COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 99 (2010) 1–24

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Biomedical question answering: A survey

Sofia J. Athenikos[a,*], *Hyoil Han*[b]

[a] *College of Information Science and Technology, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA*
[b] *Department of Computer Science, LeMoyne-Owen College, 807 Walker Avenue, Memphis, TN 38126, USA*

ARTICLE INFO

Article history:
Received 28 September 2009
Received in revised form
8 October 2009
Accepted 13 October 2009

Keywords:
Biomedical question answering
Answer/reason extraction
Semantic information extraction

ABSTRACT

Objectives: In this survey, we reviewed the current state of the art in biomedical QA (Question Answering), within a broader framework of semantic knowledge-based QA approaches, and projected directions for the future research development in this critical area of intersection between Artificial Intelligence, Information Retrieval, and Biomedical Informatics.
Materials and methods: We devised a conceptual framework within which to categorize current QA approaches. In particular, we used "semantic knowledge-based QA" as a category under which to subsume QA techniques and approaches, both corpus-based and knowledge base (KB)-based, that utilize semantic knowledge-informed techniques in the QA process, and we further classified those approaches into three subcategories: (1) semantics-based, (2) inference-based, and (3) logic-based. Based on the framework, we first conducted a survey of open-domain or non-biomedical-domain QA approaches that belong to each of the three subcategories. We then conducted an in-depth review of biomedical QA, by first noting the characteristics of, and resources available for, biomedical QA and then reviewing medical QA approaches and biological QA approaches, in turn. The research articles reviewed in this paper were found and selected through online searches.
Results: Our review suggested the following tasks ahead for the future research development in this area: (1) Construction of domain-specific typology and taxonomy of questions (biological QA), (2) Development of more sophisticated techniques for natural language (NL) question analysis and classification, (3) Development of effective methods for answer generation from potentially conflicting evidences, (4) More extensive and integrated utilization of semantic knowledge throughout the QA process, and (5) Incorporation of logic and reasoning mechanisms for answer inference.
Conclusion: Corresponding to the growth of biomedical information, there is a growing need for QA systems that can help users better utilize the ever-accumulating information. Continued research toward development of more sophisticated techniques for processing NL text, for utilizing semantic knowledge, and for incorporating logic and reasoning mechanisms, will lead to more useful QA systems.

1. Introduction

Question Answering (QA), unlike traditional Information Retrieval (IR), aims to provide inquirers with direct, precise answers to their questions, by employing Information Extraction (IE) and Natural Language Processing (NLP) techniques, instead of providing a large number of documents that are *potentially* relevant for the questions posed by the inquirers. As such, QA is regarded as involving the most

* *Corresponding author*. Tel.: +1 215 299 1299; fax: +1 215 895 2494.
E-mail address: sofia.j.athenikos@acm.org (S.J. Athenikos).
0169-2607/$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2009.10.003

critical capability required of the next generation of search engines.

While the early research on automated QA in the field of AI dates far back to the 1960s, a more recent surge of research activities involving QA within the IR/IE community has been mainly prompted by the introduction of the QA Track in TREC [1] evaluations in 1999 [2]. (See Hirschman and Gaizauskas [3] for an overview of QA as a research topic.) Since then techniques have been developed for generating answers for the three types of questions supported by TREC evaluations, namely, factoid questions, list questions, and definitional questions.

Most research development in the area of QA, as fostered by TREC and other similar evaluation venues such as CLEF [4] and NTCIR [5], has so far been focused on open-domain text-based QA. Recently, however, the field has witnessed a growing interest among researchers in restricted-domain QA. (See Mollá and Vicedo [6] for an overview of restricted-domain QA.) Also, while the earlier TREC QA systems mostly relied on a surface-level lexico-syntactic analysis in generating answers, there has been a growing research interest in the development of QA techniques that incorporate semantic knowledge.

Due to the continuous, exponential growth of information produced in the biomedical domain, and due to the crucial impact of such information upon research and upon real-world applications, there is a particularly great and growing demand for QA systems that can effectively and efficiently aid biomedical researchers and health care professionals in their information search. In order to provide information seekers with accurate answers, such systems need to go beyond surface-level lexico-syntactic analysis to semantic analysis and processing of textual, terminological, and ontological resources. Moreover, QA systems equipped with reasoning capabilities can derive more adequate answers by using inference mechanisms.

And yet, research on QA specifically directed to the information needs in the biomedical domain remains a little-explored territory. While several approaches have exploited semantic knowledge in the QA process, few approaches have explored the utility of logic representations and of inference mechanisms.

In this paper, we survey the current state of the art in biomedical QA and suggest future directions for the research development in the area. By doing so, we hope to contribute to the ongoing research in this emerging field of the intersection between AI, IR, and Biomedical Informatics.

The remainder of this paper is organized as follows: In Section 2, we briefly describe the generic architecture of QA systems for the sake of readers who are less familiar with the field. In Section 3, we present a broader classification framework involving semantic knowledge-based QA, and briefly discuss some of the open-domain or non-biomedical restricted-domain QA approaches that belong within the framework. In Section 4, we review the current research on biomedical QA, based on the framework introduced in the previous section. In Section 5, we suggest directions for the future research development in the area, based on our review. Section 6 concludes the paper.

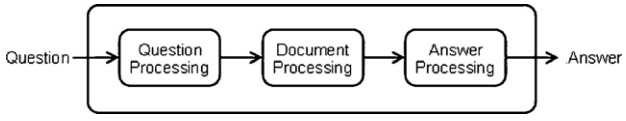

Fig. 1 – Three main processing phases of QA.

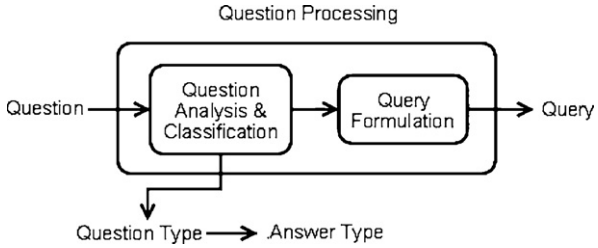

Fig. 2 – Question processing phase of QA.

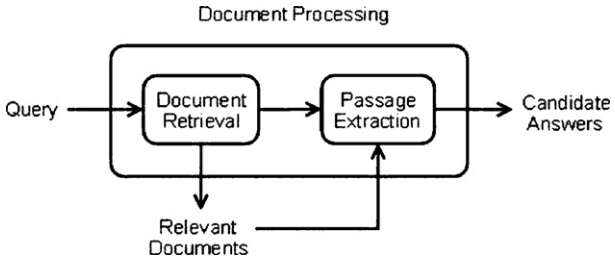

Fig. 3 – Document processing phase of QA.

2. Generic architecture of QA systems

In general (cf. Hirschman and Gaizauskas [3]), the QA processing in a QA system consists of three main processing phases, namely, question processing, document processing, and answer processing phases, as shown in Fig. 1.

The input to a QA system is a question (usually) given in natural language (NL) expressions. The question processing phase (Fig. 2) consists of question analysis & classification and query formulation. Based on the linguistic processing of the question, question analysis & classification determines the type of the question and the corresponding type of expected answer. There may be more subprocesses involved at this stage, such as named entity recognition (NER). Query formulation consists of generating a query to be input to a document retrieval engine, by transforming the question into some canonical form.

In the document processing phase (Fig. 3), the query generated in the question processing phase is fed into a search engine in order to retrieve relevant documents. The retrieved document set may be narrowed down to a smaller set of most relevant documents. Out of the retrieved and selected documents, candidate answer passages are extracted, which constitute the input for the answer processing phase. As in the question processing phase, the document processing phase will generally involve linguistic processing subprocesses.

Finally, in the answer processing phase (Fig. 4), the candidate answers generated in the document processing phase are matched against the expected answer type generated in the

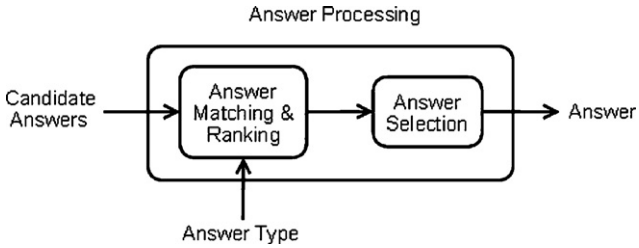

Fig. 4 – Answer processing phase of QA.

question processing phase, and are ranked according to the matching scores. Again, more sophisticated linguistic processing may be involved. The output from the QA system consists of the top-ranked answer(s) selected as the final answer(s).

3. Semantic knowledge-based QA

In this section we lay out a framework within which we will review and categorize current biomedical QA approaches in the next section.

A recent trend among QA researchers has been to incorporate semantic knowledge throughout the QA process in order to derive more accurate answers. As shown in Fig. 5, the knowledge of semantic information extracted or obtained from the textual sources (including the question text as well as the documents in the collection) and terminological/ontological resources may be fed into the QA system at each of the three main processing phases so as to improve the QA performance.

While a recent report by Lopez et al. [7] has surveyed the state of the art in semantic QA, our perspective in this paper is different from that of Lopez et al. in that we do not focus on QA over semantic metadata in structured knowledge bases (KBs) and ontologies. Rather, we use the phrase "semantic knowledge-based QA" as an overarching rubric for the category of QA techniques and approaches, both corpus-based and KB-based, which utilize semantic knowledge-informed IR/IE/NLP techniques in the QA process. We classify these approaches into three subcategories: (1) semantics-based, (2) inference-based, and (3) logic-based. In the following, we review some of the open-domain or non-biomedical-domain QA approaches that belong to each of these three categories.

3.1. *Semantics-based QA*

Most semantics-based open-domain QA approaches take advantage of the lexico-semantic information encoded in WordNet [8,9], a prominent terminological resource for the general English domain.

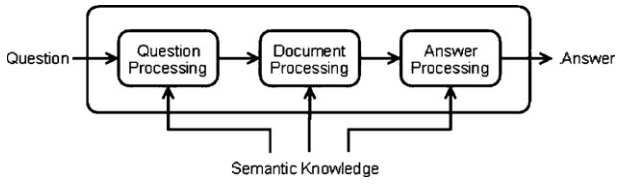

Fig. 5 – Architecture of semantic knowledge-based QA systems.

Table 1 – Semantics-based QA approaches.

Approach	Semantic feature used
Vicedo and Ferrández [10]	Semantic representation of answer context
Alfonseca et al. [11]	Semantic distance between question and answer
Hovy et al. [12]	Semantic patterns of question and answer
Fleischman et al. [13]	Semantic relations between lexical terms
Ferrés et al. [14]	Semantic environment of question; semantic constraints on answer
Punyakanok et al. [16]	Semantic distance measured by the edit distance between Q/A dependency trees
Sun et al. [17]	Semantic relation analysis based on frame representations of question and answer

Vicedo and Ferrández [10] describe a QA system that uses the semantic representation of the context of the expected answer, generated using WordNet-based tools, in ranking and selecting answers. Alfonseca et al. [11] report on a QA system whose central part is the semantic distance measuring module that makes use of all semantic relationships available in WordNet in order to estimate the semantic distance between the question and a candidate answer. Hovy et al. [12] describe semantics-based answer pinpointing that relies on a syntactico-semantic analysis of the question and candidate answers based on a hierarchical typology of semantic patterns of question and answer. Fleischman et al. [13] present an offline strategy for QA, which is based on the construction of a repository of the semantic relations between lexical terms extracted from a text collection by using POS (part-of-speech) patterns and a machine-learned classifier. Ferrés et al. [14] describe a QA approach, which uses the representation of the semantic environment of the question, based on the information obtained from EuroWordNet [15], and the specification of a set of semantic constraints to be satisfied by an answer. Answer extraction relies on iterative relaxation of the semantic constraints. Punyakanok et al. [16] present a QA technique that selects answers based on dependency tree matching. They represent the question and a candidate answer passage as dependency trees augmented with semantic information, and compute a generalized edit distance between the two representations by using an approximate tree matching algorithm. The candidate sentence that minimizes the distance is considered most semantically similar to the question and is chosen as the answer. Sun et al. [17] use syntactic dependency analysis for the sake of query expansion, and use semantic relation analysis, based on the frame-based semantic representation generated by using a shallow semantic parser, for semantic answer extraction. Answer passage selection is done via computation of frame similarity scores, based on the information in WordNet and eXtended WordNet [18]. Table 1 summarizes semantics-based QA approaches.

3.2. *Inference-based QA*

We review QA approaches that rely on some form of inference or those that involve extracting semantic relations contribut-

Table 2 – Inference-based QA approaches.

Approach	Inference method/mechanism used
Lin and Pantel [23]	Method for discovery of inferences rules
Girju [24]	Method for detection of causal relations
Beale et al. [25]	Inference on events based on ontological scripts
Harabagiu et al. [26]	Inference and reference resolution mechanisms
Narayanan and Harabagiu [27]	Probabilistic inference based on frame structure and argument-predicate structure
Narayanan and colleagues [28,31]	Probabilistic inference on events/actions based on parameterized model of events/processes
Harabagiu and Bejan [34]	Temporal inference
Shen and Lapata [35]	Method for assessment of semantic role labeling
Katz et al. [37]	Inference on inter-event relationships

ing to inference. Some use resources such as FrameNet [19,20] and PropBank [21,22] in obtaining frame or predicate-argument structures.

Lin and Pantel [23] present an unsupervised algorithm for discovering inference rules from text, which relies on finding similar paths in the dependency tree of a parsed corpus. Girju [24] proposes a method for automatic detection and extraction of causal relations from text. Beale et al. [25] describe a QA system where an ontological semantic analyzer generates frame-based semantic representations from text, and ontological scripts enable the system to infer events and states. Harabagiu et al. [26] describe a QA system that uses multiple bridging inference mechanisms and reference resolution algorithms in extracting answers to complex and contextual questions. Narayanan and Harabagiu [27] present a QA approach that uses probabilistic inference, based on frame and predicate-argument structures, a topic model, and a set of conceptual schemas. In [28], Narayanan and Harabagiu incorporate a probabilistic relational model of actions and events [29,30] in the QA system to allow it to perform inference to answer questions involving causal and temporal aspects of complex events. Within the same framework, Sinha and Narayanan [31] report on an answer selection approach that focuses on the ability of a parameterized model of events and processes, based on an action/event ontology [32,33], to improve the ranking of candidate answers. Harabagiu and Bejan [34] present a QA methodology for handling temporal inference based on the relations of the expected answer to the temporal expressions in the question or candidate answers. Shen and Lapata [35] assess the contribution of automatic semantic role labeling [36] to factoid QA, by treating semantic role assignment as a global optimization problem in a weighted bipartite graph, and answer extraction as an instance of the graph matching problem. Katz et al. [37] describe a language-driven QA approach which uses semantic decomposition of questions and resource content into lower-level assertions that provide a basis for inference on inter-event relationships. Table 2 summarizes inference-based QA approaches.

3.3. *Logic-based* QA

We review QA approaches that employ explicit logic forms (LFs) and theorem proving techniques. Most approaches adopt First Order Logic (FOL) based formalisms.

