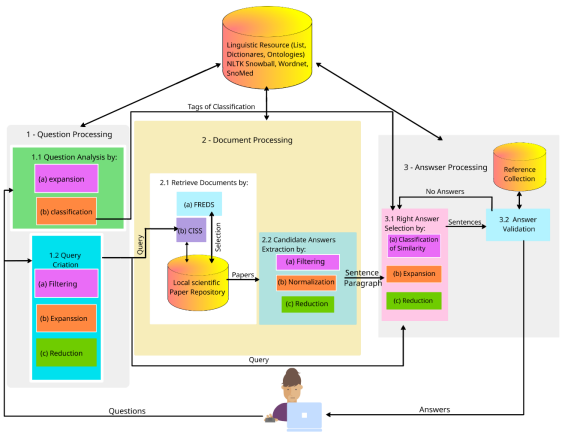

Figura 1. A Question-Answering Architecture Supporting Surveillance Information Systems on Chronic Diseases

2.1. The "Question Processing" Module

The *"Question Processing"* module converts the question a user poses in natural language to a query to search and select answers. It consists of *"Question Analysis"* and *"Query Creation"*. The first task analyzes the question by looking for implicit information, like its type (e.g., definition, a list of answers, etc.) and its domain (e.g., place, time, disease, etc.). The second task converts the question into a vector of keywords. For *"What is chronic disease?"*, the type is *"definition"* and the domain is *"disease"*, so the vector of keywords could be *(is, chronic, disease)*.

In the literature, the task *"Question Analysis"* usually considers Machine Learning, Pattern Recognition and Morphosyntactic Analysis techniques [Er and Cicekli 2013, Monz 2003, Zhang and Lee 2003]. The QASF applies a simple mechanism of Pattern Recognition and Machine Learning to handle the type and the content of the question, respectively.

The QASF analyzes the question word (e.g., *"What"*) to verify whether it is a definition question. Then, the QASF searches for the subject of the sentence and submits it to a lexical database, a dictionary and/or an ontology. The QASF exploits the Wordnet lexical database[3] and the SNOMED[4] After, the linguistic repositories return the definitions of the subject of the question. The definitions could be presented to the user, when using a Relevance Feedback approach. The technique Pattern Recognition does not manipulate other types of questions.

The second technique that the task *"Question Analysis"* exploits is Machine Learning (ML) in order to analyze the contents of the questions for the other interrogative pronouns besides the pronoun related to definition. Nowadays, the QASF supports Support Vector Machine and Naive Bayes. By using both, it is possible to build and validate the classifier on the basis of a reference collection of questions formed by classes. The algorithms analyze this collection to learn attributes from it. After that, the classifier predicts the label of an unknown question. Section 3 shows the results and the collections.

Still considering the *"Question Processing"* module, *"Query Creation"* has multiple steps. First, the user inserts the keywords of the question, which the QASF transforms into a query by disregarding the question words, the stopwords[5], and the punctuations. Next, the QASF sends the vector to the dictionary to

[3] Wordnet is an on-line English lexical database consisting of nouns, verbs, adjectives and adverbs words clustered as sets of cognitive synonyms (https://wordnet.princeton.edu/).

[4] SNOMED is a dictionary that focuses on medical terminologies [Spackman et al. 1997].

[5] Stopwords are words filtered out by processing of textual information because they refer to the most common words in a language such as preposition, pronouns and etc.

retrieve synonyms. If the dictionary does not find synonyms, the QASF requests hypernyms.

The last step of the task *"Query Creation"* is the use of stemming[6] and lemmatization[7] of the synonyms or hypernyms. Lemmatization converts verbs into their respective infinitive forms; stemming extracts the roots of all the terms. When some terms have the same root, the QASF groups them into only one root. NLTK [NLTK 2015] and Snowball [Porter and Boulton 2001] support lemmatization and stemming, respectively. The final vector consist of the terms resulting from lemmatization and stemming.

2.2. The "Document Processing" Module

The *"Document Processing"* module consists of the tasks *"Retrieve Documents"* and *"Candidate Answers Extraction"*. The first task retrieves documents that could have the answer to the user question. By using the potential documents, the *"Candidate Answers Extraction"* extracts parts of the documents that should be the answer. Nowadays, the QASF reuses two frameworks from group to compose the *"'Document Processing"* module, namely the CISS and the FREDS.

The CISS framework searches and relates information from scientific articles in epigenetics and chronic diseases to medical reports. FREDS searches and relates textual information from medical exams to exam reports of imaging tests concerning fine needle aspiration of thyroid. Further details about the relationships between data are details in the CISS [Macedo et al. 2016c, Macedo et al. 2015, Pollettini et al. 2014] and the FREDS [Macedo et al. 2016a, Pessotti 2012]. Because the frameworks employ linguistic resources, the QASF exploits them to provide the user with accurate responses. The vector of keywords generated by the *"Question Processing"* module is the input data for the frameworks to search for answers in scientific information.

