QA system to guide citizens to reliable health information

Sarah CRUCHET, Health On the Net foundation
24 rue Micheli du Crest
1211 GENEVE
Sarah.Cruchet@healthonnet.org
Arnaud GAUDINAT
Célia BOYER

Many attempts have been made in the QA domain but no system applicable to the field of health is currently available on the Internet. This paper describes a multilingual QA system adapted to the health domain and more particularly to the detection of the question's model which has a greater effect on the rest of the QA system. Our original hypothesis is that a question can be defined by two criteria: type of expected answer and medical type. These two must appear in the step of detection of the model question and thus, the corresponding answer. For this, questions were searched on the Internet and then given to experts in order to obtain classifications according to criteria mentioned above. In addition, tests of supervised and non-supervised classification were made to determine features of questions. Results of this first step were that algorithms of classification were chosen. They showed that categorizers giving the best results were the SVM. Currently, for a set of 100 questions, 84 are well categorized in English and 68 in French according to the type of expected answer. These figures fall to less than 50% for the medical type.

Evaluations have showed that the system was good to identify the type of expected answer and could be enhanced for the medical type.

It leads us to use an external source of knowledge: UMLS. A future improvement will be the usage of UMLS semantic network to better categorize a query according to the medical domain.

Introduction

The development of Internet places at our disposal, an increasing quantity of information and the medical domain is widely represented on the Web [1]. Along with this grows the interest and concern the average person shows towards information regarding his health, thus creating a bigger demand for information on the Internet [2, 3].

Actually, the main problem is in the access and evaluation of this web information. How much time is needed for you to find the precise answer to your question? How can you be sure it is reliable or not? Indeed, search engines optimised for query keywords, only provide a long list of Web documents, lists that one necessarily has to browse to find an accurate answer.

The answer to these difficulties is a question/answering (QA) system applied to the health domain [4-6] with only trustworthy web documents. The goal of such a system is to extract precise answers to a question submitted in natural form rather than a list of documents. Its role is clearly to help users as quickly as possible towards access to required information instead of having to look and evaluate through all documents proposed by search engines.

The Health On the Net Foundation (HON) is the leading organization in promoting and guiding the deployment of useful and reliable online medical and health information and was awarded the eEurope Award in 2004 in the eHealth category. HONcode is the oldest and the most used ethical and trustworthy code for medical and health related information available on Internet [7]. The HONcode is designed for three target audiences: the general public and the web publisher and the health professional. It is composed by 8 principles [8] which are respected by all certified websites. For facilitating quick access to the most relevant and up-to-date medical discoveries, HON become involved in the development of such a QA system. So, by default the research is done in the database of certified websites as the quality of the available information is paramount, far more so than the quantity.

This research activity is done within the European project PIPS (Personalized Information Platform for life and health Services). The PIPS Project has the main aim to create a new Health and Life Knowledge and Services Support Environment for protecting the health of the Individual. PIPS is an advanced platform for health personalized management using only quality knowledge sources. QA system is an alternative to usual search engine and can fill the gap of the PIPS ontology server and semantic engine which is limited to few domains and not exhaustive.

There are no QA systems applied to health available on the Internet but only studies found in scientific literature [4-6, 9-10]. However, by the 1960s, people were already attempting to come up with a direct answering system to a question submitted in a natural form. Two of the most famous QA systems are BASEBALL (questions on the US baseball league) and LUNAR (on geological analysis of rocks). Nowadays, general QA systems are available on the Web like AnswerBus or BrainBoost [11].

Most QA systems use pattern matching to classify sets of questions [12, 13]. However machines learning is known to be capable of identifying data an expert would not have thought of, whatever the language [4, 14]. The questions classification is the main issue in QA system using machine learning.

