The New Navigators: from Professionals to Patients
R. Baud et al. (Eds.)
IOS Press, 2003

Towards a Medical Question-Answering System: a Feasibility Study

Pierre Jacquemart, Pierre Zweigenbaum

STIM/DSI, Assistance Publique – Hôpitaux de Paris, France

Abstract

Question-answering (QA) systems, as have been presented and evaluated in several TREC conferences, are the next generation of search engines. They combine 'traditional' Information Retrieval (IR) with Natural Language Processing (NLP) and Knowledge Engineering techniques to provide shorter, more precise answers to natural language questions. We study here the feasibility of such a system for French in the health care domain. In this purpose, we collected a corpus of student questions in oral surgery. We examined two enabling conditions: on the IR side, how to select the right keywords in a question to identify relevant material on the Web for answering this question, a prerequisite for success; and on the NLP side, whether the contents of the questions fit the conceptual model of an existing QA prototype, a favorable condition for rapid implementation. A manual Web search enabled us to devise automatable principles for building IR queries for these questions. Besides, we could design a semantic model, using UMLS Semantic Network relations, which is consistent with our prototype and covers 90% of the questions. However, the high specialization of the domain and the clinical orientation of the questions, joined with the more limited resources online in the French language, may restrain the quantity of Web material available for answering these questions.

Keywords:
Natural Language Processing; Information Retrieval; World Wide Web; Language; France; Algorithms

1 Introduction

The Internet has made available a tremendous wealth of information, very little of which would be accessible without information retrieval (IR) search engines [1]. In these, users express their information needs in the form of a few keywords, and the system returns a ranked set of documents. It is however the responsibility of the user to choose keywords that are both relevant (specific of the information need) and non-ambiguous (polysemy is a source of noise), and to use boolean combinations as best suited. Besides, since full documents are returned, the user must also look for the passages which contain the expected information.

Question-answering (QA) systems (*e.g.*, http://trec.nist.gov/presentations/TREC10/qa/, [2]) propose an evolution over conventional IR systems. They take as input a natural language question (*e.g.*, *"Which is the highest volcano in Europe?"*) and must return a short passage or even the precise words (*e.g.*, *"Mt Etna"*) that provide the answers. They combine techniques from the fields of IR, Information Extraction and more broadly Natural Language Processing (NLP). A common architecture for a QA system includes the analysis of the question, the identification of appropriate keywords for the search (through a conventional IR engine) for potentially relevant documents, and the selection and ranking of document passages containing the answer. QA research has been stimulated by the QA tracks hosted by the Text Retrieval Evaluation Conferences (TREC) since 1999. TREC-QA

involves open-domain questions, with an evolution towards shorter, more precise answers (250 then 50 characters then exact words), drawn from newspaper article collections and increasingly from the Web. Questions are mainly factual or require simple reasoning, but will also evolve towards more complex, speculative types and contextual interaction, which are still difficult to handle [2].

Medicine cannot stay aside of this evolution in information technology. This topic though, to our knowledge, has not yet been explored. We therefore started to investigate the feasibility of QA in the medical field. QA can help access precise information amidst long Web pages and facilitate the compilation of results coming from different sources. This could bring significant time savings to students, clinicians and even the general public, often rebuked by the complexity of IR.

Turning to medical QA will bring along medical specificities. The medical field uses a specific language, with specialized terminology and possibly specific syntax. Special care will have to be given to the reliability of information, using quality criteria such as NetScoring (http://www.chu-rouen.fr/netscoring/netscoringeng.html) or directly pooling from manually checked site directories such as CISMeF (http://www.chu-rouen.fr/cismef/) [3]. And as in more traditional IR, the user should be able to ask for specific types of documents (*e.g.*, textbooks, practice guidelines, information to the general public; see, *e.g.*, CISMeF's list of resource types).

However, a first issue we want to address is whether relevant documents can be found for questions in the medical domain. Besides, preliminary work has been conducted in our team and has resulted in a prototype QA system [4]. This prototype relies on a limited set of predetermined question types, which can be modeled as semantic triples [A]–(R)–[B] (we illustrate this notation below). We also wanted thus to know whether medical questions could be modeled in this framework.

