Answering Clinical Questions by Identifying Roles in Medical Texts

Yun Niu, Graeme Hirst *, Gregory McArthur,

Patricia Rodriguez-Gianolli

Department of Computer Science

University of Toronto

Toronto, Ontario, Canada M5S 3G4

*Corresponding author. Fax: +1 4169781455. Email addresses: yun, gh, gregm, prg@cs.toronto.edu

Abstract

If high-quality evidence is available rapidly at the point of care, clinicians will use it in their decision-making. In the context of a broader project of medical text retrieval, we have developed an approach to medical question-answering with natural language analysis. We analyze the limitations in the medical domain of the technologies that have been developed for general question-answering systems, and describe an alternative approach whose organizing principle is the identification of semantic roles in both question and answer texts that correspond to the fields of PICO format.

Keywords: Natural language processing; Question-answering; Semantic roles; Evidence-based medicine; Information retrieval

1 Motivation

In every aspect of patient treatment, questions arise for which a search of the published medical evidence is appropriate, as it is likely that the answer has already been found from the work of other clinicians. For example:[1]

Q: In a child with asthma, do increased doses of inhaled corticosteroids lead to a decrease in growth?

A: Growth was significantly slower in the group receiving higher dose inhaled steroids (3.6 cm, 95% CI 3.0 to 4.2 with double dose beclometasone v 5.1 cm, 95% CI 4.5 to 5.7 with salmeterol v 4.5 cm, 95% CI 3.8 to 5.2 with placebo). [3]

Studies have shown that searching in the literature can help clinicians in answering questions generated in patient treatment [7, 4, 12]. It has also been found that if high-quality evidence is available in this way at the point of care—for example, the patients' bedside—clinicians will use it in their decision making, and it frequently results in additional or changed decisions [16, 19]. But speed is very important. Investigation of potential end-users has shown that physicians need access to the information within 30 seconds, and that if the search takes longer, it is likely to be abandoned [20].

The practice of using the current best evidence to help clinicians in making decisions on the treatment of patients is called Evidence-Based Medicine (EBM). Finding relevant evidence is a typical question-answering (QA) problem in the medical area. There are many QA systems that have achieved some success in other domains, such as finding the answer to "factoid" questions in a corpus of general news stories, as in the question-answering track of recent Text Retrieval Conferences [21]. However, we have found that there are large differences between general QA (GQA) and medical QA (MQA) (see section 3 below). This paper analyzes the challenges of applying QA technology to answer clinical questions automatically, and then describes our on-going work on the problem.

[1] All the examples in this paper are taken from a collection of questions that arose over a two-week period in August 2001 in a clinical teaching unit at the University of Toronto.

2 The EPoCare Project

2.1 Background

Our work is part of the EPoCare project ("Evidence at Point of Care") at the University of Toronto. The project aims to provide fast access at the point of care to the best available medical information. Clinicians will be able to query sources that summarize and appraise the evidence about the diagnosis, treatment, prognosis, etiology, and prevalence of medical conditions. In order to make the system available at the point of care, the question-answering system will be accessible using hand-held computers. The project is an interdisciplinary collaboration that involves research in several disciplines. Project members in Industrial Engineering and Cognitive Psychology are investigating the design of the system through a user-centered design process, in which requirements are elicited from end users who are also involved in the evaluation of prototypes. Project members in Knowledge Management and Natural Language Processing aim to ensure that the answers to queries are accurate and complete. And project members in Health Informatics will test the influence of the system on clinical decision-making and clinical outcomes.

The approach taken to building the question-answering system revolves around the use of key concepts. Though this approach began with the particular application in Evidence-Based Medicine that we have described, it can be seen as a general methodology for building a question-answering system. In this methodology, key concepts in the domain of the application, which may come from an ontology for the domain, drive the system's understanding of both queries and data sources.

In the system as it is currently implemented, queries and retrieval are keyword-based, thus it is similar to the Query-By-Example query language for databases. Queries consist of several fields, each with a name and a value to be filled in by the user. The names are key concepts in the domain, and the user is intended to fill the values of the fields with instances of the concept given by the field category.

Information sources are understood in terms of the same key concepts. Information sources may come in free-text format, or they may be semi-structured or structured. As described in section 2.3, the current implementation focuses on semi-structured data sources in XML format.

Presently, the identification of key concepts is done manually. The goal of the work that we will report in the later sections of the paper is to automate and improve this process through the use of natural language understanding techniques to automatically identify instances of the key concepts in queries and in information sources. This will allow the system to accept questions in natural language[2] enable the application of this methodology to free-text data sources.

2.2 System architecture

There are two main components in the system. The *evidence information* is stored in an XML document database. The *EPoCare server* uses this database to provide answers to the queries that are posed by clinicians.