The task *"Candidate Answers Extraction"* is one of the most complex because there are questions in different semantic levels of complexity and answers in different levels of granularity (sentence, paragraph, document, etc). For example, a question could be a simple definition such as *"What is chronic disease?"* or a complex question such as *"Is it legitimate to use fenfluramine and phentermine longer than six months to control obesity? Is it legal?"*. Analysis of the syntax tree, dependence graphs and linear proximity are the main strategies to manipulate semantic levels of information [Allam and Haggag 2012, Gupta and Gupta 2012, Monz 2003], but we do not apply to the QASF yet. QASF just manipulates the size of the answers in terms of documents, paragraphs, and sentences.

The current QASF extracts the candidate answers from the scientific information in two steps. In the first step, the QASF retrieves and manipulates all the words of the sentences in the abstracts (a set of grouped sentences (or sometimes paragraphs) labeled as the metadata "abstract") of scientific articles. Then, the QASF converts the letters of the words from uppercase to lowercase, to normalize the words considered in the linguistic processing. After, the lemmatization and stemming respectively convert verbs to their infinitive forms and extract the roots of the words. The next step of "Answer Processing" matches the root of the words of the candidate answer and the root of the words of the vector of the user question. Considering a threshold t of similar words, the potential candidate has a higher of similar words than the threshold. Consequently, they are considered the potential right answers.

2.3. The "Answer Processing" Module

The *"Answer Processing"* module processes all the potential right answers and classifies them according to a similarity value. The QASF considers the cosine as the similarity value. Hence, the user is shown the n first answers. Many Q&A systems return just one answer. The *"Answer Processing"* module comprises: *"Right Answers Selection"* and *"Answer Validation"*.

The task *"Right Answer Selection"* calculates the similarity between the user question and each potential right answer in the set of candidate answers. To this end, the QASF calculates the cosine between the user question and each candidate answer and converts them into two vectors in a vector space. Next, the cosine value of the angle between the two vectors is measured. The task *"Right Answer Selection"* also sorts the results of the measurements in descending order. This task selects n higher cosines and shows the

[6] Stemming is the process that reduces inflected or derived words to their stem, base or root form—generally a written word form.

[7] Lemmatization is the process that groups the different inflected forms of a word together so the QASF can analyze them as a single item.

corresponding candidate answer(s) to the user. Nowadays, the QASF returns the five first answers, but it is possible to adjust this value.

The second approach to measure the similarity between the user question and the candidate answer is more complex because the user question and the candidate answers require pre-processing before calculation of the cosine. The steps include: (i) stemming and lemmatization, (ii) changing the synonymous words from the candidate answer to keywords that represent the user question, and (iii) measuring the similarity on the basis of cosine. The two first steps intend to improve matching and similarity because many words have the same meaning. For example, the words "*thyroidal*" and "*thyroid gland* can be changed to "*thyroid*.

In the literature, the ranking of answers exploits some criteria such as (i) the number of words present in both the user question and each candidate answer, (ii) the number of keywords in the length of the user question keywords compared to each candidate answer, or (iii) the number of words in the candidate answer differing from the number of keywords in the user question [Allam and Haggag 2012, Gupta and Gupta 2012]. The task *"Answer Validation"* verifies if each answer returned by the Q&A system is right for the user question. When the answer is not right, the task discards it and analyzes the next answer.

3. Evaluation

After creating the QASF, we evaluated all its modules by considering classic measures (precision, recall, F-measure, etc) and reference collections. In the Q&A area, reference collections consist of questions and right answers, or questions and tags. Conferences such as TREC [Voorhees and Tice 2000] and CLEF [Magnini et al. 2004, Olvera-Lobo and Gutiérrez-Artacho 2015] provide some reference collections. However, finding a useful reference collection in the QA area is laborious, but discovering a reference collection with a closed domain; i.e., with questions and answers on a specific subject such as chronic diseases, is still more complicated in the case of the QASF. Hence, evaluation was one of the most challenging steps of the developing of the QASF. To overcome the challenges of the evaluation, we used three reference collections and assessed different approaches exploited by the QASF.

3.1. The Reference Collections

The **AskHERMES** reference collection consists of 4,654 questions classified into device, diagnosis, epidemiology, etiology, history, management, pharmacology, physical finding, procedure, prognosis, test and treatment & prevention. In terms of balancing, the quantity of questions per topic differs considerably; whereas some topics have over than 500 questions, others have fewer than 100 questions. Some examples of questions are: *"What is the protein you can test to see if someone has diabetes type 1 or type 2?"* for the test topic, and *"Is diabetes a risk factor for carpal tunnel syndrome?"* for the diagnosis topic. Reference [Cao et al. 2011] presents the AskHERMES reference collection for download. We have used this collection to classify user questions to specific topics during the task *"Question Analysis"*.

