# Medical Question Answering: Translating Medical Questions into SPARQL Queries

Asma Ben Abacha
LIMSI-CNRS
BP 133, 91403 Orsay Cedex, France
abacha@limsi.fr

Pierre Zweigenbaum
LIMSI-CNRS
BP 133, 91403 Orsay Cedex, France
pz@limsi.fr

## ABSTRACT

Designing question answering systems requires efficient and deep analysis of natural language questions. A key process for this task is to translate the semantic relations expressed in the question into a machine-readable representation.

In this paper we tackle question analysis in the medical field. More precisely, we study how to translate a natural language question into a machine-readable representation. The underlying transformation process requires determining three key points: (i) What are the main characteristics of medical questions? (ii) Which methods are the most fitted for the extraction of these characteristics? and (iii) how to translate the extracted information into a machine-understandable representation?

We present a complete question analysis approach including medical entity recognition, semantic relation extraction and automatic translation to SPARQL queries. Our study supports the fact that SPARQL can represent a wide range of natural language questions in a question-answering perspective. Experiments on a corpus of real questions show that we obtain encouraging results in medical entity recognition and relation extraction. The obtained results also show that the output SPARQL queries correctly represent more than 60% of the original questions.

## Categories and Subject Descriptors

INFORMATION STORAGE AND RETRIEVAL [**Information Search and Retrieval**]: Query formulation

## General Terms

Algorithms, Languages

## Keywords

Medical Question Analysis, Information Extraction, Question Answering, Machine Learning, SPARQL, RDF


*IHI'12,* January 28–30, 2012, Miami, Florida, USA.


## 1. INTRODUCTION

A question-answering system is a specific kind of information-retrieval engine which aims at automatically answering natural language questions. The contribution of such systems is to provide a fast access to the searched knowledge, which is a crucial point in the medical domain for both practitioners and patients. Their performance is often evaluated by measuring their precision and recall over the returned answers (e.g. *CLEF*[1]: the Cross-Language Evaluation Forum).

The question answering task has two reference inputs: the corpora to be used to extract the relevant answers and the question itself. Each of these inputs must be analyzed in a manner that makes the question-answer matching semantically relevant, easy to understand and potentially traceable. This matching process implies that we represent both questions and candidate answers (or whole corpus) in a homogeneous semantic representation that can be processed by information systems. In the last decade, several meta-languages such as RDF(S)[2] and OWL[3] have been standardized by the W3C in order to formalize the representation of meaning on the Web. These languages provide a high level of expressivity and are more and more used in semantically-enabled applications and supported by efficient storage systems and APIs (e.g. Sesame[4], Jena[5], Virtuoso[6]).

In this work, we focus on the analysis of natural language questions in the medical field. We discuss the main characteristics of medical questions and present our approaches for medical entity recognition, semantic relation extraction and translation of the extracted information into SPARQL[7] queries. SPARQL (SPARQL Protocol and RDF Query Language) is the standard query language for RDF data. RDF and SPARQL allow a high expressivity by representing and interrogating data as instances of concepts and relations defined in a reference ontology. We evaluate each of the proposed unitary information extraction methods and the overall question analysis process on a real question corpus collected from the Journal of Family Practice (JFP)[8].

[1] http://clef.isti.cnr.it/
[2] http://www.w3.org/TR/rdf-syntax/
[3] http://www.w3.org/TR/owl-ref/
[4] http://www.openrdf.org
[5] http://jena.sourceforge.net/
[6] http://virtuoso.openlinksw.com/dataspace/dav/wiki/Main/VOSRDF
[7] http://www.w3.org/TR/rdf-sparql-query/
[8] http://www.jfponline.com

In the next section we provide background information about question analysis. In section 3 we discuss our information extraction methods. Then we present our question translation approach for medical question analysis in section 4. Finally, in section 5, we present our experiments and the obtained results.

# 2. BACKGROUND

## 2.1 Related Work

Several research efforts tackled automatic question analysis. For example, Verberne *et al.* [13] present an approach to automatically answering why-questions using syntactic categorization and show that answer type determination can be very effective for the analysis of a subset of why-questions. Fan *et al.* [9] proposed a question analysis approach for Chinese question answering. Their approach takes into account questions with multiple types and proposes a multi question type identification method based on semantic information. This information is extracted by a chunk annotation method which classifies question-level information into five types according to their semantic role. In a close topic, Duan *et al.* [6] proposed a method for question analysis aimed at retrieving similar questions. The proposed approach identifies the focus and the topic of input questions using an MD-based tree cut model. Similar questions are then retrieved by matching the extracted topic and focus.

The medical field is characterized by a rich and evolving terminology. It is also supported by several resources such as the UMLS (Unified Medical Language System) which encompasses a semantic network of medical concepts and relationships, a Metathesaurus including 2 million concepts and 7 million concept names, and a Specialist Lexicon. Another important resource is Medline which contains more than 18 million medical article citations.