Harabagiu et al. [38] discuss a QA system that uses a theorem prover, based on FOL Logic Form Transformation (LFT) of question/answers. Moldovan and Rus [39] discuss conversion of WordNet glosses into axioms via LFT in the context of eXtended WordNet (XWN). Moldovan et al. [40] report on the implementation of the COGEX logic prover, which takes in Q/A LFs and WXN/NLP axioms and selects answers based on the proof score. In [41] Moldovan et al. discuss enhancing the capabilities of COGEX by incorporating semantic and contextual information during LF generation. Mollá et al. [42] present ExtrAns, a QA system applied to the Unix domain, which uses Minimal Logical Forms (MLFs) that are converted to Prolog facts/queries. In [43] Mollá compares MLFs with grammatical relations as the overlap-based similarity scoring measures for answer ranking. Benamara [44] presents a QA system applied to the tourism domain, called WEBCOOP, which operates on a deductive KB that contains facts, rules, and integrity constraints encoded in Prolog, and a set of texts indexed via FOL formulae. Waldinger et al. [45] discuss QUARK, a QA system which uses the FOL theorem prover SNARK [46] that generates answers by linking domain axioms with factual knowledge

Table 3 – Logic-based QA approaches.

Approach	Logic formalism/reasoning mechanism used
Harabagiu et al. [38]	FOL theorem prover using Q/A LFs
Moldovan et al. [39–41]	COGEX theorem prover using Q/A LFs and XWN/NLP (and semantic/ontological) axioms
Mollá et al. [42,43]	ExtrAns QA system using MLFs converted into Prolog facts and queries
Benamara [44]	WEBCOOP QA system using Prolog-encoded KB and FOL-indexed text set
Waldinger et al. [45]	QUARK QA system using FOL theorem prover SNARK operating on domain axioms and KBs
Curtis et al. [47]	QA system using Cyc KB encoded in CycL
Clark et al. [49]	FOL representation of contextual knowledge
Tari and Baral [50]	AnsProlog for representation and reasoning
Baral et al. [52]	AnsProlog plus Constraint Logic Programming
Bobrow et al. [53]	Textual Inference Logic

from multiple sources. Curtis et al. [47] describe a system which uses the Cyc KB [48] to support deductive QA. Clark et al. [49] present a layered approach to the FOL representation of contextual knowledge, coupled with reasoning mechanisms, to enable contextual inference and default reasoning for QA. Tari and Baral [50] propose a QA system that uses AnsProlog for representation and reasoning, Link Grammar [51] for fact extraction, and WordNet for disambiguation. Baral et al. [52] present a QA system that combines AnsProlog and Constraint Logic Programming, to enable textual inference on events, actions, and temporal relations. Bobrow et al. [53] describe Textual Inference Logic (TIL) as a representation language for QA. Table 3 summarizes logic-based QA approaches.

4. State of the art in biomedical QA

Based on our discussions so far, in this section we embark on the main task in this paper: a detailed review of the current state of research on biomedical QA.

4.1. *Biomedical QA: background*

4.1.1. *Open-domain QA vs. restricted-domain QA*

Most semantic knowledge-based QA systems, techniques, and approaches that we reviewed in the previous section deal with open-domain QA, while others concern restricted-domain QA in domains other than the biomedical domain. As Mollá and Vicedo [6] point out in their recent overview of QA in restricted domains, there are important factors that distinguish restricted-domain QA from open-domain QA. Those factors include: (1) size of the data, (2) domain context, and (3) resources. The size of the data available for general open-domain QA tends to be quite large, which justifies the use of redundancy-based answer extraction techniques. In the case of restricted-domain QA, however, the size of the corpus varies from domain to domain, and redundancy-based techniques would not be practical for a domain with a small corpus size. In restricted-domain QA, the domain of application provides a context for the QA process. This involves domain-specific (meanings of) terminologies and domain-specific types of questions, which also differ between domain experts and non-expert users. Finally, a major difference between open-domain QA and restricted-domain QA exists in the availability of domain-specific resources and the incorporation of domain-specific information in the QA process in the latter.

4.1.2. *Characteristics of biomedical QA*

Several biomedical QA researchers have discussed the differences between general open-domain QA and domain-specific QA, and the peculiar characteristics and challenges of the biomedical domain as an application domain for QA. Yu and Sable [54], for example, note that restricted-domain QA can exploit domain-specific knowledge resources for deeper question processing and that it may take advantage of a domain-specific typology of questions in order to develop answer extraction strategies appropriate for each question type. Similarly, Rinaldi et al. [55] observe that restricted-domain QA can exploit deeper text analysis/processing, taking advantage of domain-specific formatting and style conventions as well as domain-dependent terminology. They also point out the fact that, in contrast to open-domain QA for which generic NER is a major concern, in restricted-domain QA domain-dependent terminology plays a major role and presents a major challenge. In his brief overview of QA in the biomedical domain, Zweigenbaum [56] also notes that biomedical QA, in contrast to open-domain QA, is challenged with a more acute need to cater for specialized terminological variation, and that the gap in technicality between non-expert user questions and target documents may be larger than in other restricted domains. Along a similar line, Niu et al. [57] discuss the differences between general QA and medical QA in particular. Most recently, Zweigenbaum [58] notes the role of (domain-specific) knowledge and reasoning for restricted-domain QA, such as (bio-)medical QA, versus open-domain QA. As he notes, knowledge and reasoning may be both more necessary and more manageable for the former compared to the latter, due to the relative specificity or difficulty of questions and due to the relatively limited scope of questions.

Some of the characteristic features of QA in the biomedical domain, particularly in terms of the aforementioned three major factors that distinguish restricted-domain QA, namely, dataset size, domain-dependent context, and domain-specific resources, may be summarized as follows:

1. Large-sized corpora (described below).
2. Highly complex domain-specific terminology.
3. Domain-specific lexical, terminological, and ontological resources (described below).
4. Tools and methods for exploiting the semantic information embedded in the above resources (described below).
5. Domain-specific format and typology of questions (medical QA, described below).

4.1.3. *Resources for biomedical QA*

While the biomedical domain poses a particular challenge for QA, with a huge amount of literature and highly complex domain-specific terminology, it also provides various resources that can be exploited for QA, as Zweigenbaum [56] also notes in his overview. Here we review some of the well-known and oft-used knowledge resources in the biomedical domain.

The primary corpora for text-based QA in the biomedical domain are accessible through PubMed and PubMed Central. PubMed, a service provided by the National Library of Medicine (NLM), under the U.S. National Institutes of Health (NIH), contains over 17 million citations from MEDLINE, a bibliographic database (DB) of biomedical literature, and other biomedical and life science journals dating back to the 1950s. It is accessible through the National Center for Biotechnology Information (NCBI). PubMed Central (PMC) is a digital archive of full-text biomedical and life science articles, maintained and updated by the NIH. Also available in the biological domain is a semantically annotated corpus of MEDLINE abstracts involving protein reactions, developed by the University of Tokyo within the GENIA project [59].

Due to the acute need of systematically organizing, updating, and cross-referencing its highly complex domain-specific terminology, and of making it machine-readable and machine-understandable, the biomedical domain has

developed various lexical, terminological, and ontological resources.

Medical Subject Headings (MeSH), the NLM's controlled vocabulary thesaurus, consists of medical subject descriptors in a hierarchical taxonomic structure which allows search at various levels of specificity.

Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT), originally created by the College of American Pathologists (CAP), is the most comprehensive clinical terminology available, which was formed by the merger of SNOMED RT (Reference Terminology) and the U.K. National Health Service (NHS) Clinical Terms. SNOMED CT is a current U.S. standard for electronic health information exchange, and is accessible through NLM and the National Cancer Institute (NCI).

Arguably, the semantic knowledge resources that are most frequently exploited in biomedical QA are those of the Unified Medical Language System (UMLS), provided by NLM. There are three UMLS knowledge resources: the Metathesaurus, the Semantic Network, and the SPECIALIST Lexicon.

The UMLS Metathesaurus is a comprehensive, multilingual biomedical vocabulary database that contains information on biomedical and health-related concepts, their various names, and the semantic relationships between them. The Metathesaurus is constructed from various source vocabularies, and retains the concept meanings, names, and relationships from the original sources, even in the presence of conflicts among different source vocabularies. As such, it is not meant to represent a coherent biomedical ontology with a single, consistent view.

The UMLS Semantic Network contains information on the set of semantic types and the set of semantic relationships that may hold between these semantic types. The Semantic Network provides a consistent categorization of the concepts in the Metathesaurus, by means of the semantic types that can be assigned to those concepts and the semantic relationships that can be defined between them. The latest release of the Semantic Network contains a total of 135 semantic types and 54 semantic relationships, both of which are hierarchically organized.

The UMLS SPECIALIST Lexicon is a general English lexicon whose coverage includes both common English words and biomedical terms. The Lexicon has been developed to provide the lexical, syntactic, and morphological information needed by the SPECIALIST NLP System.

Somewhat similar to NLM's UMLS, NCI's Enterprise Vocabulary Services (EVS) include the Thesaurus, the Metathesaurus, and the Terminology Browser. The NCI Thesaurus contains definitions, synonyms, and other information on terms related to cancer and biomedical research. The NCI Metathesaurus is a large biomedical terminology database which contains the vocabularies from the UMLS Metathesaurus as well as other cancer-related vocabularies. The NCI Terminology Browser provides access to lexical, terminological, and ontological resources, including SNOMED CT and the Gene Ontology (GO).

While various terminological resources in the biomedical domain provide quasi-ontological functions, the domain also has resources that are specifically intended as structured formal ontologies.

The OpenGALEN ontology, a European project, rigorously defines complex medical concepts in terms of primitive concepts and roles (relations) based on a Description Logic (DL) formalism [60]. GO consists of three structured ontologies for genomics, which describe gene products in terms of associated cellular components, biological processes, and molecular functions. The GENIA corpus is based on the GENIA ontology, which is intended as a formal model specifically of cell signaling reactions in human.

Besides the aforementioned resources, there are also oft-used tools and techniques that have been developed to help exploit the semantic information contained in those resources. MetaMap [61] (its implementation as MetaMap Transfer (MMTx)) and SemRep [62] are often used by researchers working on biomedical NLP and Text Mining, including biomedical QA researchers, in order to map terms in free text to UMLS Metathesaurus concepts and UMLS Semantic Network semantic relationships, respectively.

The ClinicalQuestions Collection maintained by the NLM is a growing repository of questions that have been collected from healthcare providers in clinical settings across the US. The repository includes 4049 questions collected by Ely et al. [63,64] and 605 questions collected by D'Alessandro et al. [65]. The questions can be searched by keywords and by specific criteria, such as disease/condition, physician specialty, patient age, patient gender, etc.; they can also be browsed by disease category, question source, and question elements. (Table 4).

4.2. *Medical QA*

4.2.1. *Introduction to medical QA*

In order to discuss QA for the medical domain, we need to first cover some background information.

A dominant paradigm in the medical/clinical field is that of Evidence-Based Medicine (EBM) [66,67], which refers to the use of the best evidence obtained from scientific research in making clinical decisions. Within the EBM framework, physicians are urged to ask questions in order to find the best available evidences.

There have been several investigations [63,68–78], often conducted by medical researchers and practitioners, concerning the usage and effectiveness/efficiency of the online biomedical resources in answering medical/clinical questions. While those studies have validated the usefulness of various resources to a certain extent, they have also revealed serious problems in the medical QA process. For example, Ely et al. [63] have found that physicians spend on average 2 min or less in seeking an answer, while Hersh et al. [72] have found that it takes more than 30 min on average for a health care professional to search for an answer. As a result, many clinical questions go unanswered. Studies investigating the obstacles to finding answers to medical/clinical questions [79,80] have found physicians' doubt about the existence of an answer, excessive time required for search, difficulty of formulating an answerable question, uncertainty about an optimal search strategy, and failure of the selected resource to provide a synthesized answer, among the main factors.

The EBM framework recommends a specific frame for formulating a clinical question in searching for the best available evidence, namely, Problem or Patient/Population,

Table 4 – Resources for biomedical QA.

Resource	Source
PubMed	http://www.ncbi.nlm.nih.gov/sites/entrez
PubMed Central	http://www.pubmedcentral.nih.gov/
GENIA	http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/
MeSH	http://www.nlm.nih.gov/mesh/meshhome.html
SNOMED	http://www.nlm.nih.gov/snomed/
UMLS	http://www.nlm.nih.gov/research/umls/
NCI EVS	http://www.nci.nih.gov/cancerinfo/terminologyresources
OpenGALEN	http://www.opengalen.org/
Gene Ontology	http://www.geneontology.org/
MetaMap/SemRep	http://skr.nlm.nih.gov/
ClinicalQuestions Collection	http://clinques.nlm.nih.gov/About.html

Intervention, Comparison, Outcome (PICO) [81,82]. The PICO framework can be expanded to PICOTT, adding information about the Type of question asked or type of task involved, i.e., therapy, diagnosis, prognosis, etiology, etc., and the Type of study design for that particular task/question [83]. Studies [83–85] have found the usefulness of the PICO/PICOTT frame in facilitating answer discovery.

Within the EBM framework, medical/clinical domain-specific taxonomies of questions have also been developed. Bergus et al. [84] have developed a taxonomy of medical questions according to the PICO elements of the questions and the categories of clinical tasks involved in the questions. Ely et al. [63,79,86] have developed a generic taxonomy of common clinical question types and an "Evidence Taxonomy" of clinical questions, from their studies with primary care doctors. On the top level of the Evidence Taxonomy, questions are classified into Clinical vs. Non-clinical. The Clinical questions are divided into General vs. Specific. The General questions are classified into Evidence vs. No Evidence. The Evidence questions are further classified into Intervention vs. No Intervention categories. Ely et al. have concluded that only the Evidence type questions are potentially answerable. Ely et al.'s [79] Evidence Taxonomy is shown in Fig. 6. Table 5 shows 10 most common types of generic clinical questions as identified by Ely et al. [63,86].

4.2.2. *Medical QA systems and approaches*

In this section, we review current research efforts directed toward QA in the medical (clinical) domain.

Table 5 – Ten most common clinical question types [86].

Question type
What is the drug of choice for condition X?
What is the cause of symptom X?
What test is indicated in situation X?
What is the dose of drug X?
How should I treat condition X (not limited to drug treatment)?
How should I manage condition X (not specifying diagnostic or therapeutic)?
What is the cause of physical finding X?
What is the cause of test finding X?
Can drug X cause (adverse) finding Y?
Could this patient have condition X?

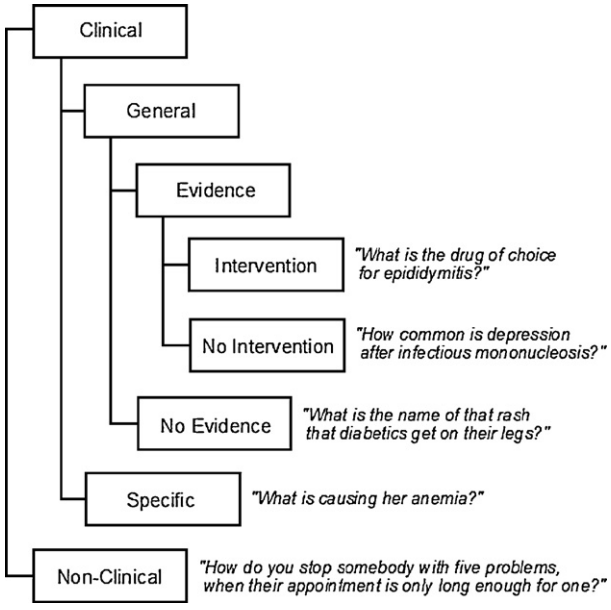

Fig. 6 – Ely et al.'s evidence taxonomy of clinical questions.

