BioASQ: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering

George Tsatsaronis
Michael Schroeder
Biotechnology Center
Technische Universität Dresden
Dresden, Germany

Georgios Paliouras*
Yannis Almirantis
NCSR "Demokritos"
Athens, Greece

Ion Androutsopoulos
Department of Informatics
Athens University of Economics and Business
Athens, Greece

Eric Gaussier
Université Joseph Fourier
Grenoble, France

Patrick Gallinari
Thierry Artieres
Université Pierre et Marie Curie
Paris, France

Michael R. Alvers
Matthias Zschunke
Transinsight GmbH
Dresden, Germany

Axel-Cyrille Ngonga Ngomo
Universität Leipzig
Leipzig, Germany

Abstract

This article provides an overview of BIOASQ, a new competition on biomedical semantic indexing and question answering (QA). BIOASQ aims to push towards systems that will allow biomedical workers to express their information needs in natural language and that will return concise and user-understandable answers by combining information from multiple sources of different kinds, including biomedical articles, databases, and ontologies. BIOASQ encourages participants to adopt semantic indexing as a means to combine multiple information sources and to facilitate the matching of questions to answers. It also adopts a broad semantic indexing and QA architecture that subsumes current relevant approaches, even though no current system instantiates all of its components. Hence, the architecture can also be seen as our view of how relevant work from fields such as information retrieval, hierarchical classification, question answering, ontologies, and linked data can be combined, extended, and applied to biomedical question answering. BIOASQ will develop publicly available benchmarks and it will adopt and possibly refine existing evaluation measures. The evaluation infrastructure of the competition will remain publicly available beyond the end of BIOASQ.

Introduction

Question answering (QA) is one of the oldest research areas of AI and Computational Linguistics (Woods, Kaplan, and Webber 1972; Hendrix et al. 1978; Warren and Pereira 1982). QA systems for databases were a particularly popular research topic in the 1980s and early 1990s (Copestake and Jones 1990; Androutsopoulos, Ritchie, and Thanisch 1995). From the mid 1990s, the focus of QA research has shifted to systems that attempt to find answers in document collections or the entire Web, with the TREC QA track providing a significant thrust towards that direction (Voorhees 2001). Unlike information retrieval systems and Web search engines, which typically return lists of relevant documents, QA systems for document collections aim to return exact answers (e.g., names, dates) or snippets (e.g., sentences) that contain the answers sought, typically by applying further processing to the user's question and the relevant documents that an information retrieval engine has returned. In recent years, QA research for document collections has been moving away from simple factoid questions (e.g., questions asking for a name) and towards more difficult questions (e.g., definition questions, like "What is thalassemia?"). This brings us to query-focused summarisation (Jurafsky and Martin 2009), where the input is a question or topic description along with a cluster of relevant documents, and the desired output is a *set* of snippets jointly providing the requested information.

Most QA systems for document collections are intended to be *open domain*. There is a recent trend, however, to develop *closed domain* QA systems in domains with large-scale specialised resources (e.g., subject taxonomies, ontologies), and the biomedical domain is a prime target (Zweigenbaum 2003; Molla and Vicedo 2007).[1] The advent of the Semantic Web (Berners-Lee, Hendler, and Lassila 2001; Shadbolt, Berners-Lee, and Hall 2006) has also revived interest in QA systems for structured data, such as RDF graphs and OWL ontologies, mostly because end-users find it very difficult to understand the formal semantic underpinnings of the Semantic Web (Lopez et al. 2007; Kaufmann and Bernstein 2010; Ferrandez et al. 2011).

Although research on biomedical QA has boomed in recent years (Athenikos and Han 2010; Cairns et al. 2011; Cao et al. 2011), current systems focus on particular resources; for example, MedQA (Lee et al. 2006) uses MeSH and the Gene Ontology only. By contrast, biomedical knowledge workers need to synthesise and filter information from multiple, extremely large and fast-growing sources. Existing search engines (e.g., PUBMED[2], GOPUBMED[3], EBIMED[4]) only partially address this need. They focus on a limited range of resources (e.g., only PUBMED articles and concepts from the GENE ONTOLOGY or MESH), whereas multiple

*Corresponding author. E-mail: paliourg@iit.demokritos.gr

[1] IBM also recently announced that biomedical QA is a prime target for its Watson technology; see http://www-03.ibm.com/press/us/en/pressrelease/36989.wss.