The architecture of the system is presented in Figure 1. A clinical query is passed to the front controller to form a database query of keywords. The query is sent by the retriever to the XML document database to retrieve relevant documents and XML paths in the data sources, using late present keyword matching. The results are then passed to the query–answer matcher to find the matching answers.

The current evidence information includes the reviews of experimental results for clinical problems that are published in *Clinical Evidence* (CE) (version 7) [3] and *Evidence-based On Call* (EBOC) [1]. The texts are stored with XML mark-up in the database. The XML database is manipulated by ToX, a repository manager for XML data [2]. Distributed XML documents may be stored in a file system, a relational database, or remotely on the Web. ToX supports document registration, collection management, storage and indexing choice, and queries on document content and structure.

[2]Here and throughout the paper, we make the conventional distinction between *query* and *question*; the former is a keyword-based string or structure, and the latter is in natural language. A query may represent a question, and vice versa.

2.3 PICO-format queries

At present, the system accepts queries in a format known in Evidence-Based Medicine as *PICO format* [17]. In this format, a clinical question is represented by a set of four fields that correspond to the basic elements of the question:

P: a description of the patient (or the problem);
I: an intervention;
C: a comparison or control intervention (may be omitted);
O: the clinical outcome.

For example, the sample question in section 1 can be represented in a simple PICO format as follows:

P: asthma
I: inhaled corticosteroids
C: —
O: growth

A more-complete PICO representation of the same question is this:

P: child with asthma
I: increased doses of inhaled corticosteroids
C: —
O: decrease in growth

This representation contains more information; but neither of the two expresses the complete semantics of the natural-language question. Thus, the PICO format is limited in its ability to represent the meaning of questions. Especially in the case of *yes–no* questions, the point of the question is likely to be unclear. However, the PICO format indicates the basic semantics of the question, and it is commonly used in question representation in EBM. Thus it was used as a starting point in the development of the system.

The present keyword-based retrieval procedure is composed of three steps:

Retrieving. For each query keyword, XML paths in which the keyword appears are found.

Filtering. For each PICO category in the question, some XML contexts are meaningful for it while others are not. Paths that are not meaningful contexts for the PICO category of the keyword are filtered out. For example, a chapter title is meaningful (and valuable) context for an instance of patient population in the keyword matching, but titles of cited references are not.

Building answers. In the filtered paths, the system identifies cases in which all the key concepts in the question have been found, in context, in such a way that an answer pattern is satisfied. Then it returns the related segment of text in XML format so that the user can view it with a browser. A set of answer patterns were constructed for this matching process; each answer pattern consists of a set of XML paths for each of the four PICO categories. To identify a path as relevant, all four components should find a match in it.

While this searching strategy is based on the PICO format, it is not confined to it. Within the domain of Evidence-Based Medicine, the patterns can be extended so that additional fields are included. In addition, this searching strategy can be applied to other domains entirely.

Figure 2 shows an example of a clinical question with the corresponding EPoCare query and the segment of text that was retrieved from *Clinical Evidence* in response. The segment that was retrieved is clearly relevant to the question, but it has too much irrelevant data.

3 QA in medicine: The problem

We will now discuss medical question-answering, with the goal of refining the current EPoCare system by accepting natural-language questions and better identifying answers in the data sources.

In this section, we examine the difference between general and medical QA from the perspective of the three main research problems of QA: question processing, question–answer matching, and answer extraction. For each problem, we describe features that current QA technology is not appropriate for, and features that are not addressed by existing technology.

3.1 Question processing

For a question to be answered correctly, a QA system first has to understand what the question is asking about. This is an important task of question processing. Most current QA systems address it by identifying the *type* of answer sought. As GQA systems focus on *wh-* questions, many of which have named entities (NEs) as their answer, they usually classify answers according to different types of NE, such as product, organization, person, and so on. This classification is not appropriate in the medical domain, in which questions often ask about the treatment for a disease, outcome of a treatment, possible disease, and so on. As a result, the method of identifying an answer type must be different in MQA from GQA.

Even for the same answer type, there may be a different understanding. For example, *when* questions ask for the time that an event happens. In GQA systems, they are usually answered by an absolute date, e.g., *15 May 1932*. However, in the medical area, *when* questions are usually answered by relative time, e.g., *two hours after the onset of chest pain*. Sometimes the answers are not even a time but a clinical condition, e.g., in response to *When should antibiotics be applied?*

Some problems of MQA are not addressed at all by current QA technologies:

Question focus. Sometimes, the answer type is not enough to determine what a question is about. Other information contained in the question is needed to understand its goal. This information is defined as the *focus* of the question [13]. Although different systems use different names for the idea of question focus, it is regarded to be very important in question processing. However, there is still no special technique to tackle this problem.