We therefore set a usage context: the case of student questions in the domain of oral pathology, for which we collected a corpus of questions in French. We checked the existence of answers on the Web, studying methods to select appropriate search keywords for sending IR queries. We then examined the ability of these question types to integrate our current QA prototype. The proportion of questions that can be modeled in its framework provides a feasibility estimate for their automatic processing.

2 Material and Methods

We first describe the preparation of the corpus of questions, then the strategy for building IR queries from questions and evaluating the presence of answering material on the Web, and finally the semantic modeling of question contents.

2.1 Corpus of Questions

The majority of our questions were collected from clinical students; for better coverage of the variety of the domain, we completed the sample with questions derived from student textbooks for preparing class exams in stomatology [5]. The formulation of the questions collected is more or less explicit. For instance, *"How does one recognize a lichen planus?"* (for ease of understanding, we present English translations of our French questions) is more explicitly stated as *"Which signs characterize lichen planus?"*. This lead us to convert each question into a canonical form. This form is meant to help derive general question patterns, to which we return below (2.3). Complex sentences were simplified when possible into more direct questions (e.g., *"My patient is diabetic; can I use vasoconstrictors?"* becomes *"Does diabetes contraindicate vasoconstrictors?"*). Some context-dependent questions

were made more precise by instanciating their context, *e.g.*, *"How does one recognize a lichen planus?"* becomes *"How does one recognize an oral lichen planus?"*. Finally, we put aside nested questions and other context-dependent questions, *such as "Is this image a cyst?"*.

2.2 Looking for Answering Material on the Web

Keyword selection To elicit keywords, we first collect 'named entities': in open-domain QA, these are names of persons, of places, of companies, quantities and dates. Here, they are mainly 'technical' names: diagnoses, signs, treatments, etc. To allow for flexibility and augment recall, we use morphologically derived words (*"diabetes"*, *"diabetic"*), synonyms and hypernyms (more general terms). We also add terms which describe the semantic relation implied in the question, and some terms implied by the domain. For instance, in the question *"Why does one put dental implants?"*, the term *"indication"* is added to keywords *"implant"* and *"dental"*. The result corresponds to a question of the type *"What is the indication of treatment T?"*

Query formulation strategy Our heuristic for building queries consists in varying the number of keywords then in applying the preceding method, trying named entities and relation name first, then derivation and synonyms, and finally domain-context terms. Search engines provide operator AND and sometimes NEAR (Alta Vista), which is useful to constrain the proximity of search terms in the answer (the default behavior of search engines is to work with 'bags of words' [6], discarding order and distance information). We did not use the disjunction operator OR, which might be useful to factor several term variants, because it is not available in many search engines; we preferred to keep tighter control of the queries, evaluating their results and refining the quantity and quality of answers. An excess in answering documents is handled by grouping words into 'frozen expressions' (within quotation marks), then by adding a conjunct term (AND operator). A lack of answers may be solved by using derived words, synonyms and hypernyms then by deleting from the query the term judged 'least strategic', keeping only the key terms.

This kind of procedure has been used by Moldovan and Harabagiu in their QA system [7], using WordNet [8] for synonyms and hypernyms. It can thus be automated provided that equivalent resources are available for French. For terms in the medical domain, hierarchical terminologies such as MeSH, ICD-10 and SNOMED provide such resources for French.

Using search engines The queries were submitted to several search engines, using the above-mentioned heuristic. We tried to obtain about 30 documents for each page (a quantity often deemed relevant by librarians) and at least one relevant document in the top five documents (the first TREC-QA evaluations allowed five answers to be returned). Result documents were checked manually to find a passage answering the question in a consistent context.

2.3 Design of a model for question contents

We modeled the form of each question as syntactico-semantic patterns so as to identify regularities and model their semantic contents. These patterns were obtained by generalizing the above-mentioned canonical forms to generic categories of the domain. Signs are generalized into an abstract category as OBS, so that our example question *"Which signs characterize lichen planus?"* yields the pattern *"Which OBS diagnoses a PATO?"* (pathology). We finally reach a semantic model for the question by identifying the relevant semantic relation; we draw here from the UMLS Semantic Network relations [9]. In the above example, relation *"diagnoses"* (R5.6, child of *"conceptually-related-to"*) may represent the meaning of *"characterize"*. The compatibility of this relation, as defined

in the UMLS, with the semantic categories of the question, must also be checked. The semantic model obtained for question *"How does one recognize a lichen planus?"* is finally represented as a triple [OBS]–(R5.6)–[PATO]; the specific focus of the question, which is on the OBS, can be symbolized in the triple as [which OBS]–(R5.6)–[PATO].