Medical question analysis processes can take benefits of the medical domain resources but are also different from open domain question analysis due to some specificities of the medical domain (e.g. role of interrogative pronouns, specific domain relations between entities). For example, in an open domain question 'When' refers to a time where in medical questions it may refer to a specific condition (e.g. When should you suspect community-acquired MRSA?). Also, several taxonomies for medical questions were proposed. Ely *et al.* [8] propose a taxonomy of medical questions which contains the 10 most frequent question categories among 1396 collected questions. We list here the first 5 models (which represents 40% of the set of questions).

1. What is the drug of choice for condition x?
2. What is the cause of symptom x?
3. What test is indicated in situation x?
4. What is the dose of drug x?
5. How should I treat condition x (not limited to drug treatment)?

Ely *et al.* [7] proposed another taxonomy which classifies questions into *Clinical vs Non-Clinical*, *General vs Specific*, *Evidence vs No Evidence*, and *Intervention vs No Intervention*. Yu et al. [14] used Ely et al.'s taxonomy [7] to automatically classify medical questions with supervised machine learning. Their approach showed that the use of the question taxonomy with a SVM classifier and UMLS metathesaurus led to the highest performance.

Jacquemart and Zweigenbaum [10] collected 100 clinical questions on oral surgery and used them to propose another taxonomy of medical question models. Three models account for 90 out of 100 questions. These models are:

1. Which [X]-(r)-[B] or [A]-(r)-[which Y]
2. Does [A]-(r)-[B]
3. Why [A]-(r)-[B]

However, question taxonomies have some expressivity limits. For instance, Ely *et al.* [8]'s question taxonomy provides only some forms of expression for each question category, when in the real world we may often retrieve several other expressions for the same categories. Another example is the clinical questions taxonomy proposed by Jacquemart and Zweigenbaum [10] where the semantic relations are not fully expressed.

Niu et al. [11] analyze the limitations of general-purpose question-answering ontologies in the medical domain. They present an alternative approach based on the identification of semantic roles in the question and answer's texts using the PICO format as a reference representation (P: Population/disease, I: Intervention or Variable of Interest, C: Comparison, O: Outcome).

The PICO format has been studied on 100 medical questions in [5]. The main observations on the adequacy of the format on the evaluated questions show that the PICO framework is best suited for capturing therapy questions, and considerably less-suited for diagnosis, etiology, and prognosis questions. However, the framework is unable to reconstruct the original question. For example, does the frame `[Problem: hypomagnesemia, Intervention: ?]` correspond to "What is the most effective treatment for hypomagnesemia?" or "What are causes of hypomagnesemia?". This is mainly due to the inability of encoding fine-grained relations between medical entities. Other limitations include ambiguity of the format (for example, the standard PICO frame represents problem and population by the same "P" element) and the inability to capture anatomical relations.

## 2.2 Characteristics of Medical Questions

Several, possibly different, question characteristics are addressed in question analysis approaches. We try here to group together the most basic and conventional ones.

1. Question Type: WH Question or a Yes/No Question. A WH Question can be a Definition question (e.g. What is Depression?), List (e.g. What are the symptoms of blood cancer?), a complex question which needs a detailed answer and not just a medical entity, etc.
2. Expected Answer Type. For a WH question, the type of the expected answer is important: a treatment (e.g. What is the best treatment for Psoriasis?), a medical test (e.g. Colon Cancer: Which screening test should I have?), etc.
3. Focus. The focus of the question is the medical entity closest to the expected answer. For example, "pyogenic granuloma" is the focus of the question "What's the best treatment for pyogenic granuloma?".

4. Main relation. For WH questions, the main relation of a question is the semantic relation that links the expected answer with the focus. In Yes/No questions, the main relation is the most important relation between two medical entities (focus).

5. Medical Entity Recognition (MER). Each word in the question is important and especially medical entities. Recognizing each entity (e.g. headache) and its category (e.g. Medical Problem) helps to find the precise answer. It allows us to determine the focus but also other medical entities in the question. MER is a crucial step towards efficient question analysis as it deals with problems such as the high terminological variation in the medical domain (one word can have synonyms, abbreviations, etc.) and the evolution of entity naming (new drugs, new diseases, etc.).

6. Semantic relations. Correctly extracting the main relation but also contextual ones (e.g. patient-level information: age, family history) is the key point for relevant question analysis.

A key observation in our work is that the definition of one focus and/or one answer type limits the range of questions that can be fully parsed. In Section 4, we present our approach which allows dealing with multi-focus questions as well as more complicated cases (e.g. more than one answer type, context-bearing questions). This approach relies on previous extraction steps where medical entities and relationships are detected. In the following section we present our methods for these information extraction steps. We note that other information extraction tools could be used for these tasks which are independent from the final translation step.

# 3. PROPOSED METHODS FOR INFORMATION EXTRACTION

We present an extension of our methods for medical entity recognition and semantic relation extraction ([2], [4], [3]) where we address more medical categories and semantic relations. In this section, we quickly recall the main characteristics of these methods and present their new results obtained on a question corpus from the Journal of Family Practice (JFP) in section 5.

## 3.1 Medical Entity Recognition

Medical Entity Recognition (MER) consists in two main steps: (i) detection and delimitation of phrasal information referring to medical entities and (ii) classification of located entities into a set of predefined medical categories. For instance, in the sentence: *High blood pressure may cause Kidney Disease*, this task allows recognizing that "High blood pressure" and "Kidney Disease" are two "Medical Problems". We target 7 medical categories: Problem, Treatment, Test, Sign or Symptom, Drug, Food and Patient. These medical categories have been chosen according to an analysis of different medical question taxonomies. Table 1 presents some medical categories and their corresponding UMLS Semantic Types.