***Yes–no* questions.** As mentioned, most current QA systems focus on *wh-* questions; *yes–no* questions are still left untouched. However, we have found that they are very common in our collection of clinical questions that arose in patient treatment. Efficient processing of *yes–no* questions is an important task in MQA.

3.2 Question–answer matching

The matching of question and answer is the process that most GQA systems put great effort into. Different methods are applied according to different views of the problem. The approaches can be classified into two categories: *knowledge-intensive* and *data-intensive.* Knowledge-intensive approaches try to find the correct match between a question and the answer by using effective natural language processing techniques that combine linguistic and real-world knowledge. Typical systems include those of Paşca and Harabagiu [15] and Hovy, Hermjakob, and Lin [9]. Data-intensive approaches explore information embedded in the data sources to extract the evidence that supports a good answer. They can be further divided into *information extraction–based* [18], *redundancy-based* [5, 6], and *statistical* QA [10]. Many systems contain elements of both approaches.

Although there have been many technologies developed for matching the answer with the question, they are not applicable to the medical area directly for the following reasons.

Knowledge taxonomy. WordNet is the main knowledge base that most current GQA systems use in analyzing relationships among words when calculating the similarity of a question and a candidate answer. However, as a general-purpose knowledge base, it is not possible for WordNet to cover all the concepts in any particular domain, such as medicine. A domain-specific knowledge base is needed. For example, it may be important to know that *metoprolol* is an instance of *β-blocker* in order to locate the correct answer. A good complement to WordNet is the Unified Medical Language System (UMLS) [11], developed by the National Library of Medicine. UMLS contains three knowledge sources: the Metathesaurus, the Semantic Network, and the Specialist Lexicon. The Metathesaurus represents biomedical knowledge by organizing concepts according to their relationships and meanings. It will be very helpful in tasks such as query expansion and answer-type identification in MQA.

Named-entity identification. As the types of NE in the medical area are different, the method of identifying them must be changed accordingly. For example, an MQA system must be able to distinguish medication from diseases. Medical terminology plays an important role in NE identification, as before a concept can be classified, the corresponding terminology has to be recognized to make sure that the correct concept is found. Different phrases can be used to refer to the same medical concept. For example, a drug may be referred to by its abbreviation, its common name, or its formal name (*ASA, Aspirin, acetylsalicylic acid*). Also, different medical concepts may have the same abbreviation, which will lead to ambiguities in concept understanding.

Information source. A medical information source is often organized in accordance with a hierarchy of medical concepts. For example, *Clinical Evidence* [3] groups clinical information according to disease categories. A part of the structure is shown in Figure 3.

The positive aspect of such well-organized information is that once the set of candidate answers is found, it is very likely that it includes the correct answer. However, it is unlikely that the answer for a question will appear redundantly in many different places in the data source. This is different from GQA systems, which usually require a relatively large number of redundant answer candidates to support good performance by the system.

Context. In current GQA systems, a correct answer to a question is often independent of its context. This is not the case in the medical text, in which the context containing a candidate answer may be important to the question–answer matching. The context usually explains a conclusion, provides more evidence, or even presents contrary evidence. A correct answer may be missed or the incorrect answer may be extracted if the context is not considered in the matching process.

Complex constraints. Clinical questions often contain a very specific description of the patient conditions, as shown in the following examples:

Q: Should β-blocker (metoprolol) be used to continue treatment for *a male with hypertension and coronary artery disease* even though *he has Type 2 diabetes mellitus?*

Q: Do *patients surviving an AMI and experiencing transient or ongoing congestive heart failure (CHF)* have reduced mortality and morbidity when treated with an ACE inhibitor (ex. Ramipril)?

The detailed description of the patient acts as a constraint in matching with candidate answers. As the complexity of questions increases, more-sophisticated techniques are needed to find a matching answer.

3.3 Answer extraction

An MQA system should be able to answer clinical questions in the course of patient treatment. Hence the format of the answer is important, and this will affect the answer extraction process. For the two types of questions—*wh-* questions and *yes–no* questions—the EPoCare study of user requirements shows that both a short answer and a long answer should be prepared.

Short answer. The short answer provides accurate and concise information to the physicians so that they can make the decision quickly. For *yes–no* questions, the answer can be just *yes* or *no*. However, if the answer candidates contain some contradictory evidence, or both benefits and harms, they must be included even in the short answer. For example, the answer to the question in Figure 2 should contain the following two outcomes as both of them may affect the decision making.