The application developed by HON is available in English, French and Italian. In addition, the user is able to choose the domain of research i.e. Websites certified by HON or all the websites, regardless of certification. **Figure1.** presents the architecture of the QA system. It is composed of five modules. The module “QuestionAnalyser” takes place at the beginning of the QA system. **Figure1.** shows the importance of this module for the continuation of the treatment of a question. Indeed, all modules describe in **Figure1**. use its results to carry out their treatment to seek and retrieve the answer.

Our hypothesis is that a medical question is characterized by its medical type and its type of expected answer. The questions classification is the main issue in QA system using machine learning [9, 14].

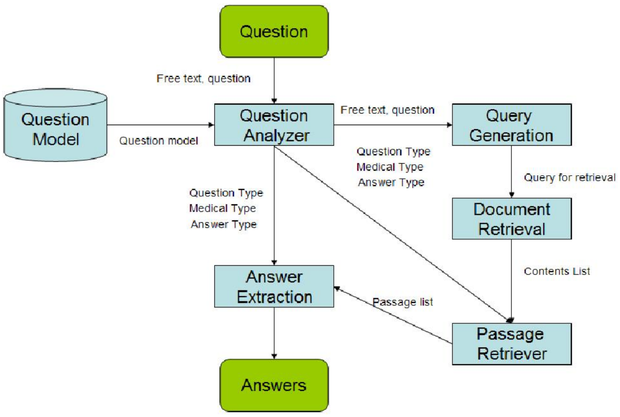

Figure 1. Architecture of the questions/answering system

Methods & Results

In a first step we preferred to classify automatically questions, with no information about their categories. For this task we used an environment for machine learning.

Initially, the questions were presented in the untreated form (called unigram form). Several algorithms of clusterisation were tested (K-Means, EM, DBSCAN, Kménoids, and Kernel-KMeans).

Following this, questions were proposed with bigrams of words or co-occurrences. The bigrams of words allow identifying the expressions composed. The co-occurrences associate the distant words within a sentence belonging to the same linguistic unit. It is considered that there is co-occurrence when the presence of a word in a text gives an indication on the presence of another word. The interest to apply such work to the set of questions is to locate the expressions, associations of words, which are recurring on the entire corpus.

Non supervised classification did not emphasize new classes of questions. However, we could observe that the software creates clusters according to the type of interrogation term. For instance, all questions beginning by “What is” were in the same cluster. Consequently, this work was not useless and has prepared the step of conception by providing distinctive structures of sentences and useful expressions.

Two corpora were used to define the types of questions: one of 136 questions in English and one of 140 questions in French. In both cases, the questions were collected from Frequently Asked Questions (FAQ) found in forums specialized in health as well as in discussions on the Internet. We tried to retrieve questions covering as much as possible of the wide spectra of the subject within the topic while respecting patients’ interrogations. These sets of questions were then given for classification by experts

Table1. and Table2.: Question repartition for the medical types and the types of answer expected.

	En	Fr	It
Causes	8	9	13
Diagnostic	6	2	1
Diet	18	14	14
Disease	25	12	0
Evolution	12	11	6
Physiology	15	11	10
Prevention	6	8	4
Routine	10	34	22
Symptoms	13	8	3
Treatment	32	55	23

	En	Fr	It
Boolean	53	65	26
Causal	17	5	3
Definition	47	12	5
Factoid	6	13	23
Duration	4	3	0
List	7	15	14
Moment	4	6	4
Person	7	4	1
Place	9	4	1
Procedure	4	30	5
Quantity	15	7	14

Table1. and **Table2.** give this manual classification. It has brought out 10 medical types and 11 answer types, categories for both English and French. Italian has been added later and existing categories were used.

According to human experts, whatever the language considered, questions most usually asked are related to the treatment of a disease. For the type of expected answer, **Table1**. and **Table2**. show that the most representative categories are “Boolean” and “Definition” for English and French, “Boolean” and “Factoïd” for Italian. It is due to the data-gathering which is based on questions most frequently asked with no consideration of proportion (actually categories were unknown during the data-gathering for English and French). The characterization according to the type of answer expected found in our research matches the one described in the literature [15-17]. However we decide to re-define categories because of the specificity of our study which is based on medical questions only. We assumed that the medical type of the question provides the context, context which will be helpful to identify answers in documents.