[2] http://www.ncbi.nlm.nih.gov/pubmed/

[3] http://www.gopubmed.com/web/gopubmed/

[4] http://www.ebi.ac.uk/Rebholz-srv/ebimed/

sources (e.g., specialised drug databases and ontologies) often need to be combined. Semantic indexing, i.e., annotating resources with concepts from established semantic taxonomies or, more generally, ontologies, provides a means to combine multiple sources and facilitates matching questions to answers. Current semantic indexing, however, is largely performed manually, and needs to be automated to cope with the vast amount of new information that becomes available daily. At the same time, both current semantic indexing and QA methods require a significant push to reach a level of quality and efficiency acceptable by biomedical experts. BIOASQ[5], a new competition funded by the European Commission, is intended to push towards that direction: integrating efficient and effective semantic indexing and QA methods for the biomedical domain, and establishing an evaluation framework and benchmarks for biomedical QA.

Motivating Example

To illustrate the challenges that a modern biomedical QA system faces, we present below a case study, which is part of a larger scenario from the PONTE project[6]. The larger scenario, called THIRST, pertains to the design of a Clinical Trial Protocol (CTP) regarding the safety and feasibility of synthetic thyroid (TH) replacement therapy with a triiodothyronine analogue (Liotir) in patients with ST-Elevation Myocardial Infarction (STEMI), both in the acute (in-hospital period) and chronic phase (after hospital discharge) of coronary artery disease and its association with cardiac function and outcome. In addition, THIRST examines the effects of TH replacement therapy on the clinical outcome in terms of major (cardiac and non cardiac death, reinfarction) and minor (recurrence of angina, coronary revascularization, and hospital re-admission) events. During the THIRST scenario, the Principal Investigator (PI) of the CTP design formulates a "hypothesis", in effect a target to be proven, based on which a new clinical trial can potentially start. The target is *"Evaluate the safety and the effects of TH treatment in patients with acute myocardial infarction"*. The target requires concrete answers to several questions; we show below two of the questions (Q1, Q2). The questions are produced by the PI and his/her colleagues, and in effect capture their information needs using natural language, their preferred interaction medium.

Q1 What is the role of thyroid hormones administration in the treatment of heart failure?

Q2 What is the relation of thyroid hormones levels in myocardial infarction?

Unfortunately, the questions cannot be submitted directly to current bibliographic databases (e.g., PUBMED). To retrieve the scientific articles that contain the answers to the questions, the PI and his/her team have to translate the questions to collections of terms, in effect concepts from the taxonomy used by the curators of the bibliographic database (e.g., MESH headings in the case of PUBMED). The terms (concepts) that correspond to Q1 and Q2 are shown below as T1 and T2, respectively. Note that this translation process is not trivial, as the original questions may, for example, contain synonymous terms of the taxonomy's concepts. Furthermore, additional terms (concepts) may have to be added, e.g., topically related terms, hypernyms etc., to increase the recall of document retrieval (find more relevant documents) and its precision (i.e., avoid documents corresponding to different concepts, e.g., due to sense ambiguity).

T1 heart failure thyroid hormone treatment

T2 myocardial infarction thyroid hormone

T1 and T2 are submitted to a document retrieval engine as queries, and the engine returns a (possibly long) list of documents. Finally, the PI and his/her team have to study these documents to find snippets that contain information answering their questions. As an example of the advantages, but also of the limitations that current state-of-the-art biomedical semantic search engines offer, Figure 1 shows a screenshot from GOPUBMED, where the terms of T1 are used as MeSH filters. The user still gets 217 documents, which he/she has to read to manually extract the answer to Q1. The major advantage of such systems is, of course, their ability to filter the 21 millions of MEDLINE documents using the specified concepts, reducing the search space of the human reader to just a few hundreds, yet the engine cannot directly produce answers to questions like Q1 and Q2.