> The overview found that thrombolysis versus placebo significantly *reduced short term mortality.*
> The overview found that thrombolytic treatment versus control significantly *increased the risk of stroke.*

It is possible that answers for some questions cannot be found in the evidence information. For such *no-answer*[3] questions, the system should indicate its failure explicitly as its short answer.

Long answer. Sometimes clinicians want to read a long answer that may contain an explanation of an evidence or other results of related experiments. A long answer can be obtained by searching in information about medical concepts having "is-a" relations with the medications (diseases) mentioned in the question. Because these explanations are closely related to the evidence of interest, they usually are contained in the adjacent text. In *Clinical Evidence*, these explanations are grouped under small topics in each section.

Some other cases that may be included in a long answer are shown in the following examples:

> We found little evidence on the balance between benefits and harms from thrombolysis in acute ischaemic stroke. One systematic review has found that thrombolysis ... reduces overall risk of death and dependency in the long term, but that this benefit was achieved at the cost of an increased short term risk of fatal intracranial haemorrhage.

The first sentence does not describe what actually the "benefits" and "harms" are; these are given by the second sentence.

[3] A no-answer question should not be confounded with a *yes–no* question for which the answer happens to be *no*.

> The systematic review found that the earlier thrombolytic treatment was given (with respect to the onset of symptoms), the greater the absolute benefit of treatment. For each hour of delay in thrombolytic treatment, the absolute risk reduction for death decreased by 0.16%.

In this example, the second sentence specifies “how much” the benefit is.

> Thrombolysis reduced the risk of death or dependency at the end of the studies (ARR 4.2%, 95% CI 1.2% to 7.2%; RRR 7%, 95% CI 3% to 12%; NNT 24, 95% CI 14 to 83).

Numbers are listed here to justify the claim in this sentence.

For the no-answer questions, clinicians may expect to read at least some relevant information. Again, the “is-a” relation described above may be considered as *relevant* here.

3.4 Evaluation metrics

Evaluation of QA systems in the medical area is different from current evaluation methods for general QA systems. The Text Retrieval Conference uses the Mean Reciprocal Rank (MRR) as an evaluation metric. In this method, a system may return an ordered list of up to five different candidate answers to a question, and the score received is $1/n$, where n is the position in the list of the correct answer (if it appears at all); for example, if the correct answer is fourth in the list, the system receives a score of 0.25 for that test item. This metric cannot be applied here, since returning a list of alternative candidate answers to a question, each of which must then be further verified, is not acceptable for a clinical question that is posed on site.

Correctness should not be the only focus of the evaluation. Different answer formats have different emphasis, and so they should be evaluated separately. Short answers have to be concise. As explained in the previous sub-section, a concise answer should include only the best matching answer and any contradictory result. The other focus of evaluating a long answer is completeness. Various types of relations such as explanation and specification mentioned above can be considered as part of a long answer. What a complete answer is seems to depend on the task itself.

Partial answers should be considered in the evaluation. If part of the correct answer is included in the system output, it should be evaluated according to the importance of the correct information. A partial answer that contains more crucial information should obtain a higher score. Similarly, if an answer helps make a wrong decision, it should be punished in the evaluation.

4 Locating answers with role identification

From the discussion in the previous section, we can see that MQA poses new challenges for QA research that require new approaches. We have found that the use of *roles* and *role identification* is effective, and we take them as an organizing principle for MQA that goes beyond the use of named entities in GQA. This section will explain the principle. In this approach, the four roles represented by the PICO format will first be located in both the natural-language question and the candidate answer texts obtained by the retrieval phase. For example, PICO roles would be identified in these candidate answers as shown by the labelled bracketing.

> One RCT found [*no evidence that (low molecular weight heparin)*$_I$ is superior to *(aspirin alone)*$_C$]$_O$ for the treatment of *(acute ischaemic stroke in people with atrial fibrillation)*$_P$.

> *(Thrombolysis)*$_I$ *(reduces the risk of dependency, but increases the risk of death)*$_O$.

> We found *(no evidence of benefit)*$_O$ from *(surgical evacuation of cerebral or cerebellar haematomas)*$_I$.

In the matching process, the roles in the question will be compared with the corresponding roles in the answer candidates to determine whether a candidate is a correct answer.

4.1 Why roles?

In GQA systems, as mentioned, in the question–answer matching process, usually the answer candidates are first checked to see if they contain the expected answer type, in order to rule out irrelevant candidates. This is shown to be efficient, as indicated by Harabagiu et al. [8]: systems that did not include NE recognizers performed poorly in the TREC evaluations. The effectiveness of this method depends on successfully recognizing NEs in the answer candidates. However, for questions that cannot be answered by named entities, the QA task is more complex, as it will be more difficult to recognize the corresponding answer type in the answer candidates.

