A Study of the Knowledge Base Requirements for Passing an Elementary Science Test

Peter Clark, Phil Harrison
Allen Institute for Artificial Intelligence
505 Fifth Avenue South
Seattle, WA 98104
{peterc,philipha}@vulcan.com

Niranjan Balasubramanian
Turing Center
Dept CS & Engineering
University of Washington, Seattle, WA 98195
niranjan@cs.washington.edu

ABSTRACT
Our long-term interest is in machines that contain large amounts of general and scientific knowledge, stored in a "computable" form that supports reasoning and explanation. As a medium-term focus for this, our goal is to have the computer pass a fourth-grade science test, anticipating that much of the required knowledge will need to be acquired semi-automatically. This paper presents the first step towards this goal, namely a blueprint of the knowledge requirements for an early science exam, and a brief description of the resources, methods, and challenges involved in the semi-automatic acquisition of that knowledge. The result of our analysis suggests that as well as fact extraction from text and statistically driven rule extraction, three other styles of automatic knowledge-base construction (AKBC) would be useful: acquiring definitional knowledge, direct "reading" of rules from texts that state them, and, given a particular representational framework (e.g., qualitative reasoning), acquisition of specific instances of those models from text (e..g, specific qualitative models).

Categories and Subject Descriptors:
I.2.7 Natural Language Processing; I2.4 Knowledge Representation Formalisms and Methods

General Terms: Algorithms

Keywords:
Knowledge acquisition; knowledge base construction.

1. INTRODUCTION
There has been substantial advances in knowledge extraction technology in recent years, e.g., ontology learning [2], fact extraction [3,16], and the creation of proposition stores [20]. Our interest is in pulling these various techniques and resources together, and extending them where necessary, in order to perform a specific task, namely passing a fourth grade science test. The test provides a clear and easily measurable performance task to ensure that the acquired knowledge is useful, i.e., to connect the techniques of knowledge extraction with a particular use case. In addition, fourth grade science poses a variety of challenges in simple, commonsense knowledge and reasoning, challenges that any intelligent system would be expected to overcome; thus we are also interested in whether existing AKBC techniques are adequate for this challenge, and if not, where the gaps are.

AKBC'13, October 27-28 2013, San Francisco, CA, USA.

ACM 978-1-4503-2411-3/13/10…$15.00..

This project is only in its preliminary stages, and so this paper presents just the first step of this endeavor, namely an analysis of a particular fourth grade test (the New York Regents Science Test) and its knowledge requirements, and an analysis of how well current AKBC methods can meet those requirements.

2. OVERALL ARCHITECTURE
Before describing the analysis, we briefly describe the system that we are building, currently existing as a partial prototype. As shown in Figure 1, there are three main components: question interpretation, reasoning, and a library of knowledge resources:

- **Question interpretation:** As we will describe, there are a variety of question types, requiring a variety of solution strategies (problem-solving methods, PSMs). The role of the question interpretation component is to identify and instantiate the appropriate PSM to use, using evidence from a combination of NLP techniques (syntactic parsing, matching with question patterns, and keywords). Currently we have a small catalog of 12 PSMs for different question types (e.g., what-is-a-x, find-a-value, how-many, similarity/differences, how/why, etc.).
- **Reasoning:** Rather than a universal reasoning engine, the architecture contains several reasoning modules, each capable of answering certain classes of questions with varying degrees of reliability. During question-answering, the selected PSM calls the appropriate module(s) to answer its various subgoals.

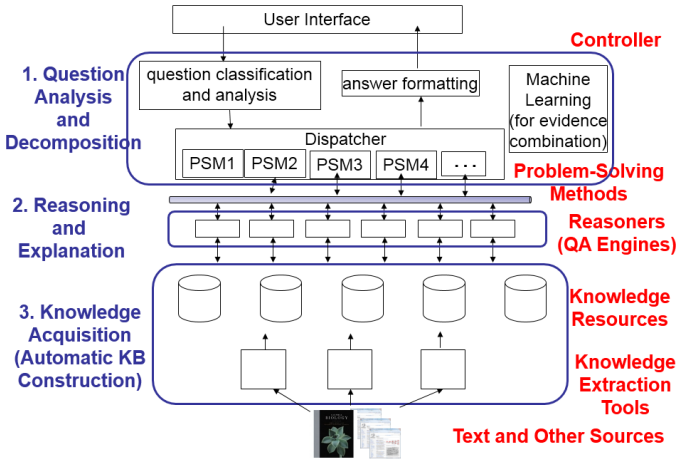

Figure 1: Overall architecture of the system

In cases where more than one module can answer a subgoal, an overall confidence to an answer is computed as a weighted combination of evidence from the different modules, the weights learned using machine learning (simulated annealing) trained with gold-standard question-answer pairs.