EDRMrSa (*Epigenetics Data Related between Medical Records and Scientific Articles*) is a reference collection that relates scientific articles on epigenetics and chronic diseases to medical reports. EDRMrSa has been gradually created to evaluate CISS in references [Macedo et al. 2015, Pollettini et al. 2012, Pollettini et al. 2014]. It has been collaboratively developed with the Children's Institute of the Hospital (ICr) of the Medical School of the University of São Paulo (HCFMUSP). The ICr group has conducted a study of two thousand children to identify risk factors during their development, with a view to preventing chronic disease in adulthood. As a result, this group has compiled a list of medical terms (expressions) frequently found in the medical reports of these children. Considering a blank medical report and the terms of the filled ICrs reports, our research group has simulated thirty medical reports with the aid of the Mathematica software [8]. Two ICr specialists have related and classified six of the thirty simulated medical reports to a set of 226 scientific articles on epigenetics and chronic diseases selected from Pubmed by considering the subjects epigenetics and chronic diseases. Our group has defined the classes "Strongly Relevant", "Relevant", "Weakly Relevant", and "Irrelevant" to classify the relationship between a paper and a patient's medical report. Because the QASF looks for answers based on a user query, we have remodeled the simulated medical report for it to become a user question. Consequently, EDRMrSa should provide papers to questions based on medical reports. We have used this collection to evaluate the task concerning reformulation of the query, the *"Document Processing"*, and the selection of right answers of the QASF.

[8] www.wolfram.com/mathematica

BioASQ is a challenge that aims to index biomedical data semantically. The 5_{th} year of the BioASQ challenge has three tasks. To evaluate the QASF, we have exploited the task 5b, about Biomedical Semantic QA, because it contains biomedical training and testing questions together with references (golden standard), exact responses, and "ideal" answers. Currently, the BioASQ5b provides 1,799 training questions, along with their references and the golden standard answers. We have specifically used BioASQ5b to evaluate the task *"Answers Validation"* of QASF because EDRMrSa provides answers in the level of paragraph or abstract. BioASQ5b contains exact responses in terms of sentences to validate a Q&A system.

3.2. Results

We have evaluated the modules of QASF – the *"Question Processing"*, the *"Document Processing"*, and the *"Answers Processing"* – to measure the efficiency of the QASF in terms of quantitative and qualitative measures. We evaluated the three modules separately, but a system built over our proposal will able to integrate them. Nowadays, the QASF manipulates the domain of epigenetics and chronic diseases by considering a "question + topic (for classification tasks) + answer" structure.

We used the AskHERMES collection to examine the *"Question Processing"* module, mainly the task *"Question Analysis"*. We exploited the EDRMrSa collection during evaluation of all the three modules because this collection manipulates scientific papers on epigenetics and chronic diseases. EDRMrSa returns answers in the level of papers or abstracts, whereas the BioASQ5b collection evaluates the *"Answer Processing"* module. This collection has golden standards based on sentences. To analyze the results, we have applied the t-Student test with a significance level of 5% to reject the null hypothesis H_0 with a confidence level of 95%. H_0 assumes that all approaches perform equally.

3.2.1. Question Processing

The *"Question Processing"* module (composed by the tasks *"Question Analysis"* and *"Query Creation"*) analyzes the user question and extracts implicit and complementary information to create a query that represents the user question. We have assessed the task *"Question Analysis"* by the technique "Pattern Matching", with the EDRMrSa reference collection, and by "Machine Learning", with the AskHERMES reference collection. We have also assessed the task *"Query Creation"* by using the EDRMrSa collection.

To evaluate the use of the technique "Pattern Matching", we have created six user questions from the medical records of the EDRMrSa reference collection, segmented into terms composed of unigram, bigram, trigram and four-gram terms[9]. The QASF sends the terms to the dictionaries Wordnet and SNOMED, and returns the definitions to the QASF.

Considering Wordnet, just the unigram terms return some result. For all the patients' medical records, the number of returned terms is smaller than the number of submitted terms. On the other hand, SNOMED returns results for unigram, bigram, trigram and some four-gram terms, except for patients "p3" and "p18". SNOMED focuses on medical terms, so it provides a better number of matches than Wordnet, which is composed of general subjects. This experiment illustrates the utility of employing a closed-domain dictionary to analyze the question of providing the QASF with more information to expand the query. In the near future, we are going to evaluate other dictionaries and ontologies together with the SNOMED.

Still regarding the evaluation of the task *"Question Analysis"*, we tested the "Machine Learning" technique by using the reference collection AskHERMES. We considered the reference collection as (i) a whole (4,654 questions classified into 12 topics) and as (ii) a whole but ignoring the topics with fewer than 100 questions (4,258 question classified into 7 topics: 475 questions about diagnosis; 578 about test; 319 about treatment & prevention; 189 about pathological findings; 172 about etiology; 1,380 about management; and 1,145 about pharmacology). Considering the $k-$fold cross-validation with $k = 10$, we measured the performance of the classifiers. In Table 1, the two first columns show the results for situation (i), with the whole collection, and the last two columns concern situation (ii). On the basis of the t-Student test, all the measures of precision, recall, accuracy and F-measure are larger than 0.5. In situation (ii), precision and accuracy are almost 0.6. In both situations (i) and (ii), the Support Vector Machine [Bergman 1970] classifier is significantly better than the Naive Bayes classifier (p-value $\simeq 0.04$ and p-value < 0.01, respectively).

