Information Processing and Management 51 (2015) 570–594

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies

Asma Ben Abacha [a,*], Pierre Zweigenbaum [b]

[a] LIST, Luxembourg
[b] LIMSI-CNRS, France

ARTICLE INFO

Article history:
Received 15 August 2014
Received in revised form 16 April 2015
Accepted 22 April 2015
Available online 11 June 2015

Keywords:
Question answering
Natural language processing
Semantic search
Medical informatics

ABSTRACT

The Question Answering (QA) task aims to provide precise and quick answers to user questions from a collection of documents or a database. This kind of IR system is sorely needed with the dramatic growth of digital information. In this paper, we address the problem of QA in the medical domain where several specific conditions are met. We propose a semantic approach to QA based on (i) Natural Language Processing techniques, which allow a deep analysis of medical questions and documents and (ii) semantic Web technologies at both representation and interrogation levels. We present our Semantic Question-Answering System, called MEANS and our proposed method for "Answer Search" based on semantic search and query relaxation. We evaluate the overall system performance on real questions and answers extracted from MEDLINE articles. Our experiments show promising results and suggest that a query-relaxation strategy can further improve the overall performance.

1. Introduction

The increasing knowledge accessible via internet affects our habits to find information and to obtain answers to our questions. According to an american health survey[1] published in January 2013, 35% of U.S. adults state that they have gone online specifically to try to figure out what medical condition they or someone else might have. Asked about the accuracy of their initial diagnosis, *41%* of "online diagnosers" (who searched for answers on the internet) say a medical professional confirmed their diagnosis, but *35%* say they did not visit a clinician to get a professional opinion. *18%* say they consulted a medical professional and the clinician either did not agree or offered a different opinion about the condition. 77% say that they start looking for health information using a search engine. However, while these search engines contribute strongly in making large volumes of medical knowledge accessible, their users have often to deal with the burden of browsing and filtering the numerous results of their queries in order to find the precise information they were looking for. This point is more crucial for practitioners who may need an immediate answer to their questions during their work.

Ely et al. (1999) presented an observational study in which investigators visited family doctors for two half days and collected their questions. The 103 doctors saw 2467 patients and asked 1101 questions during 732 observation hours. Each doctor asked an average of 7.6 questions during the two half days (3.2 questions per 10 patients).

* Corresponding author at: Luxembourg Institute for Science and Technology (LIST), 29 avenue John F. Kennedy, L-1855 Kirchberg, Luxembourg.
E-mail addresses: asma.benabacha@list.lu (A. Ben Abacha), pz@limsi.fr (P. Zweigenbaum).

[1] Pew Research Center's Internet & American Life Project: http://pewinternet.org/Reports/2013/Health-online.aspx.

http://dx.doi.org/10.1016/j.ipm.2015.04.006
0306-4573/

Covell, Uman, and Manning (1985) studied information needs of physicians during office practice. In their study, information needs were obtained by self-reports from 47 physicians who raised a total of 269 questions during their half-day practice. The raised questions were very heterogeneous in terms of topics and highly specific to the patients. On average only 30% of their information needs were met during the patient visit, most often by other physicians having different subspecialities. As shown in their study, print sources were not used for several reasons such as inadequate indexation of books and drug information sources, age of the available textbooks, lack of knowledge about the relevant sources or the time needed to access the required information.

In this context, we need tools such as question answering (QA) systems in order to respond to user queries with precise answers. Such systems need deep analysis of both user questions and documents in order to extract the relevant information. At the first level of this information come the medical entities (e.g. diseases, drugs, symptoms). At the second, more complicated level comes the extraction of semantic relations between these entities (e.g. treats, prevents, causes).

Within an overall common framework, QA systems aim to provide precise answers to natural language questions. The answer can be a piece of text extracted from a document collection (Demner-Fushman & Lin, 2006) or the Web (Lin & Katz, 2003) as well as data retrieved from a database (Popescu, Etzioni, & Kautz, 2003) or a knowledge base (Rinaldi, Dowdall, & Schneider, 2004). In more rare cases, the returned answers are multimedia information (Katz, 1999). A question answering system can be composed of three main tasks: (i) analysis of the user question, (ii) analysis of the documents used to find the answers and (iii) answer retrieval and extraction. The second task is not required for systems that use databases or knowledge bases as answer sources. Methods used to analyze questions and/or documents can be semantic, surface-level or hybrid.

In this paper, we address the problem of answering English questions formulated in natural language. We consider several types of questions, but we focus on two main types: (i) factual questions expressed by WH pronouns and (ii) boolean questions expecting a yes/no answer. An answer can be (i) a medical entity for factual questions or (ii) Yes or No for boolean questions. Moreover, for each answer extracted from a corpus, we associate a justification[2] including the line containing the answer, the two previous sentences and the two following sentences. We focus on searching and extracting answers from scientific articles and clinical texts. However, the proposed approach can be extended to consider other resources like websites, textual corpora, Linked Open Data and ontologies.

There are three main contributions in this paper:

1. We propose an original system for medical QA combining: (i) NLP methods which allow a deep analysis of medical questions and corpora used to find answers and (ii) semantic Web technologies which offer a high level of expressiveness and standards for data sharing and integration.
2. We introduce a novel query relaxation approach for QA systems that deals with errors or weaknesses of NLP methods in some cases (e.g. implicit information, need for reasoning).
3. We experimentally evaluate our system, called MEANS, with a benchmark (Corpus for Evidence Based Medicine Summarisation) and we discuss the obtained results.

The remainder of the paper is organized as follows. Section 2 introduces related work and discussion about the main QA approaches with a particular focus on the medical domain. Section 3 describes the overall architecture of the proposed approach and its main three steps: offline corpora annotation using NLP methods (described in Section 4), online question analysis (described in Section 5) and answer retrieval based on semantic search and query relaxation (described in Section 6). Section 7 presents our experiments on a standard corpus and the results of our QA system MEANS. In Section 8, we discuss the combined use of NLP methods and semantic technologies, then we analyze the error cases for the boolean and factual questions. Finally, the conclusions are made in Section 9.

2. Related work

BASEBALL (Green, Wolf, Chomsky, & Laughery, 1961) and LUNAR (Woods, 1973) are among the first known question answering systems. BASEBALL was able to answer questions about dates, locations and American baseball games. LUNAR was one of the first scientific question-answering systems. It was conceived to support the geological analysis of the rocks brought by the Apollo mission. In its evaluation, it answered correctly 90% of the questions asked by human users. Both BASEBALL and LUNAR exploited knowledge bases that were written manually by domain experts.

Nowadays, searching precise answers for natural language questions became one of the major challenges in modern information retrieval. With the exponential growth of digital documents and the continuous specialization of domain knowledge, the development of question-answering systems followed different tracks according to three main dimensions. The first dimension is related to the *answer source* which can be unstructured documents or structured databases. The second dimension is related to the considered *domain of application* which can be specific (e.g. medicine, sports, arts) (cf. see Mollá & Vicedo (2007) for an overview of the methods and applications used in restricted domains) or open (e.g. news, cross-domain encyclopedias). The third dimension is related to the complexity of *question analysis methods* (i.e. shallow/deep natural language processing, statistical methods, semantic methods).

[2] We define a **justification** for an answer as the source sentence found by the question-answering system.

In a related note several challenges have been organized to promote research and benchmarking on medical question answering and medical information retrieval:

- BioASQ challenges[3] on biomedical semantic indexing and question answering (Tsatsaronis et al., 2012);
- TREC[4] tracks such as TREC medical tracks[5] in 2011 and 2012 and Clinical Decision Support Track[6] in 2014 and 2015;
- CLEF challenges and tasks such as the ShARe/CLEF eHealth Evaluation Lab[7,8] and CLEF QA4MRE Alzheimer's task[9] (Morante, Krallinger, Valencia, & Daelemans, 2012).

In this section we will discuss the main question-answering approaches according to the three dimensions with a particular focus on the medical domain.

2.1. Question analysis

Basic features of question analysis are the determination of the question type, the expected answer type and the question focus. The focus of a question can be defined as the main information required by the interrogation (Moldovan et al., 1999). We define the *focus* of a question as the medical entity that is the most closely linked to the answer according to the information provided in the question itself. For example, "pyogenic granuloma" is the focus of the question "What is the best treatment for pyogenic granuloma?".The focus is linked to the expected answer by a relation, that we call *principal relation of the question*. Extracting the question focus and related entities requires Named Entity Recognition (NER) methods. More advanced analyses extract the relations that link the different entities referred in the natural language question.

The results of this first language-level analysis can then be translated to formal representations that allows querying structured databases or document annotations or, furthermore, performing automated reasoning to expand the question and its potential answers.

Several formal representation languages are proposed for question analysis. While some are standard languages such as SQL (Popescu et al., 2003) or SPARQL (Cimiano, Haase, Heizmann, Mantel, & Studer, 2008), other are ad hoc languages proposed for the open domain (e.g. Minimal Logical Form (Rinaldi et al., 2004)) or restricted domains (e.g. PICO for the medical field (Niu, Hirst, McArthur, & Rodriguez-Gianolli, 2003)).

Niu et al. (2003) analyzed the limitations of general-purpose question-answering ontologies in the medical domain. They presented 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) (Sackett, Straus, Richardson, Rosenberg, & Haynes, 2000).

The PICO format has been studied on 100 medical questions in Demner-Fushman and Lin (2005). 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.

Another important aspect in question analysis is the classification of the question itself. Several taxonomies for medical questions were proposed. Ely et al. (2000) 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. (2002) proposed another taxonomy which classifies questions into *Clinical vs Non-Clinical*, *General vs Specific*, *Evidence vs No Evidence*, and *Intervention vs No Intervention*. Yu, Sable, and Zhu (2005) used Ely et al.'s taxonomy (Ely

[3] http://www.bioasq.org/.
[4] http://trec.nist.gov/.
[5] http://trec.nist.gov/data/medical.html.
[6] http://www.trec-cds.org/.
[7] http://sites.google.com/site/shareclefehealth/.
[8] http://clefehealth2014.dcu.ie/.
[9] http://celct.fbk.eu/QA4MRE/.

et al., 2000) 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 (2003) 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 expressiveness limits. For instance, Ely et al. (2000)'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 (2003) where the semantic relations are not fully expressed. This allows to cover a larger base of questions which corresponds to the main purpose of modeling, but requires additional work to match questions and answers with specific relations within question-answering systems.

2.2. Answer extraction from document collections

Analyzing textual corpora for precise answers retrieval starts from NLP methods similar to those employed for question analysis, i.e. NER and relation extraction. For instance, the LASSO system (Moldovan et al., 1999) answers open-domain questions by returning a short and long answers. Lasso includes three modules: Question Processing (identification of the question type, the expected answer type, the question focus, and question keywords), Paragraph Indexing and Answer Processing. A question type classification containing 25 question types was manually constructed from the analysis of the TREC-8 training data. LASSO uses 8 ordered heuristics to extract questions keywords and a set of other heuristics to recognize named entities (persons, organizations, locations, dates, currencies and products). The extraction of the answer is based also on a set of heuristics. To evaluate the correctness of each answer candidate, 7 scores are calculated. The answer extraction is performed by choosing the answer candidate with the highest combined score.