The proposed MER approach uses a combination of two methods to recognize medical entities: MetaMap Plus (Section 3.1.1) and BIO-CRF-H (Section 3.1.2).

| Category | UMLS Semantic Type |
|---|---|
| **Medical Problem** | Virus, Bacterium, Anatomical Abnormality, Congenital Abnormality, Acquired Abnormality, Sign or Symptom, Pathologic Function, Disease or Syndrome, Mental or Behavioral Dysfunction, Neoplastic Process, Cell or Molecular Dysfunction, Injury or Poisoning |
| **Treatment** | Medical Device, Drug Delivery Device, Clinical Drug, Steroid, Pharmacologic Substance, Antibiotic, Biomedical or Dental Material, Therapeutic or Preventive Procedure |
| **Medical Test** | Laboratory Procedure, Diagnostic Procedure. |
| **Sign or Symptom** | Sign or Symptom |
| **Drug** | Clinical Drug, Pharmacologic Substance, Antibiotic |
| **Food** | Food |
| **Patient** | Patient or Disabled Group |

**Table 1: Our medical categories and corresponding UMLS semantic types**

### 3.1.1 MetaMap Plus

MetaMap Plus (MM+) uses the MetaMap tool [1] which maps Noun Phrases (NP) in texts to the best matching UMLS concepts and assigns them matching scores. This method proposes an enhanced use of MetaMap through the following steps: (i) Chunker-based noun phrase extraction: we use Treetagger-chunker as it outperforms other tools for this task [4], (ii) Noun phrase filtering with a stop-word list, (iii) Search for candidate terms in specialized lists of medical terms, (iv) Use of MetaMap to annotate NPs with UMLS concepts and semantic types and (v) Filter MetaMap results with a list of common errors and the selection of only a subset of semantic types to look for. We use this method for all the target medical categories.

### 3.1.2 BIO-CRF-H

This method identifies simultaneously entities boundaries and categories using (i) a CRF classifier (we use CRF++[9]) and (ii) the B-I-O format. The B-I-O format (B: beginning, I: inside and O: outside) represents entity tagging by individual word-level tagging. For instance, a problem-tagged entity is represented as: (i) first word tagged "B-P" (begin problem), (ii) other (following) words tagged "I-P" (inside a problem) and (iii) Words outside entities are tagged with the letter "O". The task consists then in a word classification process into $2n + 1$ classes (where $n$ is the number of target medical categories). We use the following set of features to train and test the CRF classifier:

- Word Features: the word itself, 2 words before, 3 words after, lemmas;
- Morphosyntactic Features: POS tags of these words (using TreeTagger);

[9] `http://crfpp.sourceforge.net/`

- Orthographic features: (i) the word contains hyphen, plus sign, ampersand or slash, (ii) the word is a number, letter, punctuation, sign or symbol, (iii) the word is in uppercase, capitalized or lowercase, (iv) prefixes and suffixes of different lengths (from 1 to 4), etc.;

- Semantic Features: semantic category of the word (provided by MM+).

The MM+ method does not need any particular preparation while the second method needs an annotated corpus with medical entities. Such a resource is not always available. One important annotated medical corpus is the i2b2 2010 corpus. This corpus was built for the i2b2/VA 2010 challenge[10] in NLP for Clinical Data [12]. The corpus is fully manually annotated for concept, assertion, and relation information. It contains entities of 3 different categories: Problem, Treatment and Test. We use a part of this corpus to train our CRF classifier and another part for testing.

Results of medical entity extraction are represented in first order logic with the predicates `category`, `name` and `position`. For example the set of predicates

```
{ category(#ME1,TREATMENT)
∧ name(#ME1,aspirin)
∧ position(#ME1,3) }
```

indicates that the third token of the question, "aspirin", is a medical entity, and that its category is TREATMENT.

## 3.2 Relation Extraction

This step is very important for relevant question analysis. We target 7 semantic relations (chosen according to an analysis of medical question taxonomies):

1. *treats*. Treatment improves or cures medical problem
2. *complicates*. Treatment worsens medical problem
3. *prevents*. Treatment prevents medical problem
4. *causes*. Treatment causes medical problem
5. *diagnoses*. Test detects, diagnoses or evaluates medical problem
6. *DhD*. Drug has dose
7. *P_hSS*. Problem has signs or symptoms.

Figure 1 presents an excerpt from our reference ontology which describes the target medical categories and semantic relation types, as well as textual information that we will be using in question-answering and more particularly in the question translation step.

To extract semantic relations, we use a combination of two methods: a pattern-based method (Section 3.2.1) and a machine-learning method based on a SVM-classifier (Section 3.2.2). We also extract attributes specific to patients: (i) Patient Sex, (ii) Patient Age, (iii) Patient Age Group (Adult, Adolescent, Child, Toddler, Infant, newborn).