4.2.2.1. *Preliminary approaches to medical QA.* Among medical QA researchers, Huang et al. [87], Yu et al. [54,88,89], and Kobayashi and Shyu [90] have investigated question classification as a first step toward developing medical QA systems.

Huang et al. [87] examined the adequacy and suitability of PICO as a representation framework for clinical questions posed in NL, by means of manual classification of primary care clinical questions. The study has reaffirmed the value of the PICO framework overall, but also found that PICO is primarily centered on therapy type questions and less suitable for other types of questions. It has also noted that many UMLS semantic types show strong associations with specific PICO elements, while other semantic types can be mapped to more than one PICO slot.

Yu and Sable [54] developed a question filtering component that automatically determines whether or not a question is answerable, based on the Evidence Taxonomy of Ely et al. They used various supervised machine learning (ML) algorithms, with bag-of-words features and semantic features consisting of UMLS concepts and semantic types. The results have shown that incorporating semantic features in general moderately enhances the performance of question classifi-

cation. The results have identified a probabilistic indexing algorithm to be the best performer, with an accuracy rate of 80.5%. In a follow-up study, Yu et al. [88] focused on the harder task of automatically classifying questions into the specific categories in the Evidence Taxonomy. The results of the evaluation, conducted in a setting similar to that in [54], have shown that Support Vector Machine (SVM) outperforms all other systems in most cases. The results have also revealed that including UMLS concepts and semantic types as additional features can enhance results in most cases. More recently, Yu and Cao [89] explored supervised ML approaches using SVM to automatically classifying ad hoc clinical questions into general topics, and both supervised approaches – Logistic Regression and Conditional Random Fields (CRF) – and unsupervised approaches – Inverse Document Frequency (IDF) model and Domain Filtering – to automatically extracting keywords from ad hoc clinical questions. The results of the evaluation, using the NLM ClinicalQuestions Collection, have again shown that matching of question terms to UMLS concepts and semantic types, using MMTx, resulted in the highest performance increase for question classification using SVM and for the unsupervised domain filtering approach to keyword extraction. The results have also shown that both supervised approaches outperformed unsupervised ones for keyword extraction that, between the two supervised approaches, CRF outperformed logistic regression.

Kobayashi and Shyu [90] examined the performance in classifying clinical questions using alternative representations of questions generated from using different parsing methods and augmented (or not) with the information on UMLS concepts and semantic types. They used questions labeled with Ely et al.'s taxonomic category information as well as other questions. The results have shown that using UMLS semantic types improves classification performance.

All of the above studies concerning question classification have thus found the usefulness of the semantic information obtained from the UMLS resources in performing the question classification task.

Slaughter et al. [91] investigated the semantic patterns of health consumers' questions and physicians' answers, by manually identifying semantic relationships in the question–answer pairs obtained from medical information Web sites. Identification of the semantic relationship instances within the question/answer texts was based on the semantic relations in the UMLS. The results have indicated that identification of causal relationships is fundamental to QA. The study has also found that a direct correspondence between the semantic representations of the questions and those of the answers exists only about 30% of the time. The study suggests that semantic relationships extracted from real-life questions and answers can lead to clues for creating semantics-based QA techniques.

In contrast to the above studies investigating various approaches to medical question classification or identifying common patterns found in questions and answers, as preliminary steps toward building effective medical QA systems, Cao et al. [92] recently evaluated the relative effectiveness of different kinds of answer presentation provided by a medical QA system. The results of their evaluation have suggested that, while sentence-based presentation is effective for some types of questions, generally passage-based presentation is more effective, thanks to the fact that it provides richer context and that it matches relevant answers across sentences.

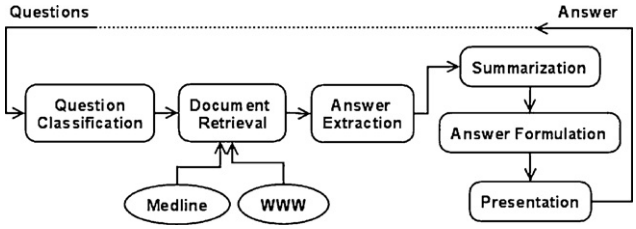

Fig. 7 – Yu et al.'s MedQA system architecture [94].

4.2.2.2. *Non-semantic-knowledge-based medical QA systems and approaches*. Yu et al. [93,94] and Sang et al. [95] have employed medical QA approaches that do not extensively exploit domain-specific semantic knowledge.

Yu et al. [93,94] describe their implemented medical QA system, MedQA, which generates paragraph-level answers from both the MEDLINE collection and the Web. The system in its current implementation deals with definitional questions (i.e., "What is X?"). MedQA incorporates text summarization in the answer processing phase of the QA process (see Fig. 7).

Question classification in the MedQA system is done using the approaches developed by Yu et al. in [54,88], described above. For query formulation and document retrieval, Yu et al. use a shallow syntactic parser and a standard IR engine. For answer extraction, they employ multiple strategies to identify relevant sentences, including the document zone detection method for biomedical articles [96,97], sentence categorization using cue phrases [98], and identification of lexico-syntactic patterns that comprise definitional sentences. For text summarization, MedQA uses hierarchical clustering [99] and centroid-based summarization [100] techniques. Yu et al. recognize the need for using a robust and accurate domain-specific parser. They also note that the current implementation of MedQA does not capture semantic information which plays an important role for both answer extraction and text summarization. They plan to incorporate the results from their previous work [101], which found statistical correlation and dependence relation between the semantic types of a definitional term and the semantic types and lexico-syntactic patterns of definitional sentences.

In [94,102], Yu et al. present cognitive evaluation of MedQA against three state-of-the-art internet search engines – Google, OneLook, and PubMed – for answering definitional questions posed by physicians. The results indicate that Google is an effective search engine for medical definitions, but that MedQA exceeds search engines insofar as it provides direct answers to user questions. Yu et al. suggest the desirability of combining an effective search engine, such as Google, with a domain-specific QA system.

Sang et al. [95] describe ongoing work in developing a Dutch-language medical QA system. They use two different offline strategies for information extraction, one exploiting the regular layout of a Dutch medical encyclopedia, and the other using syntactic patterns based on dependency relations for extracting semantic tuples. Their analysis of the evalua-

Table 6 – Semantic models of medical questions [73].

Semantic model
[Which X]–(R)–[B]
[A]–(R)–[which Y]
Does [A]–(R)–[B]
Why [A]–(R)–[B]
[Which X,Y]–(R)–[B]
[Which X]–(R)–[B,C]
Duration [A]–(precedes)–[B]
Define [A]
Which specific precaution if [A]–(R)–[B]

tion results suggests that lack of coverage is the main source of error and that ontological knowledge of the domain would be very useful in improving the performance of the QA system. In this regard, they cite the lack of available non-English semantic knowledge resources as a challenge.

4.2.2.3. *Semantics-based medical QA systems and approaches.* Jacquemart and colleagues [73,103], Niu et al. [57,104–106], Demner-Fushman et al. [107–111], and Weiming et al. [112] have explored semantics-based medical QA approaches.

Jacquemart and colleagues [73,103] present semantics-based approaches toward the development of a French-language medical QA system.

In their study on the feasibility of the medical QA system, Jacquemart and Zweigenbaum [73] examined the issues as to whether documents relevant to medical questions can be found through Web search and as to whether medical questions can be semantically modeled and categorized in the conceptual framework of their prototype QA system. For the purpose of the study, they used 100 clinical questions on oral surgery, each of which was converted into a canonical form, by simplifying complex questions into more direct questions or by instantiating the context of context-dependent questions.

Concerning the first issue of their focus, Jacquemart and Zweigenbaum have found Google to be the best Web search engine for the task. However, considering that out of 100 questions only 60% obtained relevant results in the top five hits, they note that the high specialization of the medical domain and the clinical orientation of the questions, coupled with the more limited online resources available for the French language, may restrict the quantity of material available for answering the questions.

Concerning the second issue, Jacquemart and Zweigenbaum modeled the forms of the 100 medical questions as syntactico-semantic patterns, in order to identify regularities and to capture their semantic content. These patterns were obtained by generalizing the canonical forms to the generic domain-specific categories. They then constructed semantic models of the questions, in the form of a semantic triple [Concept]–(Relation)–[Concept], by identifying the relevant semantic relations in the UMLS Semantic Network. They obtained 66 distinct syntactico-semantic patterns which were categorized into eight generic semantic models. Three of these semantic models, which accounted for 90 out of 100 questions in the collection, fit the semantic triple representation [A]–(R)–[B] with a modality "which", "does", or "why" (see Table 6).

Jacquemart and Zweigenbaum note that automating the conversion of questions into canonical forms needs more research. Exploiting UMLS semantic relations for this task requires one to find a good fit between NL terms and those relations. They thus consider using UMLS semantic types for concepts as a natural follow-up task.

In a follow-up study, Delbecque, Jacquemart, and Zweigenbaum [103] explore the use of UMLS concepts and semantic types for medical domain-specific NER (named entity recognition), and of semantic relations for answer extraction for specific types of medical questions.

They present an experiment in semantically tagging a French-language medical text collection, obtained from various health-related Web sites, with UMLS concepts, semantic types, and semantic relations, in the context of a QA system. The system first tags the corpus with POS (part-of-speech) information, and uses the POS patterns to locate noun phrases. It then tags the noun phrases with Metathesaurus concepts and associated Semantic Network semantic types. After the noun-processing phase, the system also uses POS patterns to locate clauses, roughly structured in the form of [Subject]–[Verb]–[Complement], in order to detect co-occurrences of semantic types. In the case of a co-occurrence of semantic types that are linked by a semantic relation, the clause is tagged with that semantic relation, which completes the tagging process.

Delbecque et al. evaluated the quality of the tagging process by identifying and examining missing tags and false tagging. They also used ascending hierarchical classification to investigate the match between the meaning of semantic relations and the meaning of tagged clauses. More importantly, they evaluated the usefulness of treating semantic relations as NEs (named entities) in a medical QA context, by using the relation "treats" as a criterion for selecting clauses as answers to specific types of questions, such as "What is the treatment for…?", and by examining the precision of answers thus obtained.

In concluding the study, Delbecque et al. note the importance of taking into account the individual origin of the documents in which search is done, when using semantic relations as NEs. They also recognize that further work needs to be done to improve the tagging process, by using more sophisticated linguistic tools.

Niu et al. [57,104–106] have reported on their work in progress on NL analysis for medical QA within the EPoCare (Evidence at Point of Care) project. The EPoCare system is based on keyword-based query/retrieval. The goal of Niu et al.'s work is to allow the system to accept questions in NL and to better identify answer from NL data sources, the latter of which is their initial focus.

The two main components of the EPoCare system are the XML document DB containing the data and the EPoCare server that uses the DB to provide answers to queries (see Fig. 8). The current data sources include Clinical Evidence (CE) [113] and Evidence-Based on Call (EBOC) [114]. The XML DB is manipulated by a repository manager for XML data, called ToX Engine [115]. The system accepts queries in the PICO format. The front controller takes a clinical question and formulates a keyword query. The retriever sends the query to the XML document DB to retrieve relevant documents using keyword matching.

The query-answer matcher finds answer candidates from the retrieved results. The best answer is selected and returned to the user. The keyword-based retrieval is based on a preconstructed set of answer patterns which consists of XML paths for each of the four PICO categories. The system is supposed to identify XML paths in the data that contain all the keywords in the question, filtering out those paths that do not constitute meaningful contexts for the PICO categories corresponding to those keywords, so as to find those paths that satisfy answer patterns.

In [104], Niu et al. propose a QA approach that locates answers by means of identification, in both question and answer texts, of semantic roles which correspond to the four fields in the PICO frame. The approach is based on first identifying the four roles represented by PICO in the texts of the NL question and the candidate answers and then comparing the roles in the question and the corresponding roles in the candidate answers in order to determine whether or not a candidate is a correct answer. In order to apply the role-based method in QA, Niu et al. consider the problems of detecting PICO roles in text, determining the textual boundary of each role, and identifying the relationships between different roles, with a focus on therapy-related questions. They note that Outcome is the most difficult non-NE role to detect.

In a follow-up study, Niu and Hirst [57] focus on identifying semantic classes in the medical text. For identifying the semantic classes Disease and Medication, they used some training text to find the mapping between UMLS semantic types and the two semantic classes, using MetaMap. For the more complicated task of identifying Outcome, they collected cue words from CE, obtained POS and phrase information from the text, and then identified the boundary of clinical outcomes for each POS category of cue words. In analyzing the relations between semantic classes, they have found that these relations can also be identified by a set of cue words. They note that in a specific domain such as medicine, some default relations often hold between semantic classes, e.g., the cause-effect relation holding between the semantic classes Medication, Disease, and Outcome. Niu and Hirst suggest that semantic classes and their relations have important roles for medical QA. In addition, they consider the task of identifying the polarity of outcomes from medical text.

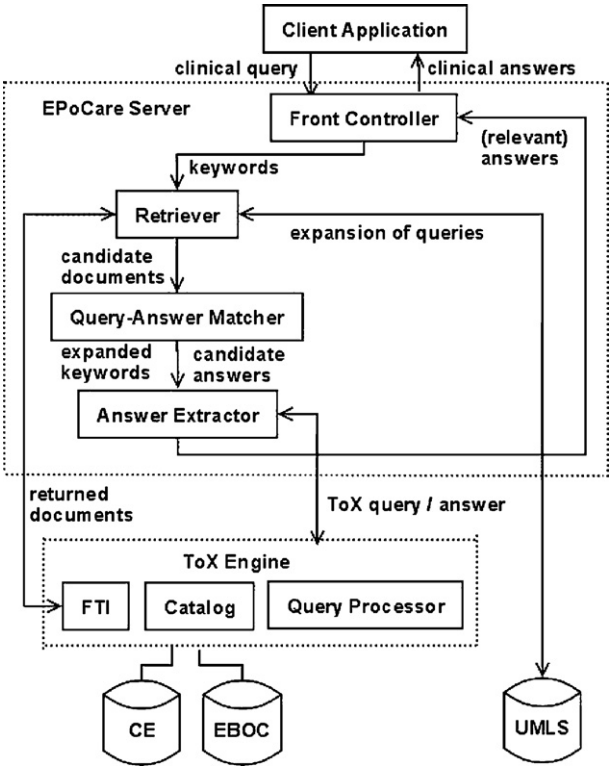

Fig. 8 – Niu et al.'s EPoCare system architecture [104].

Continuing in [105], Niu et al. focus on the problem of automatically detecting and classifying clinical outcomes in the medical text. They applied NLP and ML techniques to detect four possible classes of outcome polarity: no outcome and positive, negative, or neutral outcome. With 1509 sentences collected from CE as the data set, Niu et al. used SVM (Support Vector Machine) to perform the classification task. The results revealed that the highest accuracy (79.42%) was achieved by combining linguistic features, encoding context information, with domain knowledge, using information on the UMLS semantic types.