We note that Moldovan et al. (1999) defined a focus as a word or a sequence of words which define the question and disambiguate it in the sense that it indicates what the question is looking for, or what the question is all about. For example, for the question "What is the largest city in Germany?", the focus is largest city. This definition differ from our definition and use of the term focus.

In the medical domain, Demner-Fushman and Lin (2006) proposed a hybrid approach to question answering in the clinical domain that combines techniques from summarization and information retrieval. They targeted a frequently-occurring class of questions having the form "What is the best drug treatment for X?".

The system identified the drugs under study from an initial set of MEDLINE citations. MEDLINE abstracts (as the source for answers) are clustered using semantic classes from UMLS. for each abstract, a short extractive summary is generated to populate the clusters consisting of three elements: (i) the main intervention, (ii) the title of the abstract (iii) and the top-scoring outcome sentence[10] calculated by an outcome extractor based on supervised machine learning techniques (Demner-Fushman & Lin, 2005).

Terol, Martínez-Barco, and Palomar (2007) used Ely et al. (2000)'s question taxonomy and targeted the ten most frequent questions (e.g. *What is the drug of choice for condition x?*, *What is the cause of symptom x?*). They proposed a medical QA system able to answer NL questions according to these ten generic medical questions giving priority to Precision rather than Recall. The proposed QA system has 4 main components: (i) Question Analysis, (ii) Document Retrieval, (iii) Relevant Passages Selection and (iv) Answer Extraction. The system is based on logic forms derived through the dependency relationships between the words of a sentence. Dependency relationships are obtained using MINIPAR (Lin, 1998). Other NLP resources are used by the system like the WordNet lexical database (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990) to extract the similarity relationships between the verbs and the UMLS metathesaurus (Humphreys & Lindberg, 1993) to recognize medical entities.

Regardless of the NLP methods used for document analysis, answer extraction from a collection of documents has the advantage of providing a justification[11] besides the answer. However, focusing on a specific type of textual documents reduces the scalability of these methods (e.g. classifier overfitting, handcrafted rules/patterns with limited coverage). Also, approaches for answer extraction from document collections do not often propose strategies for cumulative acquisition of knowledge, like using common references (e.g. knowledge bases, thesaurus) that allow solving co-references by associating them to a same element (e.g. a thesaurus concept, a knowledge base individual).

[10] Here, outcome sentences state the efficacy of a drug in treating a particular disease.
[11] The fragment of text containing the answer.

2.3. Answer extraction from databases

The QA issue was highlighted within the context of Natural Language Interfaces (NLI) allowing the search of answers for NL questions from relational databases or knowledge bases. Therefore, many efforts tackled the transformation of NL question into SQL queries. For instance, the PRECISE system analyze NL questions and translate them into SQL queries (Popescu et al., 2003). The system refers to a set of semantically tractable[12] question categories and translates user questions that belong to at least one of them (tractable questions) to output the associated SQL query (or queries). In its evaluation PRECISE provided correct answers to over 80% of the questions in the domains of geography, jobs and restaurants. But it is only effective on semantically tractable questions (which presents 77.5% in the geography database and 97% in restaurant database).

Other approaches are based on the sense of query words to find answers from pre-built knowledge bases. For instance, the START system (Katz et al., 2002) answers questions about geography with a precision of 67% on 326 000 queries. START uses knowledge bases enriched manually to find triples having the form subject-relation-object. Answer extraction is based on the comparison of the user question with the annotations of the knowledge bases. However, binary relations cannot represent all questions. Also, the acquisition of scientific knowledge could be a limitation of the system as only domain experts are able to add knowledge and increase the system coverage.

AquaLog (Lopez & Motta, 2004) is an ontology-driven query answering system which allows users to choose an ontology as input and to ask questions in natural language. The system returns answers from the available semantic markup by combining several techniques in order (i) to make sense of NL queries with respect to the universe of discourse covered by the ontology and (ii) to map them to semantic markup. AquaLog is a portable question-answering system since the configuration time required to customize the system for a specific ontology is negligible. Also, AquaLog includes a learning component ensuring that the performance of the system improves over time.

Another example of ontology-based system is ORAKEL (Cimiano et al., 2008). ORAKEL is a question-answering system for specialized domains able to answer factoid questions starting with wh-pronouns (except why or how). It parses the user question and constructs a query in logical form formulated with respect to domain-specific predicates. ORAKEL is based on the explicit representation of a lexicon which maps NL constructions to ontological structures. To adapt the system to a specific domain, ORAKEL allows domain experts (or lexicon engineers) to create domain-specific lexicons.

In a related topic, a series of challenges has been organized for question answering over linked data[13] (QALD[14]) since 2011. Lopez et al. presented the issues related to this task and discussed the results of the first and second evaluation campaigns QALD-1 and QALD-2 (Lopez, Unger, Cimiano, & Motta, 2013).

In the biomedical domain, Rinaldi et al. (2004) applied the question-answering system ExtrAns (Mollá, Schwitter, Hess, & Fournier, 2000) to a genomics application using a knowledge base. The knowledge base was constructed from the analysis of documents from medical corpora that were converted into clauses in "Minimal Logical Form". During the querying phase, the user question is analyzed with the same method, then the system tries to align the semantic representation of the user question with the knowledge base. If a matching with clauses in the knowledge base is found, the sentences that originated these clauses are returned as a possible answer.

More generally the performance of the methods used for answer extraction from databases remains variable according to the available resources and remains weak for domains having low data coverage. However these methods have the great advantage of being able to integrate easily new knowledge/data, notably through their standardized data format. A main challenge for methods based on databases is also to provide a justification to each extracted answer. The relevance of these justifications is particularly essential since some data formats are not readable/understandable by the general public (e.g. RDF triples, tuples of a relational database).

3. Proposed approach

In this paper we propose a semantic approach to medical question-answering from document corpora. Fig. 1 presents the main steps of our approach which are: corpora annotation (detailed in Section 4), question analysis (described in Section 5) and answer search (Section 6). We apply NLP methods to analyze the source documents used to extract the answers and the users questions expressed in natural language (NL).

We exploit this first NL analysis to build RDF annotations of the source documents and SPARQL queries representing the users questions. SPARQL is a standard language recommended by the W3C to query RDF data sources. Using SPARQL implies annotating the textual corpora from which the answers are extracted in RDF according to the same reference ontology (cf. Section 3.1). More particularly, the selection of SPARQL as a machine readable language aims to avoid the loss of expressiveness in the query construction phase (e.g. identifying only one focus or one expected answer type while the NL question contains more).

The annotation process relies on the same basic information extraction methods used in question analysis (i.e. named entity recognition and relation extraction). It stores the RDF annotations of the documents in separate files and keeps the

[12] (Popescu et al., 2003) refer to a sentence as *semantically tractable* if the sentence tokenization contains only distinct tokens and at least one of its value tokens matches a wh-value.

[13] http://www.linkeddata.org.

[14] http://greententacle.techfak.uni-bielefeld.de/cunger/qald/.

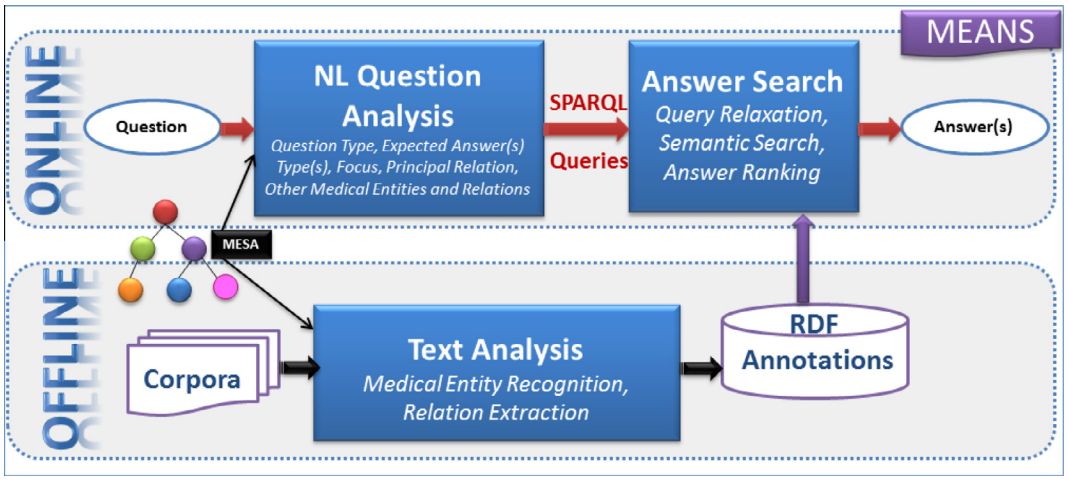


Fig. 1. Overall architecture of the question-answering system MEANS.

links between each sentence and its annotations. On the question analysis side, MEANS constructs several *relaxed SPARQL queries* for the same question in order to reach more answers by dealing with the potential annotations errors and gaps. The final step in the MEANS pipeline is to project the graph patterns expressed by the SPARQL queries on the semantic graphs expressed by the RDF annotations in order to select the RDF nodes that represent the answer. NL answers are then extracted using the annotation links between the RDF nodes and the source texts.

Semantic graphs can suit the formal representation of NL questions at several levels. Semantic graph nodes can represent entities, concepts and literal values and the arcs represent the semantic relations linking them. In the particular case of question formalization, empty graph nodes can also be used to represent the expected answer and its relations with the known entities in the user question. A wide range of NL questions could thus be covered by such graphs. One noticeable shortcoming is the fact that unary relations could not be represented, this, however, could be easily managed when designing the schema or more precisely the ontology that defines the set of concepts and relations to be used. In the same way, semantic graphs are an efficient form of representation for document annotations, except that we are not supposed to have unknown/empty nodes when annotating documents.

Before going into further details we define precisely what will be considered in terms of reference ontology, questions, answers and resources.

3.1. 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 NL questions. MESA 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. patient data). Fig. 2 presents the MESA ontology.

The MESA ontology actually represents 6 medical categories organized hierarchically through the *rdfs:subClassOf* relation. MESA also represents 7 medical relations linking the defined categories. To each medical entity *ME*, we associate:

1. a UMLS concept (by the *mesa:umls_concept* property)
2. a UMLS semantic type (by the mesa:umls_semanticType property)
3. the path of the file containing the medical entity *ME* (by the *mesa:filepath* property)
4. the number of the line containing the medical entity *ME* in the file (by the *mesa:line* property). In our corpora (used to find answers), files contain one sentence per line.