[10] `http://www.i2b2.org/NLP/Relations/`

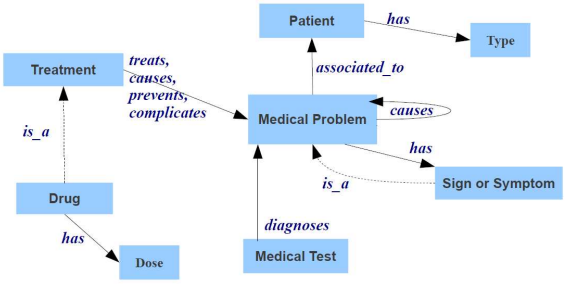



**Figure 1: Domain Model (Excerpt)**

### 3.2.1 Pattern-based Method

Semantic relations are not always expressed with explicit words such as *treat* or *prevent*. They are also frequently expressed with combined and complex expressions which makes building patterns with high coverage more difficult. However, the use of patterns is one of the most effective methods for automatic information extraction from textual corpora if they are efficiently designed. A benefit of pattern-based methods is also that we do not need an annotated corpus or training data to extract semantic relationships.

In this method, we manually constructed a set of patterns from abstracts of medical articles extracted from Medline. We list here two simplified examples of patterns (** represents a limited sequence of characters):

- TREATMENT ** for prophylaxis (against|of) ** PROBLEM
- TEST ** ordered to evaluate ** PROBLEM

In order to extract attributes information (e.g. drug dose, patient age) we use regular expressions that involve specific keywords (e.g. age, g/l) in order to extract the attribute type and value. For doses notations we used a list of units of measure collected from online resources[11].

### 3.2.2 SVM-based Method

Pattern-based methods have the disadvantage of the pattern construction process which is relatively time consuming compared to machine learning. On the other hand, statistical approaches are very efficient in extracting semantic relations if sufficient training examples are available.

We proposed [3] a machine-learning approach based on a SVM classifier which is trained on the i2b2 challenge's corpus. This method uses a set of lexical, morphosyntactic and semantic features for each couple (E1,E2) of medical entities:

1. Lexical Features: include words of the source entity (E1) and the target entity (E2), words between E1 and E2, 3 words before E1 and 3 words after E2 and also lemmas of these words;
2. Morphosyntactic Features: parts-of-speech (POS) of each word (with TreeTagger);
3. Verbs between E1 and E2, the first verb before E1, and the first verb after E2;

[11] `http://www.hc-sc.gc.ca/dhp-mps/prodpharma/notices-avis/abbrev-abrev/unitsmeasure\_unitesmesure-eng.php`

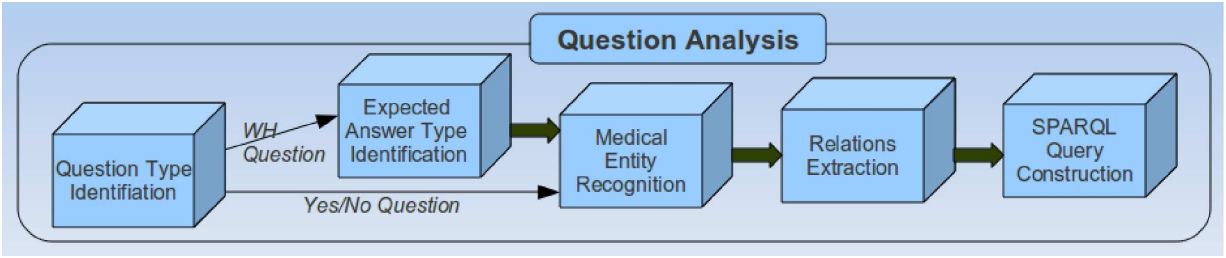



**Figure 2: Medical Question Analysis - most important steps**

4. Semantic Features: semantic types of E1, E2 and medical entities between E1 and E2.

Results of semantic relation extraction are represented in first order logic with several predicates indicating the name of the relation (e.g. treats, causes). For example the set of predicates

```
{ treats(#ME1,#ME2)
∧ patientHasProblem(#ME3,#ME2) }
```

indicates that three extracted medical entities are linked with two semantic relationships: treats and patientHasProblem.

# 4. QUESTION TRANSLATION TO SPARQL QUERIES

The selection of SPARQL as a machine readable language aims at avoiding the loss of expressivity in the query construction phase (e.g. identifying only one focus and/or EAT while the natural language question contains more). SPARQL is a standard language recommended by the W3C to query RDF data sources. Using SPARQL implies to annotate the textual corpora from which the answers are extracted in RDF according to a reference ontology.

In the scope of this paper we do not present the corpus annotation process in details. The annotation process uses the same information extraction methods presented in section 3. It stores the RDF annotations of the documents in separate files and keeps the links between each sentence and its annotations.

In this section we describe our question translation algorithm which uses a First Order Logic (FOL) representation of the extracted information and logical rules in order to obtain SPARQL reformulations of the user queries.

## 4.1 Approach Description

We propose a 6-fold method (cf. Figure 2) which consists in:

1. Identifying the question type (e.g. WH, Yes/No, Definition)
2. Determining the Expected Answer(s) Type(s) (EAT) for WH questions
3. Constructing the question's affirmative and simplified form (new form)
4. Medical Entity Recognition based on the new form of the question
5. Relation Extraction based on the new form
6. SPARQL Query Construction (cf. Section 4.2)

Table 2 presents the output of each step on two examples.