The BioASQ Challenge

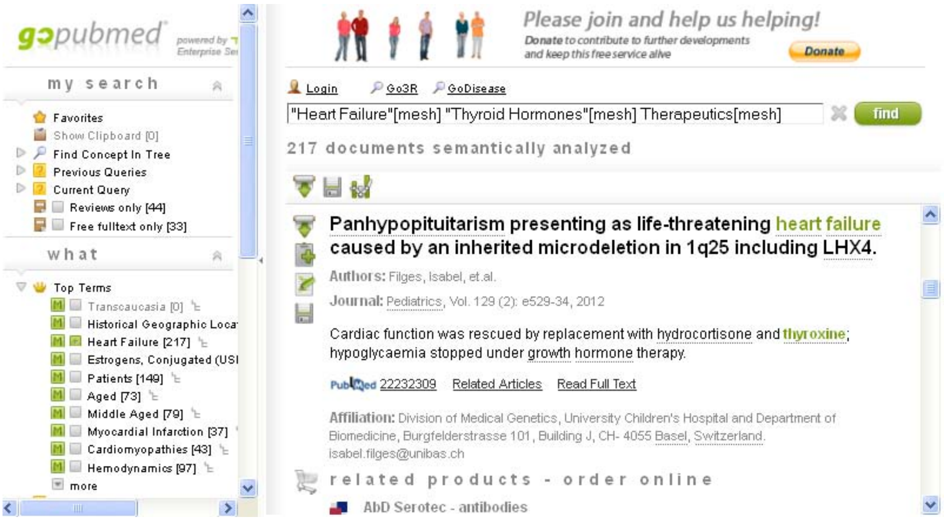
BIOASQ aims to push towards solutions to the problems illustrated in the previous scenario. It will set up a challenge on biomedical semantic indexing and QA, which will require the participants to semantically index content from large-scale biomedical sources (e.g., MEDLINE) and to assemble data from multiple heterogeneous sources (e.g., scientific articles, ontologies, databases) in order to compose informative answers to biomedical natural language questions. More precisely, the BIOASQ challenge will evaluate the ability of systems to perform:

- large-scale classification of biomedical documents onto ontology concepts, to automate semantic indexing,
- classification of biomedical questions on the same concepts,
- integration of relevant document snippets, database records, and information (possibly inferred) from knowledge bases, and
- delivery of the retrieved information in a concise and user-understandable form.

Benchmarks containing development and evaluation questions and gold standard (reference) answers will be developed during the project. The gold standard answers will be produced by a team of biomedical experts from research teams around Europe. Established methodologies from QA, summarisation, and classification will be followed to produce the benchmarks and evaluate the participating systems.

[5] http://www.bioasq.org/

[6] http://www.ponte-project.eu/



"Heart Failure"[mesh] "Thyroid Hormones"[mesh] Therapeutics[mesh]

217 documents semantically analyzed

Panhypopituitarism presenting as life-threatening heart failure caused by an inherited microdeletion in 1q25 including LHX4.

Authors: Filges, Isabel, et.al.

Journal: Pediatrics, Vol. 129 (2): e529-34, 2012

Cardiac function was rescued by replacement with hydrocortisone and thyroxine; hypoglycaemia stopped under growth hormone therapy.

PubMed 22232309 Related Articles Read Full Text

Affiliation: Division of Medical Genetics, University Children's Hospital and Department of Biomedicine, Burgfelderstrasse 101, Building J, CH- 4055 Basel, Switzerland. isabel.filges@unibas.ch

related products - order online

AbD Serotec - antibodies

Figure 1: Using GoPubMed to find the answer to question Q1.

BIOASQ Methodology and Infrastructure

The biomedical bibliographic database PUBMED alone currently comprises approximately 21 million references and was growing at a rate often exceeding $20,000$ articles *per week* in 2011. The number and size of non-textual biomedical data sources also increase rapidly. For example, the MESH[7] thesaurus, which is used to index relevant articles, grew by 454 new descriptors (subject headings) in 2011, an increase of approximately 2%. At the same time, new specialised data sources appear, which are complementary to generic ones (e.g., LinkedCT for clinical trials, and Orphanet for rare diseases[8]) or combine several specialized resources (e.g., BioPortal[9]). Many of these data, along with related biomedical ontologies, are increasingly available on the Linked Open Data (LOD) network,[10] making biomedicine one of the best represented areas of LOD.

Producing sufficient and concise answers from this wealth of information is a challenging task for traditional search engines, which largely rely on term (keyword) indexing. Obtaining the required information is made even more difficult by non-standard terminology and the ambiguity of the technical terms involved. Therefore, indexing at the semantic (concept) level, rather than at the level of keywords only, is particularly important. Biomedical concept taxonomies or, more generally, ontologies are abundant and they provide concept inventories that can be used in semantic indices. Hierarchical classification algorithms (Silla and Freitas 2011) can classify documents and questions onto the concepts of these inventories, facilitating the matching of questions, documents, and also structured data (e.g., RDF triples) that already have explicit semantics based on the same concepts.