Medical entities and categories will be presented in Section 4.1. Semantic relations linking these entities will be presented in Section 4.2.

4. Offline corpora annotation using NLP methods

NLP methods exploiting medical and semantic resources (e.g. UMLS) are well suited to process textual corpora and to annotate them. We use NLP methods to extract medical entities and semantic relations.

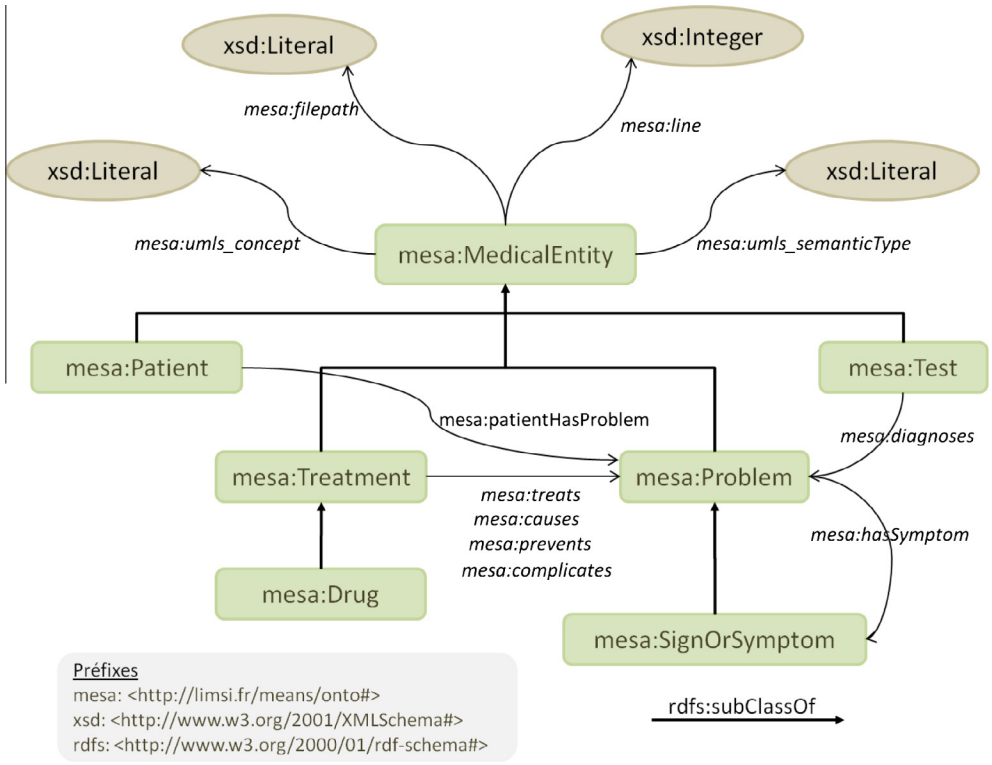


Fig. 2. MESA, the defined reference ontology for the question-answering system MEANS (excerpt).

4.1. 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 defined in MESA: 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 (Ely et al., 2000; Ely et al., 2002; Jacquemart & Zweigenbaum, 2003; Yu et al., 2005). Table 1 presents our medical categories and their corresponding UMLS Semantic Types.

The proposed MER approach uses a combination of two methods (presented and evaluated in Ben Abacha & Zweigenbaum (2011)):

- A rule based method using the MetaMap tool (Aronson, 2001), domain knowledge and a set of filters: *MetaMap Plus* (cf. Section 4.1.1).
- A statistical method using a CRF classifier with a set of lexical, morphosyntactic and semantic features: *BIO-CRF-H* (cf. Section 4.1.2).

4.1.1. MetaMap plus

MetaMap Plus (MM+) uses the MetaMap tool (Aronson, 2001) 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 (Ben Abacha & Zweigenbaum, 2011), (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.

4.1.2. BIO-CRF-H

This method identifies simultaneously entities boundaries and categories using a Conditional Random Fields (CRF) classifier (Lafferty, McCallum, & Pereira, 2001). We use the CRF++ tool[15] and the B–I–O format. The B–I–O format (B: beginning, I:

[15] http://crfpp.sourceforge.net/.

Table 1
Our medical categories and corresponding UMLS semantic types.

Category	Associated UMLS semantic type(s)
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 2
Results per setting on the i2b2 corpus. *R* = recall, *P* = precision, *F* = *F*-measure.

Setting	*P*	*R*	*F*
MM	15.52	16.10	15.80
MM+	48.68	56.46	52.28
TT-SVM	43.65	47.16	45.33
BIO-CRF	70.15	83.31	76.17
BIO-CRF-H	**72.18**	**83.78**	**77.55**

Bold font denotes the best results.

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);
- *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[16] in NLP for Clinical Data (Uzuner, South, Shen, & Duvall, 2011). 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 a subset of first order logic with the predicates `category`, `name` and `position`. For example the set of predicates

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

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

4.1.3. Performance of our MER Methods

We studied MER with three different methods: (i) a semantic method relying on MetaMap (MM+), (ii) chunker-based noun phrase extraction and SVM classification (TT-SVM) and (iii) a last method using supervised learning with CRF (BIO-CRF), which is then combined with the semantic method MM+ by using the results of MM+ as semantic features (BIO-CRF-H). With these methods we particularly studied two processing directions: (i) pre-extraction of noun phrases with specialized tools, followed by a medical classification step and (ii) exploitation of machine-learning techniques to detect simultaneously entity boundaries and their categories. All three methods were experimented on the i2b2/VA 2010 challenge corpus of clinical texts (Uzuner, 2010). Our study showed that hybrid methods achieve the best performance w.r.t machine learning approaches or domain knowledge-based approaches if applied separately (Ben Abacha & Zweigenbaum, 2011).

[16] http://www.i2b2.org/NLP/Relations/.

Table 3
Results per setting and per category on the i2b2 corpus.

Setting	Category	*P*	*R*	*F*
MM+	Problem	60.84	53.04	56.67
	Treatment	51.99	61.93	56.53
	Test	56.67	28.48	37.91
TT-SVM	Problem	48.25	43.16	45.56
	Treatment	42.45	50.86	46.28
	Test	57.37	35.76	44.06
BIO-CRF-H	Problem	**82.05**	**73.14**	**77.45**
	Treatment	**83.18**	**73.33**	**78.12**
	Test	**87.50**	**68.69**	**77.07**

Bold font denotes the best results.

Table 2 presents the results obtained by each method (*R*: Recall, *P*: Precision and *F*: *F*-measure). Recall is the proportion of correctly detected entities among the reference entities. Precision is the proportion of correctly detected entities among those output by the system. *F*-measure is the harmonic means of recall and precision. BIO-CRF and BIO-CRF-H obtained the best precision, recall and *F*-measure. MM+ comes next, followed by TT-SVM and MetaMap alone.

Table 3 presents the obtained results per each medical category (i.e. Treatment, Problem and Test) for three configurations. Again, BIO-CRF-H obtains the best results for all metrics and all categories.

Our experiments showed that performing the identification of entity boundaries with a chunker in a first step limits the overall performance, even though categorization can be performed efficiently in a second step. Using machine learning methods for joint boundary and category identification allowed us to bypass such limits. We obtained the best results with a hybrid approach combining machine learning and domain knowledge. More precisely, the best performance was obtained with a CRF classifier using the BIO format with lexical and morphosyntactic features combined with semantic features obtained from a domain-knowledge based method using MetaMap.

Our MER system uses a combination of an enhanced version of the methods **MM+** and **BIO-CRF-H**. A hybrid method allows balancing the shortcomings of each method if applied alone. Our hybrid method combines both rule-based and statistical methods by taking into account the number of training examples available for each medical category.

4.2. Semantic relation extraction

Relation Extraction plays a central role for relevant questions and texts analysis. This task is not trivial since:

- Different relations exist between two medical entities (e.g. in the UMLS Semantic Network, 5 relations exist between a Medical Problem and a Treatment which are: 'treats', 'prevents', 'complicates', 'causes', and 'associated_with'.)
- A same relation can be expressed in different manners (e.g. several expressions for the same relation "treats or improves" linking a Treatment *TX* and a Problem *PB*: *TX significantly reduces PB*, *TX is safe and effective for PB* or *PB requiring TX*, etc.)

We target 7 semantic relations: *treats*, *complicates*, *prevents*, *causes*, *diagnoses*, *DhD* (Drug has Dose) and *P_hSS* (Patient has Sign or Symptom). These medical categories have been chosen according to an analysis of different 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.

Fig. 3 describes the target medical categories and semantic relation types.

To extract semantic relations, we use a combination of two methods (presented and evaluated in Ben Abacha & Zweigenbaum (2011)):

- A pattern-based method using a set of manually constructed patterns (Section 4.2.1),
- A statistical method using a SVM-classifier with a set of lexical, morphosyntactic and semantic features (Section 4.2.2).

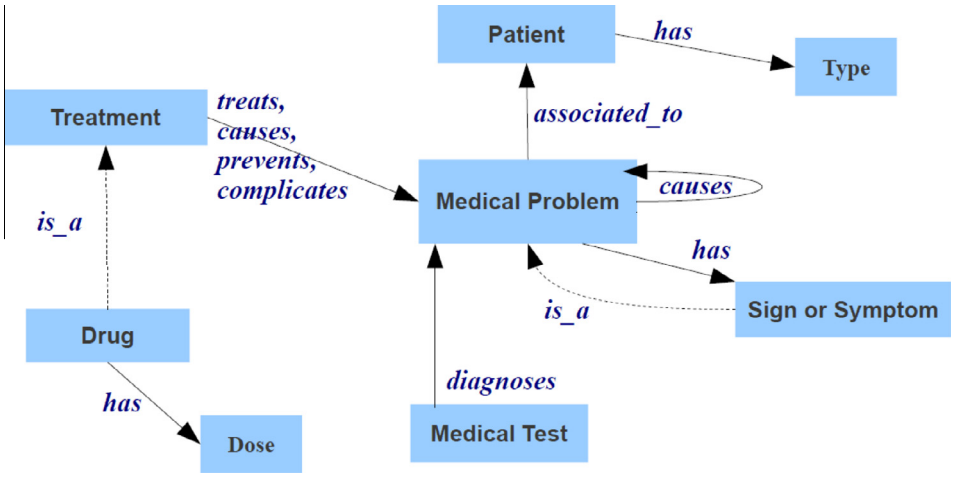


Fig. 3. Domain model (excerpt).

We also extract patient-specific features: (i) Patient Sex, (ii) Patient Age, (iii) Patient Age Group (Adult, Adolescent, Child, Toddler, Infant, Newborn). For this task, we use manually constructed patterns.

4.2.1. 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 (Cohen et al., 2011; Kilicoglu & Bergler, 2011). A benefit of pattern-based methods is also that we do not need an annotated corpus or training data to extract semantic relations.