In [106], Niu et al. use a multi-document summarization approach to finding answers to clinical questions about effects of using a medication for disease treatment. They collected 197 MEDLINE abstracts that were cited in CE, and annotated each sentence with information on the presence of clinical outcomes and their polarity. The results of evaluation of outcome classification using SVM again showed that combining context information and domain knowledge leads to the best performance. For the task of identifying sentences that are important in answering clinical questions, Niu et al. used additional features that had been shown to be effective in text summarization, such as sentence position, sentence length, presence of numbers in a sentence, and maximal marginal relevance (MMR) [116]. While the results of evaluation using ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [117] showed little difference in performance between different feature combinations, those from sentence-level evaluation using SVM showed that detecting the presence and polarity of outcomes is helpful in answer extraction. The additional benefit from determining the polarity, not just the presence, of clinical outcomes, however, was shown to be small, pointing to the need of a more accurate polarity detection system.

Demner-Fushman et al. [107–111] have pursued a similar line of research as Niu et al., with their view of PICO frame as the core organizing knowledge structure for a clinical medical QA system, and of clinical QA as a matter of semantic unification between the PICO frame of the query and that of answers.

Demner-Fushman and Lin [107] describe semantic knowledge extractors that they developed as a component of a clinical QA system to identify PICO frame elements from MEDLINE abstracts and to classify their evidence grade level. As the basis for determining the quality of evidence, they use the Strength of Recommendations Taxonomy (SORT) as developed by Ebell et al. [118], which grades evidences as A-, B-, and C-level according to their objective validity and strength.

Demner-Fushman and Lin extensively use MetaMap [61] and SemRep [62] in order to identify UMLS concepts and

semantic relations from text. In addition, they use more coarse-grained semantic groups [119] in order to capture higher-level generalizations. They also take advantage of discourse markers present in some abstracts.

Demner-Fushman and Lin consider the task of identifying clinical outcomes from primary literature sources, as vs. from secondary sources used by Niu et al., to be a harder task. Their Population, Problem, and Intervention/Comparison extraction modules use manually constructed pattern-matching rules; the Outcome extractor module, in contrast, employs supervised ML techniques to perform a classification task at the sentence level using an ensemble of classifiers including a semantic classifier. Demner-Fushman and Lin note that identification of entities at the semantic class level simplifies extraction of PICO elements.

In [108], Demner-Fushman et al. present a scheme for annotating clinically relevant elements in MEDLINE citations, along the categories of Background, Population, Intervention, Statistics, Outcome, Supposition, and Other, and discuss the evaluation of the supervised ML-based automatic Outcome extractor developed in [107] in detail. They assessed automatic outcome identification in terms of the accuracy and positive predictive value and in terms of the effectiveness of the outcome-based ranking of MEDLINE search results obtained through PubMed Clinical Queries. The results showed the accuracy of automatic outcome identification to be 88–93%, with the positive predictive value of outcome sentences ranging from 30% to 37%. The outcome-based document ranking was shown to improve precision in retrieval by up to 389%, suggesting the potential validity of the EBM/PICO model approach to clinical QA.

In [111], Lin and Demner-Fushman introduce a generic framework for semantic knowledge-based conceptual retrieval and discuss its instantiation in the clinical domain. They note that three categories of knowledge necessary for conceptual retrieval are readily accessible in the clinical domain, two of which are provided by the EBM paradigm: (1) knowledge about the problem structure (PICO frame), (2) knowledge about user tasks (four types of clinical tasks and strength of evidence considerations), and (3) knowledge about the domain (UMLS resources).

According to the proposed framework, a clinical QA system requires three components: (1) knowledge extractors for extracting PICO elements from free-text abstracts, (2) semantic matcher for scoring and ranking citations, and (3) answer generator that produces responses for physician users. Lin and Demner-Fushman use the first component (knowledge extractors) as developed in [107], and here focus on the citation scoring algorithm for the second component (semantic matcher). The algorithm computes the relevance of a MEDLINE citation as a weighted linear combination of a few factors: matching PICO frames, the strength of evidence in the citation, and associated MeSH terms that indicate appropriateness for clinical tasks. In the evaluation, the citation scoring algorithm was shown to dramatically outperform a state-of-the-art baseline.

In [110], Demner-Fushman and Lin discuss a clinical QA system based on the knowledge extractors in [107] and the citation scoring algorithm in [111], which employ a combination of knowledge-based and statistical techniques. Instead of using NL questions, they use a structured PICO query frame as the input to the QA system, obviating the need for linguistic analysis of NL questions. In accordance with their conception of a semantic knowledge-intensive clinical QA system based on conceptual retrieval, the system architecture consists of a query formulator, knowledge extractors, a semantic matcher, and an answer generator (see Fig. 9).

The query formulator converts a clinical question (in a PICO query frame) into a PubMed search query. PubMed returns a list of MEDLINE abstracts, which is then analyzed by knowledge extractors. The input to the semantic matcher is the query frame and annotated MEDLINE abstracts. The answer generator takes a reranked list of citations, and extracts textual answers to physicians' questions, presenting the title of the abstract and the top three outcome sentences.

Demner-Fushman and Lin used 50 clinical questions collected from two online sources, Journal of Family Practice [120] and Parkhurst Exchange [121], and performed three different evaluation tasks—one focusing on the accuracy of knowledge extractors, one evaluating the document reranking task, and one manual evaluation by two physicians. The results have shown that the system significantly outperforms the PubMed baseline.

In [109], Demner-Fushman and Lin explore a hybrid approach to clinical QA, which combines techniques from IR and text summarization. They tackle a frequently occurring class of questions in the form of "What is the best drug treatment for X?" (cf. Table 5).

Given an initial set of MEDLINE abstracts retrieved through PubMed, the system first identifies the drugs under study, using the Intervention extractor in [107]. The system groups the retrieved MEDLINE abstracts into semantic clusters based on the main interventions identified in the abstract text, and uses a variant of hierarchical agglomerative clustering algorithm [122] which utilizes UMLS semantic relations in order to compute similarities between interventions. For each MEDLINE abstract, the system generates a short extractive summary consisting of the main intervention, the title of the

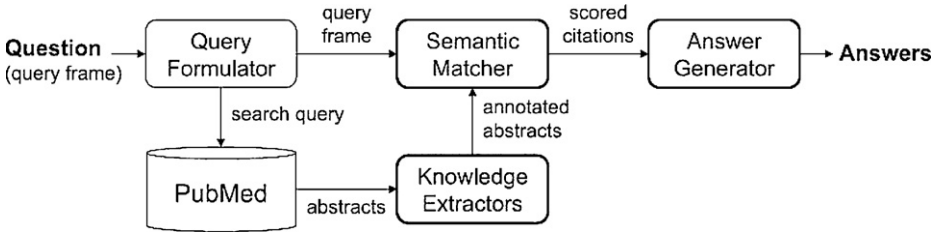

Fig. 9 – Demner-Fushman et al.'s clinical QA system architecture (taken from [110]).

abstract, and the top-scoring outcome sentence identified by the Outcome extractor in [107,108].

Demner-Fushman and Lin conducted two evaluations so as to assess the quality of the system output—a manual evaluation focused on the cluster labels (i.e., drug categories), and an automatic evaluation of the retrieved abstracts using ROUGE. They randomly selected 30 diseases from the CE in order to generate the question sets. They also collected MEDLINE citations associated with each disease, and used them as the reference summaries for the ROUGE-based automatic evaluation. The results of both evaluations have shown that the system outperforms the PubMed baseline, once again demonstrating the value of semantic resources in the QA process.

Weiming et al. [112] propose a clinical QA approach, which incorporates semantic clustering, based on the semantic representations of questions and documents using UMLS concepts, semantic types, and semantic relations.

In the question analysis phase, Weiming et al.'s system parses a question using MetaMap Transfer (MMTx) and SemRep to identify medical concepts and relations. The system uses noun keywords and concept mapping rules for interpreting the semantic relationships in the question and documents. The medical concepts generated in the question analysis phase, together with their synonyms, acronyms, and abbreviations, are used to retrieve relevant documents and to select candidate sentences. In the answer extraction phase, phrase-level answers are generated from candidate sentences by mapping the semantic types and relations in these candidates and those in the question. In the semantic clustering phase, answers are clustered based on the hierarchical relationships in the UMLS. The system lists three kinds of information for each answer: semantic type, associated concepts, and the sentence from which the answer originates.

Weiming et al. evaluated their system on a set of 200 documents concerning different aspects of headache, i.e., medication, therapy, etiology, etc., with factoid questions and complex questions. The results showed that the system achieved 78% recall and 94% precision on factoid questions, and 75% recall and 86% precision on complex questions. The average recall and precision were 77% and 92%, respectively.

4.2.2.4. *Logic-based medical QA systems and approaches*. Terol et al. [123] have explored a logic-based approach, in adapting a generic restricted-domain QA system to the medical domain. The medical QA system is designed to answer NL questions that belong to the 10 most frequent generic medical question types in Ely et al.'s generic taxonomy of clinical questions (see Table 5). The QA processing in the system is based on the derivation of LFs (logic forms) from texts through the application of NLP techniques and on the complex treatment of the derived LFs.

Terol et al.'s medical QA system consists of four main processing modules, as shown in Fig. 10.

The four main QA processing stages rely upon sentence preprocessing and LF derivation as well as upon medical NER and question pattern generation.

The LF of a sentence is derived by applying NLP rules to the dependency relationship of the words in the sentence. Terol et al. use the broad-coverage parser Minipar [124] in order to obtain the dependency relationships. Once the dependency

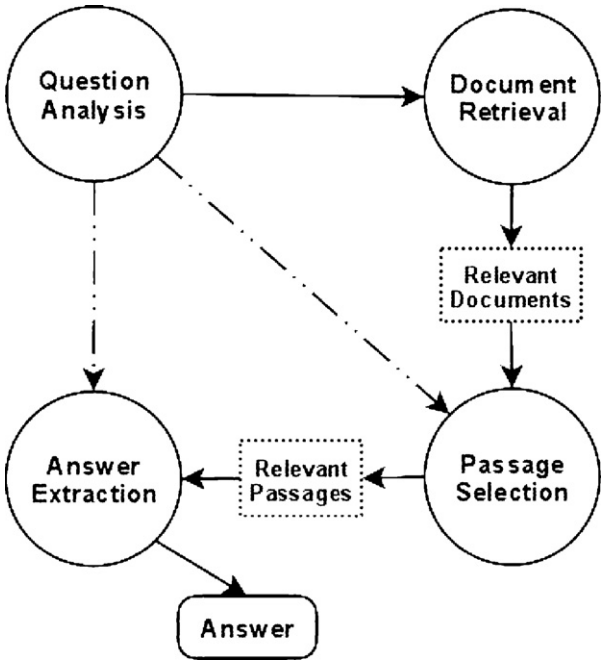

Fig. 10 – Terol et al.'s medical QA system architecture [123].

relationships are obtained, the LF is derived by applying two kinds of NLP rules to the dependency tree, starting in the leaves, continuing through the ramifications of the tree, and ending in the root.

Terol et al. note that their technique for deriving LFs is different from other techniques such as the one used in Moldovan et al.'s [40] COGEX system, which takes as input the parse tree of a sentence, or that in Mollá et al.'s [42] ExtrAns system, which employs the flat form as an intermediate step between the sentence and the LF. Both in Moldovan et al.'s system and in Terol et al.'s system, the identification of predicates in the LFs is based on the format specified by the Logic Form Transformation of eXtended WordNet [18,39], whereas in Mollá et al.'s system it is accomplished through a more complex terminology based on logic treatment. Terol et al. used the technique for deriving LFs both in the question analysis and answer extraction stages, similarly as in other systems. Terol et al.' medical domain-specific QA system, however, uses inference rules deeper than those applied by Moldovan et al. and Mollá et al. in the logic form treatment task.

The medical domain-specific NER task is accomplished by retrieving from UMLS Metathesaurus information on the concepts and semantic types corresponding to the arguments in the LFs.

The offline task of question pattern generation consists of defining the patterns that identify each generic question type. Terol et al. describe schemes for manual pattern generation and supervised automatic pattern generation, which involve setting lower and upper thresholds for the number of medical entities.

The core of Terol et al.'s medical QA system is the module for question analysis. The question analysis phase consists of question classification and question analysis. The question classification task consists of derivation of the LF of the

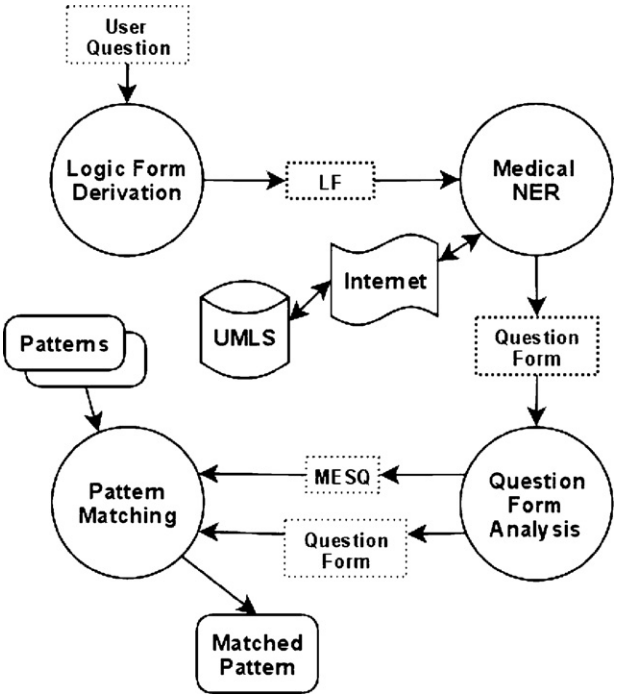

Fig. 11 – Process of question classification in Terol et al.'s logic-based medical QA system [123].

question, extraction of the main verb of the LF, medical NER, computation of medical entities score in question (MESQ) through question form analysis, matching of the main verb and medical entities in the question LF with the verbs and medical entities in the LFs of the patterns of generic questions, and finally selection of the pattern that best meets the criteria (see Fig. 11). Question analysis consists of capturing the semantics of the question using WordNet and UMLS Metathesaurus, recognition of the expected answer type according to the classification of answer types corresponding to the 10 generic question types, and identification of keywords through applying heuristics to the predicates and the relationships between predicates in the question LF.

The document retrieval module in Fig. 10 uses Google to retrieve relevant documents according to a predefined classification of medical Web sites. In the passage selection stage in Fig. 10, the system extracts sentences in the documents that contain at least one question keyword.

Finally, the answer extraction process consists of derivation of the LF of a candidate answer sentence, identification of the main verb in the LF, comparison of the main verb with the set of verbs that correspond to the generic question, recognition of medical NEs in the LF, verification of whether or not the medical NEs match the ones expected by the question, and analysis of the predicates relating the candidate answer, the main verb, and the rest of the medical NEs in the answer LF.

Terol et al. evaluated the question analysis module, obtaining 94.4% precision on 250 questions.

4.2.3. *Summary of medical QA*

A brief summary of the medical QA approaches reviewed is presented in Table 7.