[9] $n-gram$ is a way to represent a contiguous sequence of n items from a given sequence of text. For QASF, the items are words.

The "Pattern Matching" and "Machine Learning" techniques are simple and complementary. The former technique expands the user queries by means of dictionaries, whereas the latter technique exploits a classifier to analyze questions according to topics of health and medical information. We have evaluated both techniques, but the QASF does not use the topics generated by the classification during the next steps. The QASF uses closed-domain collections, which dismiss analysis/filtration of answers according to tags of classification. We have used the "Pattern Matching" technique to conduct the task *"Query Creation"*.

Tabela 1. Precision, Recall, F-Measure and Accuracy for classification of two sets of questions and topics by Naive Bayes and Support Vector Machine.

	4654 questions & 12 topics		**4265 questions & 7 topics**	
	Naive Bayes	2***SVM**	**Naive Bayes**	2***SVM**
Precision	0.48	0.51	0.48	0.59
Recall	0.42	0.53	0.46	0.58
F-Measure	0.32	0.51	0.38	0.56
Accuracy	0.43	0.55	0.47	0.59

We have evaluated the task *"Query Creation"* by using the EDRMrSa questions in two steps. First, we have ignored interrogative pronouns, stopwords, and punctuations; so that the remaining words consist basically of nouns, adjectives, and verbs. We have searched for synonyms and hypernyms of the remaining words in a dictionary and applied the stemming and lemmatization techniques to extract the roots of words. All the resulting vectors are larger than the input vectors. For example, questions of the collection with approximately 70 terms have 500 terms after the task *"Query Creation"*. However, the terms can change notably if we consider the number of question words and synonyms related to each question. The addition of synonyms, hypernyms, and roots of words during *"Question Analysis"* supports the *"Query Creation"* and aids the search for potentially right answers if we consider different words with the same meaning. Therefore, the two tasks highlight that efficient linguistic resources are important during the manipulation of queries that are to be processed by tasks within the *"Question Processing"* module.

After evaluation of "Question Processing", the best scenario of the QASF should include a final query vector composed of unigram terms from medical records obtained by using dictionaries, by eliminating specific non-useful terms and by adding synonymous and hypernyms. We have re-used this configuration in the next steps of the evaluation.

3.2.2. Document Processing

Given the query by the *"Question Processing"* module, the *"Document Processing"* module retrieves documents that should have the answer and also searches for potential right answers in these documents. It includes the *"Document Retrieval"* and *"Candidate Answers Extraction"*.

The task *"Document Retrieval"* recovers the documents that could contain potentially right answers for user questions. The QASF reuses the CISS [Pollettini et al. 2014] and FREDS [Pessotti 2012] frameworks to perform the document retrieval task. In the present evaluation, the QASF exploits CISS because it aims to measure the recall and the precision of the retrieval of scientific papers after the submission of queries consisting of medical records as queries by the *"Question Processing"* module. Figure 2 contains a plot of precision versus recall measures for the retrieved scientific papers. This plot presents average precision and average recall if we consider that six medical records have been submitted as query, and that the maximum quantity of scientific papers that the QASF retrieves varies. In [Pollettini et al. 2014], the CISS relates scientific papers to medical records. For the present evaluation, the QASF transforms medical records into questions and carries out the experiment presented here.

Figure 2 shows that the number of retrieved scientific papers influences the performance of the QASF. For a set of "two hundred retrieved scientific articles" (line plotted with "X" markers), the precision measure is almost 0.9, and the recall measure is approximately 0.3. However, when the precision is between

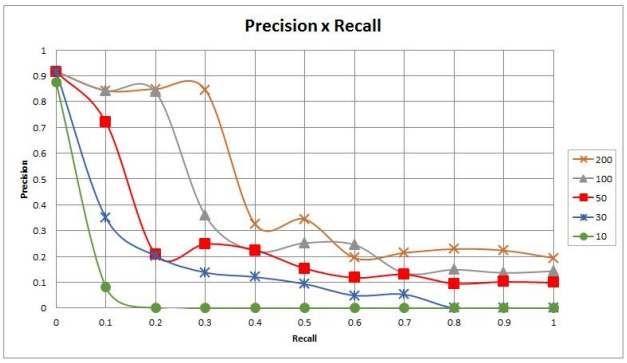

Figura 2. Precision and recall after the max quantity of retrieved papers varied.

