Knowledge Representation for Question Answering

Deborah L. McGuinness
Knowledge Systems Laboratory
Stanford University, Stanford, CA
dlm@ksl.stanford.edu

Question answering systems that produce "knowledgeable" answers are highly desirable. As web use and, in particular, web search, continues to grow, there is increasing demand for information that can be accessed in what appears to be a more intelligent manner (e.g., [Hagen, et al, 2000]). There are many fields from which to access techniques that can be leveraged to improve question answering. In this paper we briefly address some ways that the field of knowledge representation has been and may be leveraged to impact applications that retrieve answers.

Question answering for many people has moved beyond simple use of information retrieval techniques of typing in a query and retrieving a ranked ordering of matching documents containing query elements. Useful answers will contain the answer to the question preferably without including a lot of superfluous or unrelated information. These answers may contain information from sources that are current and authoritative, include use of terms in the same sense as the query, include links to relevant related information such as a justification for the answer, provenance information concerning sources used, typical follow-up questions asked, etc. Question answering systems may also provide support for creating better queries – they may identify incoherent questions, and questions that are too general (i.e., retrieve too many answers), questions that are over constrained (i.e., retrieve few or no answers). All of these notions can be supported by knowledge representation.

There are a number of techniques that may be leveraged to improve question answering. One approach attempts to improve the source information from which answers will be retrieved. Some applications add meta-information to the data so that a question answering system may search the meta-information in addition to the content. If a question answering system knows something about a document from which it is retrieving answers, it can appear more intelligent. One simple source of information is provenance – author, input date, source, authoritativeness ranking, etc. This may be contained in markup information in a document or in a centralized or distributed source concerning data sources. Other sources of markup information include tagging portions of a document with a shared domain vocabulary. For example, a researcher's home page could be marked up so that their name is associated with RESEARCHER, their university could be tagged with EMPLOYER and UNIVERSITY, their email address could be tagged with EMAIL-ADDRESS, etc.) Then if someone asks for researchers who work for Stanford, just the portion of the document containing the researcher name could be retrieved. Fairly expressive markup could be done in any of the markup languages today such as W3C's OWL [Dean.et.al, 2002] or DAML+OIL [Connolly,et.al, 2001]. Examples of such work can be seen in the DAML ontology library[1] and the DAML crawler[2]. Simple markup information could be captured in languages such as XML. Moving beyond just markup of free text documents, source information can be accessed from semi-structured or structured form. Among other things, this allows applications accessing the information to understand things such as domain and range for terms (thus facilitating integrity checking), it facilitates connection with reasoners that might use inference to make implicit information explicit, and it also facilitates better information integration. The structured or semi-structured information makes use of some notion of a database schema or a knowledge base of terms and their inter-relationships thereby allowing some form of object manipulation.

The previous uses of knowledge representation techniques touch the data and may be viewed as manipulating the information that is told and derived as source information. Knowledge representation can also be used to manipulate what is asked of a system. Information retrieval approaches have included a notion of query expansion – expanding a query to include more terms than are initially input. Thus a query, such as "car", may be augmented with synonyms such as "auto" or augmented with specializations of the query term, such as "roadster". Background ontologies containing class and subclass relationships may be used for this expansion as seen in efforts such as FindUR [McGuinness, 1998] or TAP's Semantic Search[3]. By using background information to help improve recall and sometimes precision, FindUR, for

[1] http://www.daml.org/ontologies/
[2] http://www.daml.org/crawler/
[3] http://tap.stanford.edu/tap/ss.html

example was able to take an electronic yellow pages application and improve its recall over 300 percent.

Ontologies may benefit queries in ways other than just query expansion; they can also be used to help analyze and refine queries. Queries that have terms from a background ontology may be analyzed to determine their interrelationships (e.g., IMACS [Brachman, et.al, 1993]). For example, a query for "sports cars" can be determined to be more specific than a query for "autos" assuming a background ontology containing the synonyms "auto" and "car" and the "sports car"-"car" subclass relationship. If a query is entered that is too general, such as "auto" on a car web site, subclasses of car may be offered to the user to help refine a query. Similarly a query that is over-constrained such as sports cars that cost less than three thousand dollars could be analyzed and generalizations may be offered such as "cars that cost less than three thousand dollars" or "sports cars that cost less than x" where x is noticeably greater than 3000. Ontologies may be used in a number of other ways as well to help refine queries, select portions of answers or documents to return, help crawl free text to generate semi-structured text, etc. [McGuinness, 2003].

Answers may also be improved with some knowledge representation techniques. Answers may include the ability to return identified information that satisfies a query specification. Document retrieval systems typically do not identify the portion of the document (i.e. the document object(s)) that contains the answer. Today's agent or human user who asks for universities in Santa Clara County for example, does not just want documents including "Stanford University" or "UCSC" (or the university name and its address), they want the identified object representing Stanford University or UCSC, along with the option of accessing object properties such as the address, the county of the object, the source of that information, relationships between the object and other objects such as the university's student population, etc.