We identify the question type (e.g. WH, Yes/No, Definition) by applying a set of simple rules on user questions (e.g. firstWordOf(Q) = (How|What|Which|When) indicates that question Q is a WH question).

For WH questions, we determine EATs by matching the natural language questions with manually built lexical patterns. A set of patterns is constructed fo each question type. These patterns use interrogative pronouns, syntactic analysis and generic words in order to identify a set of matching questions. It is often the case that a question has more than one EAT. In this context, we keep the results obtained from all matching patterns even if the set of matching patterns belongs to more than one answer type (e.g. Treatment, Drug, Medical test).

In a second step we construct the questions' affirmative forms by replacing interrogative pronouns by the 'ANSWER' keyword. This form will be used in the relation extraction step. We also construct a simplified form of the question where the sequence of words indicating the EAT are replaced by the 'ANSWER' keyword. This allows avoiding noise when extracting medical entities. For example, in the question "What is the best treatment for headache?" the MER system will return treatment and headache as medical entities which is not an efficient input for relation extraction. Simplifying the question to "ANSWER for headache" produces more effective relation extraction results: identifying relations only between ANSWER (which is a treatment) and headache.

In a last step before the SPARQL query construction, relation extraction is performed. The process is straightforward for Yes/No questions as all medical entities are completely identified. For WH questions, as pointed out earlier, we may have more than one EAT. In such cases, we need a more generic method.

### *Multiple EATs.*

To take into account the case of multiple EATs, we construct as many questions as expected answers. The process includes:

1. Identifying the question type
2. Identifying Expected Answer Types EATs for WH questions ($m$ EATs)
3. Constructing the question's affirmative and simplified form (new form)

| Question Analysis - Information Extraction | Examples | |
|---|---|---|
| | *WH Question* | *Y/N Question* |
| | What treatment works best for constipation in children? | Does spinal manipulation relieve back pain? |
| **Expected Answer Type Identification** | EAT = Treatment | — |
| **Question Simplification and Transformation in Affirmative Form** | new_Q = ~~What treatment~~ *ANSWER* works best for constipation in children? | new_Q = ~~Does~~ spinal manipulation relieve back pain. |
| **Medical Entity Recognition *(using new_Q)*** | ANSWER works best for <PB> constipation </PB> in <PA> children </PA>. | <TX> spinal manipulation </TX> relieve <PB> back pain </PB>. |
| **Semantic Relation Extraction *(using new_Q)*** | treats(ANSWER,PB), with EAT = Treatment, patientHasType(children) | treats(TX,PB) |

**Table 2: Medical Question Analysis - Examples (EAT: Expected Answer Type, PB: Problem, PA: Patient, TX: Treatment)**

4. Medical Entity Recognition using the new form
5. *for ($x = 1$, $x{+}{+}$, $x \leq m$)*
   - Extraction of semantic relations *[Input: (i) $EAT_x$, (ii) Medical Entities, and (iii) new question form)]*
   - Construction of SPARQL Query $x$

**Example.** *How to diagnose and treat Anxiety Disorder?*

- EATs:

  EAT1 = Test, EAT2 = Treatment

- Affirmative, simplified question form:

  *ANSWER to diagnose and treat Anxiety Disorder*

- Medical Entity Recognition:

  <PB>Anxiety Disorder</PB>

- $x = 1$: (i) Relation(s): diagnoses(EAT1,PB) is the only possible relation according to the EAT. The output SPARQL query will then be:

```
SELECT ?answer WHERE {
   ?answer mesa:category mesa:Test
.  ?answer mesa:diagnoses ?e1
.  ?e1 mesa:category mesa:Problem
.  ?e1 mesa:name 'Anxiety Disorder' }
```

- $x = 2$: (i) Relation(s): treats(EAT2,PB) is the only possible relation according to the EAT. The output SPARQL query will then be:

```
SELECT ?answer WHERE {
   ?answer mesa:category mesa:Treatment
.  ?answer mesa:treats ?e1
.  ?e1 mesa:category mesa:Problem
.  ?e1 mesa:name 'Anxiety Disorder' }
```

In this example, we return only the expected medical entity (which is the IRI of a RDF resource in this case), though we could also return the sentence of the answer as in the example of figure 3.

## 4.2 SPARQL Query Construction

Once the medical entities and semantic relationships are extracted from the natural language questions, the last step consists in constructing an equivalent SPARQL query. SPARQL defines four different query forms: SELECT, DESCRIBE, ASK and CONSTRUCT[12]. The CONSTRUCT form aims at generating new RDF graphs from available RDF data, where the DESCRIBE form is just informative (i.e. return a random selection of the answers). In our SPARQL query construction approach, we will be using the ASK and SELECT forms in order to represent respectively Yes/No questions and WH questions (Definition questions are considered as WH questions as they aim at retrieving a text fragment containing the searched definition).