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 attribute 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.[17]

4.2.2. 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 use a machine-learning method based on a SVM classifier which is trained on the i2b2 2010 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;
4. *Semantic features*: semantic types of E1, E2 and medical entities between E1 and E2.

Results of semantic relation extraction are represented in a subset of 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 relations: treats and patientHasProblem.

[17] http://www.hc-sc.gc.ca/dhp-mps/prodpharma/notices-avis/abbrev-abrev/unitsmeasure_unitesmesure-eng.php.

4.2.3. Hybrid method

This method combines the two preceding methods to compute a global result according to the confidence index associated to the result of each method. Our hybrid method then depends on the influence, or weight, granted to each method. We rely on the number of training examples of a relation to compute the influence of the supervised learning approach on the extraction procedure. This weight is noted $\mu_S(R)$ for a given relation R.

The influence of the pattern-based approach is computed with two different weights: a global weight $\mu_P(R)$, which is the complement of μ_S for a given relation $R : \mu_P(R) + \mu_S(R) = 1$, and a finer-grained weight for each extracted relation occurrence, which takes into account the confidence index computed for this relation occurrence. A pattern-extracted relation only has influence when (i) its confidence index is greater than a given threshold I_{min} and (ii) its global weight is greater than or equal to the weight of the supervised learning method for the same relation: $\mu_P(R) \geqslant \mu_S(R)$.

4.2.4. Performance of our relation extraction methods

We used the corpus of Rosario and Hearst (2004). This corpus is extracted from MEDLINE 2001 and annotated with 8 semantic relations between diseases (DIS) and treatments (TREAT). We chose three relation types: Cure, Prevent and Side Effect. We split the initial corpus into equally-sized training and test corpora for each target relation. We used the training corpus to (i) design sentence patterns for relation extraction with the pattern-based method and to (ii) train our SVM-based method. All methods were then tested on the test corpus.

Table 4 presents the 5 experimented settings. Multi-class machine learning (ML1) uses only one model for all relation types. It provides a multiple classification of sentences into three classes (one class per relation type). Mono-class machine learning (ML2) uses 3 different models, each associated to only one target relation. It provides a binary classification for each relation type (positive and negative). We use the classical measures of Precision, Recall and *F*-measure. Precision is the number of correct relations extracted divided by the number of returned relations. Recall is the number of correct relations extracted divided by the number of reference relations. *F*-measure is the harmonic mean of recall and precision.

Table 5 presents the results for each relation type. Table 6 presents the overall recall, precision and *F*-measure values computed on all extracted relations.

The Hybrid methods effectively outperform the pattern-based and machine learning approaches in terms of *F*-measure. The contribution of both hybrid approaches on the "prevent" and "side effect" relations is important w.r.t the results obtained for these relations by the machine learning approach. In the same way, their contribution is important on the "cure" relation w.r.t the pattern-based technique.

Several semantic relation extraction approaches detect only whether a relation type *T* occurs in a given sentence or not. In our approach we tackle the extraction of medical relationships between specific medical entities in the sentences. We can therefore extract many relations in one sentence (e.g. from the sentence: "*E1 treats E2 but increases the risk of E3*" we can extract cure(E1,E2) and side_effect(E1,E3)).

A second aspect of our work is the contribution of hybrid approaches w.r.t pattern-based approaches and machine learning approaches for relation extraction. In our experiments we observed that pattern-based methods offer good precision values but can be weak when faced with heterogeneous vocabulary and sentences complexity. On the other hand, machine-learning techniques can be very efficient but need enough training examples to obtain good results. The combination of pattern-based and machine learning approaches allows us to take advantage of both techniques. The proposed hybrid approach obtained good results on the "cure" relation extraction because an important number of training examples was available in the corpus (524 sentences in the training corpus). For the other two relations (*Prevent* and *Side Effect*), few

Table 4
Relation extraction: experimental settings.

Pat	Pattern-based method
ML1	Multi-class machine learning
ML2	Mono-class machine learning
H1	Pat + ML1
H2	Pat + ML2

Table 5
Precision *P*, Recall *R* and *F*-measure *F* of each relation for each setting.

Config.	Cure			Prevent			Side effect		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Pat	95.55	32.45	48.44	89.47	51.51	65.37	65.21	53.57	58.63
ML1	90.44	**100**	94.98	15.15	15.15	15.15	0	0	0
ML2	**99.43**	91.97	95.55	90	27.27	41.86	**100**	7.14	13.33
H1	95.07	98.30	96.66	**90**	**54.54**	**67.92**	65.21	53.57	58.82
H2	95.42	98.30	**96.84**	**90**	**54.54**	**67.92**	68.00	**60.71**	**64.15**

Bold font denotes the best results.

Table 6
Precision *P*, Recall *R* and *F*-measure *F* on all relations.

Setting	Precision (%)	Recall (%)	*F*-measure (%)
Pat	91.89	34.51	50.17
ML1	90.52	90.52	90.52
ML2	91.96	91.03	91.49
H1	93.73	93.73	93.73
H2	94.07	94.07	**94.07**

Bold font denotes the best results.

training examples are available (respectively 43 and 44) and machine learning performance was largely diminished (cf. Table 5). However this lack was compensated by (i) the use of manually-constructed patterns and (ii) the weighting of both approaches which takes into account the number of training examples available.

4.3. Offline corpora analysis and annotation

NLP methods exploiting medical and semantic resources (e.g. UMLS) are well-suited to process textual corpora and to annotate them. We use NLP methods to extract medical entities (cf. Section 4.1) and semantic relations (cf. Section 4.2). The next step consists in using the extracted information (i.e. medical entities and relations) to generate RDF triples. The following example presents the generated triples related to the entity "acute pelvic inflammatory disease".

⟨114354_2_6⟩ ⟨http://limsi.fr/means#value⟩ "acute pelvic inflammatory disease".
⟨114354_2_6⟩ ⟨http://limsi.fr/means#category⟩ "problem".
⟨114354_2_6⟩ ⟨http://limsi.fr/means#umls_semanticType⟩ "Disease or Syndrome".
⟨114354_2_6⟩ ⟨http://limsi.fr/means#umls_concept⟩ "Acute pelvic inflammatory disease NOS".
⟨114354_2_6⟩ ⟨http://limsi.fr/means#file⟩ "114354.sn".
⟨114354_2_6⟩ ⟨http://limsi.fr/means#line⟩ "2".
⟨114354_2_6⟩ ⟨http://limsi.fr/means#startPosition⟩ "6".
⟨114354_2_6⟩ ⟨http://limsi.fr/means#endPosition⟩ "9".

5. Online question analysis

5.1. Question classification

We propose a classification of medical questions into 10 categories:

1. *Yes/No questions* (e.g. "Can Group B streptococcus cause urinary tract infections in adults?")
2. *Explanation, Reason or "why" questions* (e.g. "Why do phenobarbital and Dilantin counteract each other?")
3. *Condition, case or the greatest number of "when" questions* (e.g. "When would you use gemfibrozil rather than an HMG (3-hydroxy-3-methylglutaryl) coenzyme A inhibitor?")
4. *Manner, some "how" questions* (e.g. (i) "How are homocysteine and folic acid related to hyperlipidemia?", (ii) "How can you do a screening motor exam of the hand?")
5. *Definition* (e.g. "What is seronegative spondyloarthropathy?")
6. *Factoid* (expected answer is a medical entity, a named entity or a specific information)
   - Type1; expected answers correspond to medical entities (most frequent or at least that we treat), expressed in general with "what", "which" and "how" (e.g. "What is the complete workup for a meconium plug in a newborn?", "How should you treat osteoporosis in a man, caused by chronic steroid use?", "Which test (culdocentesis or pelvic ultrasound) would be best for diagnosis of ovarian cyst in this case?", "Which medication is causing neutropenia in this baby?")
   - Type 2; other expected answer types; expressed in general with when (for Time in this case), where, who, some questions with how (e.g. "When will inhaled insulin be available?", "Where do I go to work up a jaw mass?", "Where would a stroke be located that involved a right facial palsy and dysarthria?", "How much Delsym cough syrup do I give?", "How often should someone have tonometry as a screen for glaucoma?", "How old should a patient be before I quit doing prostate specific antigens (PSA's)?")
7. *List* (e.g. (i) "What are the symptoms of Alport's Syndrome?", (ii) "What are the causes of hypomagnesemia?", (iii) "What are the clinical features and prognosis of post-concussion syndrome?")

8. *Questions with patient description* (e.g. "74-year-old female with memory loss and a negative workup. What is the dose of Aricept?")
9. *Chained questions* (e.g. (i) "What is serum sickness and what are the causes?"/ (ii) "What is cerebral palsy? How do you diagnose and treat it? What is the etiology? What is the pathogenesis?")

5.2. Question analysis and translation into SPARQL queries

We use a generic method for question analysis consisting in (i) extracting question characteristics such as medical entities and semantic relations and (ii) constructing equivalent SPARQL queries. Our method takes into account the case of multiple Expected Answer Types (EAT) and constructs as many queries as expected answers. We focus on medical question of 4 types: Factoid (type 1), Yes/No, Definition and List. 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.

The proposed 6-fold method 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 (m EAT, $m \geqslant 1$ for WH questions and $m = 0$ form Y/N questions)
3. Constructing the question's affirmative and simplified form (new form)
4. Medical Entity Recognition based on the new form of the question
   for $(x = O, x++, x \leqslant m)$
5. Extraction of semantic relations based on the new form *[Input: (i) EAT$_x$ or NIL (for Y/N questions), (ii) Medical Entities, and (iii) new question form]*
6. Construction of SPARQL Query x

We note that other information extraction tools could be used for the first 4 steps (which are independent from the final translation step).

Table 7 presents the output of each step on two examples.

For the first 3 steps, we use manually constructed patterns from a large number of clinical questions. We distinguish 3 pattern categories:

- *Definition question*. This type of question is detected with patterns like: What is X? What does X mean? (X is a noun phrase). In this case, X is the focus of the question. The simplified form of the question will be X. Question analysis will consist only of medical entity recognition and SPARQL query construction.
- *Yes/No question*. This type of question is recognized using patterns that detect the absence of a WH pronoun and check the specific structure of Yes/No questions. Simple transformations are needed to transform the question into the affirmative form. Questions analysis will consist of Medical entity recognition, relation extraction and SPARQL query construction.
- *Factoid question*. This type of question is recognized using a set of simple rules on user questions (e.g. firstWordOf(Q) = (How/What/Which/When) indicates that question Q is a WH question) and patterns which allow also determining expected answer type (and whether it is a list question) and constructing the affirmative and simplified form. Next steps will consist of medical entity recognition, semantic relation extraction and SPARQL query construction.

For Factoid questions, we determine EAT by matching the NL questions with manually built lexical patterns. A set of patterns is constructed for 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 case, 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, Test).

Table 7
Medical question analysis – examples (EAT: Expected Answer Type, PB: Problem, PA: Patient, TX: Treatment).