Also, answers can be enhanced with optional justifications including information such as why Stanford is retrieved, where the data came from, if inferences were made, how they were deduced, etc. Inference Web [McGuinness-Pinheiro da Silva, 2003], for example, uses a background portable proof specification that allows information sources and reasoners to provide optional justifications for any answers that they return in a portable, machine-understandable, and distributed "proof". Then the inference web browser allows other programs to access those proofs and also allows humans to use a browser to view the justifications. Thus humans or agents can obtain access to information such as author, recency, and deductive process used in order to help decide if the information should be trusted and how it should be used. This can be particularly useful when agents or humans receive conflicting answers and want to know which to trust or if humans are attempting to reuse past answers and they want to examine the assumptions and sources used before relying on the information again.

Answers may also be pruned for presentation if an answer is too long or complicated. Pruning or matching languages such [Baader, et.al, 1999] have been used to ask for answers with a number of constraints needing to be satisfied. It is simultaneously used to filter out pieces of an answer that are considered irrelevant. The KR language allows variables in the query and then bindings for those variables are returned subject to pruning specification information stored in an ontology.

Customized answers may also be constructed based on knowledge of the query and/or the answer. One simple way to enhance answers is to provide optional background information in a consistent manner. The background information may be simple definitional information such as simple ontology descriptions or previously constructed web pages for common term information. Hyper-linking to definitional information can be seen today in products such as Sentius' Richlink[1]. It allows an application to maintain background information associated with words, phrases, or meta-information to help decide which link to use for hyper-linking or additional link presentation. The links can be used with simple controlled vocabularies or more structured ontologies. Links can also be chosen for presentation taking context into account. This approach can be used for background information (optional definitional links) or for related links. For example, if a search is done for a term that is known to be a performing artist, links can be supplied for the preferred page for performing artist as well as considering optional links for concerts in a near future time period and local geographic location if location is known.

Initial work on the approach including related links can be seen with many search engines (e.g., Google[2], AltaVista[3]) when they return matching documents in the main portion of the screen and provide selected (typically sponsored) links on the right with related information such as albums that can be purchased by the performing artist or books about or by the performer. Tap's Activity-based search[4] takes this a step further adding links and content on the right side of the display screen from a knowledge base containing structured information concerning activities that the term participates in. It then chooses right side links from activities in which the term participates instead of choosing links according to what terms have been purchased. For example, activities

[1] http://www.sentius.com/
[2] http://www.google.com/
[3] http://www.altavista.com/
[4] http://tap.stanford.edu/tap/ss.html

associated with performers include concert schedules, albums, posters, and biographies and all may be provided as optional portions for hyper-linking use with returned answers.

The talk accompanying this paper will provide examples of the knowledge-enhanced techniques used in applications such as search, data mining, configuration, ontology environments, and explanation systems. We address ways KR can help in storing, accessing, presenting, justifying, and filtering answers as well as in analyzing and improving queries.

References

[Baader,et.al, 1999] Baader, Borgida, Kuesters, and McGuinness. Matching in Description Logics. In Journal of Logic and Computation – Special Issue on Description Logics. Volume 9, number 3, June 1999.

[Brachman,et.al., 1993] Brachman, Selfridge, Terveen, Altman, Borgida, Halper, Kirk, Lazar, McGuinness, Resnick. ``Integrated Support for Data Archaeology.'' In *International Journal of Intelligent and Cooperative Information Systems*, **2**:2 1993, pages 159--185.

[Connolly,et.al, 2001] Connolly, van Harmelen, Horrocks, McGuinness, Patel-Schneider, Stein, eds. DAML+OIL Reference Description, 2001. http://www.w3.org/TR/
daml+oil-reference.

[Dean,et.al, 2002] Dean, Connolly, van Harmelen, Hendler, Horrocks, McGuinness, Patel-Schneider, Stein. Eds. Web Ontology Language (OWL) Reference Version 1.0, 2002 http://www.w3.org/TR/2002/WD-owl-ref-20021112/

[Hagen,et.al, 2000] Hagen, Manning, and Paul. "Must Search Stink" Forrester Research TechStrategy Briefing, June 2000.

[McGuinness, 1996] McGuinness. "Explaining Reasoning in Description Logics". Ph.D. Thesis, Rutgers University, 1996. Technical Report LCSR-TR-277.

[McGuinness, 1998] McGuinness. ``Ontological Issues for Knowledge-Enhanced Search''. In the *Proceedings of Formal Ontology in Information Systems*, June 1998. Also in Frontiers in Artificial Intelligence and Applications, IOS-Press, Washington, DC, 1998.

[McGuinness, 2003]. McGuinness. ``Ontologies Come of Age''. In Dieter Fensel, Jim Hendler, Henry Lieberman, and Wolfgang Wahlster, editors. Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential. MIT Press, 2002. http://www.ksl.stanford.edu/people/dlm/papers/ontologies-come-of-age-abstract.html .

[McGuinness-Pinheiro da Silva, 2003] McGuinness and Pinheiro da Silva. Inference Web: Portable and Shareable Explanations for Question Answering ". In Proc. AAAI Spring Symposium on New Directions for Question Answering. Stanford, CA. March 2003.