- **Knowledge Resources:** The knowledge resources provide the information required by the various reasoning modules, and include both off-the-shelf resources and custom-built resources constructed from various text sources. We currently use: a parse-derived logical form database; an OpenIE triple (proposition) store [6]; The Bio101 ontology [5]; WordNet [7]; the PPDB paraphrase database [4], derived from a large corpus of parallel texts using bilingual pivoting; and the DIRT paraphrase database [15].

This architecture was originally developed to answer simple questions about the narrower domain of cell biology. We are now retargeting it towards the more commonsense-level fourth grade science test, expanding all three of the main components, but with particular emphasis on automatically constructing the required knowledge resources, as we now describe.

3. AKBC for a Fourth Grade Science Exam

Elementary grade science tests are interesting as they test a wide variety of commonsense knowledge and skills that adults largely take for granted, although to an elementary student are challenging. The particular test we are now targeting is the New York Regents' Exam [17] for which a relatively large number of tests (10 years worth) are publically available. The Regents exam is mainly multiple choice, but also includes a direct answer section (that frequently involving diagrams and other graphical components - something we are not currently tackling). For the multiple choice, we are interested in generating answers that are also explainable, rather than just performing "smart guessing" based on statistical word correlations.

Based on an analysis of the 2004, 2005, and 2006 exams (multiple choice parts), it is clear that question-answering in this domain is not a uniform task. Rather, there is a significant variation in the types of questions asked, and the types of knowledge required to answer them, ranging from simple "isa" questions to those requiring more structured models of the world. The questions can be loosely gathered into six groups, based on the types of knowledge and reasoning required to (explainably) answer them. This helps provide a blueprint for the AKBC tasks required to address the challenge. Figure 2 shows a pie chart of these groups and their relative frequency. Although these groups are loosely defined and overlap to some degree, they nevertheless are useful to provide some structure to the challenge. We now describe these categories and discuss how the required knowledge may be acquired. For our purposes here, we focus on the knowledge requirements and AKBC tasks, rather than details of question interpretation and reasoning.

3.1 Taxonomic Knowledge

The simplest questions involve simple application of taxonomic ("isa") knowledge, for example:

2004-2: Sleet, rain, snow, and hail are forms of:
(A) erosion
(B) evaporation
(C) groundwater
(D) precipitation

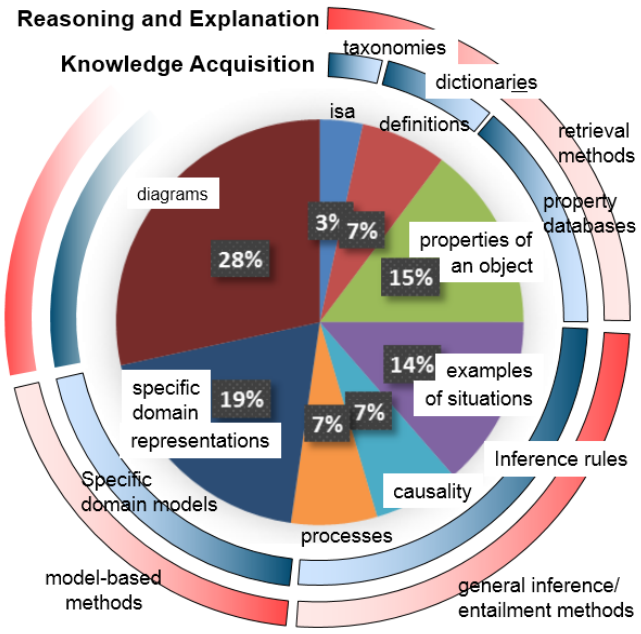

Figure 2: Relative frequency of different question types (pie chart), plus knowledge and reasoning requirements (rings).

Various taxonomies already exist to help answer such questions, e.g., WordNet, YAGO, the Wikipedia topic hierarchy, and "isa" information extractable from dictionaries using sentence patterns.

3.2 Definitions

A second class of questions tests a students' knowledge of terminology, and his/her ability to map definitions to a provided description. For example:

2006-4: The movement of soil by wind or water is called
(A) condensation
(B) evaporation
(C) erosion
(D) friction

Dictionary resources (e.g., Collins, available from LDC; Wiktionary; the WordNet glosses) provide definitions, which can be used for this, reducing these questions to a task of textual entailment, namely determining which definition most plausibly entails the question phrase. (e.g., to what degree does the definition "erosion: the gradual wearing away of land surface material by the action of water or wind" entail the question phrase "The movement of soil by wind or water"?). Although less explored in recent years, there is a wealth of work in extracting semantics from machine-readable dictionaries from earlier years that can be drawn on, e.g., MindNet [18], Extended WordNet [10], and more recently by Allen [1].