4.3. *Biological QA*

4.3.1. *Introduction to biological QA*

In contrast to the medical domain, where the EBM paradigm provides a structured framework that can be exploited for QA, we find, to our knowledge, no typology/taxonomy of biological domain-specific questions, and, correspondingly, fewer approaches that have explored biological domain-specific QA. (In this review we exclude the approaches that belong to the TREC Genomics Track, which remain at the level of passage extraction.)

4.3.2. *Biological QA systems and approaches*

In this section, we review current research efforts directed toward QA in the biological (or biomedical) domain.

4.3.2.1. Preliminary approaches to biological QA. Yu and Lee [125,126] propose to build a biological QA system that provides experimental evidences as answers. As a first step, they explore NLP techniques for identifying sentences that summarize the images appearing in full-text articles. Their study is based on a few assumptions: (1) that the content of images appearing in a full-text article can be summarized by the sentences in the abstract of the article, (2) that the content of an image corresponds to the text description associated with the image, and (3) that there are lexical similarities between the descriptive text associated with each image and the corresponding sentence(s) in the abstract.

Yu and Lee used hierarchical clustering techniques to cluster abstract sentences and images, based on the lexical similarities. Specifically, they explored three strategies for linking abstract sentences to images: (1) per-image, which clusters each image caption with abstract sentences, (2) per-abstract-sentence, which clusters each abstract sentence with image captions, and (3) mix, which clusters all image captions with all abstract sentences. The results of evaluation showed that both per-image and per-abstract-sentence options outperformed mix, and that per-image significantly outperformed per-abstract-sentence, thereby suggesting the relative importance of the features in abstract sentences as compared to those in image captions. One of the best systems in the evaluation achieved a precision of 100% and a recall of 4.6%. The user interface developed by Yu and Lee, BioEx, which allows biologists to access images from abstract sentences, was favored over two baseline systems (PubMed and SummaryPlus [127]) by 87.8% of 40 biologists who evaluated the interfaces.

4.3.2.2. Semantics-based biological QA systems and approaches. Semantics-based approaches to biological QA include those by Takahashi et al. [128], Lin et al. [129], and Shi et al. [130].

Takahashi et al. [128] propose a semantics-based biological QA system, which utilizes UMLS, as well as gene/protein/compound name and family name dictionaries [131] and other thesauri, for resolving synonyms and handling semantic classes. The semantic class information is used in both question analysis and answer extraction. They used approximately six million MEDLINE entries, after removing the entries having only titles. In the preprocessing phase, each

Table 7 – Summary of medical QA approaches.

Approach	(Category); task concerned; technique used
Huang et al. [87]	Question classification by manual evaluation. Evaluation of the PICO framework.
Yu et al. [54,88,89]	Question filtering and question classification using ML algorithms (w/ and w/o semantic information).
Kobayashi and Shyu [90]	Question classification using different parsing methods (w/ and w/o semantic information).
Slaughter et al. [91]	Analysis of semantic patterns of Q/A, based on UMLS semantic relations.
Cao et al. [92]	Evaluation of different answer presentation formats.
Yu et al. [93,94,102]	Non-semantic-knowledge-based. MedQA system for definitional questions. Incorporation of text summarization in QA.
Sang et al. [95]	Non-semantic-knowledge-based. Offline strategies for Dutch medical QA system. Use of document layout and syntactic patterns based on dependency relations.
Jacquemart and colleagues [73,103]	Semantics-based. Semantic modeling of medical questions. Use of UMLS concepts, semantic types and relations for NER and answer extraction.
Niu et al. [57,104–106]	Semantics-based; PICO frame-based. EPoCare system. Identification of PICO roles in medical text. Identification of semantic classes and relations. Detection and classification of clinical outcome. Incorporation of text summarization in QA.
Demner-Fushman et al. [107–111]	Semantics-based; PICO frame-based. Semantic knowledge extractors for identifying PICO elements from medical text. Scheme for annotating MEDLINE abstracts w/ information on clinically relevant elements. Scoring algorithm for semantic matching. Use of semantic clustering and summarization.
Weiming et al. [112]	Semantics-based. Semantic interpretation of Q/A. Incorporation of semantic clustering in QA.
Terol et al. [123]	Logic-based. Logic form transformation of Q/A. Medical NER based on UMLS concepts and semantic types. Question classification using pattern matching.

biological term in the document set was assigned with an ID and semantic class.

The main QA process in the proposed system consists of four phases (see Fig. 12). In the question analysis phase, the question is parsed and corresponding answer types are specified, based on a predefined set of answer types, created by utilizing UMLS and other thesauri. In the next phase, the system first expands the query terms, based on the resources and heuristics, to formulate queries for a full-text search. The system uses the MySQL full-text search function to retrieve relevant documents. In the next phase, the system selects terms from the retrieved documents, whose semantic class or superclass corresponds to the answer types determined in the question analysis phase. If those terms and keywords in the query are described in the sentences in certain relationships, such as subject-object, or are positioned closely in the abstract, the system selects them as answer candidates. The output from this phase includes the candidate terms, their IDs, and supporting evidences, including the sentences in which they appear. In the final phase, the system selects the answer from candidate terms, using a voting mechanism, if necessary. The system presents the answer with the evidential sentence and abstracts.

Takahashi et al. mention that they are evaluating the system by several methods.

Lin et al. [129] propose a QA system that uses syntactic and semantic feature matching, which focuses on answering factoid questions about biomolecular events, such as gene–protein interactions, the corresponding answers to which are biomedical named entities (NEs).

The QA process in the proposed system consists of four stages: Question Processing, Passage Retrieval, Candidate Extraction and Feature Generation, and Answer Ranking. The question processing stage involves named entity recognition (NER), semantic role labeling (SRL), question classification, and query modification. The NER step extracts NEs from the question. The SRL step then extracts predicates and corresponding arguments. The question classification step uses hand-crafted patterns to identify the target answer NE type. The query modification step uses WordNet and Longman's dictionary [132] to expand queries, which consist of the remaining phrases from the SRL step excluding stop words, by including a list of synonyms and other tenses for the main verb in the question. In the passage retrieval stage, queries are sent to Google

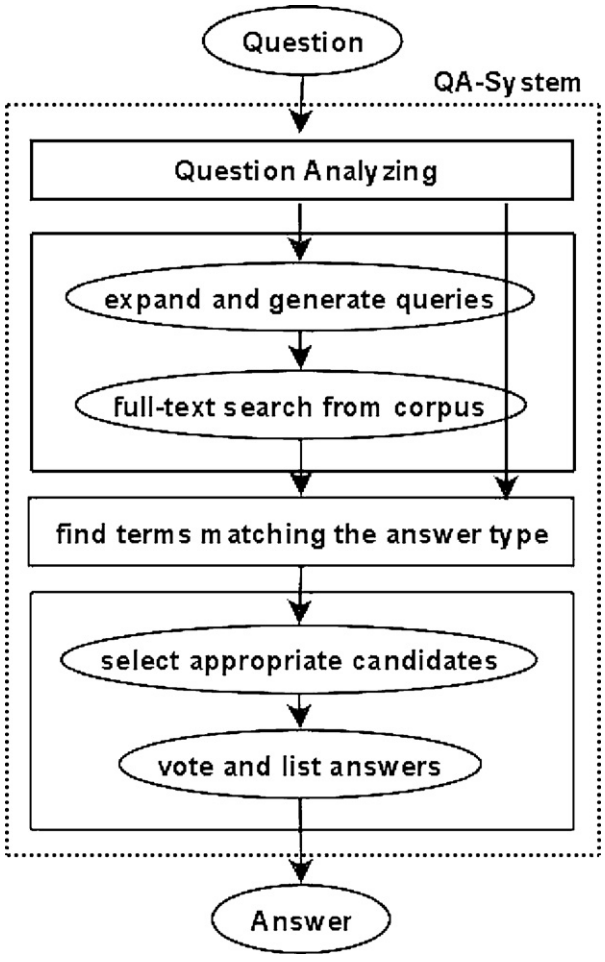

Fig. 12 – Takahashi et al.'s biological QA system architecture [128].

and Web pages are retrieved exclusively from Google's index of the PubMed database. In the candidate extraction and feature generation stage, NER and SRL are used to extract candidate NEs and corresponding features. The system uses the GENIA Tagger [133] to identify four types of NE: protein, DNA, RNA, and cell. It also extracts biomolecular events expressed in nominal form in which the relevant NEs are involved. The SRL component generates semantic features for answer ranking by recognizing the predicate and corresponding argument phrases from a sentence (see Table 8). The SRL step also checks whether or not answer candidates generated from the NER step correspond to the expected type. In the answer ranking stage, the system treats each NE extracted during the previous stage as an answer candidate, and calculates a score for each candidate using a linear function of the weighted sum of the candidate's features. The features considered include both syntactic and semantic ones, such as verb match, argument match, NE match, NE similarity, keyword similarity, argument similarity, consecutive word match, and Google reciprocal rank.

Noting that QA system evaluation measurements may suffer from the same score problem, Lin et al. propose an improved mean reciprocal rank (MRR) measurement, called mean average reciprocal rank (MARR), as well as a formula to reduce the computational complexity of MARR. The evaluation of the proposed QA system has shown that, when using all eight features and tuning feature weights, the system achieves a top-1 MARR of 74.11% and top-5 MARR of 76.68%, thereby exhibiting 16.17% and 18.61% respective increases compared to the baseline system configuration without using features.

Table 8 – Argument types for semantic role labeling [129].

Type	Description
Arg0	Agent
Arg1	Direct object/theme/patient
Arg2-5	Not fixed
ArgM-NEG	Negation marker
ArgM-LOC	Location
ArgM-TMP	Time
ArgM-MNR	Manner
ArgM-EXT	Extent
ArgM-ADV	General-purpose
ArgM-PNC	Purpose
ArgM-CAU	Cause
ArgM-DIR	Direction
ArgM-DIS	Discourse connectives
ArgM-MOD	Modal verb
ArgM-REC	Reflexives and reciprocal
ArgM-PRD	Marks of secondary predication

Shi et al. [130] present BioSquash, a semantics-based QA-oriented multi-document summarization system for the biomedical domain, built upon a general-purpose summarizer, Squash [134].

The BioSquash system consists of four main components (see Fig. 13). The annotator module annotates the documents and the question text with syntactic and shallow semantic information, utilizing the GENIA ontology for biomedical NER and using automatic semantic role labeling [35]. The conceptualizer (or concept similarity) module obtains biomedical and general concepts and their semantic relations by utilizing UMLS and WordNet. The synthesizer (or extractor) module constructs a semantic graph, based on the semantic role labeling and the semantic information on the documents and the question text. It performs sentence selection based on the selection of a sub-graph in the semantic graph. The module also generates sentence clusters related to the question, based on the measurement of sentence redundancy. The editor module performs sentence ordering upon the sentence clusters, eliminates irrelevant content, and generates the final summary.

Shi et al. evaluated the system by using MEDLINE abstracts as the benchmark in lieu of expert-written summaries. Evaluation results have suggested the usefulness of the summarization system for QA.

4.3.2.3. Inference-based biological QA systems and approaches. Kontos et al. [135,136] present an inference-based approach to biomedical QA, in describing construction of an NL grammar and analysis of textual rhetoric relations for their model-based biomedical QA system AROMA.

The AROMA system is designed to provide model-based explanations as answers to non-factoid questions, with the use of rhetoric relation recognition and causal knowledge

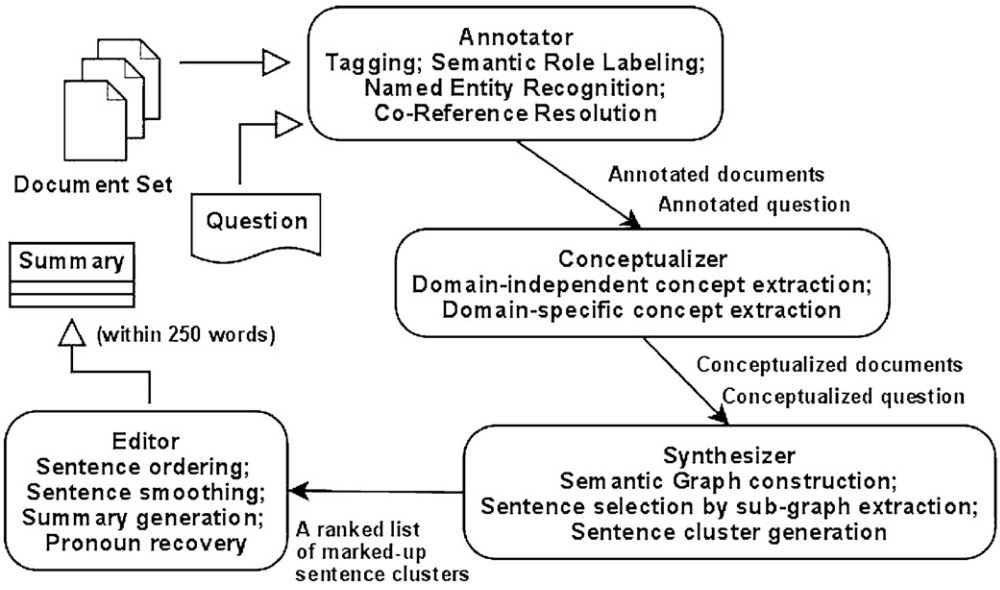

Fig. 13 – Shi et al.'s BioSquash system architecture [130].

extraction, automatically generating textual descriptions of the behavior of a biomedical model.

Kontos et al. compare two possible methodologies for applying deductive reasoning on texts. The first methodology, employed by those approaches that we categorize in this paper as "logic-based", relies on the translation of texts into formal representations of their content, upon which deduction is performed. The advantage of this methodology consists in the simplicity and availability of the required inference engine. Its disadvantage concerns the cost of reprocessing all the texts and restoring their formal representations in the case of changes. The second methodology eliminates the need for translating the texts into formal representations by using an inference engine capable of performing deduction "on the fly", i.e., directly from the texts. The disadvantage of this second methodology is that it needs a more complex inference engine. Its advantage is that it avoids the translation into a formal representation. The AROMA system is intended to perform causal reasoning "on the fly", based on a representation-independent syllogistic text analysis method [137].

The AROMA system consists of three subsystems. The knowledge extraction subsystem extracts knowledge including rhetoric relations from biomedical texts, and integrates partial causal knowledge extracted from multiple texts. The operation of this subsystem is based on the recognition of noun phrases, verb groups, and their relations. The extracted sentences are automatically converted into Prolog facts. The causal reasoning subsystem generates answers and model-based explanations of the answers, by applying inference rules over the linguistic knowledge and domain knowledge manually entered as Prolog facts, combined with the causal knowledge extracted by the first subsystem. Finally, the simulation subsystem generates time-dependent numerical values for a model, which are compared with experimental data.

Kontos et al. focus on causal knowledge involving the interaction of entities and events/processes. The AROMA system is intended to answer causal questions such as those shown in Table 9.

Table 9 – Questions involving causal knowledge [135].

Question	Abstract specification
Causal antecedent	What caused some event to occur? What state/event causally led to an event/state?
Causal consequence	What are the consequences of an event/state? What causally unfolds from an event/state?
Enablement	What object or event enables an agent to perform an action?
Instrumental/Procedural	What instrument or body part is involved when a process is performed?