Figure 2 provides an overview of the biomedical semantic indexing and QA architecture adopted by BioASQ. To the best of our knowledge, this architecture subsumes all the existing relevant approaches, but no single existing biomedical search system currently instantiates all the components of the architecture. Hence, the architecture can be seen as a broader description of the future systems that BioASQ hopes to push towards. Starting with a variety of data sources (lower right corner of the figure), semantic indexing and integration brings the data into a form that can be used to respond effectively to domain-specific questions. A semantic QA system associates ontology concepts with each question and uses the semantic index of the data to retrieve the relevant pieces of information. The retrieved information is then turned into a concise user-understandable form, which may be, for example, a ranked list of candidate answers (e.g., in factoid questions, like *"What are the physiological manifestations of disorder Y?"*) or a collection of text snippets, ideally forming a coherent summary (e.g., in *"What is known about the metabolism of drug Z?"*).

BIOASQ Tasks

BIOASQ will run for 24 months, from October 2012. Within this period it will organise two challenges (which can also be thought of as two editions of a single challenge) on large-scale biomedical semantic indexing and QA. The first challenge will comprise two tasks:

Task 1a: Large-scale biomedical semantic indexing

This task will be based on the standard process followed by PUBMED curators, who manually index biomedical articles. The participants will be asked to classify new abstracts, written in English, as they become available online, before PUBMED curators annotate (in effect, classify) them manually; at any point in time there is usually a backlog of approximately $10,000$ non-annotated abstracts. The classes (concepts) will come from MeSH; they will be the subject headings currently used to manually index the abstracts. As

[7] http://www.ncbi.nlm.nih.gov/mesh/

[8] See http://linkedct.org/, and http://www.orpha.net/.

[9] http://bioportal.bioontology.org/

[10] See http://linkeddata.org/ and http://www.w3.org/wiki/HCLSIG/LODD/Data.

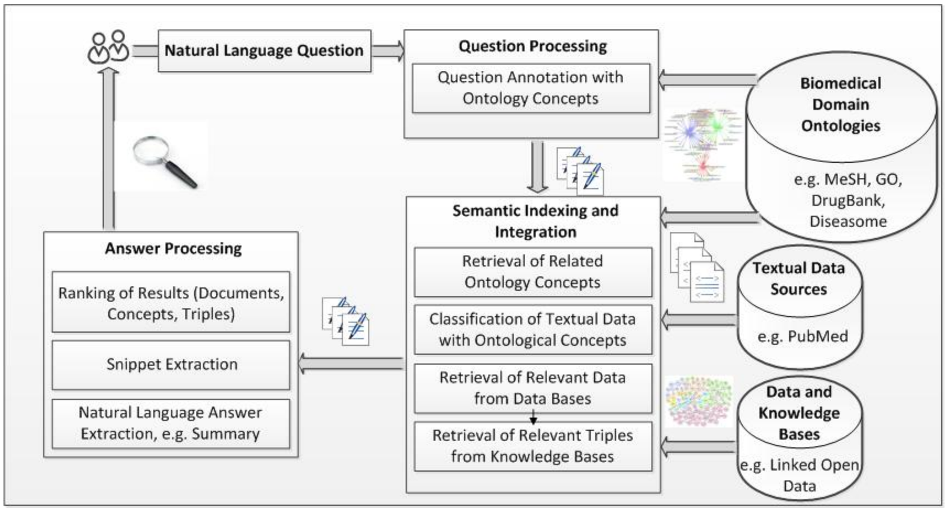

Natural Language Question
Question Processing
Question Annotation with Ontology Concepts
Biomedical Domain Ontologies
e.g. MeSH, GO, DrugBank, Diseasome
Semantic Indexing and Integration
Retrieval of Related Ontology Concepts
Classification of Textual Data with Ontological Concepts
Retrieval of Relevant Data from Data Bases
Retrieval of Relevant Triples from Knowledge Bases
Textual Data Sources
e.g. PubMed
Data and Knowledge Bases
e.g. Linked Open Data
Answer Processing
Ranking of Results (Documents, Concepts, Triples)
Snippet Extraction
Natural Language Answer Extraction, e.g. Summary

Figure 2: Overview of semantic indexing and question answering in the biomedical domain.

new manual annotations become available, they will be used to evaluate the classification performance of participating systems (which will classify articles before they are manually annotated) using standard IR measures (e.g., precision, recall, accuracy), as well as hierarchical variants of them (Brucker, Benites, and Sapozhnikova 2011). The participants will be able to train their classifiers, using the whole history of manually annotated abstracts.