The same problem occurs in MQA. The important information in medical text usually corresponds to the basic PICO fields. For example, therapy-related text describes the relationships among four elements: the status of the patient, the therapy, the comparison therapy, and the clinical outcome. Descriptions of the diagnosis process often consist of the patient status, the test method, and the outcome. These elements are the key concepts of understanding medical text. They act as different *roles*, which together construct the meaning of the text. While some of the roles correspond to NEs, others do not. For example, in answering a therapy-related question, the patient status and the therapy can often be treated as NEs, but the clinical outcome often cannot be. In a description of diagnosis, the test process often is not represented by an NE. While medical NEs can be expected to be recognized by applying terminology techniques with the support of UMLS, the recognition of non-NE roles in the answer candidates, on the other hand, becomes the main challenge.

Thus, it is not sufficient to use information-extraction techniques, as in some GQA systems [15, 18], in which patterns are matched against the text to fill in the roles in the template. In such systems, the coverage of the pattern set is quite limited; it is very time-consuming to manually construct a large set of suitable patterns, especially for complicated phrasings; and the patterns are very specific: specific words or phrases are usually required to occur at a fixed location in each pattern, making it applicable only to expressions phrased in exactly the same way. While we will need to look for some specific words, we need much greater flexibility than is afforded by simple pattern-matching to identify the PICO roles in the text. This can be done by analyzing the different roles and their relationships.

4.2 Understanding the data

To apply a role-based method in MQA, we need to deal with the following problems:

1. Identifying the roles in text.
2. Determining the textual boundary of each role.
3. Analyzing the relationships among different roles.
4. Determining which combinations of roles are relevant to the queston.

Our work currently focuses on therapy-related questions. We manually analyzed 170 sentences from the Cardiovascular Disorders section of *Clinical Evidence* to obtain a better understanding of these problems. Among the sentences, 141 contained at least one role that we are interested in. For therapy-related questions, we found that often if an outcome role appeared in a sentence, then the sentence provided some interesting information related to clinical evidence. But clinical outcome is the most difficult non-NE role to locate.

4.2.1 Identifying clinical outcomes

In our analysis, we found that the lexical identifiers of clinical outcome may belong to any of the three primary open-class part-of-speech categories: noun, verb, and adjective. For example:

> Thrombolysis *reduces* the risk of *dependency*, but *increases* the risk of *death*.

> Lubeluzole has also been noted to have *adverse outcome*, especially at higher doses.

Some words that identify outcomes are listed below:

Nouns: *death, benefit, dependency, effect, evidence, outcome.*
Verbs: *improve, reduce, prevent, produce, increase.*
Adjectives: *responsible, negative, adverse, slower.*

Clinical outcomes must be carefully distinguished in the text from the outcomes of clinical trials themselves. We refer to the latter as *results* in the following. A result might or might not include a clinical outcome. They often involve a comparison of the effects of two (or more) interventions on a disease. Sometimes a result will state that an outcome did *not* occur:

> One RCT found *evidence* that hormone treatment plus radiotherapy *versus* radiotherapy alone *improved survival* in locally advanced breast cancer.

> In the systematic review of calcium channel antagonists, indirect and limited *comparisons* of intravenous *versus* oral administration found no significant *difference* in *adverse* events.

> We found *no evidence* of benefit from surgical evacuation of cerebral or cerebellar haematomas.

Examples of the identifiers of results are these:

Nouns: *evidence, difference, comparison.*
Verbs: *compare.*

Adjectives: *superior, versus.*

4.2.2 Determining the textual boundary of clinical outcomes and results

In determining the textual boundary of a clinical outcome, the three groups of words (nouns, verbs and adjectives) are treated separately. Our finding is that for the noun identifiers, the noun phrase that contains the nouns will be an outcome. For the verb identifiers, the verb and its object together constitute an outcome. For the adjective identifiers, usually the adjective itself is an outcome. If several identifiers occur in one sentence, the outcome is all of the text indicated by one or more of the identifiers.

Determining the textual boundary of the results of clinical trials is more complicated. If a result is a comparison of two or more interventions, it will contain the interventions, words that indicate a comparison relationship, and often the aspects that are compared. In the first of the previous group of examples, the elements of the results are *evidence, hormone treatment plus radiotherapy versus radiotherapy,* and *improved survival.* However, if the interventions can be identified as NEs, it will not be too difficult to determine the boundary.

We tested these rules manually on 50 sentences from *Clinical Evidence* on the topic of acute otitis media. Out of 54 outcomes (including both clinical outcomes and clinical trial results), 45 were identified correctly, and 40 correct textual boundaries were found.

