Supervised approach to recognize question type in a QA system for Health

Sarah CRUCHET, Arnaud GAUDINAT, Célia BOYER
Health On the Net Foundation, Geneva, Switzerland

Abstract

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 bilingual French/English question answering system adapted to the health domain and more particularly the detection of the question's model. Indeed, the Question Analyzer module for identifying the question's model has a greater effect on the rest of the QA system.

Our original hypothesis for the QA is that a question can be defined by two criteria: type of answer expected and medical type. These two must appear in the step of detection of the model in order to better define the type of 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 such as type of question and type of medical context as mentioned above. In addition, tests of supervised and non-supervised classification were made to determine features of questions. The result of this first step was that algorithms of classification were chosen.

The results obtained 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 answer expected. This figures fall to less than 50% for the medical type. Evaluations have showed that the system was good to identify the type of answer expected 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.

Keywords
Model detection, Supervised classification, Medical Type

1. Introduction

The development of Internet places, an increasing quantity of information at our disposal. The main problem resides in the access to this information. How much time is needed for you to find the precise answer to your question? 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 alternative to this difficulty is a question/answering system. The goal of such 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 through all documents proposed by search engines.

With the progress in the medical field and better access to healthcare, the average life span is constantly growing. 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. The medical domain is widely represented on the Web but how can a common user judge the relevance of medical documents?

It is to answer this double problem that the Health On the Net Foundation (HON), a leading organization in promoting and guiding the deployment of useful and reliable online medical and health information became involved in the development of a question/answering system specific to the medical domain, which is a complex domain requiring the limitation of the framework of research. 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 [1, 2, 3, 4, 5]. 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 [6].

Most QA systems use pattern matching to classify sets of questions [7, 8]. However machines learning is known to be capable of identifying data an expert would not have thought of, whatever the language [3, 9]. The questions classification is the main issue in QA system using machine learning. Annually competitions like TREC or CLEF aim to evaluate such systems.

The application developed by HON is available in both English and French. In addition, the user is able to choose the domain of research i.e. Websites accredited by HON or all the websites, regardless of accreditation. By default the research is based on the accredited sites as the quality of the available information is paramount, far more so than the quantity. **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.
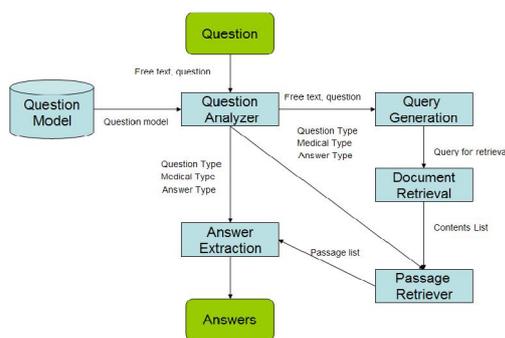

Figure 1. Architecture du système de questions/réponses

The following sections present the material and methods used to classify medical questions and the results obtained for the supervised method.

2. Materiel and Methods

2.1 Corpora

Two corpora were used for this study: the first was composed of 136 questions in English and the second, 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 discussions on the Internet. We tried to retrieve questions covering as much as possible of the wide spectra of the subject within the topic. The medical field being very vast, it was agreed to focus only on the topic diseases. The project PIPS targeting the diabetic's patients, and the module of analysis of questions carried out was also targeted on this disease. We make the assumption that this module is independent of the type of medical subject. The choice to study only one topic makes it possible to only categorize the relevant classes and not the medical topic.

2.2 Non supervised method

In a first step we preferred to classify automatically questions, with no information about their categories. For this task we used the software Rapid Miner which is an environment for machine learning. It consisted of automatically detecting clusters only from the corpora according to the characteristics of the questions. Moreover, according to the algorithm used, it is possible to fix or remove the number of classes of exits.

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 to 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.

2.3 Classification by an expert

The two sets of questions used for this study were provided by two physicians who classified them in two different ways: according to the type of answer expected and

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

	En	Fr
Causes	8	9
Diagnostic	6	2
Diet	18	14
Disease	25	12
Evolution	12	11
Physiology	15	11
Prevention	6	8
Routine	10	34
Symptoms	13	8
Treatment	32	55

	En	Fr
Boolean	53	65
Causal	17	5
Definition	47	12
Factoid	6	13
Last	4	3
List	7	15
Moment	4	6
Person	7	4
Place	9	4
Procedure	4	30
Quantity	15	7

according to the medical type of the question. For the medical questions, the medical expert classified the questions according to the medical language used, while focusing on the medical meaning of the words. The classification by categories suggested corresponds to topics proposed in medical text books.

Table1. and **Table2.** give this manual classification. It has brought out 10 medical types and 11 answer types, categories for both English and French.

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". 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).

The characterization according to the type of answer expected found in our research matches the one described in the literature [10, 11, 12]. However we decide to re-define categories because of the specificity of our study which is based on medical questions only.

2.4 Supervised method and criteria of judgment

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 [3]: SVM, NaiveBayes, Knn, ROCCHIO, Decision Trees.

This method was carried out in two phases: according to the medical categories and the type of expected answers. In addition, criteria of judgment were used to apply cross validation. The corpora were divided into two sets: the training base (90% of the corpus) and the one of tests (10%). Cross evaluation have been realised. This was done to allow us to calculate a score of performance of the automatic tools and its relevance for the task. 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

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

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

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).

3. Results and Discussion of the supervised classification

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 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 [13], 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 the machine learning task is more difficult 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 of 84% for English and 68% for French (macro-precision). It can be explained by the linguistic differences between French and English (e.g. conjugation).

4. Conclusion and 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 both English and French. 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] P. ZWEIGENBAUM, Question Answering in Biomedicine, in DE RIJKE M., WEBBER B., Eds., ACL, p. 1–4, 2003.
[2] M. LEE & AL, Beyong information retrieval-Medical Question Answering, AMIA, 2006.
[3] H. YU & AL, Classifying Medical Questions based on an Evidence Taxonomy, AAAI, 2005.
[4]T. DELBECQUE & AL, Indexing UMLS Semantic Types for Medical Question-Answering, ENMI, 2005.
[5]Y. NIU, Question Answering in Medicine, Technical Report for the Oral Qualification Exam, Department of Computer Science, University of Toronto 2003.
[6] AnswerBus: http://www.answerbus.com/index.shtml; BrainBoost: http://www.brainboost.com/.
[7]P. JACQUEMART & AL, Towards a Medical Question-Answering System: a Feasibility Study, Stud Health Technol Inform. 2003.
[8]E. ALFONSECA & AL, A prototype Question Answering system using syntactic and semantic information for answer retrieval, TREC, 2002.
[9] R. MAY & AL, Building a Question Classifier for a TREC-Style Question Answering System, The Stanford Natural Language Processing Group, Final Projects 2004.
[10]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.
[11]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.
[12]O. FERRET & AL, How NLP can improve Question Answering. Knowledge Organization, Vol.29, n°3-4, 2002.
[13]H. YU & AL, Classifying Medical Questions based on an Evidence Taxonomy, AAAI, 2005.