3 Results

We collected one hundred questions involving pathology, procedures, treatments, examinations, indications, diagnosis and anatomy.

3.1 Web search

We performed a manual search for answering documents for each question according to the above-mentioned strategy, using the following search engines: AltaVista, Google, AlltheWeb, Lycos, Mirago, Kartoo, MetaCrawler, Vivissimo, QueryServer and MedHunt. We found it easier to obtain answers to this question set with Google. Although it has a very large index, its actual coverage of the Web is not known, nor is its specific capability to find medical pages. It is, though, the search engine from which the CISMeF health catalog receives the largest number of referred users [10]. A query size of 2 to 5 keywords enabled us to reach the best results: among one hundred questions, 60% obtain relevant results in the top five pages. However, for the remaining 40%, thorough manual search proved unable to obtain relevant answering documents within the top five hits.

3.2 Categorization of questions

We obtained 66 different syntactico-semantic patterns (many are very similar) which can be categorized into 8 broad semantic models. Three of these models fit the semantic triple representation [A]–(R)–[B] with a modality *"which"*, *"does"* or *"why"* (see table 1). These three models account for 90 out of 100 questions in our corpus.

Table 1: Semantic models of questions

	Semantic model	# synt-sem patterns	# questions
1	[which X]–(r)–[B]	24	39
	[A]–(r)–[which Y]	17	29
2	does [A]–(r)–[B]	12	17
3	why [A]–(r)–[B]	4	5
4	[which X, Y]–(r)–[B]	1	1
5	[which X]–(r)–[B, C]	3	3
6	duration [A]–(precedes)–[B]	2	2
7	define [A]	1	2
8	which specific precaution if [A]–(r)–[B]	1	1
	total	66	100

Model 1 has two reciprocal formulations; it is expressed through 41 syntactic patterns, which cover 68 questions out of 100. It corresponds to the search for an unknown term in the triple. For instance, *"The Reed-Sternberg cell evokes which disease?"* has syntactico-semantic pattern *"An OBS diagnoses which PATO?"*; its semantic relation is *"diagnoses"* (R5.6) and its semantic model [OBS]–(diagnoses)–[which PATO]. Model 2 checks the validity of the relation in the triple and accounts for 12 patterns found in 17 questions. Model 3 asks for an explanation about a triple, which models a treatment for a given pathology in all 5 questions involved. These first three models neatly fit our triple-based representation. The rest involve more complex constructs: multiple concepts as in models 4 and 5 (*e.g.*, *"Which difference between granuloma pyogenicum, 'diapneusie', epulis?"*),

specifically focussed questions as in 6 and 8 (associations of events), definition as in 7 (*e.g.*, *"How does the TNM classification work?"*).

Schematically, we can then categorize the questions into two classes: those that fit our simple, triple-based representation easily, and those for which a more complex model is needed. The first class, which accounts for 90% of our sample, qualifies for being handled by our prototype QA system, whereas the rest will need more development.

4 Discussion

This study examined two key aspects involved in a question answering system: Web search and question categorization. It showed that given a sample of 100 student questions, a large proportion should be amenable to automatic processing within our model. Several points must however be emphasized.

It was necessary for an initial, manual study to start with a manageable set of questions; however, one hundred questions can only be a starting point for such a study, and cannot cover all question types. Note though that the size of a typical set of questions in TREC-QA is 500, not many times more than our study set.

Web search for our sample questions only managed to obtain 60% relevant documents in the top five hits. A reason may be that the conjunction of our two requirements, oral pathology specialty and French-language documents, is insufficiently represented on the Web. English documents may be more numerous on this topic, which suggests that cross-language question-answering might be specifically useful here. The absence of a NEAR function in some search engines is also a limitation, since documents with all keywords grouped together may not be ranked in the top hits. Looking into the top 200 hits, as is often done in QA systems [2], might relax constraints and enable to select more relevant documents; this can be done when these queries are tested with our prototype. Besides, questions may have several answers and their validity depends on context; furthermore, expert judgement does not always match user opinion [11]. Finally, we may note that on the full task of open-domain QA on collections of press articles, TREC-QA systems obtained 70% accuracy [7, 12].