Task 1b: Introductory biomedical semantic QA

This task aims to be an introductory step towards biomedical semantic QA for state-of-the-art generic IR and QA systems. It will be based on benchmarks created specifically for BIOASQ with the help of biomedical experts. The task will take place in two phases:

Annotate questions, retrieve relevant snippets and triples. In the first phase of Task 1b, the participants will be provided with questions written in English and will be asked to (i) semantically annotate the questions with concepts from a particular set of ontologies (from the LOD cloud), and (ii) retrieve data (text snippets from PUBMED articles written in English, knowledge base triples, etc.) that are relevant to the questions (possibly as revealed by the semantic annotations of Task 1a and the semantic annotations of the questions) from designated sources. The system responses will be compared against gold responses provided by the human experts, using standard IR measures.

Find and report answers. In the second phase of Task 1b, the questions and gold responses of the first phase will be provided as input and the participants will be asked to report answers found in the input snippets and triples. In effect, this phase assumes a perfect first-phase system, which is available to obtain relevant snippets and triples. The competing systems will be required to output ranked lists of candidate answers (e.g., names or numbers) in the case of factoid questions, or sets of text snippets and/or triples in the case of questions that ask for summaries. The answers of the systems will be compared against gold answers constructed by biomedical experts, using evaluation measures from QA and summarisation, such as mean reciprocal rank (Voorhees 2001), ROUGE (Lin 2004), Basic Elements (Tratz and Hovy 2008), and other automatic summary evaluation measures (Giannakopoulos et al. 2009). Systems that opt to provide partial responses (e.g., report only triples and no snippets) will be evaluated partially.

Task 2a: Large-scale biomedical semantic indexing

This task will be the same as Task 1a, improved according to the feedback that will have been collected.

Task 2b: Biomedical semantic QA

This task aims to combine the two phases of Task 1b. The participants will be provided with a fresh set of questions and will be asked to (i) annotate them with concepts and retrieve relevant data (snippets and triples) from designated sources, as in the first phase of Task 1b; and (ii) find and report answers, as in the second phase of Task 1b, but now without assuming that a perfect first-phase system is available to obtain relevant snippets and triples, i.e., no gold responses for (i) will be provided. Again, systems that opt to provide partial responses will be evaluated partially.

Related Previous Competitions

Since the late 1990s, QA research has benefited significantly from competitions organised in the context of large conferences, such as the Text Retrieval Conference (TREC) (Voorhees and Harman 2005).[11] TREC's QA track (Voorhees 2001) initially focused mostly on factoid questions. More recent research, however, has also explored questions that ideally require a summary of the most important information from a set of relevant documents, gradually bringing QA for document collections closer to text summarisation. This trend is also evident in the Text Analysis Conference (TAC), which has included challenge tasks such

[11] http://trec.nist.gov/data/qamain.html

Track	Description	Last ran in
Medical Records	Promoting research on content-based access to free-text fields of electronic medical records.	TREC2011
Chemical IR	Promoting chemical IR. The datasets comprise over $100,000$ full-text chemical patents and $45,000$ research papers (Lupu et al. 2009).	TREC2011
Entity	Promoting entity-related search (e.g., finding entities and their properties) on Web data, a kind of search that is not well supported by typical document search.	TREC2011
Web	Promoting Web retrieval technologies. It uses a billion-page dataset.	TREC2011
Genomics	Retrieval tasks for genomics data, including gene sequences and supporting documentation, such as research papers and lab reports (Hersh and Voorhees 2009).	TREC2007
QA	Open-domain QA for document collections (Voorhees 2001).	TREC2007
Terabyte	Scaling retrieval methods for larger datasets.	TREC2006
HARD	High accuracy retrieval of documents.	TREC2005
Novelty	Finding new, non-redundant information.	TREC2004
Interactive	User interaction with text retrieval systems.	TREC2003
Summarisation	Currently multi-document topic-based summarisation of newswire articles, with predefined categories of topics; and tasks for automated evaluation of summaries.	TAC2011
Knowledge population	Discovering information about named entities in corpora, and adding the discovered information into knowledge bases.	TAC2011
Recognising Textual entailment	Recognising if a sentence from a document entails a given hypothesis, or if a document entails a knowledge-base relation.	TAC2011
QA track	QA for document collections, with a focus on opinion questions.	TAC2008

Table 1: Related TREC and TAC tracks that ran in the past.

as query-focused (or topic-based) summarisation.[12] Table 1 lists some of the most relevant current and previous TREC and TAC tracks.