0.2 and 0.3, the recall measure is 0.6. The ideal number of articles used to retrieve most of the answers to the evaluated questions is 100 papers, plotted with triangular markers in Figure 2. The task *"Candidate Answer Extraction"* searches for potential right answers in the abstracts of the scientific papers retrieved in the set of "one hundred scientific articles retrieved" during the task *"Document Retrieval"*. The candidate answers are sentences of the abstracts of the papers classified as "strongly relevant" or "relevant" to the candidate answer. Candidate answers correspond to the abstracts broken into sentences. Then, the next module compares the vector constructed by the task *"Query Creation"* with each potential right answer filtered from the papers after a threshold is considered. If the potential answer has a similarity value greater than the threshold, the potential right answer is considered a candidate answer. After experimentation, we defined the QASF threshold by using the following equation:

$$t = \frac{N}{log_{10}^{N}}$$

whose N is the number of words obtained during the task *"Query Creation"*, and the threshold (t) is N divided by log_{10}^{N}

Briefly in the *"Document Processing"* module, the input data is the vector of words generated by the *"Question Processing"* module, and this version of the *"Document Processing"* retrieves papers by exploiting the CISS. We have considered that a set of retrieved scientific articles consists of potential right answers. The QASF breaks the abstract of this set of papers into sentences. In terms of results, the precision measure is almost 0.9, whereas the recall measure is almost 0.3 when the task *"Retrieve Documents"* is evaluated. We have evaluated the right answer in terms of sentence in the next processing step by using the BioASQ collection because the EDRMrSa relies on a document as a unit of information retrieval.

3.2.3. Answer Processing

In the *"Answer Processing"* module, the QASF ranks the candidate answers according to the similarity measures calculated between the user question and the candidate answers. The *"Answer Processing"* module consists of the tasks *"Right Answer Selection"* and *"Answer Validation"*. The *"Right Answer Selection"* accounts for the ranking and selection of answers on the basis of measures of similarity that consider the sentences submitted by the *"Document Processing"* module. The QASF selects the most similar answers as answers. The *"Answer Validation"* tries to evaluate whether the answers are valid. Next, the QASF presents the results to the user.

The QASF carries out the task *"Right Answer Selection"* by calculating the cosine similarity between each sentence of the abstract and the query in two ways: (i) *cosine 1* – cosine measured between the question and the candidate answers by using the roots of the words extracted through stemming and lemmatization approaches, the synonym words, and hypernyms; and (ii) *cosine 2* – cosine between the question and the candidate answers measured in their original grammatical form.

In terms of *"Answers Validation"*, one of the best ways to evaluate returned answers is to use a reference collection. The QASF used the EDRMrSa represented by tuples of questions and answers with

questions formulated from medical reports to obtain answers from scientific papers. The unit of retrieval was abstract of papers.

Figure 3 presents the average precision of the ten candidate answers provided by the QASF by varying the patient and the amount of analyzed papers. Patient "p24" provided more relevant answers.The medical record of this patient did not have the highest number of medical terms used in the user question. On the other hand, the medical records of patients "p3" and "p18" contained more medical terms used in the user question and led to less precise relevant answers. Therefore, the precision of the answers is not associated with the amount of medical terms in the user question, but it is related to the quality of the terms used to compose the query. The average precision for *cosine 1* and *cosine 2* is 0.56 and 0.57, respectively, which is considered similar. According to Figure 1, if the QASF does not find right answers for the user question, it returns to the task *"Right Answer Selection"* and looks for another set of candidate answers.

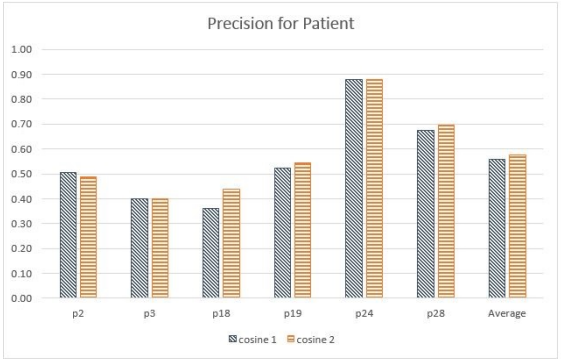

Figura 3. The average precision calculated for patients.

To evaluate the unit of retrieval as a sentence, we have conducted QASF evaluations by using a golden standard as a reference answer, namely BIOASQ. Thirty questions and their respective correct answers as well as the abstracts of the Pubmed papers where these answers are were randomly selected and manipulated by the QASF. Then, the database where the QASF searches for the abstracts of the papers with the possible answers is updated with the articles corresponding to the 30 questions selected. To evaluate the QASF, we have to enter all the 30 questions into the framework one by one and to compare the returned responses to the correct answer provided by the BioAsk. Table 2 shows the results and the thresholds that help to select the abstracts of the articles with the possible correct answer and to choose the correct answer are modified between 0.07 and 0.17 and between 3 and 4, respectively. These evaluations aid verification of the best combination of thresholds and allow the return of a greater number of correct answers.

Tabela 2. Results of combinations of thresholds of CISS and QASF.