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)

In a second step we construct the question's affirmative form. 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.

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. Medical entity recognition is performed using our hybrid method described in Section 4.1. In a last step before the SPARQL query construction, relation extraction is performed using our hybrid method for relation extraction described in Section 4.2.

In the SPARQL query construction step, we use the ASK and SELECT query forms in order to represent Yes/No questions and WH questions respectively. 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). The body of the output SELECT or ASK query contains the basic graph pattern constructed using the medical entities and relations extracted from the NL question.

The final translation of the user query consists in one or several SPARQL queries (if we have more than one EAT) constructed by assembling the unitary translations of each predicate obtained from the medical entity recognition and the relation extraction steps.

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

- Expected Answer Types:

  EAT1 = Test, EAT2 = Treatment

- Affirmative, simplified question form:

  ANSWER to diagnose and treat Anxiety Disorder

- Medical Entity Recognition:

  ⟨PB⟩Anxiety Disorder⟨/PB⟩ (*the focus of the question).*

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

```
SELECT ?answer1 WHERE {
?answer1 mesa:category mesa:Test
.?answer1 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 ?answer2 WHERE {
?answer2 mesa:category mesa:Treatment
.?answer2 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 8 (i.e. 'Select ?answer1 **?text1**').

Table 8 shows an example of a translated WH question.

5.3. Performance of our question analysis method

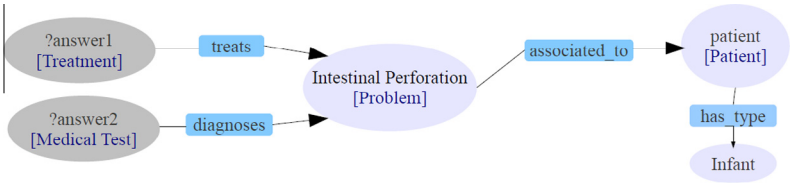
We used 100 questions extracted from the Journal of Family Practice JFP[18] (latest questions from 11/1/2008 to 4/1/2011) to evaluate our question analysis method (Ben Abacha & Zweigenbaum, 2012). This set of questions contained 64 WH questions and 36 Yes/No questions. We tested our 2 MER methods on 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 9 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.

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

Table 8
Example of a translated WH question.

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⟩
  ?patient mesa:hasProblem ?focus
  ?patient mesa:category ⟨Infant⟩
  OPTIONAL{
  ?text1 mesa:contains ?answer1
  ?text1 mesa:contains ?patient
  ?text1 mesa:contains ?focus}}
```

SPARQL Query 2

```
Select ?answer2 ?text2
where {
  ?answer2 mesa:category ⟨MedicalTest⟩
  ?answer2 mesa:diagnoses ?focus
  ?focus mesa:name 'intestinal perforation'
  ?focus mesa:category ⟨Problem⟩
  ?patient mesa:hasProblem ?focus
  ?patient mesa:category ⟨Infant⟩
  OPTIONAL{
  ?text2 mesa:contains ?answer2
  ?text2 mesa:contains ?patient
  ?text2 mesa:contains ?focus}}
```

Table 9
Evaluation of MER on our question corpus (categories: Treatment, Problem and Test).

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

Bold font denotes the best results.

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.

6. Answer search

Ontology-based information retrieval has several benefits such as (i) handling synonymy and morphological variations and (ii) allowing semantic query expansion or approximation through subsumption and domain relations. For example, if we search the term "ACTH stimulation test", the semantic search engine can find documents containing "cosyntropin test" or "tetracosactide test" or "Synacthen test" as they are synonyms. Also, if we search treatments for "Cardiovascular disease" we can also find, for instance, treatments for "Myocardial infarction" or "congestive heart failure" or "Endocarditis" since they are sub-classes of Cardiovascular Diseases.

6.1. Query relaxation approach

The annotation and translation processes presented above allow a semantic search based on the MESA ontology. However it is common that the linguistic and semantic annotation processes produce errors or miss existing information either in the user questions or in the source documents used to extract the answer. The answer search module of MEANS implements a query relaxation approach that we defined in order to tackle annotation errors.

As described in Section 5, for each user question, MEANS constructs one or several initial queries according to the number of expected answers. These initial queries correspond to a precise representation of the most important characteristics of the question (i.e. medical entities, semantic relations, information about patient, etc.). In the answer search phase, we associate to each initial SPARQL query one or several less precise queries containing less constraints (e.g. only medical entities and relations, or only medical entities, etc.), which can be considered as *relaxed forms* of the initial query. These relaxed forms are constructed dynamically by truncating one more triple pattern of the user query at each step.

To begin with a concrete example, for the question:

(1) "What is the best treatment for oral thrush in healthy infants?"[19]

We construct a first precise SPARQL query (cf. Fig. 4a), then we generate less precise queries. Fig. 4 presents the less precise SPARQL query associated to the question. The presented SPARQL queries in Fig. 4a and b are, among others, generated automatically by our question-answering system MEANS.

We define three query-relaxation levels for an initial query. Each level includes one or several queries that we rank according to their precision (i.e. their closeness to the initial question):

- **LEVEL 1**: This level includes the initial query and then a relaxed form of this query obtained by deleting the values of named entities. For instance, in the question (1), we will look for sentences containing "healthy infants" but also sentences containing medical entities having as UMLS concept *Infant* without precising the value of the entity. This first relaxation type can lightly decrease the precision (we keep the UMLS concept which guarantees to look for the appropriate medical entity), but will highly increase the recall.
- **LEVEL 2:** At this level, we delete medical entities one by one, but we keep (i) the expected answer and (ii) the question focus.
- **LEVEL 3:** The last step consists in deleting the principal relation(s). A principal relation is defined as a relation having the expected answer as subject or object.

6.2. Answer ranking and presentation

Constructed queries are executed in order to interrogate RDF triples generated on the corpus-annotation step. Two answers are considered as **identical** (or equivalent) if the medical entities returned as answers have the same UMLS CUI[20] (e.g. the corticosteroid injection, corticosteroid injections: C2095490). Justifications are then grouped by entity/CUI and counted.

Then, answers are ranked according to **two criteria**:

(1) According to the first criteria (applied to all questions), answers are ranked according to the queries rank (as queries are ranked from the most precise query to the less one).
(2) An additional ranking is performed on answers for factual questions since several answers can be extracted by the question-answering system. Our second ranking criteria for factual questions takes account of the number of justifications. In fact, for N answers, we select an answer A1 rather than an answer A2 if there are more justifications for A1 than there are for A2.

[19] This question is extracted from our evaluation set of questions.
[20] Concept Unique Identifier.

```
SELECT ?concept1 ?file ?line WHERE {

?concept1 <http://limsi.fr/means#file> ?file .
?concept1 <http://limsi.fr/means#line> ?line .

?concept2 <http://limsi.fr/means#file> ?file .
?concept2 <http://limsi.fr/means#line> ?line .

?concept3 <http://limsi.fr/means#file> ?file .
?concept3 <http://limsi.fr/means#line> ?line .

?concept1 <http://limsi.fr/means#concept_category> "Treatment" .

?concept2 <http://limsi.fr/means#concept_category> "Problem" .
?concept2 <http://limsi.fr/means#umls_semanticType> "Disease or Syndrome" .
?concept2 <http://limsi.fr/means#umls_concept> "Oral candidiasis" .
?concept2 <http://limsi.fr/means#value> "oral thrush" .

?concept3 <http://limsi.fr/means#concept_category> "Patient" .
?concept3 <http://limsi.fr/means#umls_semanticType> "Age Group" .
?concept3 <http://limsi.fr/means#umls_concept> "Infant" .
?concept3 <http://limsi.fr/means#value> "healthy infants" .

?concept1 <http://limsi.fr/means#treats> ?concept2 .
?concept3 <http://limsi.fr/means#patient_has> ?concept2 .

}
```

(a) The most precise query

```
SELECT ?concept1 ?file ?line WHERE {
?concept1 <http://limsi.fr/means#file> ?file .
?concept1 <http://limsi.fr/means#line> ?line .

?concept2 <http://limsi.fr/means#file> ?file .
?concept2 <http://limsi.fr/means#line> ?line .

?concept1 <http://limsi.fr/means#concept_category> "Treatment" .

?concept2 <http://limsi.fr/means#concept_category> "Problem" .
?concept2 <http://limsi.fr/means#umls_concept> "Oral candidiasis" .
}
```

(b) The less precise query

Fig. 4. SPARQL queries automatically generated by the MEANS system for the question "What is the best treatment for oral thrush in healthy infants?".

Table 10
Questions vs. answers *[ME: Medical Entity (e.g. Cancer, Depression), REL: the Principal Relation (e.g. Treats, Prevents)].*

Question types	Examples of question models	Simplified forms	Examples of answers
Definition	What is *ME*?	*ME*	*ME* is . . .*(a definition)*
Yes/No	Can *ME*1 REL *ME*2? *(e.g. Can Vitamin D Deficiency cause High Blood Pressure?)*	*ME*1 REL *ME*2	Yes or No
Factual or List: Known EAT & REL	(1) How can you REL *ME*2? *(e.g. How can you prevent Metabolic Syndrome?)* (2) What is *ME*1 REL *ME*2? *(e.g. What is the best Antidepressant to treat Bipolar Disorders?)*	ANSWER? REL *ME*2 (we delete *ME*1)	*ME*1.1? REL *ME*2 with: *ME*1.1 is a *ME*1 (e.g. Lithium [*ME*1.1] is an Antidepressant [*ME*1])
Factual or List: Unknown EAT & REL	What is *ME*1 REL *ME*2?	(1) *ME*1 REL *ME*2 (2) ANSWER? REL *ME*2	*ME*1.1? REL *ME*2 (with, *ME*1.1 is a *ME*1)

Several answer types can be returned by the question-answering system *MEANS* according to question type:

- *Definition question*: Answer type = a sentence
- *Boolean question*: Answer = Yes or No
- *Factual question*: Answer = named entity (e.g. a treatment) or a precise information (e.g. a drug dose)
- *List question*: Answer = a list of medical entities (e.g. list of symptoms)

Table 10 presents some examples of question models and their corresponding answers.

As defined before, a **justification** for an answer is the sentence containing the found answer by our question-answering system. It is important to point out that many justifications are understandable only in their context (sentences before and/or after) (Cao, Ely, Antieau, & Yu, 2009). We therefore present for each question: (i) the answer, (ii) the justification, (iii) 2 sentences before the justification and (iv) 2 sentences after.

7. Evaluation of the question-answering system MEANS

7.1. Evaluation criteria

Several QA evaluation campaigns such as TREC[21] (English), CLEF[22] (multi-language), NTCIR[23] (Japenese) and Quaero[24] (French, English) have been conducted in open domain. In the medical field, English QA challenges are rare. The Genomics task of the TREC challenge can be seen as an related track, even if it was not officially introduced as such.