As already noted, the semantic indexing task of BioASQ will ask for documents and questions to be annotated with concepts from biomedical hierarchies. The following hierarchical classification challenges are, hence, also relevant:

- *The Large Scale Hierarchical Text Classification challenges* (LSHTC) (Kosmopoulos et al. 2010), which were organised by members of the BioASQ consortium, provided benchmarks (based on Wikipedia and the *DMOZ* Open Directory Project), as well as a common evaluation framework for hierarchical classifiers.[13]

- *JRS data mining competition*: A competition on topical classification of biomedical articles, based on MeSH concepts automatically assigned to articles by the organisers.[14]

Finally, the Special Interest Group on Biomedical Natural Language Processing (SIGBIOMED) of the Association for Computational Linguistics (ACL) organises the BioNLP annual workshops, which focus mostly on information extraction from biomedical documents (e.g., recognising named entities, particular relations between named entities, or biological events mentioned in scientific articles). The 2009 and 2011 BioNLP workshops featured shared tasks in these areas.[15] Similar biomedical information extraction and data mining challenges are organised in the context of BioCreative, with a recent additional emphasis on helping curators of biomedical document collections (e.g., to prioritise articles to be manually curated).[16]

Overall, although there have been several competitions relevant to BioASQ, which required searching large-scale document collections or structured datasets, question answering or text summarisation, and hierarchical classification or semantic indexing, very few of them were directly concerned with biomedical data. Furthermore, to the best of our knowledge, none of the previous competitions required participants to answer biomedical natural language questions by searching in both structured data (e.g., databases, ontologies, and the LOD cloud) and unstructured data (e.g., biomedical articles), and none of them pushed at the same time towards matching questions to answers at the conceptual level. Several research organizations and companies that are active in areas directly relevant to BioASQ have already stated their intent to participate in the BioASQ challenges.

Summary

We have provided an overview of BioASQ, a new challenge on biomedical semantic indexing and QA. We have highlighted the problems that biomedical workers face when attempting to obtain information from multiple, extremely large and fast-growing sources. BioASQ aims to push towards systems that will allow biomedical workers to express their information needs in natural language and that will return concise and user-understandable answers by combining information from multiple sources of different kinds, includ-

[12] http://www.nist.gov/tac/

[13] http://lshtc.iit.demokritos.gr/

[14] http://tunedit.org/challenge/JRS12Contest

[15] http://www.bionlp-st.org/

[16] http://www.biocreative.org/

ing biomedical articles, databases, and ontologies. We have also highlighted a broad semantic indexing and QA architecture that BioASQ adopts. The architecture subsumes current relevant approaches, even though no current system instantiates all of its components. Hence, the architecture can also be seen as our view of how work in the fields of information retrieval, hierarchical classification, question answering, ontologies, and linked data can be combined, extended, and applied to biomedical question answering. BioASQ encourages participants to adopt semantic indexing as a means of combining multiple information sources and facilitating the matching of questions to answers. Participants, however, may choose to take part in all or only a subset of BioASQ's challenges, in effect instantiating all or only particular components of BioASQ's architecture. BioASQ will develop publicly available benchmarks and it will adopt and possibly refine existing evaluation measures. The evaluation infrastructure of the competition, including web services that organizations will be able to use to evaluate their systems, will remain publicly available beyond the end of BioASQ.

References

Androutsopoulos, I.; Ritchie, G. D.; and Thanisch, P. 1995. Natural language interfaces to databases – an introduction. *Natural Language Engineering* 29–81.

Athenikos, S. J., and Han, H. 2010. Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine* 1–24.

Berners-Lee, T.; Hendler, J.; and Lassila, O. 2001. The Semantic Web. *Scientific American* 34–43.