A pilot experiment using the GENIA corpus, however, revealed some problems with automatic induction of grammar rules meant to capture causal relations.

4.3.2.4. Logic-based biological QA systems and approaches. Rinaldi et al. [55] have explored a logic-based approach to biological QA, in adapting Mollá et al.'s [42] ExtrAns system to the genomics domain.

The ExtrAns system, as we briefly discussed in the review of semantic knowledge-based open-domain and non-biomedical-domain QA approaches, is a logic-based restricted-domain QA system, originally developed for the Unix manual domain. ExtrAns answers domain-specific questions by exploiting linguistic knowledge extracted from the documents and terminological knowledge about the given domain. In an offline phase, the system analyzes the sentences in the documents, transform them into formal semantic representations, i.e., MLFs (Minimal Logical Forms), and store the MLFs in a KB. In an online phase, the system

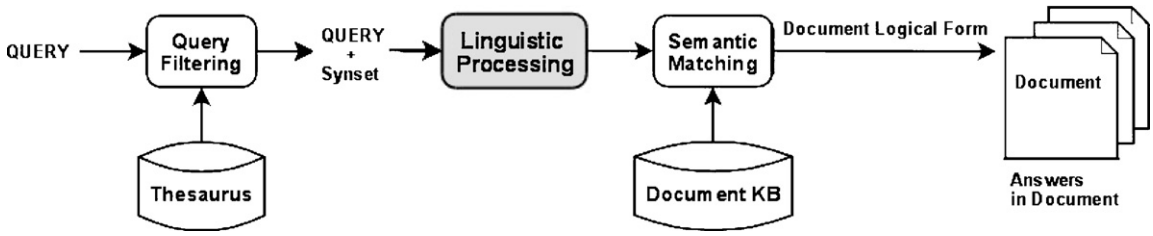

Fig. 14 – Rinaldi et al.'s biological QA system architecture [55].

processes the question using the same basic mechanism. The MLF of the question is proved by deduction over the MLFs of document sentences in the KB. In case there is a direct match, the original sentence is presented as answer. Otherwise, relaxation of the proof criteria is applied in a stepwise manner.

The QA process in the ExtrAns system as described above depends upon the tokenization and terminological preprocessing of documents and the question. The system relies on a domain-specific terminological KB, a WordNet-like handcrafted thesaurus, which assigns a synset (synonym set) identifier to each set of synonyms representing an identical concept. If a term in a question or document sentence belongs to a synset in the terminological KB, the term is replaced by the synset identifier in the LF. This effectively yields a canonical form, whereby each term in the question can be matched to all its synonyms. Rinaldi et al. note that this approach amounts to an implicit terminological normalization for a given domain. In this regard, they also note that the impact of domain-specificity in the ExtrAns system involves only the construction of the thesaurus and the preprocessing of input texts. They thus consider an advantage of ExtrAns' MLFs as consisting in the fact that they can be produced with minimal domain knowledge, thus making the approach easily portable to different domains.

Rinaldi et al.'s QA system for the genomics domain has the same basic architecture as the original ExtrAns system, as shown in Fig. 14.

In adapting the ExtrAns system to the genomics domain, Rinaldi et al. worked with two domain-specific document collections: (1) GENIA corpus, and (2) 'Biovista' corpus consisting of full-text journal articles, generated from MEDLINE using two seed term lists concerning genes and pathways.

The process of analyzing and processing the document sets consisted of a series of steps, including preliminary document zone detection, terminological processing, deep-syntactic parsing, and derivation of MLFs.

Rinaldi at al. first used an XML-based filtering tool in order to identify document zones that need to be processed in a specific manner.

For terminological processing, which is not needed for the annotated GENIA corpus, Rinaldi et al. first marked up the terms in the Biovista corpus using additional XML tags. Next they chunked the entire corpus using the shallow syntactic chunker LT Chunk [138]. They then expanded the corpus terms to the boundary of the phrasal chunk in which they appear. In detecting the relations between terms, Rinaldi et al. focused their attention on the relations of synonymy and hyponymy, and gathered those relations in a thesaurus, using the WordNet synset as the organizing unit. Rinaldi et al. mention that one of the most difficult problems that they have encountered in working on restricted-domain QA concerns the syntactic ambiguity generated by multi-word units, especially those involving technical terms. The solution that they adopted was to parse multi-word terms as single syntactic units.

In consideration of the advantages of parsing over shallow processing methods as well as the disadvantages of probabilistic parsers compared to deep-linguistic, formal grammar-based parsers, Rinaldi et al. adapt ExtrAns to use a broad-coverage deep-syntactic parser, Pro3Gres [139], which is comparable to a probabilistic parser in speed but is more deep-linguistic than the latter. The deep-linguistic parser generates functional dependency structures representing sentence-level syntactic relations, which lend easily to predicate-argument structure-based shallow semantic representations such as MLFs.

The final stage of document processing concerns construction of MLFs. Rinaldi et al. note that the construction of MLFs is simplified by the deep-linguistic dependency-based parsing done in the previous stage, thanks to the relatively direct mapping between labeled dependencies and surface semantic representations.

Rinaldi et al. have not reported any formal evaluation of the QA system adapted to the genomics domain.

4.3.3. *Summary of biological QA*

A brief summary of the biological (or biomedical) QA approaches reviewed is presented in Table 10.

5. Future directions

In this section we briefly recapitulate the overall current trends in biomedical QA research, and project directions for the future research development in the area.

First, we discuss the matter from the point of view of the three main phases of QA processing.

5.1. *Question processing*

In the case of medical QA, we have seen that the EBM paradigm provides a useful framework for question analysis and classification, with PICO frame and Ely et al.'s taxonomy of clinical questions serving as bases for question analysis and classification. However, we have also noted that current medical QA approaches have limitations in terms of the types and for-

Table 10 – Summary of biological QA approaches.

Approach	(Category); task concerned; technique used
Yu and Lee [125,126]	NLP techniques for identifying summary sentences for biological images. BioEx interface for accessing biological images.
Takahashi et al. [128]	Semantics-based. Use of UMLS, lexicons, and thesauri for semantic and terminological information for QA.
Lin et al. [129]	Semantics-based. Use of semantic role labeling for extraction of predicate and arguments. Use of semantic features for answer ranking.
Shi et al. [130]	Semantics-based. BioSquash system for QA-oriented summarization of biomedical documents.
Kontos et al. [135,136]	Inference-based. AROMA system for model-based biological QA. NL grammar for capturing causal relations. Text analysis of rhetoric relations involving causal knowledge.
Rinaldi et al. [55]	Logic-based. ExtrAns system adapted to genomics domain. Construction of MLFs. Terminological processing based on WordNet.

mats of questions that they can process. Yu et al.'s MedQA system, for example, can only handle definitional questions. Niu et al.'s EPoCare system operates on PICO-format queries. Similarly, Demner-Fushman et al.'s clinical QA system accepts structured queries in the PICO format, not NL questions. As the ultimate goal of QA systems, including biomedical QA systems, is to be able to accept a variety of NL questions and to generate appropriate NL answers, further research needs to be done so as to enable more sophisticated analysis and classification of medical questions as well as their conversion into canonical forms. Jacquemart and Zweigenbaum's semantic modeling of medical questions and Terol et al.'s LF- and NER-based analysis/classification of medical questions seem to point to promising directions. In the case of biological QA, research needs to be done also toward the development of domain-specific typology and taxonomy of questions.

5.2. *Document processing*

The search engine used by a biomedical QA system for document retrieval depends on whether the given system is Web-based or corpus-based. We have seen that Jacquemart and Zweigenbaum's and Terol et al.'s systems, for example, use Google to retrieve documents from the Web. Other systems, e.g., Weiming et al.'s, use standard IR engines, such as Lucene, to retrieve documents from the text collection. Still other systems, e.g., Demner-Fushman et al.'s, use biomedical domain-specific document query systems, e.g., PubMed, to retrieve documents from the MEDLINE collection. Yu et al.'s MedQA system uses both Google and Lucene to retrieve documents from the Web and from MEDLINE. In this regard, it seems worthwhile to consider Yu et al.'s suggestion for a possible combination of a state-of-the-art search engine with a biomedical domain-specific QA functionality. In addition, more research on the methods of utilizing semantic knowledge, both domain-dependent and question-specific, in the document retrieval process seems to be in order. The phase of passage extraction, compared to that of document retrieval, can benefit even more from incorporation of semantic knowledge. In this regard, we have seen that, in addition to other strategies for extracting relevant sentences, such as document zone detection, cue-phrase-based sentence categorization, and identification of lexico-syntactic patterns of sentences that correspond to question types (e.g., Yu et al.), semantic tagging and annotation of text (e.g., Delbecque et al., Demner-Fushman et al.) has been used. Besides question analysis/classification, passage extraction has a crucial impact upon the quality of answers generated by a QA system. As such, biomedical QA researchers need to continue to refine techniques for exploiting semantic knowledge in extracting passages for answer selection.

5.3. *Answer processing*

Most biomedical QA approaches that we have reviewed rely on some form of semantic matching between question and candidate answers, exploiting knowledge about the concepts involved, their semantic types, and the relations between those concepts and semantic types, mostly obtained from UMLS resources, in ranking and selecting answers. In addition to the utilization of semantic knowledge, we have also seen several approaches (e.g. Yu et al., Niu et al., Demner-Fushman et al., Weiming et al.) that incorporate (semantic) clustering-based text summarization techniques in the answer processing phase. In this regard, we have also reviewed a QA-oriented extractive summarization system (Shi et al.) that generates multi-document summaries on the basis of semantic role labeling, semantic graph construction, and semantic clustering. It seems that incorporation of text summarization in biomedical QA points to a promising direction, given the need of generating synthesized answers to biomedical questions that require clinical or experimental evidences as answers. In this regard, research needs to be done as to how to properly synthesize potentially conflicting evidences in generating answers. More research is also in order as to the appropriate format of answer presentation.

Next, we discuss the matter from the point of view of the framework of semantic knowledge-based QA.

5.4. *Utilization of semantics*

Most biomedical QA approaches reviewed in this paper more or less utilize domain-specific semantic information throughout the question processing, document processing, and answer processing phases of the QA process, as expected, given the fact that a major characteristic of restricted-domain QA concerns utilization of domain-specific semantic knowledge resources. Without repeating our discussion above, it may be said that continued research on the effective incorporation of semantic knowledge in the QA process is both required and envisioned.

5.5. *Use of logic and reasoning*

Perhaps the one aspect that has been least researched in biomedical QA concerns the incorporation of logic and reasoning mechanisms. We have encountered only a few approaches that have been attempted in this regard. As we have seen, Kontos et al. worked on analysis of textual rhetoric relations with a view to enabling causal inference, without using logic form transformation of texts. Terol et al. and Rinaldi et al. worked on adapting logic-based restricted-domain QA systems to the medical domain and the genomics domain, respectively. Given the fact that answering biomedical questions involves finding supporting or denying evidences, it seems fitting to explore the methods of deriving indirect evidences, beyond directly and explicitly stated answers, by utilizing inference mechanisms. Furthermore, the fact that some of the terminological and ontological resources available in the biomedical domain are structured and accessible in logic-based formalisms, suggests the relevance and feasibility of exploring logic-based approaches to biomedical QA.

To sum up, we envisage the following tasks ahead for biomedical QA research:

1. Construction of domain-specific typology and taxonomy of questions (biological QA).
2. Development of more sophisticated techniques for NL question analysis and classification.
3. Development of effective methods for answer generation from potentially conflicting evidences.
4. More extensive and integrated utilization of semantic knowledge throughout the QA process.
5. Incorporation of logic and reasoning mechanisms for answer inference.

As our own way of contributing to the development in the field as envisioned above, we have begun our work on a *LO*gic-based *Q*uestion-Answering *S*ystem for the *Med*ical domain (LOQAS-Med) [140,141], which uses Description Logic [142] as the formalism for knowledge representation and reasoning.

The proposed logic-based QA system aims to provide answers to medical questions based on explicitly stated facts as well as to derive hypothesis-confirming or hypothesis-denying evidences by utilizing inference. As a first step toward building the proposed system, we have devised our own semantic categorization scheme to re-classify Ely et al.'s 10 most common generic clinical questions (see Table 5) and some of their variations. Our categorization scheme classifies medical questions along four hierarchical levels of classification: (1) semantic relations, (2) semantic classes, (3) target specificity, and (4) context specificity. Based on the scheme, we have semantically analyzed each question category. More specifically, we have constructed question and answer patterns as semantic triples in the form of subject–predicate–object, based on the identification of the semantic types and semantic relations in the UMLS Semantic Network that correspond to the arguments and predicates contained in each question type.

In contrast to Niu et al.'s and Demner-Fushman et al.'s semantics-based medical QA approaches, which are mainly based on the identification and extraction of PICO elements, our approach aims at identification and extraction of specific semantic types and relations that directly correspond to the arguments and predicates. While Jacquemart, Zweigenbaum, and Delbecque have also investigated semantic modeling of medical questions in the form of concept–relation–concept triples, our semantic analysis of medical question/answer patterns is based on our own classification of medical questions as described above. Finally, while Terol et al.'s logic-based QA system is also intended to handle Ely et al.'s 10 most common clinical questions, the Q–A pattern matching used by their system is mainly based on the matching of the number of medical entities of corresponding semantic types, whereas the Q–A pattern-matching process in our proposed system involves direct matching of the semantic types corresponding to the arguments and the semantic relations corresponding to the arguments between question and candidate answers.

As we have already contributed to the field with our semantic analysis, modeling, and classification of medical questions, we believe that LOQAS-Med, with its extensive utilization of semantic knowledge and with its incorporation of logic and reasoning mechanisms for answer inference from potentially conflicting evidences, will further contribute to moving the field forward.

6. Conclusion

In this paper, we have reviewed the current state of the art in biomedical QA research, with a focus on semantic knowledge-based QA approaches. Corresponding to the growth of biomedical information, there is a growing need of QA systems that can help better utilize the ever-accumulating information. While the biomedical domain poses particular challenges for QA, with highly complex terminology, it also provides QA researchers with a variety of domain-specific semantic knowledge resources that can be exploited in the QA process. It is envisioned that continued research toward development of more sophisticated techniques for processing NL text, for utilizing semantic knowledge, and for incorporating logic and reasoning mechanisms, will lead to more useful QA systems.

Conflict of interest statement

None declared.

Acknowledgement

We would like to thank the anonymous reviewers for their helpful feedback.