The performance of QA systems is often evaluated using the MRR (Mean Reciprocal Rank), the precision and the recall of the returned answers, under the assumption that the QA system gives an ordered list of answers, potentially with a fixed maximum number of answers.

- *MRR*: Mean Reciprocal Rank of the first correct answer (i.e. 1 if the first answer is correct, 0.5 if the second answer is the first correct answer, etc.)
- *Recall*: Proportion of correct answers returned by the system w.r.t. the number of all possible correct answers.
- *Precision*: Proportion of the returned correct answers among all the answers returned by the system.

A relevant application of these measures requires to define precisely what is a "correct" answer and what criteria must be taken into account in the expected answers?

A first answer to these two questions can be in the following key points:

1. **Granularity:** In open domain, a question like "Where is the Louvre museum?" has several possible answers:
   - *The Louvre museum is in France.*
   - *The Louvre museum is located in Paris.*
   - *The Louvre museum is located in an old royal palace built by François premier on the old fortress of king Philippe Auguste.*
   - *The Louvre is on the opposite side of the Opéra avenue and it can be reached after a ten-minute walk from the Horset Opera hotel.*

   We can then distinguish several evaluations of the answers that can be asserted to be, for instance, correct, complete, incomplete or false. In the example above, an important ambiguity point comes from the determination of the type of the expected answer. It can be a neighborhood, a city, a country, a complete address or even travel hints. In the medical domain, this problem arises in open questions (e.g. "To what extent can we say that cell phones are harmful?") but it is more restricted in closed questions (e.g. factual questions, lists) where the expected answer type is explicit or in Yes/No questions.
2. **Answer justification:** The justification is the text (e.g. sentence) containing the extracted answer. This aspect leads to different case studies, such as: (i) the answer can be correct but the justification is false or (ii) the answer can be wrong at present time even if the justification is correct. For instance, In the question "Who is the Secretary General of the United Nations?" the answer "Boutros Boutros-Ghali" is false, even if the justification "*Dr. Boutros Boutros-Ghali, before his selection as sixth Secretary General of the UN, had been a Foreign Minister of Egypt, a professional diplomat, a lawyer.*" is correct.

In addition to the previously mentioned performance aspects, another important aspect for question-answering systems is **quickness**. More particularly, in the medical field, a study made by Takeshita, Davis, and Straus (2002) shows that doctors need to access to the information in less than 30 s and cancel their search beyond.

[21] http://trec.nist.gov.
[22] http://clef.isti.cnr.it.
[23] http://research.nii.ac.jp/ntcir.
[24] http://www.quaero.org.

Table 11
Number of boolean questions answered correctly and error types per category.

Type	Number of questions	L1	L1, L2, L3	Error types
Treats	9	5	5	E1, E2, E4
Prevents	4	0	3	E3, E4
Diagnoses	1	1	1	–
Others	6	3	4	E3, E4
Total	20	9	12	–

7.2. Evaluation data

In the scope of our evaluation we used a "standard" corpus[25] (Mollá, 2010; Mollá & Santiago-Martínez, 2011). This corpus, called Corpus for Evidence Based Medicine Summarisation, consists of 50 questions randomly collected from the Journal of Family Practice (JFP).[26] More precisely, the questions are collected from the clinical questions section and the answers are extracted from MEDLINE articles. The corpus uses XML markup and is annotated with the following elements:

- The clinical question,
- The answer(s) to the question,
- Text passages justifying the answers extracted from MEDLINE
- Reference to the articles containing the passages with their PubMed identifiers.

This sequence of 50 questions contains 39 questions having semantics (i.e. medical entities and relations) that can be expressed by our ontology, including 20 Boolean questions and 19 factual questions. The 11 remaining questions involve semantics that cannot be expressed by our ontology.

In our evaluation, an answer is considered as *correct* if:

- the medical entity (for factual questions) or the Boolean value (for Yes/No questions) is correct, and
- the justification returned by the system is a correct justification.

In the XML corpus used for the evaluation, answers were developed manually from MEDLINE articles. We used these reference answers to evaluate manually the MEANS system. An automatic evaluation is not possible since justifications also have to be evaluated.

7.3. Results for boolean questions

As mentioned above, two elements are taken into account in the evaluation of the answers. In the case of Boolean questions it is: (i) the answer value (Yes/No) and (ii) the returned justification. We measure the precision of MEANS according to these two criteria.

Table 11 presents the obtained results for the 20 Boolean questions. $L1, L2$ and $L3$ (i.e. level 1, level 2 and level 3) represent the relaxation level that is applied. For boolean questions, we evaluated the answers returned by level 1 then the answers returned by all 3 levels.

We denote by "others", the boolean questions whose main relations differ from the main relations covered by our system, for instance the relation *have a role* ("Does routine amniotomy have a role in normal labor?") or the relation *tolerated* ("Are any oral iron formulations better tolerated than ferrous sulfate?").

The obtained precision for boolean questions is **45%** if we use only level 1 of query relaxation and **60%** if we use level 3. Precision and recall have the same value here as (i) either the system answers "no" to a question instead of the correct answer "yes" and there will be no justifications to evaluate (counts for 1 automatically-detected error) or (ii) the system answers correctly "yes" and we did not have erroneous justifications in these cases.

After an analysis of the system errors, we observed 4 types of error:

- **E1**: The answer does not exist in the collection of articles we are using (some files contain only the title of the article).
- **E2**: Answers/justifications are sometimes described over 2 sentences or more (e.g. description of experimental results).
- **E3**: A main entity/relation was not detected. This is the case, for instance, for the following questions:
  - *"Does reducing smoking in the home protect children from the effects of second-hand smoke?"*
  - *"What is the appropriate use of sunscreen for infants and children?"*
  - *"Do preparticipation clinical exams reduce morbidity and mortality for athletes?"*

[25] We thank Dina Demner-Fushman who led us to the selection of this corpus.
[26] http://www.jfponline.com.

Table 12
Answers to factual questions: MRR and precision at 5 answers (in %).

Type	Quest. Nbr	L1		L1 + L2 + L3	
		MRR	P@5	MRR	P@5
Treats	8	0.625	70.58	1	62.5
Prevents	1	0	–	1	60
Diagnoses	5	0	–	0.432	25
Causes	1	0	–	1	20
Manages	3	0.66	100	0.5	80
2 EAT or more	1	0	–	1	66.66
Total	39	**0.42**	**85.71**	**0.77**	**57.47**

- **E4**: Questions needing external knowledge or inference, as:
  - *"Does heat or cold work better for acute muscle strain?"* "Cold" here does not refer to a medical problem but rather to "Cold Therapy" or "Cryotherapy", the same goes for the word "heat" ("heat therapy").
  - *"Does psychiatric treatment help patients with intractable chronic pain?"* Treatments mentioned in the articles can be subtypes of "psychiatric treatment " such as "Cognitive therapy".
  - *"Do antiarrhythmics prevent sudden death in patients with heart failure?"* The potential answers can be present in the articles as sub-types, for instance "beta blockers" are "antiarrhythmic drugs".

For some questions, the answer is not detected but, in fact, the articles supposed to contain it are provided only with a title and without content. This problem impacted the results of the questions of type "treats" even if they were correctly analyzed overall. Another point for some of the "treats" questions is the answer had to be read on several sentences. For instance, for the question "Does yoga speed healing for patients with low back pain?", the returned justification is "Yoga for women with hyperkyphosis: results of a pilot study" whereas in the corpus the full justification is: *"In a case series, 21 women aged >60 years (mean age, 75) with hyperkyphosis, participated in twice-weekly 1-hour sessions of hatha yoga for 12 weeks. Measured height increased by a mean of 0.52 cm, forward curvature diminished, patients were able to get out of chairs faster, and they had longer functional reach. Eleven patients (48%) reported increased postural awareness/improvement and improved well-being; 58% perceived improvement in their physical functioning."*.

For E3 errors, query relaxation enhanced the results (e.g. for the "prevents" questions). Indeed, relaxation allowed MEANS to find further elements answering the SPARQL queries, which consequently led to positive (yes) answers. These positive answers were correct answers, while, without relaxation, the system returned negative "no" answers due to the lack of information. We did not have noisy answers in our relaxation experiments, mainly due to the fact that the strongest relaxation keeps at least two main entities of the question (this heuristic is restrictive for some questions but it was used to guarantee a minimum of precision in the scope of the QA task).

In some cases, even if the question is correctly analyzed and the associated SPARQL query is correct, the system did not retrieve the correct answer because of the lack of external knowledge or inference methods. For instance, in the question "Do antiarrhythmics prevent sudden death in patients with heart failure?", the existing answers were, for instance, (i) *Beta-blockers to reduce mortality in patients with systolic dysfunction: a meta-analysis* or (ii) *Beta-blockers are particularly effective in people with a high sympathetic drive (i.e., high pulse rates) to lower blood pressure and reduce cardiovascular risk*. In order to retrieve them it was necessary (i) to infer that "systolic dysfunction" is a type of "heart failure", that "Beta-blockers" are "antiarrhythmics" and that the relations "reduce mortality" and "prevent (sudden) death" are sufficiently similar.

7.4. Results for factual questions

According to the definition of correct answers, answers to factual questions are assessed as "correct" if the good (correct) medical entity is returned and the justification is correct. Thus, incomplete medical entities (e.g. lacking one word) or entities extracted with noise are considered as false, and the answers returned with correct entities but false justifications are considered as false. Table 12 presents the obtained results for the 19 factual questions of the reference QA corpus. $L1, L2$ and $L3$ represent the relaxation levels 1, 2 and 3.

The final precision on all factual questions is **85.71%** without relaxation with a MRR of 0.42. The MRR increased by 0.35 with the relaxation that allowed MEANS to find more answers. However, relaxation also decreased the overall precision. This was expected, however the loss of 0,28 precision points is alleviated by the increase in number of answers and, most importantly, the enhancement of the MRR. The results obtained for each factual question are detailed in Table 13.

For a concrete example, we present the results obtained for the question Q1 of Table 13. The question Q1 is "What is the best treatment for oral thrush in healthy infants?". Without relaxation, two correct answers are retrieved by MEANS for Q1: "Nystatin" and "Nystatin Suspension".[27] With the first level of relaxation (i.e. deletion of the exact textual values) MEANS retrieves one supplementary and correct answer: "Fluconazole". The second level of relaxation (i.e. deletion of the entity "healthy infants") allows retrieving another supplementary correct answer "Gentian Violet". However, the third level of

[27] Nystatin and Nystation suspensions have two different CUIs in UMLS, they concretely considered as two different medical entities.

Table 13
Detailed results by question and category.