Our triple-based model ([concept]–(relation)–[concept]) accounts for a vast majority of questions. Automatizing the conversion of questions into a canonical form needs more research; this might prove difficult for some of the questions in their present form, *e.g.*, *"My patient is diabetic; can I use vasoconstrictors?"*, since this would require much domain knowledge and reasoning. Regularizing questions into syntactico-semantic patterns then into semantic models corresponds to a generalization, thus to some reduction of information; but it should help to identify answering material with more flexibility. Relying on UMLS relations requires to find a good fit between natural language terms and these relations. Using UMLS semantic types for concepts would be a natural follow-up, since it would enable us to take advantage of the UMLS relations usage constraints, which are expressed as triples.

The questions which fit this triple-based model can be integrated into our prototype [4]. For the rest, we shall need to study how to extend the model or to resort to other QA methods. Further work will also be needed to identify the question types for which answers can be found on the Web depending on domain and technicity.

5 Conclusion

The future generation of search engines will probably integrate question-answering facilities. Work around TREC-QA shows promising solutions. To study such a system in a medical domain, we collected a corpus of student questions in oral surgery. On the one hand, we were able to fit most of these questions into a simple domain model which will facilitate their handling by our QA prototype. On the other hand, experimenting with a search methodology which can serve as the basis for automatization of the document retrieval part of the QA process, we found that French medical documents for answering these specific questions were scarce on the Web. This phenomenon might occur with other medical domains and contrasts with the general domains (*e.g.*, general knowledge, tourism) mainly addressed in the TREC evaluations. In QA as in IR, one cannot search for every kind of information on the Web, since they are not equally represented.

Bibliography

[1] Baeza-Yates R and Ribeiro-Neto B. *Modern Information Retrieval.* Addison-Wesley, New York, 1999.

[2] Harabagiu S and Moldovan D. Tutorial on open-domain textual question answering. In: Proc 19th COLING, Taipei, Taiwan. 2002.

[3] Darmoni SJ, Leroy JP, Thirion B, et al. CISMeF: a structured health resource guide. *Methods Inf Med* 2000;39(1):30–5.

[4] Dalmas T and Rivoallan R. Système de question-réponse dans le domaine médical. Projet de DESS, Intelligence Artificielle, Université Paris 6, 2002.

[5] Achard JL. *Tests – 740 questions réponses.* CdP, Paris, 1990.

[6] Salton G. *Introduction to Modern Information Retrieval.* Mc Graw Hill, Singapore, 1987.

[7] Moldovan D, Harabagiu S, and Surdeanu M. Performance issue and error analysis in an open-domain question answering system. In: Proc 38 ACL, Philadelphia, PA. ACL, 2002.

[8] Fellbaum C, ed. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA, 1998.

[9] McCray AT and Nelson SJ. The semantics of the UMLS knowledge sources. *Methods Inf Med* 1995;34(1/2).

[10] Zweigenbaum P, Darmoni SJ, Grabar N, Douyère M, and Benichou J. An assessment of the visibility of MeSH-indexed medical web catalogs through search engines. *J Am Med Inform Assoc* 2002;8(suppl):954–8.

[11] Voorhees EM. Variations in relevance judgments and the measurement of retrieval effectiveness. In: Croft WB, Moffat A, van Rijsbergen CJ, Wilkinson R, and Zobel J, eds, Proc 21th ACM SIGIR, 1998.

[12] Soubbotin MM and Soubbotin SM. Patterns of potential answer expressions as clues to the right answers. In: Proc 10th Text Retrieval Conference, Gaithersburg, MD. NIST, 2001:175–82.

Address for correspondence

Pierre Jacquemart, Mission de recherche en Sciences et Technologies de l'Information Médicale (STIM), DSI, Assistance Publique–Hôpitaux de Paris, 91, boulevard de l'Hôpital, 75634 Paris Cedex 13, France
E-mail: pja@biomath.jussieu.fr Url: http://www.biomath.jussieu.fr/