3.3 Property Databases

Some questions ask about "basic" properties of an object, e.g., the parts of an object, or the function of an artifact, or the steps in a process. For example:

2004-24: Which part of a plant produces the seeds?
(A) flower
(B) leaves
(C) stem
(D) roots

While there is some subjectivity about which properties are "basic", the intention here is that there are some properties that are repeatedly used in question-answering, and thus it is worth investing extra effort in creating and curating databases of those properties. The property that we are currently targeting is the parts of an object, being constructed using supervised learning, specifically using MultiR [11] run over Wikipedia text, described in [14]. In a separate analysis of RTE problems, LoBue and Yates [15] identify several other categories of basic facts needed for RTE, including geography, public entities, organizational membership, and parts.

3.4 Inference Rules

While the first three question types can be considered as "lookup" questions, a large number of the science questions appeal to simple, general rules and principles about the world, including reasoning about particular situations (scenarios), e.g.,:

2006-9: Which object is the best conductor of electricity?
(A) metal fork
(B) rubber boot
(C) plastic spoon
(D) wooden bat

2005-1: Which example describes an organism taking in nutrients?
(A) A dog burying a bone
(B) A girl eating an apple
(C) An insect crawling on a leaf
(D) A boy planting tomatoes in the garden

causality, e.g.,:

2005-20: What is one way to change water from a liquid to a solid?
(A) decrease the temperature
(B) increase the temperature
(C) decrease the mass
(D) increase the mass

and (simple) processes, e.g.,:

06-21. One way animals usually respond to a sudden drop in temperature is by
(A) sweating
(B) shivering
(C) blinking
(D) salivating

For explainable answers, the first example appeals to the knowledge that metal objects conduct electricity, and the second example appeals to the general knowledge that if an animal eats then it takes in nutrients (along with other taxonomic knowledge, described earlier). Such knowledge might be encoded and reasoned with using a variety of forms and degrees of formality; our architecture allows for multiple approaches, with the question-answering method learning the best weighted combination of answers.

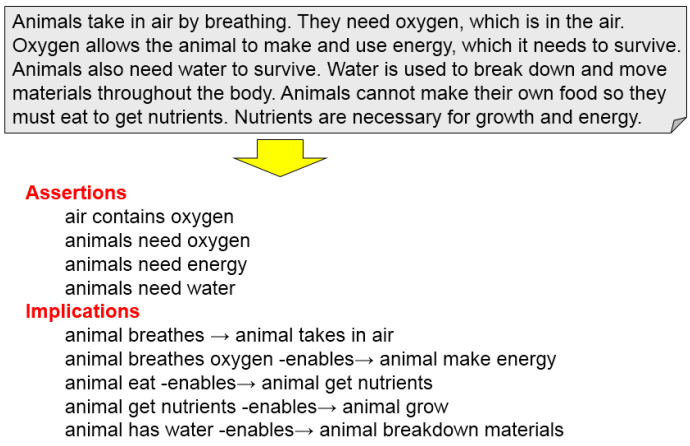

Figure 3: A paragraph from the Barrons Study Guide for the Regents Exam, and some plausibly achievable extractions from the text.

We can loosely characterize these questions as requiring short (eg., 1 step) inferences from the explicit facts, and thus need to acquire rules encoding the inference knowledge. One approach to acquiring such rules is to induce them from examples, e.g., NELL's earlier use of N-FOIL to induce Horn clauses [12], or DIRT's use of distributional similarity [13]. However, another somewhat less explored approach is to directly "read" them (using NLP) from texts that state them explicitly. One genre of texts we have found useful for this are course study guides (rather than textbooks), which tend to summarize just the key knowledge in a rather dry way, without the prose typically used in textbooks to hold a reader's attention - boring for people but ideal for machines. An example paragraph from the Barron's Study Guide (for the Regence exam) along with some plausible extractions is shown in Figure 3. Even though the language is simple, the challenges are still formidable, but at least plausibly addressable from this style of writing.