Category	Q	P@5 (x)		MRR		P@5/category		MRR/category	
		$L1$	$L1,2,3$	$L1$	$L1,2,3$	$L1$	$L1,2,3$	$L1$	$L1,2,3$
Treats	Q1	100 (3)	100 (5)	1	1	70.58	62.5	0.625	1
	Q2	na (0)	40 (5)	1	1				
	Q3	60 (5)	60 (5)	1	1				
	Q4	na (0)	100 (5)	0	1				
	Q5	na (0)	60 (5)	0	1				
	Q6	100 (2)	60 (5)	1	1				
	Q7	80 (5)	80 (5)	1	1				
	Q8	na (0)	100 (5)	0	1				
Diagnoses	Q9	na (0)	20 (5)	0	0.33	na (0)	0.25	0	0.432
	Q10	na (0)	0 (5)	0	0				
	Q11	na (0)	40 (5)	0	0.5				
	Q12	na (0)	20 (5)	0	1				
	Q13	na (0)	50 (4)	0	0.33				
Prevents	Q14	na (0)	60 (5)	0	1	na (0)	60	0	1
Manages	Q15	na (0)	na (0)	0	0	100	80	0.66	0.5
	Q16	100 (3)	80 (5)	1	1				
	Q17	100 (3)	80 (5)	1	0.5				
Causes	Q18	na (0)	20 (5)	0	1	na (0)	20	0	1
2 TRA	Q19	na (0)	66.66 (3)	0	1	na (0)	66.66	0	1

P@5 (x): precision on the first 5 answers and (x) number of retrieved answers (maximum number fixed to 5).
RR:
na: "no answers", precision cannot be calculated.
Li: relaxation level *i*.

Table 14
Comparison of the results of MEANS on factual questions and a keyword-based IR system.

System	Evaluation of	P@5	MRR
MEANS	Returned documents	$\geqslant$0.574	$\geqslant$0.77
	Extracted answers	0.574	0.77
Keyword-based IR	Returned documents	0.488	0.726
	Extracted answers	$\leqslant$0.488	$\leqslant$0.726

relaxation (i.e. deletion of the main relation, here *treats*) brings 5 new answers containing only 1 correct answer: "Miconazole Gel". Let's note here that Table 13 presents only the results for the first 5 answers and consequently, it does not show all the answers of this example, but only the first 5, which are obtained with the first answer of the level 3 of relaxation.

7.5. Keyword-based information retrieval vs. semantic question answering

To compare our results to the results of baseline keyword approaches, we implemented a keyword-based IR system using the Terrier IR platform[28] for indexing and retrieving documents (Ounis et al., 2006; Ounis & Lioma, 2007). The indexing process includes tokenization, stop-words removal and stemming using the Porter algorithm. The same process is used for query analysis. A query is processed by removing stopwords and applying stemming. If a query expansion is applied, an appropriate term weighting model is selected and the most informative terms from the top ranked documents are added to the initial query. Terrier includes a wide variety of models such as the classical TF-IDF weighting scheme, the Okapi BM25 probabilistic ranking formula and the LGD model.

We compared different IR models and selected BM25 which gave the best results on our corpus. We applied query expansion choosing the top 30 terms from the top ranked 20 documents.

Implementing a question-answering system based on this IR module requires adding a second module for answer extraction using different NLP techniques. We limit our efforts for this experiment to developing the first IR module. The results of this module can give an idea about the general performance of a "good" question answering system based on keyword IR.

We ran the system on the same 50 questions used for the evaluation of MEANS. Table 14 compares the results of the keyword-based IR system and MEANS. Table 15 details the results (returned documents) of the keyword-based IR system.

MEANS obtained better results (MRR = 0.77 and P@5 = 0.574 on factual questions) than the IR system (MRR = 0.726 and P@5 = 0.488), even if MEANS deals with a more challenging problem which is finding "precise answers" rather than the "documents" containing the answers. We explain this performance by two main factors: (i) query relaxation through 3 levels (in

[28] http://terrier.org/.

Table 15
Evaluation of the returned documents by the keyword-based IR system: detailed results for each question/query.

Question	P@5	MRR	Question	P@5	MRR
1	0.6	1	26	0.4	0.3
2	0	0	27	0.4	1
3	0.4	1	28	0.8	1
4	0.8	1	29	1	1
5	0.4	1	30	0.6	1
6	0.8	1	31	0.4	1
7	0.8	1	32	0.2	1
8	1	1	33	0	0
9	0.2	0.5	34	0	0
10	1	1	35	0.8	1
11	0.6	0.5	36	0	0
12	1	1	37	0	0
13	0	0	38	0.6	0.5
14	0.4	1	39	0.8	1
15	0.4	0.5	40	0	0
16	1	1	41	0.6	0.5
17	0.2	1	42	0.2	1
18	0.6	1	43	0.2	1
19	1	1	44	0.2	0.5
20	1	1	45	0.2	0.25
21	1	1	46	0.2	0.25
22	0	0	47	1	1
23	0.4	1	48	0.2	1
24	0.6	1	49	0.4	1
25	0.6	1	50	0.4	0.5

order to control the precision/recall ratio semantically rather than statistically) that improved the MRR and (ii) semantic analysis of queries and documents based on medical entity recognition and semantic relations in order to improve the precision.

8. Discussion

8.1. NLP methods

We experimented NLP methods on different corpora (e.g. clinical texts, scientific articles). Our experiments confirm the fact that:

- Rule or pattern based methods have average performances, even weak, on large corpora with heterogeneous documents.
- Statistical methods can be very robust, but their performance diminishes significantly if (i) a small number of training examples is available and/or (ii) the test corpus has different characteristics from the training corpus.

To tackle the scalability issues of rule-based and statistical methods, we used hybrid methods for information extraction from texts and questions. These hybrid methods take into account the number of training examples for each medical category/relation.

8.2. Semantic approach

We presented the medical QA system *MEANS* and evaluated its results with medical questions extracted from the QA corpus proposed by (Mollá, 2010; Mollá & Santiago-Martínez, 2011). The obtained results are promising according to both precision and MRR. *MEANS* constructs several SPARQL queries to represent the NL question of the user and ranks them in a decreasing precision (specificity) order before their execution. The obtained results show that this ranking at query level allows an efficient ranking of the returned answers. For factual questions, the relaxation led to a partial decrease of the precision measure which is however acceptable as a small counterpart w.r.t the retrieval of new answers and to the substantial enhancement of the MRR.

However, 11 of the 50 questions of the evaluation corpus have not been processed because they involve specific types of answers that are not yet processable by *MEANS*. This will be the subject of our perspectives which will tackle the support of new question/answer types.

8.3. Error analysis

The experimentation of MEANS on this first corpus shows the assets of query relaxation for the performance of the QA system: the overall precision increased for boolean questions and the MRR was enhanced for factual questions. In the first

case (i.e. boolean questions), the lack of retrieved information, due to the initial, strongly restrictive, form of the query is interpreted as a negative (no) answer, while positive and correct answers are present in the source corpus and reachable through relaxation.

Several error cases for boolean questions relate to the presence of *negations* in noun/verbal phrases (e.g. *"Another Cochrane review found no added benefit in function from combining deep transverse friction massage with ultrasound or a placebo ointment"*). Negations are not taken into account in the current implementation, their integration is among short-term perspectives.

From another side, determining the *level of certainty* plays an important role in the selection of relevant justifications. For instance, the current annotation of the sentence *"There's insufficient evidence to support specific physiotherapy methods or orthoses (braces), shock wave therapy, ultrasound, or deep friction massage"* led to a wrong justification.

The main observations of error causes in the experiments on factual questions are the failures that occur in the context of comparisons, an aspect that is not currently taken into account by *MEANS*. For instance, the question "What are the most effective treatments for *PB*?" which is "looking for" the best treatments for a disease, abstracted as *PB*, cannot have as answer a treatment *T*1 with the justification "T1 is less effective than T2". *T*1 can however be considered as a correct answer if the contrary justification exists, as sometimes studies lead to contradictory results (we had an occurrence of this case in our experiments).

9. Conclusion

In this paper, we tackled automatic Question Answering in the medical domain and presented an original approach having four main characteristics:

- The proposed approach allows dealing with different types of questions, including questions with more than one expected answer type and more than one focus.
- It allows a deep analysis for questions and corpora using different information extraction methods based on (i) domain knowledge and (ii) Natural Language Processing techniques (e.g use of patterns, machine learning). Text and question analysis includes the extraction of (i) medical entities, (ii) semantic relations and (iii) additional information about the patient (age, sex, etc.).
- It is based on Semantic Web technologies, which offer more expressiveness, standard formalization languages and makes our corpus and question annotations sharable through the Web.
- Our approach includes a novel query relaxation method which deals with errors or weakness of NLP methods in some cases.

Three main perspectives are planned:

- **Scalability**. 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 NL 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 triples (Entity1-Relation-Entity2).
- **Hybrid information-extraction methods**. Our hybrid methods are based on the number of training examples available for each entity or relation type in order to use the result of the rule-based or the statistical method. We plan to include other factors to combine these methods.
- **Exploiting linked open data**. Many medical knowledge bases adopted W3C standards to publish their data online as Linked Open Data (e.g. BioPortal). The availability of these data provides additional answer sources that can be exploited together with textual corpora. We plan to use Linked Open Data as a complementary answer source if the query cannot be answered from RDF annotations.

Appendix A. List of questions used for the evaluation of MEANS

Here, the list of 50 questions that we used for the evaluation of the question-answering system MEANS. We randomly selected these questions from the Corpus for Evidence Based Medicine Summarisation (Mollá, 2010; Mollá & Santiago-Martínez, 2011). These questions were initially collected from the Journal of Family Practice (JFP).[29]

1. How effective are nasal steroids combined with nonsedating antihistamines for seasonal allergic rhinitis?
2. Is screening for lead poisoning justified?