2*Tests	Right Answers	Wrong Answers	Without Answers
CISS 0.07 and QASF 3	11	4	15
CISS 0.17 and QASF 3	11	4	15
CISS 0.27 and QASF 3	7	1	22
CISS 0.07 and QASF 4	**14**	**7**	**9**
CISS 0.17 and QASF 4	13	4	13
CISS 0.27 and QASF 4	7	3	20

On Table 2, the best combination of thresholds is 0.07 for the selection of abstracts of articles by and 4 for the selection of correct answer. This combination of thresholds returns 14 correct answers, 7 incorrect, and 9 blank answers. This means that 14 of 30 (47%) answers are right, 7 of 30 (23%) answers are wrong, and 9 of 30 (30%) answers are blanks. Other options have the same difference between right and wrong answers, but they have more blanks.

The QASF is a Question-Answer Framework consisting of modules of the *"Question Processing"*, *"Document Processing"*, and *"Answer Processing"* modules. To find the correct answers, all the modules of QASF must run harmonically. The *"Answer Processing"* module first calculates similarity measures in two ways, even though the use of textual manipulation with linguistic resources does not lead to significant improvements in this specific task. The validation of potential answers relies on two reference collections, EDRMrSa and BioASQ, changing the unit of retrieval. The whole set of presented results shows that the QASF works satisfactorily. Our team believes that future works could improve the results.

4. Related Work

Different research teams are currently investigating Q&A systems worldwide. Q&A systems have exceeded the academic areas and are now being widely employed in smartphones to support users in their daily activities. The best-known tools are Google Now[10] (for Android) and Siri[11] (for iOS). Since 2005, IBM researchers have been developing an artificial intelligence question-answer supersystem known as Watson[12].

The scientific literature contains reports on different ways of developing Q&A systems. Moreda et al. have presented a Q&A system that uses semantic rule and/or Wordnet and focuses on discovering how semantic information influences the entity named extraction step [Moreda et al. 2011]. According to these authors, semantic information makes the retrieval of the named entity more precise. Yen et al. have proposed a Q&A framework based on Machine Learning to integrate the tasks "Question Classification" and "Answer Selection" [Yen et al. 2013]. A classifier places the user questions into classes and sends them to the ranking step, which reorganizes the information provided by "Document Processing". Ryu et al. have proposed the use of the Wikipedia[13] as a source of knowledge for a Q&A system [Ryu et al. 2014]. The architecture of the system consists of an "Answer Processing" module based on semi-structured information such as scientific papers. The authors affirm that each type of information is more appropriate to the discovery of a specific type of question. For example, documents composed by definitions such as dictionaries or even papers are more compatible with descriptive questions.

Suresh-Kumar & Zayaraz have designed an interactive algorithm to extract features and relations of documents following a seed concept based on pattern rules and decision trees [Suresh kumar and Zayaraz 2015]. They developed an ontology by using user questions in natural language and a Q&A framework in a generic domain. Cao et al. have developed the Q&A system AskHERMES [Cao et al. 2011]. AskHERMES processes long questions and answers them by using abstracts of papers from MEDLINE, PubMed, eMedicine and Wikipedia. The authors advocate that long questions are not answered well by means of syntactic and semantic rules. AskHERMES integrates the UMLS [Bodenreider 2004, Brin 1999] sources to expand user queries and uses techiquqes from Natural Language Processing, Information Retrieval and Information Extraction area. The authors state that the system performs satisfactorily in the case of long and complex questions. Ben-Abacha & Zweigenbaum have developed a Q&A system called MEANS [Be and Zweigenbaum 2015], supported by Semantic Web. This system uses semantic approaches to analyze data by considering two levels of complexity: search for medical entities (drugs, symptoms and diseases) and analysis of relationships between entities (treatments, preventions and causes). The questions are factual and Boolean.

In Brazil, the field of Q&A systems is an innovative research area. Machado-Junior has developed the Shallow Question Answering System to manipulate Brazilian news [Machado Junior et al. 2009]. Wilkens et al. have created a Q&A system called COMUNICA to process voice in Portuguese in order to promote a different way of interacting with a Q&A system [Wilkens et al. 2010]. Amorim et al. have built a generic Q&A system that searches for answers in WordNet, in a database written in AIML (Artificial Intelligence Markup Language [Wallace 2003]), or even in the Web, as the last option [Amorim et al. 2012]. Prestes has compared three Q&A systems by considering the complexity of the Portuguese and the English languages [Prestes 2011]. Recently, Arrigo et al. have developed a Q&A system by using Web to create a Brazilian annotated corpus based on semantic analysis and machine learning of patterns to classify information and to analyze the precision of the answers [Arrigo et al. 2014].

[10] https://www.google.com/intl/en/landing/now/whatisit

[11] https://support.apple.com/en-us/HT204389

[12] http://www.ibm.com/en-us/homepage-b.html

[13] https://www.wikipedia.org/

When we compare QASF with other Q&A systems, we can highlight some important features. QASF can process different domains by simply changing/expanding the linguistic resources manipulated by the processing modules. The architecture of the QASF can be reused. Another aspect is that the QASF manipulates question considering different levels of complexity and different units of retrieval. The QASF processes questions that demand different levels of processing to search for an adequate answer. Examples of questions are "What is a chronic disease?" or "When can a patient be considered hypoglycemic?".