We are using two forms of acquisition techniques for this knowledge: parse-based triple/proposition stores, using Open IE technology [6], and pattern-based extraction. Proposition stores contain large numbers of generic statements, e.g.,

["metal objects"] ["conduct"] ["electricity"]

extracted from natural language text. In many cases, such propositions express "forall-exists" rules (universal quantification over the first argument). Such rules can in principle be used for reasoning, using alignment/entailment technology to match arguments (e.g., [21]). There are many open research questions still to address, including extracting reliable propositions, filling in unstated context, and identifying the appropriate quantification pattern and degree of reliability of the rules.

In addition, we are using simple pattern-based acquisition techniques to identify and extract rules from text. Some common types of knowledge are expressed using common syntactic forms, and so identifying those forms enables that knowledge to be acquired. For example, in Figure 3, the pattern "X does Y by Z" is often used to mean "IF X does Z THEN X does Y". Similarly, "X does Y to get Z" can mean "IF X does Y THEN X gets Z". This latter pattern allows us to extract that "IF X eats THEN X gets nutrients" from the Barron's text, allowing us to answer the second example question in this subsection.

3.5 Domain Models

Some questions require a specific "modeling paradigm" (i.e., "way of thinking about the world") to answer, beyond just a "sea of rules". One of the roles of teachers is to teach students such modeling techniques. In other words, certain classes of question are answerable by a computation over a certain style of representation - for example, questions about qualitative influence (does X go up when Y does down?) are answerable by a computation over a qualitative model. Although such representational paradigms, and the question-answering algorithms that operate on them, need to be designed and implemented by hand, the domain-specific models themselves may be semi-automatically acquirable from text. For example, although a qualitative reasoning engine may need to be implemented by hand, specific qualitative models themselves may be semi-automatically extractable from statements about qualitative influences in text.

An example of a modeling paradigm from the Regents exam is energy conversion, for questions such as:

> 2005-28: When a baby shakes a rattle, it makes a noise. Which form of energy was changed to sound energy?
> (A) electrical
> (B) light
> (C) mechanical
> (D) heat

The underlying modeling technique here is to (a) identify a sequence of events then (b) tag each event with an energy type (heat, light, etc.). The model can then support answers to certain question classes, e.g., What is the initial form of energy? What is the final form of energy? What form of energy produced X? What form of energy does X turn into? Although such general techniques require hand construction (for the foreseeable future), the specific models that they operate on - here a sequence of events in the question itself - may be plausibly extracted automatically.

A second example of a common modeling paradigm is modeling processes as a sequence of events with various actors playing various roles. Computations over such representations can answer questions such as: What is the role of Entity in Process? What Entity performs Role in Event? During X, what happens after Y? etc.. The AKBC task is to then acquire representations of specific processes. Again, the modeling paradigm is implemented by hand, but the specific models it operates over are to be acquired semi-automatically. We are working with Stanford University on techniques to extract such process models from text using supervised learning [19].

3.6 Diagrams

Finally, a number of questions involve non-textual information (tables, diagrams, graphs, etc.). While there are emerging techniques for diagrammatic representation and reasoning (e.g., [9]), this remains a challenging area for reasoning and explanation.

4. DISCUSSION

One category of knowledge we have not covered, but is also relevant, is more general knowledge about predicates and categories themselves such as: transitivity/"transfers thru" relationships, mutual exclusivity, predicate cardinality, and domain/range constraints on predicates. This type of knowledge is similar to what LoBue and Yates call "form-based knowledge" [15]. It is more global in nature and applies to multiple question types, and while one might view it as containing just additional forms of "inference rules" (Section 3.4), they are structurally different and are unlikely to be stated directly in text. Different acquisition methods will be needed for this kind of knowledge, e.g., hand-coding or inductive rule learning, such as by Galarraga et al. [8].

Why construct knowledge resources such as these ahead of time, rather than just extract the required knowledge on-demand at run-time? In practice, both "pre-caching" of knowledge and run-time extraction are needed. Materializing implicit knowledge allows for its correction and refinement, to reduce errors and noise, e.g., by applying global constraints to the knowledge, by searching for and removing inconsistencies, and by manual filtering of the knowledge ("crowd-correcting"). In the end, this aspect of AKBC is the most important as individual extractions will always be noisy; it is only by refining the aggregated extractions into a more accurate ("purer") form that we can hope to obtain the quality needed for the task. In other words, AKBC is more than just extraction, it also requires an assembly and refinement process. To the extent that questions appear that are not answerable by information in the knowledge resources, additional run-time search and extraction may be needed. Also, note that our architecture (Figure 1) does not assume a single, monolithic knowledge base; rather, it accommodates a variety of resources, with elements that may be in conflict with each other, hence reducing inconsistency is desirable but not an absolute requirement. Rather, an evidential reasoning is required during question answering.