A SPARQL query has two components: a header and a body. The header indicates the query form as well as other information (e.g. prefix declarations, named graph to interrogate, variables to return for the SELECT form). Constructing the header and body requires different processes when constructing SPARQL equivalents of Yes/No questions and WH questions. We first introduce preliminary definitions of RDF graphs and RDF graph patterns as well as the reference ontology that will be used to construct the SPARQL queries.

### 4.2.1 Preliminary Definitions

**RDF Graph.** Consider the pairwise disjoint infinite sets $I$, $B$, and $L$ (IRIs,[13] Blank nodes and Literals). An RDF graph is a set of RDF triples $(s; p; o) \in (I \cup B) \times I \times (I \cup B \cup L)$. In this triple, s is the subject, p the predicate and o the object. $T(G)$ defines the triples contained in a graph $G$ (see footnote 12).

**Basic Graph Pattern.** Consider a set of variables $V$ disjoint from the sets $I$, $B$, and $L$. A triple pattern is a triple $(s; p; o) \in (I \cup V) \times (I \cup V) \times (I \cup V \cup L)$. A basic graph pattern $P$ is a set of triple patterns. A question mark $?v$ in a triple indicates that $v$ is a variable (see footnote 12).

**SPARQL Query.** In the scope of this paper we consider a subset of SPARQL queries which can be defined as union and/or conjunction of basic graph patterns upon

[12] `http://www.w3.org/TR/rdf-sparql-query/#initDefinitions`

[13] Internationalized Resource Identifiers.

which filter functions can be defined. These filters consist in the evaluation of Boolean expressions involving one or more variables of the basic graph patterns of the query. Thus, each SPARQL query $Q(G, S, F)$ will be defined by a set of basic graph patterns $G = \{G_1, .., G_n\}$, a set of selected variables $S = \{S_1, .., S_n\}$ and a set of filter expressions $F = \{e_1, .., e_n\}$.

### 4.2.2 Reference Ontology

We define the MESA ontology (MEdical queStion Answering ontology) in order to represent the concepts and relations used to construct SPARQL translations of natural language questions (cf. figure 1 which represents some of our medical categories and relations). This ontology is also used to annotate the medical corpora from which the answers will be extracted. The MESA ontology is not a full domain ontology as it encompasses concepts and relations describing the text fragments that will be returned as the final answers of our question-answering system and potentially used to look for contextual information (e.g. patients data).

### 4.2.3 Translating WH questions

The SELECT query-form of the SPARQL language is the most suited for representing WH questions. The header of such queries consists mainly in the SELECT keyword and the variables to be returned. The body of the output SELECT query contains the basic graph pattern constructed using the medical entities and relationships extracted from the natural language question.

We formalize the query-body construction process as a transformation function from first order logic expressions representing the output of the information extraction steps into a basic graph pattern (see preliminary definitions).

Extracted relationships can be defined between two medical entities or between the answer and one medical entity (e.g. treats(ANSWER,flu), treats(aspirin,headache), patientAgeGroup(patient,Infant)).

Each binary predicate is transformed into an RDF triple pattern <s,p,o> where s is the subject, p is the property and o is the object. The IRI of the property p is obtained by concatenating the ontology's namespace and the predicate name, the subject and object are defined as variables representing the arguments of the input FOL predicate. Additional triples are added to indicate the category and/or precise name of the involved medical entities through the *mesa:name* and *mesa:category* properties defined in the MESA ontology. The example below shows the translation of the predicate *causes(ANSWER, Flu)*:

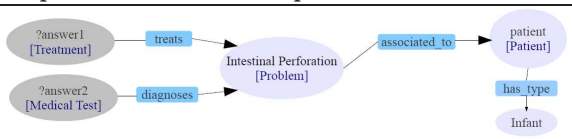
```
SELECT ?answer WHERE {
   ?answer mesa:causes ?arg2
.  ?arg2 mesa:name ?name
.  FILTER(?name='Flu') }
```

In the case where we have the EAT (e.g. category(ANSWER, TREATMENT)) or where the medical entity extraction step provides us with the medical entity categories (e.g. category(Flu,PROBLEM)), a similar translation process is applied save that the second argument (i.e. semantic type) is translated by concatenating it with the ontology namespace in order to retrieve its equivalent IRI in the MESA ontology. The next example shows the translation of the predicate *category(ANSWER,TEST)*:

**Medical Question**

*What are the current treatment and monitoring recommendations for intestinal perforation in infants?*

**Simplified Semantic Graph**




**SPARQL Query 1**

```
Select ?answer1 ?text1
where {
    ?answer1 mesa:category <Treatment>
    ?answer1 mesa:treats ?focus
    ?focus mesa:name 'intestinal perforation' }
    ?focus mesa:category <Problem>
    ?patien mesa:hasProblem ?focus
    OPTIONAL{
    ?text1 mesa:contains ?answer1
    ?text1 mesa:contains ?patient
    ?text1 mesa:contains ?focus
    }
}
```

**SPARQL Query 2**

```
Select ?answer1 ?text1
where {
    ?answer1 mesa:category < MedicalTest >
    ?answer1 mesa:diagnoses ?focus
    ?focus mesa:name 'intestinal perforation' }
    ?focus mesa:category <Problem>
    ?patient mesa:hasProblem ?focus
    OPTIONAL{
    ?text1 mesa:contains ?answer1
    ?text1 mesa:contains ?patient
    ?text1 mesa:contains ?focus
    }
}
```

**Table 3: Example of a translated wh question**

```
SELECT ?answer WHERE {
   ?answer mesa:category mesa:TEST }
```

The final translation of the user query consists in one or more SPARQL queries (if we have more than one EAT, cf. Section 4.1) constructed by assembling the unitary translations of each predicate obtained from the medical entity recognition and the relation extraction steps. Table 3 shows an example of a translated wh question.