5. Final Remarks

In the medical and biomedical areas, co-investigation of information from medical records and scientific collections can support the treatment and the prevention of diseases [Shortliffe and Cimino 2013]. However, reading of scientific papers demands time and availability. QASF supports the development of information systems that can assist healthcare professionals by offering quick search of information. QASF is an initiative to support learning healthcare systems supported by question and answering processes. Nowadays, the QASF focuses on chronic diseases, but it is possible to alter ontologies and dictionaries to other domains. The specificity is related to the manipulated information supported by linguistic resources.

The QASF is basically supported by the following modules of Q&A systems: (i) Question Processing, (ii) Answer Processing, and (iii) Document Processing, which mainly carry out textual manipulation, reduction/inflection of words, expansion of terms by dictionaries or ontologies, filtering and calculation of similarity between documents, paragraphs and sentences. A differential feature of the QASF is the fact that the tasks extensively use of linguistic resources. During evaluation, most results indicate improvements upon use of some linguistic resources.

The evaluation of the QASF indicated promising results. The set of retrieved papers had precision of almost 0.9, whereas the recall was almost 0.3. The ways used to calculate the similarity between the answer and the question showed that the similarity were close and ranged between $0,40$ and $0,46$, and they afford the same set of answers with almost the same order. The best number of articles to be retrieved by the QASF was 50 papers, with precision of 0,7. Finally, patient "24" generated more relevant answers. The retrieval of sentences also provided good results in terms of right answers during the *"Answer Validation"*. Evaluation of the QASF was difficult because the modules involved different technologies. So, we used different reference collections. We applied EDRMrSa to evaluate the QASF by considering the document or abstract as a unit of retrieval. When the unit of retrieval was sentence, we used BioASQ5b. As future work, we intend: (i) to evaluate the QASF by considering other domains, for example, thyroid applying the FREDS [Pessotti 2012], (ii) to extend the QASF by creating mobile GUI interfaces, and (iii) to include more extra semantic information by using ontologies and UMLS.

Acknowledgements

We would like to thank the CAPES, FAPESP and CNPq for financial support.

Referências

Allam, A. M. N. and Haggag, M. H. (2012). The question answering systems: A survey. *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, 2(3):211–220.

Almansa, L. and Macedo, A. (2016). Sistema de informação para perguntas e respostas em doenças crônicas. In *Anais Principais do XVI Workshop de Informática Médica*, pages 127–136, Porto Alegre, RS, Brasil. SBC.

Amorim, M. T. C. F. d., Cury, D., and Menezes, C. S. (2012). Um Sistema Inteligênte Baseado em Ontologia para Apoio ao esclarecimento de Dúvidas.

Arrigo, A. J. S., Silva, E. G., Martins, H. P., and Silva, P. P. (2014). Desenvolvimento de um Sistema de Pergunta e Resposta Baseado em Corpus. In *14o Congresso Nacional de Iniciação Científica (CONIC-SEMESP)*, pages 1–6, São Paulo, SP.

Athenikos, S. J. and Han, H. (2010). Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine*, 99(1):1 – 24.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern information retrieval*, volume 463. ACM Press New York, 1nd edition.

Barker, D. (2001). Fetal and infant origins of adult disease. *Monatsschrift Kinderheilkunde*, 149(1):S2–S6.

Be, A. and Zweigenbaum, P. (2015). MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies. *Information Processing & Management*, 51(5):570–594.

Bergman, S. (1970). *The kernel function and conformal mapping*. Number 5. American Mathematical Soc.

Bhat, S., Gijo, E., and Jnanesh, N. (2016). Productivity and performance improvement in the medical records department of a hospital: An application of lean six sigma. *International Journal of Productivity and Performance Management*, 65(1):98–125.

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270.

Brin, S. (1999). Extracting patterns and relations from the world wide web. In *Selected Papers from the International Workshop on The World Wide Web and Databases*, WebDB '98, pages 172–183, London, UK, UK. Springer-Verlag.

Cao, Y., Liu, F., Simpson, P., Antieau, L., Bennett, A., Cimino, J. J., Ely, J., and Yu, H. (2011). AskHERMES: An online question answering system for complex clinical questions. *Journal of biomedical informatics*, 44(2):277–88.

Cohen, K. B. and Hunter, L. (2008). Getting started in text mining. 4:1–20.

Er, N. P. and Cicekli, I. (2013). A Factoid Question Answering System Using Answer Pattern Matching. In *International Joint Conference on Natural Language Processing*, pages 854–858, Nagoya, Japan.

Gupta, P. and Gupta, V. (2012). A survey of text question answering techniques. *International Journal of Computer Applications*, 53(4):1–8.

Kolomiyets, O. and Moens, M.-F. (2011). A survey on question answering technology from an information retrieval perspective. *Information Sciences*, 181(24):5412 – 5434.