Our survey of the fourth grade exams suggests that there is a wide variety of question types and corresponding targets for AKBC, as we have enumerated. The result of our analysis suggests that as well as fact extraction from text and statistically driven rule extraction, three other styles of AKBC would be useful: acquiring definitional knowledge, direct "reading" of rules from texts that state them explicitly, and, given a particular representational framework (e.g., qualitative reasoning), acquisition of specific instances of those models from text (e..g, specific qualitative models). While there has been work in these areas in the past (e.g., work on processing machine readable dictionaries), further exploration is needed if machines are to have the knowledge to pass a fourth grade science exam.

5. REFERENCES

[1] J. Allen. Acquiring Commonsense Knowledge for a Cognitive Agent. In *AAAI Fall Symposium on Advances in Cognitive Systems*. 2011.

[2] P. Buitellar, P. Cimiano, G. Paliouras, M. Spiliopoulou, (organizers) *Proc. 3rd Workshop on Ontology Learning and Population*. http://olp.dfki.de/olp3/ 2008.

[3] A. Carlson, J. Betteridge, R. C., Wang, E. R. Hruschka Jr., T. Mitchell. Coupled Semi-Supervised Learning for Information Extraction. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, 2010.

[4] C. Chan, C. Callison-Burch, B. Van Durme. Reranking Bilingually Extracted Paraphrases Using Monolingual Distributional Similarity. In *Proceedings of Geometrical Models of Natural Language Semantics (GEMS-2011)*. 2011.

[5] V. Chaudhri, M. Wessel, S. Heymans. *KB Bio 101: A Repository of Graph-Structured Knowledge*. Technical Report, SRI International, CA. 2013.

[6] O. Etzioni, A. Fader, J. Christensen, S. Soderland, Mausam. Open Information Extraction: the Second Generation. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*. 2011.

[7] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

[8] L. Galárraga, C. Teflioudi, K. Hose, F. Suchanek. AMIE: Association Rule Mining under Incomplete Evidence in Ontological Knowledge Bases. In *Proceedings of the 22th international conference on the World Wide Web, WWW '13*, 2013.

[9] A. Goel, H. Jamnik, N. H. Narayanan (editors). *Diagrammatic Representation and Inference: Proceedings of the Sixth International Conference on the Theory and Application of Diagrams*, August 2010, Portland. Oregon. Lecture Notes on Artificial Intelligence #6170, Berlin: Springer. 2010.

[10] S. Harabagiu, G. Miller, D. Moldovan. WordNet 2 - A Morphologically and Semantically Enhanced Resource, *SIGLEX* 1999.

[11] R. Hoffmann., C. Zhang, X. Ling, L. Zettlemoyer, D. Weld. Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations. In *The Annual Meeting of the Association for Computational Linguistics (ACL-11)*. Portland, OR. June 2011.

[12] N. Lao, T. Mitchell, W. Cohen. Random Walk Inference and Learning in A Large Scale Knowledge Base. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP),* 2011.

[13] D. Lin, P. Pantel. *Discovery of Inference Rules for Question Answering*. Natural Language Engineering 7 (4) pp 343-360. 2001.

[14] X. Ling, D. Weld. Extracting Meronyms for a Biology Knowledge Base Using Distant Supervision. *Proc. AKBC'13*. 2013.

[15] P. LoBue, A. Yates. Types of common-sense knowledge needed for recognizing textual entailment. In *Proc 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*. 2011.

[16] J. Mayfield, J. Artiles. (organizers) *Proc. of the Knowledge-Base Population Track of the Text Analysis Conference (TAC'12)*. http://www.nist.gov/tac/2012/KBP/ 2012.

[17] Regents. *The New York Regents Examinations*. http://www.nysedregents.org/elementary.html 2013.

[18] S. Richardson, W. Dolan, L. Vanderwende. MindNet: acquiring and structuring semantic information from text. *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pp1098-1102. 1998.

[19] A. T. Scaria, J. Berant, M. Wang, C. Manning, J. Lewis, B. Harding, P. Clark. Extracting Biological Processes with Global Constraints. In *Proc. EMNLP'13*, 2013.

[20] L. Schubert, M. Tong.. Extracting and evaluating general world knowledge from the Brown corpus, *Proc. of the HLT/NAACL 2003 Workshop on Text Meaning*. 2003.

[21] X. Yao, B. Van Durme, C. Callison-Burch, P. Clark. A Lightweight and High Performance MonolingualWord Aligner. *EMNLP'13*. 2013.

.