In the last step we classified questions with a supervised approach. Questions were gathered according to the expert classification. The tool learns according to this classification and generates a model. This model was applied to the same set of questions, following which cross validations were carried out. The five algorithms tested are those found in the literature for the task of classification [9]: SVM, NaiveBayes, Knn, ROCCHIO, Decision Trees.

Supervised classifications have been realized to found which classifiers would be integrated to the system. **Table3** shows that classifiers are highly capable of performing the task of classification according to the type of answer expected. However, the research of the medical type can be enhanced using the Unified Medical Language System (UMLS) [10]. UMLS gathers medical terms in different languages and maps these terms together.

Table3. Summarization of the classifiers' results

	maR	*maP*	*maF1*	*miR*	*miP*	*miF1*	*Erreur*
English Med type (SVM W2*)	0.47717	0.52499	0.476617	0.69446	0.410362	0.410362	0.078078
English Ans type (SVM W2*)	0.80527	0.84569	0.81519	0.77891	0.91934	0.84031	0.026561
French Med type (SVM W2*)	0.47505	0.51217	0.46431	0.312687	0.62901	0.41354	0.086006
French Ans type (NB W3*)	0.7744	0.68595	0.68028	0.71381	0.63112	0.66846	0.066063
Italian Med type (SVM W2)	0.6315	0.6468	0.639	0.4367	0.736	0.5481	0.0802
Italian Ans type (SVM W2)	0.7634	0.7104	0.7359	0.5397	0.8245	0.6523	0.0577

* Wx = x grams of words, Med = Medical and Ans = Answer

Seven measures were used: macro (ma) and micro (mi) recall(R: Capacity of the system to report only relevant documents)/precision(P: Rate of relevant documents proposed by the system as compared to all the retrieved documents)/F1 and the error. Macro-measures compute values of precision/recall for each category and make a mean on these values. Micro-measures gather data of each category in a same contingency table and compute values of precision/recall according to this table. The F1-measure is a function that is maximized when the precision and the recall are near. We attached more importance to the values of precision than to those of recall because it is necessary that the composition of the categories proposed by the classifier coincide with that of the human experts. Finally, this set of measures has been calculated according to units of treatment (unigram, bigrams, trigrams and co-occurrences of words).

The tests of supervised classifications have made it possible, both to validate the categories of questions provided by the expert and to determine the best algorithm for the step of implementation. It has been realized according to both categories for medical types and categories for the type of answer expected.

The results show that a question is better classified if it belongs to a category well represented in the initial corpus. For instance, 91 % of "Definition" questions (34.3% of the English's corpus) are well-classified with the NaiveBayes (NB) classifier whereas it falls to 60% for the category "Procedure" (11.2% of the English's corpus).

Tables3. gives results for the medical type and the type of answer expected, with the best classifiers and units of treatment. As results found in literature [14], the two algorithms which are most relevant for the task of classification are NaiveBayes and Support Vector Machine (SVM) with almost 85% of micro-precision for the type of answer expected in English. Moreover, results are better for the type of answer expected. Indeed the values of recall and precision are around 80% against less than 50% for the determination of the medical type of the question (cf **Table3**). It can be explained by the lack of data for the medical types. Furthermore, the number of questions by category is better distributed for the type of answer expected. Actually some categories are very poor. For instance, "Diagnostic" is only represented by 2 questions in the French's corpus. In addition, this imbalance can be explained by the complexity to establish a model for the medical types. Two questions having the same category for the type of answer expected have a common sentence structure. It is not the case for questions belonging to the same medical category. For instance the two questions following belong to the category "Procedure" for the type of answer expected: "**How can** diabetes affect my mood?" and "**How can** I take care of myself if I have diabetes?". They have a common structure: "How can + subject + verb+ complement". Let's take two questions belonging to the medical category "Cause": "Am I at risk for diabetes?" and "How do I know if my kidneys are **affected** ". There are no commons structures. So more difficult is the machine learning task for categorizing the medical type.