Macedo, A., Pessotti, H., Almansa, L., Felipe, J., and Kimura, E. (2016a). Morphometric Information Reducing Semantic Gap on the Characterization of Microscopic Images of Thyroid Nodules. *Computer Methods and Programs in Biomedicine*, 130(162-174).

Macedo, A. A., Pessotti, H., Almansa, L. F., Felipe, J. C., and Kimura, E. (2016b). Morphometric information to reduce the semantic gap in the characterization of microscopic images of thyroid nodules. *Computer Methods and Programs in Biomedicine*, 130:162–174.

Macedo, A. A., Polettini, J., Baranauskas, J. A., and Chaves, J. (2016c). A health surveillance software framework to design the delivery of information on preventive healthcare strategies. *Submitted with minor revisions to the Journal of Biomedical Informatics*, 62.

Macedo, A. A., Pollettini, J. T., and Munson, E. V. (2015). A Chronic Illness System Using Biomedical Knowledge Sources and Relevance Feedback. In *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*, pages 244–249. IEEE.

Machado Junior, D., Foleiss, J. H., and de Souza, V. M. a. A. (2009). SQAS: Um Sistema Automático de Question-Answering para Textos Jornalísticos. In *7th Brazilian Symposium in Information and Human Language Tecnology*, pages 1–3, São Carlos, SP.

Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Penas, A., Peinado, V., Verdejo, F., and de Rijke, M. (2004). The multiple language question answering track at CLEF 2003. In *Comparative Evaluation of Multilingual Information Access Systems*, pages 471–486. Springer.

Monz, C. (2003). *From document retrieval to question answering*. Inst for Logic, Language and Computation.

Moreda, P., Llorens, H., Saquete, E., and Palomar, M. (2011). Combining semantic information in question answering systems. *Information Processing & Management*, 47(6):870–885.

NLTK (2015). Categorizing and Tagging Words. *On line http://www.nltk.org/book/ch05.html*.

Olvera-Lobo, M. D. and Gutiérrez-Artacho, J. (2015). Question answering track evaluation in TREC, CLEF and NTCIR. In Rocha, A., Correia, A. M., Costanzo, S., and Reis, L. P., editors, *New Contributions in Information Systems and Technologies*, volume 353 of *Advances in Intelligent Systems and Computing*, pages 13–22. Springer International Publishing.

Pessotti, H. (2012). Uso de Mapeamento Conceitual para Redução de Descontinuidade Semântica na Recuperação de Imagens Microscópicas de Carcinoma Tireoidiano. Master's thesis, Universidade de São Paulo.

Pollettini, J., Baranauskas, J. A., Ruiz, E. S., da Graça Pimentel, M., and Macedo, A. (2014). Surveillance for the prevention of chronic diseases through information association. *BMC Medical Genomics*, 7(1):7.

Pollettini, J., Panico, S., Daneluzzi, J. C., Tinós, R., Baranauskas, J. A., and Macedo, A. A. (2012). Using machine learning classifiers to assist healthcare-related decisions: Classification of electronic patient records. *Journal of Medical Systems*, 36(6):3861–3874.

Porter, M. and Boulton, R. (2001). Snowball. *On line http://www.snowball.tartarus.org*.

Prestes, K. V. (2011). Avaliação de métodos de seleção da resposta de um sistema de perguntas e respostas. Technical report.

Ryu, P.-M., Jang, M.-G., and Kim, H.-K. (2014). Open domain question answering using Wikipedia-based knowledge model. *Information Processing & Management*, 50(5):683–692.

Rzhetsky, A., Seringhaus, M., and Gerstein, M. (2009). Getting Started in Text Mining: Part Two. *PLoS Comput Biol*, 5(7):e1000411+.

Shortliffe, E. H. and Cimino, J. J. (2013). *Biomedical informatics: computer applications in health care and biomedicine*. Springer Science & Business Media.

Spackman, K., Campbell, K., and Côté, R. A. (1997). SNOMED RT: a reference terminology for health care. In *Proceedings of the AMIA annual fall symposium*, page 640. American Medical Informatics Association.

Suresh kumar, G. and Zayaraz, G. (2015). Concept relation extraction using Naïve Bayes classifier for ontology-based question answering systems. *Journal of King Saud University - Computer and Information Sciences*, 27(1):13–24.

Voorhees, E. M. and Tice, D. M. (2000). Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 200–207, New York, NY, USA. ACM.

Wallace, R. (2003). The elements of aiml style. *Alice AI Foundation*.

Wilkens, R., Villavicencio, A., Muller, D., Wives, L., Da Silva, F., and Loh, S. (2010). Comunica: a question answering system for brazilian portuguese. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 21–24. Association for Computational Linguistics.

Yen, S.-J., Wu, Y.-C., Yang, J.-C., Lee, Y.-S., Lee, C.-J., and Liu, J.-J. (2013). A support vector machine-based context-ranking model for question answering. *Inf. Sciences*, 224:77 – 87.

Zhang, D. and Lee, W. S. (2003). Question classification using support vector machines. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 26–32, New York, NY, USA. ACM.