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

3. Do nasal decongestants relieve symptoms?
4. Do antiarrhythmics prevent sudden death in patients with heart failure?
5. What are the most effective ways you can help patients stop smoking?
6. Does antepartum perineal massage reduce intrapartum lacerations?
7. What is the best way to screen for breast cancer in women with implants?
8. What is the best treatment for oral thrush in healthy infants?
9. How can you prevent migraines during pregnancy?
10. What is the appropriate use of sunscreen for infants and children?
11. Does treatment with donepezil improve memory for patients with mild cognitive impairment?
12. What is the appropriate evaluation and treatment of children Who are "toe walkers"?
13. Does furosemide decrease morbidity or mortality for patients with diastolic or systolic dysfunction?
14. What interventions reduce the risk of contrast nephropathy for high-risk patients?
15. Can counseling prevent or treat postpartum depression?
16. What is the best way to manage phantom limb pain?
17. What is appropriate fetal surveillance for women with diet-controlled gestational diabetes?
18. How should patients with Barrett's esophagus be monitored?
19. Does reducing smoking in the home protect children from the effects of second-hand smoke?
20. What treatment works best for tennis elbow?
21. Does routine amniotomy have a role in normal labor?
22. Does psychiatric treatment help patients with intractable chronic pain?
23. How does pentoxifylline affect survival of patients with alcoholic hepatitis?
24. Does heat or cold work better for acute muscle strain?
25. Prophylactic oxytocin: Before or after placental delivery?
26. How accurate is stress radionuclide imaging for diagnosis of CAD?
27. What causes a low TSH level with a normal free T4 level?
28. Who should get hepatitis A vaccination?
29. What is the best way to distinguish type 1 and 2 diabetes?
30. Do preparticipation clinical exams reduce morbidity and mortality for athletes?
31. What are effective medical treatments for adults with acute migraine?
32. What are the indications for bariatric surgery?
33. How best to manage the patient in term labor Whose group B strep status is unknown?
34. How often should you follow up on a patient with newly diagnosed hypothyroidism?
35. What is the best diagnostic approach to postmenopausal vaginal bleeding in women taking hormone replacement therapy?
36. What are the indications for treatment with angiotensin-converting enzyme ACE inhibitors in patients with diabetes?
37. What is the value of screening for heart disease with an exercise stress test (EST) in an asymptomatic person?
38. What are the most effective treatments for bacterial vaginosis in nonpregnant women?
39. Which tool is most useful in diagnosing bipolar disorder in children?
40. Does screening for diabetes in at-risk patients improve long-term outcomes?
41. What is the best treatment for diabetic neuropathy?
42. Do inhaled beta-agonists control cough in URIs or acute bronchitis?
43. Does yoga speed healing for patients with low back pain?
44. Is methylphenidate useful for treating adolescents with ADHD?
45. Should we screen women for hypothyroidism?
46. What is the best approach for managing recurrent bacterial vaginosis?
47. What are effective treatments for panic disorder?
48. Is there a role for theophylline in treating patients with asthma?
49. Are any oral iron formulations better tolerated than ferrous sulfate?
50. Other than anticoagulation, What is the best therapy for those with atrial fibrillation?

References

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program (Vol. 8, pp. 17–21).

Ben Abacha, A., & Zweigenbaum, P. (2011). A hybrid approach for the extraction of semantic relations from MEDLINE abstracts. In *Computational linguistics and intelligent text processing, 12th international conference, CICLing 2011, lecture notes in computer science, Tokyo, Japan* (Vol. 6608, pp. 139–150). http://dx.doi.org/10.1007/978-3-642-19400-9.

Ben Abacha, A., & Zweigenbaum, P. (2012). Medical question answering: Translating medical questions into sparql queries. In *ACM SIGHIT international health informatics symposium (IHI 2012), Miami, FL, USA.*

Ben Abacha, A., & Zweigenbaum, P. (2011). Medical entity recognition: A comparison of semantic and statistical methods. In *Actes BioNLP 2011 workshop* (pp. 56–64). Portland, Oregon, USA: Association for Computational Linguistics. <http://www.aclweb.org/anthology/W11-0207>.

Cao, Y.-G., Ely, J., Antieau, L., & Yu, H. (2009). Evaluation of the clinical question answering presentation. In *Proceedings of the workshop on current trends in biomedical natural language processing, BioNLP '09* (pp. 171–178). Stroudsburg, PA, USA: Association for Computational Linguistics.

Cimiano, P., Haase, P., Heizmann, J., Mantel, M., & Studer, R. (2008). Towards portable natural language interfaces to knowledge bases: The case of the ORAKEL system. In *Data knowledge engineering (DKE)* (Vol. 65(2), pp. 325–354).
Cohen, K. B., Verspoor, K., Johnson, H. L., Roeder, C., Ogren, P. V., Baumgartner, W. A. Jr.,, et al (2011). High-precision biological event extraction: Effects of system and of data. *Computational Intelligence, 27*(4), 681–701.
Covell, D. G., Uman, G. C., & Manning, P. R. (1985). Information needs in office practice: are they being met? *Annals Of Internal Medicine, 103*(4), 596–599.
Demner-Fushman, D., & Lin, J. (2005). Knowledge extraction for clinical question answering: Preliminary results. In *Actes AAAI 2005 workshop on question answering in restricted domains, AAAI*.
Demner-Fushman, D., & Lin, J. J., (2006). Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In *ACL*.
Ely, J. W., Osheroff, J. A., Ebell, M. H., Bergus, G. R., Levy, B. T., Chambliss, M. L., et al (1999). Analysis of questions asked by family doctors regarding patient care. *BMJ, 319*(7206), 358–361. <http://www.ncbi.nlm.nih.gov/pubmed/10435959>.
Ely, J. W., Osheroff, J. A., Ebell, M. H., Chambliss, M. L., Vinson, D. C., Stevermer, J. J., et al (2002). Obstacles to answering doctors' questions about patient care with evidence: Qualitative study. *British Medical Journal, 324*, 710.
Ely, J. W., Osheroff, J. A., Gorman, P. N., Ebell, M. H., Chambliss, M. L., Pifer, E. A., et al (2000). A taxonomy of generic clinical questions: Classification study. *British Medical Journal, 321*, 429–432.
Green, B. F., Jr., Wolf, A. K., Chomsky, C., & Laughery, K. (1961). Baseball: an automatic question-answerer. In *Papers presented at the May 9–11, 1961, western joint IRE-AIEE-ACM computer conference, IRE-AIEE-ACM '61 (Western)* (pp. 219–224). New York, NY, USA: ACM.
Humphreys, B. L., & Lindberg, D. A. (1993). The UMLS project: Making the conceptual connection between users and the information they need. *Bulletin of the Medical Library Association, 81*(2), 170–177. <http://view.ncbi.nlm.nih.gov/pubmed/8472002>.
Jacquemart, P., & Zweigenbaum, P. (2003). Towards a medical question-answering system: A feasibility study. In R. Baud, M. Fieschi, P. Le Beux, & P. Ruch (Eds.), *Actes medical informatics Europe. Studies in health technology and informatics* (Vol. 95, pp. 463–468). Amsterdam: IOS Press.
Katz, B. (1999). From sentence processing to information access on the world wide web. In *AAAI spring symposium on natural language processing for the world wide web*.
Katz, B., Felshin, S., Yuret, D., Ibrahim, A., Lin, J. J., Marton, G., et al. (2002). Omnibase: Uniform access to heterogeneous data for question answering. In *NLDB* (pp. 230–234).
Kilicoglu, H., & Bergler, S. (2011). Effective bio-event extraction using trigger words and syntactic dependencies. *Computational Intelligence, 27*(4), 583–609.
Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28–July 1, 2001* (pp. 282–289).
Lin, D. (1998). Dependency-based evaluation of MINIPAR. In *Proceedings of the workshop on the evaluation of parsing systems, Granada*.
Lin, J., & Katz, B. (2003). Question answering from the web using knowledge annotation and knowledge mining techniques. In *Proceedings of the twelfth international conference on information and knowledge management, CIKM '03* (pp. 116–123). New York, NY, USA: ACM.
Lopez, V., & Motta, E. (2004). Aqualog: An ontology-portable question answering system for the semantic web. In *Proceedings of the international conference on natural language for information systems (NLDB)* (pp. 89–102).
Lopez, V., Unger, C., Cimiano, P., & Motta, E. (2013). Evaluating question answering over linked data. *Journal of Web Semantics, 21*, 3–13.
Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography, 3*, 235–244.
Moldovan, D. I., Harabagiu, S. M., Pasca, M., Mihalcea, R., Goodrum, R., Girju, R., et al. (1999). Lasso: A tool for surfing the answer net. In *Proceedings of the eighth text retrieval conference (TREC-8)*.
Mollá, D. (2010). A corpus for evidence based medicine summarisation. In *Proceedings of the ALTA 2010, Melbourne* (pp. 76–80).
Mollá, D., & Santiago-Martínez, M. E. (2011). Development of a corpus for evidence medicine summarisation. In *Australasian language technology workshop (ALTA 2011), Australia*.
Mollá, D., Schwitter, R., Hess, M., & Fournier, R. (2000). Extrans, an answer extraction system. *Traitement Automatique de Langues, 41*(2), 495–519.
Mollá, D., & Vicedo, J. L. (2007). Question answering in restricted domains: An overview. *Computational Linguistics, 33*(1), 41–61.
Morante, R., Krallinger, M., Valencia, A., & Daelemans, W. (2012). Machine reading of biomedical texts about alzheimer's disease. In *CLEF (online working notes/labs/workshop)*.
Niu, Y., Hirst, G., McArthur, G., & Rodriguez-Gianolli, P. (2003). Answering clinical questions with role identification. In *Proceedings of the ACL 2003 workshop on natural language processing in biomedicine, BioMed '03, association for computational linguistics, Stroudsburg, PA, USA* (Vol. 13, pp. 73–80).
Ounis, C. -C. V., & Lioma, I. (2007). Research directions in terrier. In Ricardo Baeza-Yates, et al. (Eds.), *Novatica/UPGRADE special issue on web information access*, Invited Paper.
Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., & Lioma, C. (2006). Terrier: A high performance and scalable information retrieval platform. In *Proceedings of ACM SIGIR'06 workshop on open source information retrieval (OSIR 2006)*.
Popescu, A., Etzioni, O., & Kautz, H. (2003). Towards a theory of natural language interfaces to databases. In *Proceedings of the international conference on intelligent user interfaces (IUI'03)* (pp. 149–157).
Rinaldi, F., Dowdall, J., & Schneider, G. (2004). Answering questions in the genomics domain. In *Proceedings of the ACL04 workshop on question answering in restricted domains*.
Rosario, B., & Hearst, M. A. (2004). Classifying semantic relations in bioscience text. In *Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL 2004), Barcelona*.
Sackett, D. L., Straus, S. E., Richardson, W. S., Rosenberg, W., & Haynes, R. B. (2000). *Evidence-based medicine: how to practice and teach EBM*. Churchill Livingstone, Edinburgh.
Takeshita, H., Davis, D., & 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* (pp. 1409–1413).
Terol, R. M., Martínez-Barco, P., & Palomar, M. (2007). A knowledge based method for the medical question answering problem. *Computers in Biology and Medicine, 37*(10), 1511–1521.
Tsatsaronis, G., Schroeder, M., Paliouras, G., Almirantis, Y., Androutsopoulos, I., Gaussier, E., et al. (2012). Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *AAAI fall symposium: Information retrieval and knowledge discovery in biomedical text*.
Uzuner, O. (Ed.), (2010). Working papers of i2b2 medication extraction challenge workshop, i2b2.
Uzuner, O., South, B. R., Shen, S., & Duvall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *JAMIA, 18*(5), 552–556 (epub 2011 June, 16).
Woods, W. A. (1973). Progress in natural language understanding: an application to lunar geology. In *Proceedings of the June 4–8, 1973, national computer conference and exposition, AFIPS '73* (pp. 441–450). New York, NY, USA: ACM.
Yu, H., Sable, C., & Zhu, H. R. (2005). Classifying medical questions based on an evidence taxonomy. In *Proceedings of the AAAI'05 workshop on question answering in restricted domains*. <http://www.uwm.edu/hongyu/publications.html>.