Abstraction and Perspective in Question Answering

Vinay K. Chaudhri
Artificial Intelligence Center, SRI International
Menlo Park, CA 94025, USA

Introduction
Natural language question answering (QA) has served as a good surrogate for the Turing test and has effectively helped drive AI research. If we take the Watson system (Ferrucci et al. 2010) as a reference point for the state-of-the-art in QA systems, we can attribute the following characteristics to this class of systems: (1) a large fraction of questions are w-h questions, i.e., the questions that begin with who, what, where, why, which, when or how; (2) a large fraction of answers are always a single word or a short phrase (For the Jeopardy task, 94.7% of the answers were a title of some Wikipedia page (Chu-Carroll and Fan 2011).); and (3) there is only a single step interaction, which allows each question and its answer to be independent of the previous one.

To advance the current state-of-the-art, we propose a QA task that includes three question formats: describe, compare, and relate. Once a question is answered, a follow up question is permitted that asks the same question but either at a different level of abstraction or from a different perspective. Before further explaining the task, we first review our experience using a standardized test that led us to these three classes of questions.

Experience Using a Standardized Test
During Project Halo (Gunning et al. 2010), we used a standardized Advanced Placement exam as a performance task. As the scientific focus of our work was on deep knowledge representation and reasoning, the questions were initially stated by using a logical query language, and the answers were produced by a deductive reasoning system. The translation into the logical form of the query was done manually by a human, and often, was only approximate in the sense that not all elements of the input question were faithfully represented in the query. In its internal processing, Watson also has only a partial understanding of the query, except that its understanding is constructed automatically and does not require human intervention (Lally et al. 2012). In subsequent phases of the project, the focus shifted to enable knowledge construction by domain experts, which required creating a controlled English interface for posing questions.

A program for answering a multiple choice question needs to only guess a correct answer option. A random guessing strategy could potentially give a score of 20%. Because the team’s scientific focus was on knowledge representation and reasoning, the team did not believe that writing an answer guessing program was appropriate. Therefore, even for a multiple choice question, our work focused on formulating queries for each option, and building the system to answer each option. This approach presented the following issues: (a) using controlled English to state input questions was a departure from the crisp structure of a standardized test (b) the task was too difficult for an explicit representation and reasoning system; and (c) the task did not provide a natural path for transition and excitement outside AI research. We consider each of these in more detail.

In some cases, the formulation in controlled English became too different from the original question statement. In other cases, even when the question formulation was close to the original question statement, and the system correctly answered all the options of a multiple choice question, the work was criticized for not handling the question’s multiple choice format. Thus, using controlled English, as pragmatic as it may seem, was detrimental to defining a crisp evaluation criteria. A formal representation of the query, is however, essential for evaluating a knowledge representation and reasoning system.

The goal of fully understanding the question as well as the target material is too difficult for the current state-of-the-art of explicit knowledge representation. For example, Watson understands only the relevant portions of a query, and never attempts to have a complete logical understanding of the question (Lally et al. 2012). In our work, fully capturing the knowledge content of a textbook was also difficult, because knowledge representation research is not far enough along to capture even the explicit knowledge stated in a text (Chaudhri et al. 2014).

Because we wanted to ensure that our research would address an important practical need, and because textbooks and exams are used in teaching, applying the technology to education seemed to be a natural starting point. We discovered that educators and education researchers have a uniform dislike of AP questions, and more broadly, a dislike for stan-
dardized test-based assessment of students. Educators are concerned with reducing the need to memorize, improving the reading experience, and increasing engagement. Because of this, we were not able to excite the educators about a system that could answer standardized exam questions.

In summary, we found that a standardized test was not a good performance task for evaluating an explicit knowledge representation and reasoning capability that is also useful for education. As a result, we shifted the project focus to educationally useful questions, which we describe next.

Describe, Compare, and Relate

We developed a large inventory of educationally useful questions and analyzed them to identify the computations that the system would need to do to answer them. Three of those computations involve asking the system to describe an entity or a process, or to compare and relate two entities or processes. Many questions, with a variety of ways to state them in English, can be reduced to one of these three forms.

In our proposed task, the system is first asked to describe a thing, or to compare or relate two things. Next, the system is posed one or more follow up questions to refine its answer either at a different level of abstraction or from a different perspective. I illustrate this process with examples.

Biological systems have different levels of abstraction (e.g., atomic, molecular, cellular, organism, ecosystem, etc.) (Reece et al. 2011). They can also be viewed from many different perspectives (e.g., structural, functional, regulatory, evolutionary, chemical, physical, etc.). Thus, the answer to a question will vary depending on which level of abstraction and perspective is adopted. This characteristic gives us a straightforward for devising a QA dialog that requires refining an answer. Let us consider some examples.

(Q1) Describe a mitochondrion
(Q2) Describe the structure of a mitochondrion
(Q3) Describe the chemical reactions that happen in a mitochondrion
(Q4) Compare a mitochondrion and a plastid.
(Q5) Compare the structures of a mitochondrion and a plastid.
(Q6) Compare the chemical reactions in a mitochondrion and plastid.
(Q7) Relate photosynthesis and cellular respiration for a cell.
(Q8) Relate photosynthesis and cellular respiration for a cell.
(Q9) Relate photosynthesis and cellular respiration from the perspective of global warming.

Let us first comment on why these questions are interesting for AI research. A describe question is similar to the task of writing an essay, which is a recognized grand challenge for AI (Reddy 2003). A compare question requires a form of analogical reasoning (Nicholson and Forbus 2002). A relate question requires extensive search through a system's knowledge which is fundamental to AI problem solving.

If we only pose questions such as Q1 or Q4, we are likely to find that the answers have been written by a human, and can be retrieved by using a search engine. But, if we expect the program to refine its answer depending on the level of abstraction or perspective as suggested in the subsequent questions, all such refinements are unlikely to have been anticipated in advance and authored by a human. If we assume only 5000 key concepts, and ten different levels of abstraction, and ten different perspectives, then nearly one billion combinations of questions are possible. For a computer program to do well on all such combinations, it must have much deeper understanding and knowledge of the material than the current generation QA systems.

Additional work is needed to develop this proposal into an objective test for intelligence. We need to (1) provide crisp criteria for the questions that would be part of this test; (2) provide examples of both the answers that would be considered correct, and those that would be considered incorrect; and (3) provide crisp criteria for determining whether an answer should be scored correctly.

Acknowledgment This work is based on Project Halo funded by Vulcan Inc.

References