Globally, the automatic classification is better for English than for French. **Table3.** shows that for the type of answer expected the system is capable at 84% for English, 68% for French and 71% for Italian (macro-precision). It can be explained by the linguistic differences between French or Italian and English (e.g. conjugation).

Conclusion & Perspectives

For the realization of a QA system applied to health we decided to use a supervised method to characterize questions whereas most of the QA systems are using pattern matching. From set of questions and with the expert classification we have proposed a new taxonomy of clinical questions for QA systems whereas the non supervised classification of this set of questions was not efficient.

The evaluation of the Question Analyzer module for the task of classification confirms the feasibility of the supervised method for a QA system for both English and French. Indeed, SVM classifiers are highly capable of performing the task of classification according to the type of answer expected.

The careful analysis of results obtained and errors explains that the less well recognised classes are the one that are less well represented in our learning corpora. Thus we have to enhance our training base for English, French and Italian. And future research which overcome results obtained by the determination of the medical type will be focussed on the usage/integration of the Unified Medical Language System (UMLS) network. This language provides precise relations between medical terms and many synonyms for each of them.

References

[1]Hege K Andreassen & AL, European citizens' use of E-health services: A study of seven countries, BMC Public Health 2007

[2]S. Fox, http://www.pewinternet.org/PPF/r/156/report_display.asp, *Health information online*. Washington, DC: Pew Internet & American Life Project; 2005.

[3]ML. Ybarra & AL, Help seeking behaviour and the Internet: A national survey. *Int J Med Inf* 2006

[4] P. ZWEIGENBAUM, Question Answering in Biomedicine, in DE RIJKE M., WEBBER B., Eds., ACL, p. 1–4, 2003.

[5] M. LEE & AL, Beyong information retrieval-Medical Question Answering, AMIA, 2006.

[6]Y. NIU, Question Answering in Medicine, Technical Report for the Oral Qualification Exam, Department of Computer Science, University of Toronto 2003

[7]Selby M, Boyer C, Jenefski DA, Appel RD. *Health On the Net Foundation Code of Conduct for Medical and Health Websites.* MedNet 96 - European Congress on the Internet in Medicine, Brighton, U.K., Oct. 14 to 17, 1996.

[8] http://www.hon.ch/HONcode/Conduct.html

[9] H. YU & AL, Classifying Medical Questions based on an Evidence Taxonomy, AAAI, 2005

[10]T. DELBECQUE & AL, Indexing UMLS Semantic Types for Medical Question-Answering, ENMI, 2005.

[11] AnswerBus: http://www.answerbus.com/index.shtml; BrainBoost: http://www.brainboost.com/

[12]P. JACQUEMART & AL, Towards a Medical Question-Answering System: a Feasibility Study, Stud Health Technol Inform. 2003.

[13]E. ALFONSECA & AL, A prototype Question Answering system using syntactic and semantic information for answer retrieval, TREC, 2002.

[14] R. MAY & AL, Building a Question Classifier for a TREC-Style Question Answering System, The Stanford Natural Language Processing Group, Final Projects 2004.

[15]F. BENAMARA & AL, Construction de réponses coopératives. Revue Québéquoise de Linguistique, Les Éditions David, Ottawa - Canada, V. 18 N. 3, p. 34-59, septembre 2004.

[16]S. MENDES & AL, L'analyse des questions : intérêts pour la génération des réponses, TALN. Dans : TALN- Workshop question-réponses, Fes (Maroc), 22/04/04-22/04/04, Association pour le Traitement Automatique des Langues (ATALA), p. 11-22, mars 2004.

[17]O. FERRET & AL, How NLP can improve Question Answering. Knowledge Organization, Vol.29, n°3-4, 2002.