REFERENCES

[1] TREC (Text REtrieval Conference), http://trec.nist.gov/.
[2] E.M. Vorhees, The TREC question answering track, Nat. Lang. Eng. 7 (2001) 361–378.
[3] L. Hirschman, R. Gaizauskas, Natural language question answering: the view from here, Nat. Lang. Eng. 7 (2001) 275–300.
[4] A. Vallin, B. Magnini, D. Giampiccolo, L. Aunimo, C. Ayache, P. Osenova, A. Penas, M. de Rijke, B. Sacaleanu, D. Santos, R. Sutcliffe, Overview of the CLEF 2005 multilingual question answering track, in: Proceedings of the Third Cross Language Evaluation Forum (CLEF 2005), 2005.
[5] N. Kando, Overview of the fifth NTCIR workshop, in: Proceedings of the Fifth NTCIR Workshop Meeting Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, 2005.
[6] D. Mollá, J.L. Vicedo, Question answering in restricted domains: an overview, Comput. Linguist. 33 (2007) 41–61.
[7] V. Lopez, E. Motta, V. Uren, M. Sabou, State of the art on semantic question answering: a literature review. Technical Report kmi-07-03, Knowledge Media Institute, The Open University, Milton Keynes, UK, 2007.
[8] G.A. Miller, WordNet:, A lexical database for English, Commun. ACM 38 (1995) 39–41.
[9] C.E. Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, Cambridge, Mass, 1998.
[10] J.L. Vicedo, A. Ferrández, A semantic approach to question answering systems, in: Proceedings of the Ninth Text REtrieval Conference (TREC-9), NIST, 2000, pp. 511–516.
[11] E. Alfonseca, M. De Boni, J.-L. Jara-Valencia, S. Manandhar, A prototype question answering system using syntactic and semantic information for answer retrieval, in: Proceedings of the Tenth Text REtrieval Conference (TREC 2001), NIST, 2001, pp. 680–685.
[12] E. Hovy, L. Gerber, U. Hermjakob, C.-Y. Lin, D. Ravichandran, Toward semantics-based answer pinpointing, in: Proceedings of the First Human Language Tech. Conference (HLT-2001), ACL, San Diego, CA, USA, 2001, pp. 1–7.
[13] M. Fleischman, E. Hovy, A. Echihabi, Offline strategies for online question answering: answering questions before they are asked, in: Proceedings of the 41st Annual Meeting Assoc. Comp. Ling. (ACL 2003), ACL, Sapporo, Japan, 2003, pp. 1–7.
[14] D. Ferrés, S. Kanaan, E. González, A. Ageno, H. Rodríguez, M. Surdeanu, J. Turmo, TALP-QA system at TREC 2004: structural and hierarchical relaxing of semantic constraints, in: Proceedings of the 13th Text REtrieval Conference (TREC 2004), NIST, 2004.
[15] P. Vossen, EuroWordNet: a multilingual database for information retrieval, in: Proceedings of the DELOS Workshop Cross-Language Information Retrieval, 1997.
[16] V. Punyakanok, D. Roth, W-.T. Yih, Natural language inference via dependency tree mapping: an application to question answering, Comput. Linguist. 6 (2004) 1–10.
[17] R. Sun, J. Jiang, Y.F. Tan, H. Cui, T.-S. Chua, M.-Y. Kan, Using syntactic and semantic relation analysis in question answering, in: Proceedings of the 14th Text REtrieval Conference (TREC 2005), NIST, 2005.
[18] S. Harabagiu, G.A. Miller, D.I. Moldovan, WordNet2—a morphologically and semantically enhanced resource, in: Proceedings of the ACL-SIGLEX99: Standardizing Lexical Resources, 1999, pp. 1–8.
[19] C.F. Baker, C.J. Fillmore, J.B. Lowe, The Berkeley FrameNet Project, in: Proceedings of the 36th Annual Meeting of ACL and 17th International Conference on Computational Linguistics (COLING-ACL'98), 1998, pp. 86–90.
[20] C.J. Fillmore, C.R. Johnson, M.R. Petruck, Background to FrameNet, Int. J. Lexicogr. 16 (2003) 235–250.
[21] P. Kingsbury, M. Palmer, M. Marcus, Adding semantic annotation to the Penn TreeBank, in: Proceedings of the 2nd Human Language Technology Conference (HLT-2002), 2002.
[22] M. Palmer, D. Gildea, P. Kingsbury, The Proposition Bank: an annotated corpus of semantic roles, Comput. Linguist. 31 (2005) 71–106.
[23] D. Lin, P. Pantel, Discovery of inference rules for question answering, Nat. Lang. Eng. 7 (2001) 343–360.
[24] R. Girju, Automatic detection of causal relations for question answering, in: Proceedings of the ACL 2003 Workshop Multilingual Summarization and Question Answering, ACL, 2003, pp. 76–83.
[25] S. Beale, B. Lavoie, M. McShane, S. Nirenburg, T. Korelsky, Question answering using ontological semantics, in: Proceedings of the ACL-2004 Workshop Text Meaning and Interpretation, ACL, 2004, pp. 41–48.
[26] S. Harabagiu, D. Moldovan, M. Paşca, M. Surdeanu, R. Mihalcea, R. Gîrju, V. Rus, F. Lăcătuşu, P. Morărescu, R. Bunescu, Answering complex, list and context questions with LCC's question-answering server, in: Proceedings of the Tenth Text REtrieval Conference (TREC 2001), NIST, 2001, pp. 355–361.
[27] S. Narayanan, S. Harabagiu, Answering questions using advanced semantics and probabilistic inference, in: Proceedings of the HLT-NAACL 2004 Workshop Pragmatics of Question Answering, ACL, Boston, USA, 2004, pp. 10–16.
[28] S. Narayanan, S. Harabagiu, Question answering based on semantic structures, in: Proceedings of the 20th Int'l Conf. Comp. Ling. (COLING 2004), Morgan Kaufman, Geneva, Switzerland, 2004.
[29] S. Narayanan, Knowledge-based Action Representations for Metaphor and Aspect (KARMA), Ph.D. Dissertation, University of California at Berkeley, 1997.
[30] K. Murphy, Dynamic Bayesian Networks: Representation, Inference, and Learning, Ph.D. Dissertation, University of California at Berkeley, 2002.
[31] S. Sinha, S. Narayahan, Model-based answer selection, in: Proceedings of the AAAI 2005 Workshop Inference for Textual Question Answering, AAAI Press, Pittsburgh, PA, USA, 2005, pp. 42–46.
[32] S. Narayanan, Reasoning about actions in narrative understanding, in: Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI'99), 1999, pp. 350–357.
[33] S. Narayanan, S. McIlraith, Analysis and simulation of web services, Comput. Netw. 42 (2003) 675–693.
[34] S. Harabagiu, C.A. Bejan, Question answering based on temporal inference, in: Proceedings of the AAAI 2005 Workshop Inference for Textual Question Answering, AAAI Press, Pittsburgh, PA, USA, 2005, pp. 27–34.
[35] D. Shen, M. Lapata, Using semantic roles to improve question answering, in: Proceedings of the Conference

Empirical Methods Nat. Lang. Proc. and Conf. Nat. Lang. Learning Joint Meeting (EMNLP-CoNLL 2007), ACL, Prague, Czech Republic, 2007, pp. 12–21.
[36] D. Gildea, D. Jurafsky, Automatic labeling of semantic roles, Computat. Linguist. 28 (2002) 245–288.
[37] B. Katz, G. Borchardt, S. Felshin, Syntactic and semantic decomposition strategies for question answering from multiple resources, in: Proceedings of the AAAI 2005 Workshop Inference for Textual Question Answering, AAAI Press, Pittsburgh, PA, USA, 2005, pp. 35–41.
[38] S.M. Harabagiu, M.A. Paşca, S.J. Maiorano, Experiments with open-domain textual question answering, in: Proceedings of the 16th Int'l Conf. Comp. Ling. (COLING 2000), Saarbrücken, Germany, 2000, pp. 292–298.
[39] D. Moldovan, B. Rus, Logic form transformations of WordNet and its applicability to question answering, in: Proceedings of the 39th Ann. Meeting Assoc. Comp. Ling. (ACL 2001), ACL, Toulouse, France, 2001, pp. 402–409.
[40] D. Moldovan, C. Clark, S. Harabagiu, S. Maiorano, COGEX: a logic prover for question answering, in: Proceedings of the 2003 Human Lang. Tech. and North Am. Chap. Assoc. Comp. Ling. Joint Conf. (HLT-NAACL 2003), Edmonton, Canada, 2003, pp. 87–93.
[41] D. Moldovan, C. Clark, S. Harabagiu, D. Hodges, COGEX: a semantically and contextually enriched logic prover for question answering, J Appl. Logic 5 (2007) 49–69.
[42] D. Mollá, R. Schwitter, M. Hess, R. Fournier, ExtrAns, an answer extraction system, TAL 41 (2000) 495–522.
[43] D. Mollá, Towards semantic-based overlap measures for question answering, in: Proceedings of the First Australasian Language Technology Workshop (ALTW'03), 2003.
[44] F. Benamara, Cooperative question answering in restricted domains: the WEBCOOP experiment, in: Proceedings of the ACL-2004 Workshop Question Answering in Restricted Domains, ACL, Barcelona, Spain, 2004, pp. 31–38.
[45] R.J. Waldinger, D.E. Appelt, J.L. Dungan, J. Fry, J.R. Hobbs, D.J. Israel, P. Jarvis, D. Martin, S. Riehemann, M.E. Stickel, M. Tyson, Deductive question answering from multiple resources, in: M.T. Maybury (Ed.), New Directions in Question Answering, AAAI Press, 2004, pp. 253–262.
[46] M.E. Stickel, R.J. Waldinger, V.K. Chaudhri, A guide to SNARK, Technical Report, SRI International, Menlo Park, CA, 2000.
[47] J. Curtis, G. Matthews, D. Baxter, On the effective use of Cyc in a question answering system, in: Proceedings of the IJCAI'05 Workshop Knowledge and Reasoning for Answering Questions (KRAQ'05), Edinburgh, UK, 2005, pp. 61–70.
[48] Cyc, http://www.cyc.com/.
[49] C. Clark, D. Hodges, J. Stephan, D. Moldovan, Moving QA towards reading comprehension using context and default reasoning, in: Proceedings of the AAAI 2005 Workshop Inference for Textual Question Answering, AAAI Press, Pittsburgh, PA, USA, 2005, pp. 6–12.
[50] L. Tari, C. Baral, Using AnsProlog with Link Grammar and WordNet for QA with deep reasoning, in: Proceedings of the AAAI 2005 Workshop Inference for Textual Question Answering, AAAI Press, Pittsburgh, PA, USA, 2005, pp. 13–21.
[51] D.D. Sleator, D. Temperley, Parsing English with Link Grammar, in: Proceedings of the Third Int'l Workshop on Parsing Technologies, 1993.
[52] C. Baral, G. Gelfond, M. Gelfond, R.B. Scherl, Textual inference by combining Multiple Logic Programming paradigms, in: Proceedings of the AAAI 2005 Workshop Inference for Textual Question Answering, AAAI Press, Pittsburgh, PA, USA, 2005, pp. 1–5.
[53] D. Bobrow, D. Condoravdi, R. Crouch, R. Kaplan, L. Karttunen, T. King, V. dePaiva, A. Zaenen, A basic logic for textual Davidsonian inference, in: Proceedings of the AAAI 2005 Workshop Inference for Textual Question Answering, AAAI Press, 2005, pp. 47–51.
[54] H. Yu, C. Sable, Being Erlang Shen: identifying answerable questions, in: Proceedings of the IJCAI'05 Workshop Knowledge and Reasoning for Answering Questions (KRAQ'05), Edinburgh, UK, 2005, pp. 6–14.
[55] F. Rinaldi, J. Dowdall, G. Schneider, A. Persidis, Answering questions in the genomics domain, in: Proceedings of the ACL-2004 Workshop Question Answering in Restricted Domains, ACL, Barcelona, Spain, 2005.
[56] P. Zweigenbaum, Question answering in biomedicine, in: Proceedings of the EACL 2003 Workshop Natural Language Processing for Question Answering, 2003, pp. 1–4.
[57] Y. Niu, G. Hirst, Analysis of semantic classes in medical text for question answering, in: Proceedings of the ACL-2004 Workshop Question Answering in Restricted Domains, ACL, Barcelona, Spain, 2004.
[58] P. Zweigenbaum, Knowledge and reasoning for medical question-answering, in: Proceedings of the ACL-IJCNLP 2009 Workshop on Knowledge and Reasoning for Answering Questions, ACL and AFNLP, Suntec, Singapore, 2009, pp. 1–2.
[59] J.D. Kim, T. Ohta, Y. Tateisi, J. Tsujii, GENIA Corpus—a semantically annotated corpus for bio-textmining, Bioinformatics 19 (2003) i180–i182.
[60] A.L. Rector, S. Bechhover, C.A. Goble, I. Horrocks, W.A. Nolan, W.D. Solomon, The GRAIL concept modelling language for medical terminology, Art. Intell. Med. 9 (1997) 139–171.
[61] A. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap Program, in: Proceedings of the AMIA 2001 Symposium, 2001, pp. 17–21.
[62] T.C. Rindflesch, M. Fiszman, The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text, J. Biomed. Inform. 36 (2003) 462–477.
[63] J.W. Ely, J.A. Osheroff, M.H. Ebell, G.R. Bergus, B.T. Levy, M.L. Chamliss, E.R. Evans, Analysis of questions asked by family doctors regarding patient care, Br. Med. J. 319 (1999) 358–361.
[64] J.W. Ely, J.A. Osheroff, K.J. Ferguson, M.L. Chambliss, D.C. Vinson, J.L. Moore, Lifelong self-directed learning using a computer database of clinical questions, J. Fam. Pract. 45 (1997) 382–388.
[65] D.M. D'Alessandro, C.D. Kreiter, M.W. Peterson, An evaluation of information-seeking behaviors of general pediatricians, Pediatrics 113 (2004) 64–69.
[66] D.L. Sackett, W.M.C. Rosenberg, J.A.M. Gray, R.B. Haynes, W.S. Richardson, Evidence-based medicine: what it is and what it isnt, Br. Med. J. 312 (1996) 71–72.
[67] D.L. Sackett, S. Strauss, W. Richardson, W. Rosenberg, R. Haynes, Evidence-Based Medicine: How to Practice and Teach EBM, 2nd ed., Churchill Livingstone, Edinburgh, UK; New York, USA, 2000.
[68] B. Alper, J. Stevermer, D. White, B. Ewigman, Answering family physicians' clinical questions using electronic medical databases, J. Fam. Pract. 50 (2001) 960–965.
[69] B.S. Alper, D.S. White, B. Ge, Physicians answer more clinical questions and change clinical decisions more often with synthesized evidence: a randomized trial in primary care, Ann. Fam. Med. 3 (2005) 507–513.
[70] L. Berkowitz, Review and evaluation of clinical reference tools for physicians, White Paper, 2002, http://www.uptodate.com/whitepaper/UTD_WP_Internet_Tools.pdf.