4.2.3 Relationships among roles

Once the roles are identified, the relationships among them can be explored to get a better understanding of the text. Two types of relations are relevant here: those of concepts in the same role category, and those among different role categories. The first type of relationship is based on our observation that more than one medication or disease is often mentioned together in the texts. We analyzed one section of *Clinical Evidence* and found that there are six important relations: COMPARISON, ALTERNATIVE, COMBINATION, SPECIFICATION, EXPLANATION, AND PREFERENCE. Each type is represented by some cue words or symbols as shown in the following examples. The word *plus* means the combination of two or more medications. The word *or*, as well as a comma, often suggests the alternative relation.

Versus usually derives a comparison relation.

> The combination of aspirin *plus* streptokinase significantly increased mortality at 3 months.

> RCTs found no evidence that calcium channel antagonists, lubeluzole, aminobutyric acid (GABA) agonists, glycine antagonists, *or* N-methyl-D-aspartate (NMDA) antagonists improve clinical outcomes in people with acute ischaemic stroke.

> One systematic review found no short or long term improvement in acute ischaemic stroke with immediate systemic anticoagulants (unfractionated heparin, low molecular weight heparin, heparinoids, *or* specific thrombin inhibitors) *versus* usual care without systemic anticoagulants.

Some symbols are ambigious, for example, *or*, *and*, and *with*. *Or* can also suggest a comparison relation, although most time it means alternative; *and* can represent an alternative relation, and *with* can be a specification relation.

Surface relationships between concepts may be the reflection of deeper connections among them. The COMPARISON, ALTERNATIVE, and PREFERENCE relationships imply that the two (or more) medications have some common effects on the disease(s) they are applied to. The relation SPECIFICATION, on the other hand, suggests a hierarchical relation between the first medication and the following ones, in which the first medication is a higher-level concept and the following medications are in a lower level. For example, in the last example, *systemic anticoagulants* is a higher-level concept, and *unfractionated heparin, low molecular weight heparin* etc. are examples of it that lie in a lower level. The relations are very important in finding an answer for a clinical question because of the semantics embedded in them.

Some default relations exist among medical semantic roles. For example, the semantic roles in the treatment scenario are connected by a default cause-effect relation: *medication* (compared with another *medication*) applying to *disease* results in *outcome*. The relation is important in that in most cases, the scenario can be understood without looking at the contexts between the roles.

We have also found that roles are helpful in understanding the relationships between sentences. For example, if a sentence contains only the intervention role and the following sentence contains only the problem and outcome roles, then it is very likely that the combination of the two sentences represents a complete idea and the roles themselves are related.

We believe that as the work continues, more interesting relations will be found.

5 Conclusion

We have described our work in progress on adding natural language analysis to query-answering in the EPoCare project. Although this work is at a relatively early stage, we have analyzed the limitations of GQA technologies in MQA and are developing techniques whose organizing principle is the identification of the semantic roles in both question and answer texts, with our initial emphasis being on the latter.

Acknowledgements

The EPoCare project is supported by grants from Bell University Laboratories at the University of Toronto. This research is also supported by a grant from the Natural Sciences and Engineering Research Council of Canada. We are grateful to Sharon Straus, MD, and other members of the project for discussion and assistance. An earlier version of the paper was presented in [14]; we appreciate all the comments received at that presentation.

References

[1] Ball, C. M. and Phillips, R. S. (2001). *Evidence-based on Call: Acute Medicine*. Harcourt Publishers Limited, London.

[2] Barbosa, D., Barta, A., Mendelzon, A., Mihaila, G., Rizzolo, F., and Rodriguez-Gianolli, P. (2001). Tox — the Toronto XML Engine. In *Proceedings of the International Workshop on Information Integration on the Web, Rio de Janeiro, 2001*.

[3] Barton, S. (2002). *Clinical Evidence*. BMJ Publishing Group, London.

[4] Cimino, J. J. (1996). Linking patient information systems to bibliographic resources. *Methods of Information in Medicine*, 35(2):122–126.

[5] Clarke, C. L. A., Cormack, G. V., and Lynam, T. R. (2001). Exploiting redundancy in question answering. In *Proceedings of the 24th International Conference on Research and*

Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA (SIGIR-2001), pages 358–365.

[6] Dumais, S., Banko, M., Brill, E., Lin, J., and Ng, A. (2002). Web question answering: is more always better? In *Proceedings of the 25th International Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland (SIGIR-2002)*, pages 291–298.

[7] Gorman, P., Ash, J., and Wykoff, L. (1994). Can primary care physicians' questions be answered using the medical journal literature? *Bulletin of Medical Library Association*, 82(2):140–146.