Brucker, F.; Benites, F.; and Sapozhnikova, E. 2011. An empirical comparison of flat and hierarchical performance measures for multi-label classification with hierarchy extraction. In *Proceedings of the 15th International Conference on Knowledge-based and Intelligent Information and Engineering systems*, 579–589.

Cairns, B.; Nielsen, R.; Masanz, J.; Martin, J.; Palmer, M.; Ward, W.; and Savova, G. 2011. The MiPACQ Clinical Question Answering System. In *Proceedings of the AMIA Annnual Symposium*.

Cao, Y.; Liu, F.; Simpson, P.; Antieau, L. D.; Bennett, A.; Cimino, J. J.; Ely, J. W.; and Yu, H. 2011. AskHERMES: An online question answering system for complex clinical questions. *Journal of Biomedical Informatics* 44(2):277–288.

Copestake, A., and Jones, K. S. 1990. Natural language interfaces to databases. *The Knowledge Engineering Review* 225–249.

Ferrandez, O.; Spurk, C.; Kouylekov, M.; Dornescu, I.; Ferrandez, S.; Negri, M.; Izquierdo, R.; Tomas, D.; Orasan, C.; Neumann, G.; Magnini, B.; and Vicedo, J. L. 2011. The QALL-ME framework: A specifiable-domain multilingual question answering architecture. *Web Semantics* 137–145.

Giannakopoulos, G.; Karkaletsis, V.; Vouros, G.; and Stamatopoulos, P. 2009. Summarization System Evaluation Revisited: N-gram Graphs. *ACM Transactions on Speech and Language Processing* 5:1–5:40.

Hendrix, G.; Sacerdoti, E.; Sagalowicz, D.; and Slocum, J. 1978. Developing a natural language interface to complex data. *ACM Transactions on Database Systems* 105–147.

Hersh, W., and Voorhees, E. 2009. TREC genomics special issue overview. *Information Retrieval* 1–15.

Jurafsky, D., and Martin, J. H. 2009. Question answering and summarization. In *Speech and Language Processing*. Pearson, second edition. chapter 23.

Kaufmann, E., and Bernstein, A. 2010. Evaluating the usability of natural language query languages & interfaces to Semantic Web knowledge bases. *Web Semantics* 377–393.

Kosmopoulos, A.; Gaussier, E.; Paliouras, G.; and Aseervaatham, S. 2010. The ECIR 2010 large scale hierarchical classification workshop. *SIGIR Forum* 23–32.

Lee, M.; Cimino, J.; Zhu, H. R.; Sable, C.; Shanker, V.; Ely, J.; and Yu, H. 2006. Beyond information retrieval-medical question answering. In *Proceedings of the AMIA Annual Symposium*, 469–73.

Lin, C. W. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL Workshop "Text Summarization Branches Out"*.

Lopez, V.; Uren, V.; Motta, E.; and Pasin, M. 2007. AquaLog: An ontology-driven question answering system for organizational semantic intranets. *Web Semantics* 72–105.

Lupu, M.; Piroi, F.; Huang, X.; Zhu, J.; and Tait, J. 2009. Overview of the TREC 2009 chemical IR track. In *TREC 2009 Working Notes*.

Molla, D., and Vicedo, J. L. 2007. Question answering in restricted domains: An overview. *Computational Linguistics* 41–61.

Shadbolt, N.; Berners-Lee, T.; and Hall, W. 2006. The Semantic Web revisited. *IEEE Intelligent Systems* 96–101.

Silla, C. N., and Freitas, A. A. 2011. A survey of hierarchical classification across different application domains. *Data Mining Knowledge Discovery* 31–72.

Tratz, S., and Hovy, E. 2008. Summarization evaluation using transformed basic elements. In *Proceedings of the 1st Text Analysis Conference*.

Voorhees, E. M., and Harman, D. K., eds. 2005. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press.

Voorhees, E. M. 2001. The TREC question answering track. *Natural Language Engineering* 361–378.

Warren, D., and Pereira, F. 1982. An efficient easily adaptable system for interpreting natural language queries. *Computational Linguistics* 110–122.

Woods, W. A.; Kaplan, R. M.; and Webber, B. N. 1972. The Lunar Sciences Natural Language Information System: Final Report. BBN Report 2378, Bolt Beranek and Newman Inc., Cambridge, Massachusetts.

Zweigenbaum, P. 2003. Question answering in biomedicine. In *Proceedings of the EACL Workshop on NLP for Question Answering*, 1–4.