[71] J.J. Cimino, J. Li, M. Graham, L.M. Currie, M. Allen, S. Bakken, V.L. Patel, Use of online resources while using a clinical information system, in: Proceedings of the AMIA 2003 Symposium, AMIA, 2003, pp. 175–179.
[72] W.R. Hersh, M.K. Crabtree, D.H. Hickman, L. Sacherek, C.P. Friedman, P. Tidmarsh, C. Mosbaek, D. Kraemer, Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions, J. Am. Med. Inform. Assoc. 9 (2002) 283–293.
[73] P. Jacquemart, P. Zweigenbaum, Towards a medical question-answering system: a feasibility study, Stud. Health. Technol. Inform. 95 (2003) 463–468.
[74] T.Y. Koonce, N.B. Giuse, P. Todd, Evidence-based databases versus primary medical literature: an in-house investigation on their optimal use, J. Med. Libr. Assoc. 92 (2004) 407–411.
[75] F. Magrabi, J.I. Westbrook, E.W. Coiera, A.S. Gosling, Clinicians' assessments of the usefulness of online evidence to answer clinical questions, in: M. Fieschi, et al. (Eds.), MEDINFO 2004., IOS Press, Amsterdam, 2004, pp. 297–300.
[76] M.R. Patel, C.M. Schardt, L.L. Sanders, S.A. Keitz, Randomized trial for answers to clinical questions: evaluating a pre-appraised versus a MEDLINE search protocol, J. Med. Libr. Assoc. 94 (2006) 382–386.
[77] J.I. Westbrook, A.S. Gosing, E. Coiera, Do clinicians use online evidence to support patient care? A study of 55,000 clinicians, J. Am. Med. Inform. Assoc. 11 (2004) 113–120.
[78] J.I. Westbrook, E.W. Coiera, A.S. Gosling, Do online information retrieval systems help experienced clinicians answer clinical questions? J. Am. Med. Inform. Assoc. 12 (2005) 315–321.
[79] J.W. Ely, J.A. Osheroff, M.H. Ebell, M.L. Chambliss, D.C. Vinson, J.J. Stevermer, E.A. Pifer, Obstacles to answering Doctors' questions about patient care with evidence: qualitative study, Br. Med. J. 324 (2002) 710–716.
[80] J.W. Ely, J.A. Osheroff, M.L. Chambliss, M.H. Ebell, M.E. Rosenbaum, Answering physicians' clinical questions: obstacles and potential solutions, J. Am. Med. Inform. Assoc. 12 (2005) 217–224.
[81] E.C. Armstrong, The well-built clinical question: the key to finding the best evidence efficiently, W. Med. J. 98 (1999) 25–28.
[82] W.S. Richardson, M.C. Wilson, J. Nishikawa, R.S. Hayward, The well-built clinical question: a key to evidence-based decisions, ACP J. Club 123 (1995) A12–13.
[83] C. Schardt, M.B. Adams, T. Owens, S. Keitz, P. Fontelo, Utilization of the PICO framework to improve searching PubMed for clinical questions, BMC Med. Inform. Decis. Mak. 7 (16) (2007).
[84] G.R. Bergus, C.S. Randall, S.D. Sinift, D.M. Rosenthal, Does the structure of clinical questions affect the outcome of curbside consultations with specialty colleagues? Arch. Fam. Med. 9 (2000) 541–547.
[85] D. Demner-Fushman, S.E. Hauser, S.M. Humphrey, G.M. Ford, J.L. Jacobs, G.R. Thoma, MEDLINE as a source of just-in-time answers to clinical questions, in: Proceedings of the AMIA 2006 Symposium, AMIA, 2006, pp. 190–194.
[86] J.W. Ely, J.A. Osheroff, P.N. Gorman, M.H. Ebell, M.L. Chambliss, E.A. Pifer, P.Z. Stavri, A taxonomy of generic clinical questions: classification study, Br. Med. J. 321 (2000) 429–432.
[87] X. Huang, J. Lin, D. Demner-Fushman, Evaluation of PICO as a knowledge representation for clinical questions, in: Proceedings of the AMIA 2006 Symposium, AMIA, 2006, pp. 359–363.
[88] H. Yu, C. Sable, H.R. Zhu, Classifying Medical Questions based on an Evidence Taxonomy, in: Proceedings of the AAAI-05 Workshop Question Answering in Restricted Domains, AAAI, Pittsburgh, PA, USA, 2005.
[89] H. Yu, Y. Cao, Automatically extracting information needs from ad hoc clinical questions, in: Proceedings of the AMIA 2008 Symposium, 2008, pp. 96–100.
[90] T. Kobayashi, C.-R. Shyu, Representing clinical questions by semantic type for better classification, in: Proceedings of the AMIA 2006 Symposium, AMIA, 2006, p. 987.
[91] L.A. Slaughter, D. Soergel, T.C. Rindflesch, Semantic representation of consumer questions and physician answers, Int. J. Med. Inform. 75 (2006) 513–529.
[92] Y.-G. Cao, J. Ely, L. Antieau, H. Yu, Evaluation of the clinical question answering presentation, in: Proceedings of the Workshop on BioNLP, ACL, Boulder, Colorado, 2009, pp. 171–178.
[93] M. Lee, J. Cimino, H.R. Zhu, C. Sable, V. Shanker, J. Ely, H. Yu, Beyond information retrieval—medical question answering, in: Proceedings of the AMIA 2006 Symposium, AMIA, 2006, pp. 469–473.
[94] H. Yu, M. Lee, D. Kaufman, J. Ely, J.A. Osheroff, G. Hripscak, J. Cimino, Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians, J. Biomed. Inform. 40 (2007) 236–261.
[95] E.T.K. Sang, G. Bouma, M. de Rijke, Developing offline strategies for answering medical questions, in: Proceedings of the AAAI-05 Workshop Question Answering in Restricted Domains, AAAI, Pittsburgh, PA, USA, 2005.
[96] J. Lin, D. Karakos, D. Demner-Fushman, S. Khudanpur, Generative content models for structural analysis of medical abstracts, in: Proceedings of the HLT-NAACL 2006 Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis (BioNLP'06), 2006, pp. 65–72.
[97] L. McKnight, P. Srinivasan, Categorization of sentence types in medical abstracts, in: Proceedings of the AMIA 2003 Symposium, 2003, pp. 440–444.
[98] M. Light, X.Y. Qiu, P. Srinivasan, The language of bioscience: facts, speculations, and statements in between, in: Proceedings of the HLT-NAACL 2004 Workshop on Linking Biological Literature, Ontologies and Databases (BioLink 2004), 2004, pp. 17–24.
[99] M. Lee, W. Wang, H. Yu, Exploring supervised and unsupervised approaches to detect topics in biomedical text, BMC Bioinform. 7 (2006) 140.
[100] D. Radev, H. Jing, M. Budzikowska, Centroid-based summarization of multiple documents: sentence extraction utility-based evaluation, and user studies, in: Proceedings of the ANLP/NAACL Workshop on Summarization, 2000.
[101] H. Yu, Y. Wei, The semantics of a definiendum constrains both the lexical semantics and the lexicosyntactic patterns in the definiens, in: Proceedings of the HLT-NAACL 2006 Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis (BioNLP'06), 2006.
[102] H. Yu, D. Kaufman, A cognitive evaluation of four online search engines for answering definitional questions posed by physicians, Pac. Symp. Biocomput. 12 (2007) 328–339.
[103] T. Delbecque, P. Jacquemart, P. Zweigenbaum, Indexing UMLS semantic types for medical question-answering, in: R. Engelbrecht (Ed.), Connecting Medical Informatics and Bio-Informatics (ENMI 2005), 2005, pp. 805–810.
[104] Y. Niu, G. Hirst, G. McArthur, P. Rodriguez-Gianolli, Answering clinical questions with role identification, in: Proceedings of the ACL-2003 Workshop Natural Language Processing in Biomedicine, ACL, Sapporo, Japan, 2003, pp. 73–80.

[105] Y. Niu, X. Zhu, J. Li, G. Hirst, Analysis of polarity information in medical text, in: Proceedings of the AMIA 2005 Symposium, AMIA, 2005, pp. 570–574.
[106] Y. Niu, X. Zhu, G. Hirst, Using outcome polarity in sentence extraction for medical question-answering, in: Proceedings of the AMIA 2006 Symposium, AMIA, 2006, pp. 599–603.
[107] D. Demner-Fushman, J. Lin, Knowledge extraction for clinical question answering: preliminary results, in: Proceedings of the AAAI-05 Workshop Question Answering in Restricted Domains, AAAI Press, Pittsburgh, PA, USA, 2005.
[108] D. Demner-Fushman, B. Few, S.E. Hauser, G. Thoma, Automatically identifying health outcome information in MEDLINE records, J. Am. Med. Inform. Assoc. 13 (2006) 52–60.
[109] D. Demner-Fushman, J. Lin, Answer extraction semantic clustering, and extractive summarization for clinical question answering, in: Proceedings of the 21st Int'l Conf. Comp. Ling. and 44th Ann. Meeting Assoc. Comp. Ling. (COLING-ACL 2006), Sydney, Australia, 2006, pp. 841–848.
[110] D. Demner-Fushman, J. Lin, Answering clinical questions with knowledge-based and statistical techniques, Comput. Linguist. 33 (2007) 63–103.
[111] J. Lin, D. Demner-Fushman, The role of knowledge in conceptual retrieval: a study in the domain of clinical medicine, in: Proceedings of the 29th Ann. Int'l ACM SIGIR Conf. (SIGIR'06), ACM, Seattle, WA, USA, 2006, pp. 99–106.
[112] W. Weiming, D. Hu, M. Feng, L. Wenyin, Automatic clinical question answering based on UMLS relations, in: 3rd Int'l Conf. Semantics, Knowledge and Grid (SKG 2007), accepted Xi'An, China, 2007.
[113] S. Barton, Clinical Evidence, BMJ Publishing Group, London, 2002.
[114] C.M. Ball, R.S. Phillips, Evidence-based on Call: Acute Medicine, Churchill Livingstone, Edinburgh, UK, 2001.
[115] D. Barbosa, A. Barta, A. Mendelzon, G. Mihaila, F. Rizzolo, P. Rodriguez-Gianolli, ToX—The Toronto XML Engine, in: Proceedings of the Int'l Workshop on Information Integration on the Web, 2001.
[116] J. Carbonell, J. Goldstein, The use of MMR diversity based reranking for reordering documents and producing summaries, in: Proceedings of the 21st Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 1998, pp. 335–336.
[117] C.Y. Lin, ROUGE: a package for automatic evaluation of summaries, in: Proceedings of the ACL-2004 Workshop on Text Summarization Branches Out, 2004, pp. 74–81.
[118] M.H. Ebell, B.D. Weiss, S.H. Woolf, J. Susman, B. Ewigman, M. Bowman, Strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature, J. Am. Brd. Fam. Pract. 17 (2004) 59–67.
[119] A.T. McCray, A. Burgun, O. Bodenreider, Aggregating UMLS semantic types for reducing conceptual complexity, in: Proceedings of the 10th World Congress Med. Inform. (MEDINFO 2001), 2001, pp. 216–220.
[120] J. Fam. Pract., http://www.jfponline.com/.
[121] Parkhurst Exchange, http://www.parkhurstexchange.com/.
[122] Y. Zhao, G. Karypis, Evaluation of hierarchical clustering algorithms for document datasets, in: Proceedings of the 11th Int'l ACM Conf. on Information and Knowledge Management (CIKM'02), 2002.
[123] R.M. Terol, P. Martínez-Barco, M. Palomar, A knowledge based method for the medical question answering problem, Comput. Biol. Med. 27 (2007) 1511–1521.
[124] D. Lin, Dependency-based evaluation of MINIPAR, in: Proceedings of the Workshop on Evaluation of Parsing Systems, 1998.
[125] H. Yu, M. Lee, Accessing bioscience images from abstract sentences, Bioinformatics 22 (2006) e547–e556.
[126] H. Yu, Towards answering biological questions with experimental evidence: automatically identifying text that summarize image content in full-text articles, in: Proceedings of the AMIA 2006 Symposium, AMIA, 2006, pp. 834–838.
[127] SummaryPlus, http://www.info.sciencedirect.com/using/display/summaryplus/.
[128] K. Takahashi, A. Koike, T. Takagi, Question answering system in biomedical domain, Proceedings of the Genome Informatics 2004 (GIW 2004) (2004) 161–162.
[129] R.T.K. Lin, J.L.-T. Chiu, H.-J. Dai, M.-Y. Day, R.T.-H. Tsai, W.L. Hsu, Biological question answering with syntactic and semantic feature matching and an improved mean reciprocal ranking measurement, in: Proceedings of the 2008 IEEE International Conference on Information Reuse and Integration (IEEE IRI 2008), Las Vegas, NV, USA, 2008, pp. 184–189.
[130] Z. Shi, G. Melli, Y. Wang, Y. Liu, B. Gu, M. Kashani, A. Sarkar, F. Popowich, Question answering summarization of multiple biomedical documents, in: Proceedings of the 20th Canadian Conference on Artificial Intelligence (CanAI'07), 2007.
[131] A. Koike, T. Takagi, Gene/protein/family name recognition in biomedical literature, in: Proceeding of the HLT-NAACL 2004 Workshop on Linking Biological Literature, Ontologies and Databases: Tools for Users (BioLink 2004), 2004, pp. 9–16.
[132] B. Boguraev, T. Briscoe, J. Carroll, D. Carter, C. Grover, The derivation of a grammatically indexed lexicon from the Longman dictionary of contemporary English, in: Proceedings of the 25th Conf. Assoc. Comp. Ling. (ACL 1987), 1987, pp. 193–200.
[133] Y. Tsuruoka, Bidirectional inference with the easiest-first strategy for tagging sequence data, in: Proceeding of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP 2005), 2005, pp. 467–474.
[134] G. Melli, Y. Wang, Y. Liu, M. Kashani, Z. Shi, B. Gu, A. Sarkar, F. Popowich, Description of Squash, the SFU question answering summary handler for the DUC-2005 summarization task, in: Proceedings of the 5th Document Understanding Conference (DUC 2005), 2005, pp. 103–110.
[135] J. Kontos, J. Lekakis, I. Malagardi, J. Peros, Grammars for question answering systems based on intelligent text mining in biomedicine, in: Proceedings of the 7th Hellenic European Conf. Computer Mathematics and Its Applications (HERCMA 2005), Athens, Greece, 2005.
[136] J. Kontos, I. Malagardi, J. Peros, Question answering and rhetoric analysis of biomedical texts in the AROMA system, in: Proceedings of the 7th Hellenic European Conf. Computer Mathematics and Its Applications (HERCMA 2005), Athens, Greece, 2005.
[137] J. Kontos, ARISTA: knowledge engineering with scientific texts, Inform. Softw. Technol. 34 (1992) 611–616.
[138] S. Finch, A. Mikheev, A workbench for finding structures in texts, in: Proceedings of the Fifth Conference on Applied Natural Language Processing, 1997.
[139] G. Schneider, Extracting and using trace-free functional dependencies from the Penn Treebank to reduce parsing complexity, in: Proceedings of the 2nd Workshop Treebanks and Linguistic Theories (TLT 2003), 2003, pp. 14–15.
[140] S.J. Athenikos, H. Han, A.D. Brooks, Semantic analysis and classification of medical questions for a logic-based

medical question-answering system, in: Proceedings of the International Workshop on Biomedical and Health Informatics (BHI 2008) in conjunction with 2008 IEEE Conference on Bioinformatics and Biomedicine (IEEE BIBM 2008), Philadelphia, PA, USA, 2008, pp. 111–112.

[141] S.J. Athenikos, H. Han, A.D. Brooks, A framework of logic-based question-answering system for the medical domain (LOQAS-Med), in: Proceedings of the 24th Annual ACM Symposium on Applied Computing (ACM SAC'09), Honolulu, Hawaii, USA, 2009, pp. 847–851.

[142] F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi, P.F. Patel-Schneider, The Description Logic Handbook: Theory, Implementation and Applications, Cambridge University Press, West Nyack, NY, 2003.