[8] Harabagiu, S. M., Moldovan, D., Paşca, M., Mihalcea, R., Surdeanu, M., Bunescu, R., Gîrju, R., Rus, V., and Morărescu, P. (2001). The role of lexico-semantic feedback in open-domain textual question–answering. In *Proceedings of the 39th Meeting of the Association for Computational Linguistics, July 9-11, 2001, Toulouse, France (ACL-2001)*, pages 274–281.

[9] Hovy, E., Hermjakob, U., and Lin, C.-Y. (2001). The use of external knowledge in factoid QA. In *Proceedings of the Tenth Text Retrieval Conference, National Institute of Standards and Technology (NIST), Nov. 13-16, Gaithersburg, Maryland, USA (TREC 2001)*, pages 644–652.

[10] Ittycheriah, A., Franz, M., and Roukos, S. (2001). IBM's statistical question answering system – TREC-10. In *Proceedings of the Tenth Text Retrieval Conference, National Institute of Standards and Technology (NIST), Nov. 13-16, Gaithersburg, Maryland, USA (TREC 2001)*, pages 258–264.

[11] Lindberg, D. A. B., Humphreys, B. L., and McCray, A. T. (1993). The Unified Medical Language System. *Methods of Information in Medicine*, 32(4):281–291.

[12] Mendonça, E. A., Cimino, J. J., Johnson, S. B., and Seol., Y.-H. (2001). Accessing heterogeneous sources of evidence to answer clinical questions. *Journal of Biomedical Informatics*, 34:85–98.

[13] Moldovan, D. and Harabagiu, S. M. (2000). The structure and performance of an open-domain question answering system. In *Proceedings of the 38th Meeting of the Association for Computational Linguistics, Oct. 1-8, 2000, Hong Kong, China (ACL-2000)*, pages 563–570.

[14] Niu, Y., Hirst, G., McArthur, G., and Rodriguez-Gianolli, P. (2003). Answering clinical questions with role identification. In *Proceedings of the Workshop on Natural Language Processing in Biomedicine, 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, 2003*, pages 73–80.

[15] Paşca, M. A. and Harabagiu, S. M. (2001). High performance question/answering. In *Proceedings of the 24th International Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA (SIGIR-2001)*, pages 366–374.

[16] Sackett, D. L. and Straus, S. E. (1998). Finding and applying evidence during clinical rounds: the "evidence cart". *Journal of the American Medical Association*, 280(15):1336–1338.

[17] Sackett, D. L., Straus, S. E., Richardson, W. S., Rosenberg, W., and Haynes, R. B. (2000). *Evidence-Based Medicine: How to Practice and Teach EBM*. Harcourt Publishers Limited, Edinburgh.

[18] Soubbotin, M. M. (2001). Patterns of potential answer expressions as clues to the right answers. In *Proceedings of the Tenth Text Retrieval Conference, National Institute of Standards and Technology (NIST), Nov. 13-16, Gaithersburg, Maryland, USA (TREC 2001)*, pages 293–302.

[19] Straus, S. E. and Sackett, D. L. (1999). Bringing evidence to the point of care. *Journal of the American Medical Association*, 281:1171–1172.

[20] Takeshita, H., Davis, D., and Straus, S. E. (2002). Clinical evidence at the point of care in acute medicine: a handheld usability case study. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting (2002)*, pages 1409–1413.

[21] TREC (2001). *Proceedings of the Tenth Text Retrieval Conference, National Institute of Standards and Technology. Gaithersburg, MD, November 13–16.*