### 4.2.4 Translating Yes/No questions

The ASK query-form of the SPARQL language is the most suited for representing Yes/No questions. The header of such queries consists mainly in the ASK keyword. The body of the output ASK query contains the basic graph pattern constructed using the medical entities and relationships extracted from the natural language question.

The translation process is similar to WH translation except that we don't have answers or EAT to take into account. Consequently, the construction of the final RDF graph patterns consists only in converting results of the medical enti-

ties and relation extraction into a triple format as described in the previous section.

The following example shows the translation of a Yes/No question into a basic graph pattern:

*Can being on prednisone cause a high serum iron?*

```
ASK{
   ?e1 mesa:causes ?e2
.  ?e1 mesa:name ?name1
.  ?e2 mesa:name ?name2
.  FILTER(?name1='prednisone')
.  FILTER(?name2='serum iron')}
```

# 5. EVALUATION

We use 100 questions extracted from the Journal of Family Practice JFP[14] (latest questions from 11/1/2008 to 4/1/2011). This set of questions contains 64 WH questions and 36 Yes/No questions. Table 4 presents some question examples.

We tested our 2 MER methods on the i2b2 test corpus and on our question corpus. We used the i2b2 training corpus to train the BIO-CRF-H method (i2b2 training corpus of 31,238 sentences and i2b2 test corpus of 44,927 sentences). Table 5 presents the obtained results without question simplification for three categories: Treatment, Problem and Test. It is important to note that results of BIO-CRF-H on the JFP corpus are not as good as the results on the i2b2 corpus. This is mainly due to the discrepancies between the two corpora and to the fact that BIO-CRF-H was trained only on the i2b2 corpus.

The question simplification improves MER results and especially the MM+ results leading to a precision value of 75.91% and a recall value of 84.55% (79.99% F-measure) on three categories: Treatment, Problem and Test of the JFP corpus.

We also evaluated our Relation Extraction and SPARQL query construction methods. For 29 of the 100 questions, we were not able to identify semantic relations.

We obtained 62 correct translations of the analyzed 100 questions and 38 false ones (29 false translations are mainly due to errors on relation extraction and 8 false translations are mainly due to errors on the expected answer type). For example, in the question "How accurate is an MRI at diagnosing injured knee ligaments?", even though we determined correctly the medical entities and the semantic relation (diagnoses), the query was not correct because of the question type (complex question).

A more detailed error analysis revealed two main causes of error: (i) new relation types that are not defined in our ontology or treated by our extraction system (e.g. *How does pentoxifylline affect survival of patients with alcoholic hepatitis?*, relation: "affects") and (ii) other expected answer types (EAT) that are not yet treated by our system (e.g. *Which women should we screen for gestational diabetes mellitus?*, EAT: Patient).

If we study the translation process on valid medical entities and relations only (i.e. excluding extraction errors), we observe that the translation was correct for 98% of the questions. However, in real world applications, the performance of question answering systems will surely depend on information extraction techniques. The obtained results could thus be improved by enhancing the implemented information extraction systems and adding further relations and medical entity types to the reference ontology.

[14] http://www.jfponline.com

| Method | i2b2 corpus | | | JFP Qs | | |
|---|---|---|---|---|---|---|
| | P | R | F | P | R | F |
| **MM+** | 56.5 | 48.7 | 52.3 | 66.66 | **84.55** | **74.54** |
| **BIO-CRF-H** | 84.0 | 72.3 | **77.7** | **77.03** | 46.34 | 57.87 |

**Table 5: MER - Results 2 (categories: Treatment, Problem and Test)**

# 6. DISCUSSION AND CONCLUSIONS

In this section, we discuss the contributions and limitations of our approach and present future work.

## 6.1 Contributions

In this paper, we tackled automatic question analysis in the medical domain and presented an original approach having 3 main characteristics:

1. The proposed approach allows dealing with different types of questions, including questions with more than one expected answer type and more than one focus.

2. It allows a deep analysis using different information extraction methods based on (i) domain knowledge and (ii) natural language processing techniques (e.g. use of patterns, machine learning) which allow extracting medical entities but also semantic relations and even additional information about the patient (age, sex, etc.).

3. Our approach is based on Semantic Web technologies, which offer more expressivity, standard formalization languages and makes our corpus and question annotations sharable through the web.

## 6.2 Limitations and Future Work

Although our approach was aimed to be generic (i.e. w.r.t. the target question types), more specific processes are still required to deal with (i) complex questions (e.g. why, when) and (ii) questions with new semantic relations that are not defined in our reference ontology.