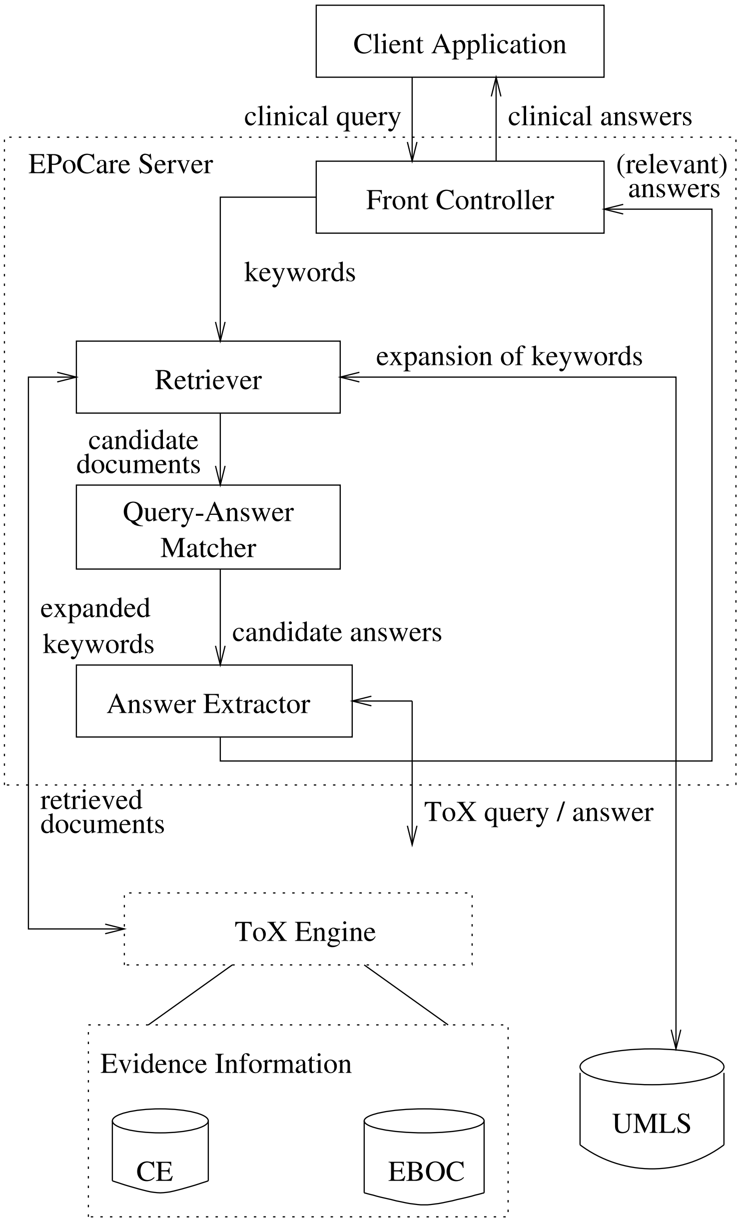

Figure 1: EPoCare system architecture.

Clinical question: In a patient with a suspected MI does thrombolysis decrease the risk of death if it is administered 10 hours after the onset of chest pain?

PICO format:

- **P:** myocardial infarction
- **I:** thrombolysis
- **C:** —
- **O:** mortality

Keywords: myocardial infarction thrombolysis mortality

Answer: Systematic reviews of RCTs have found that prompt thrombolytic treatment (within 6 hours and perhaps up to 12 hours and longer after the onset of symptoms) reduces mortality in people with AMI and ST elevation or bundle branch block on their presenting ECG.

Fifty six people would need treatment in the acute phase to prevent one additional death. Strokes, intracranial haemorrhage, and major bleeds are more common in people given thrombolysis; with one additional stroke for every 250 people treated and one additional major bleed for every 143 people treated. The reviews have found that intracranial haemorrhage is more common in people of advanced age and low body weight, those with hypertension on admission, and those given tPA rather than another thrombolytic agent.

Figure 2: Example of clinical question, with corresponding EPoCare query and answer from *Clinical Evidence*.

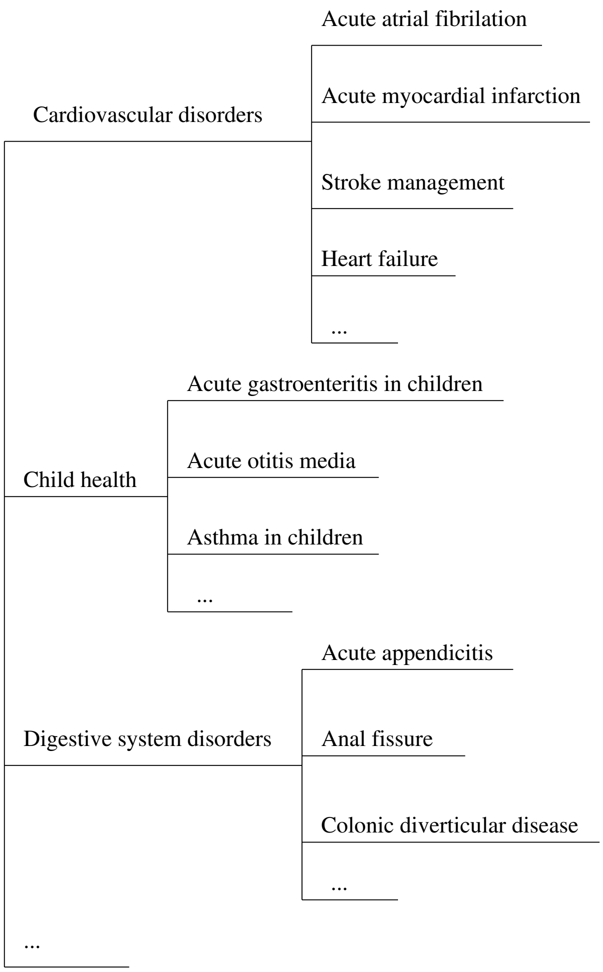

Figure 3: Disease categories in *Clinical Evidence*.