To tackle scalability on this last issue, we plan to perform a syntactic analysis of the natural language question and test the contribution of syntactic dependencies on two aspects: (i) confirmation of previously extracted semantic relations and (ii) detection of unknown relations: syntactic dependencies (Subject-Verb-Object) can replace triplets (Entity1-Relation-Entity2) if an abstract RDF property is defined in the reference ontology.

Our final goal is the conception of a question answering system for the medical domain. Figure 3 presents the planned architecture for the whole question-answering system (QAS).

# 7. ACKNOWLEDGMENTS

This work has been partially supported by OSEO under the Quaero program.

| Type | Example Question |
|---|---|
| Y/N Question | Should patients with acute DVT limit activity? |
| WH Question | What is the best approach to benign paroxysmal positional vertigo in the elderly? |
| Complex Question | When should you consider implanted nerve stimulators for lower back pain? |
| EAT = Treatment | Childhood alopecia areata: What treatment works best? |
| EAT = Drug | Which drugs are best when aggressive Alzheimer's patients need medication? |
| EAT = Test | What is the best noninvasive diagnostic test for women with suspected CAD? |
| More than one Expected Answer | When should you suspect community-acquired MRSA? How should you treat it? |

**Table 4: Example questions from our corpus (EAT = Expected Answer Type)**

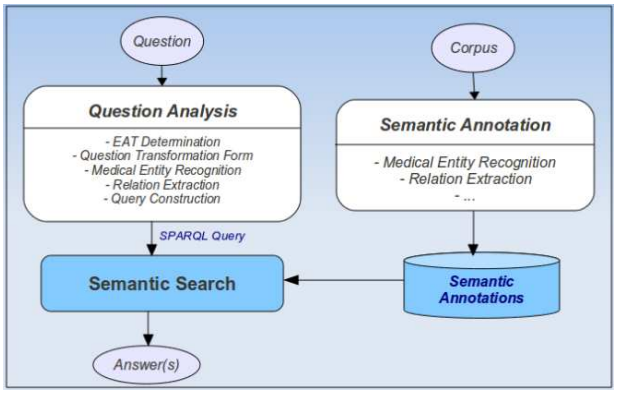



**Figure 3: QAS Architecture**

## 8. REFERENCES


[1] A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. *Journal of the American Medical Informatics Association*, 8(suppl):17–21, 2001.

[2] A. Ben Abacha and P. Zweigenbaum. Automatic extraction of semantic relations between medical entities: a rule based approach. *Journal of Biomedical Semantics*, 2011. In Press.

[3] A. Ben Abacha and P. Zweigenbaum. A hybrid approach for the extraction of semantic relations from MEDLINE abstracts. In *Computational Linguistics and Intelligent Text Processing, 12th International Conference, CICLing 2011*, volume 6608 of *Lecture Notes in Computer Science*, pages 139–150, Tokyo, Japan, 2011.

[4] A. Ben Abacha and P. Zweigenbaum. Medical entity recognition: A comparison of semantic and statistical methods. In *Proceedings of BioNLP 2011 Workshop*, pages 56–64, Portland, Oregon, USA, 2011. Association for Computational Linguistics.

[5] D. Demner-Fushman and J. Lin. Knowledge extraction for clinical question answering: Preliminary results. In *Proceedings of AAAI 2005 Workshop on Question Answering in Restricted Domains*. AAAI, 2005.

[6] H. Duan, Y. Cao, C.-Y. Lin, and Y. Yu. Searching questions by identifying question topic and question focus. In *Proceedings of ACL-08: HLT*, pages 156–164, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[7] J. W. Ely, J. A. Osheroff, M. H. Ebell, M. L. Chambliss, D. C. Vinson, J. J. Stevermer, and E. A. Pifer. Obstacles to answering doctors' questions about patient care with evidence: qualitative study. *British Medical Journal*, 324:710, 2002.

[8] J. W. Ely, J. A. Osheroff, P. N. Gorman, M. H. Ebell, M. L. Chambliss, E. A. Pifer, and P. Z. Stavri. A taxonomy of generic clinical questions: classification study. *British Medical Journal*, 321:429–432, 2000.

[9] S. Fan, X. Wang, X. Wang, and Y. Zhang. A new question analysis approach for community question answering system. *International Journal on Asian Language Processing, 19 (3): 95-108*, 2009.

[10] P. Jacquemart and P. Zweigenbaum. Towards a medical question-answering system: a feasibility study. In R. Baud, M. Fieschi, P. Le Beux, and P. Ruch, editors, *Proceedings of Medical Informatics Europe*, volume 95 of *Studies in Health Technology and Informatics*, pages 463–468, Amsterdam, 2003. IOS Press.

[11] Y. Niu, G. Hirst, G. McArthur, and P. Rodriguez-Gianolli. Answering clinical questions with role identification. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine - Volume 13*, BioMed '03, pages 73–80, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[12] O. Uzuner, B. R. South, S. Shen, and S. L. Duvall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, Jun 16 2011. [Epub ahead of print].

[13] S. Verberne. Developing an approach for why-question answering. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, EACL '06, pages 39–46, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[14] H. Yu, C. Sable, and H. R. Zhu. Classifying medical questions based on an evidence taxonomy. In *Proc. AAAI'05 Workshop on Question Answering in Restricted Domains*, 2005.