The Energy, Environmental, and Economic Apprentice: A Question Answering System

Harry E. Mynick
Princeton Plasma Physics Laboratory
Princeton University
Princeton, New Jersey
email: mynick@pppl.gov

Richard Scherl
Department of Computer and Information Science
New Jersey Institute of Technology
Newark, New Jersey
scherl@cis.njit.edu

Abstract

This paper describes our work in progress on building a Question Answering System. It is a prototype computer "apprentice system" to aid researchers and students in addressing the energy, environmental, and economic (EE&E) issues which arise in studies of integrated energy policy assessment. Much of the task of forming a reasoned opinion about such issues involves finding the relevant numbers for the questions at hand, and then combining them mathematically, to obtain estimates answering the questions at varying levels of detail and sophistication. We believe that both processes, of accessing the basic numbers and of combining them to draw new conclusions, could take less time and effort than now, and could thereby be made appreciably more efficient and productive, through a system which would enable the user to ask questions at a "high level" and have the system find the appropriate numbers and then calculate the answers.

Introduction

The field of environmental and energy policy assessment provides an advantageous target domain for the construction of a Question Answering System. The information needed derives from a wide range of fields, including climate simulation, forestry, agronomy, biochemistry, several areas of engineering, and economics. The relevant data are scattered among disparate sources, often using exotic (and sometimes inconsistent) terminology and units of measure.

Practitioners (whether scientists, policy makers, or students) must spend time gathering the data needed to answer a question before they begin to use whatever analytic techniques are appropriate. It would be useful to have a central, automated, easily accessed information repository of data and analytic resources for such a scientific community. Tools are needed to gather the information, to maintain the data, and to automate analysis of large amounts of data of heterogeneous origin. Additionally, facilities need to be provided to make the repository accessible and usable by the various categories of users. They should be able to ask questions at a high level and have the system find the appropriate numbers and perform the various analytical tasks.

As a paradigm testbed for addressing this challenge, we are developing a prototype computer "apprentice system" to aid scientists, government policy makers, and students in addressing the energy, environmental, and economic (EE&E) issues which arise in making integrated assessments of energy policy. Much of the task involved in developing new insights into such issues involves finding the relevant numbers for questions of interest, and then combining them mathematically to answer the questions at varying levels of detail and sophistication.

E3 would consist of a "knowledge base" (KB), a single, unified, easily accessed, and easily updated reservoir of EE&E data, on top of which would sit the "mill", facilities which would give the user the ability not only to query the raw data in the KB, but also to combine the information in the KB in various ways. These range from simple calculations to automating tasks (such as analyses of the economic and environmental impacts of systems under consideration) which presently must be performed 'by hand.' A major part of the task of constructing the system consists of building both a representation for and facilities to reason about the concepts underlying the EE&E field. This paper describes our current work, in progress, on this topic.

The Apprentice

There is an immense amount of data, in a wide variety of sources, relevant to EE&E issues. Researchers and government officials in that domain must spend a good part of their time searching through these various sources and handling the heterogeneous nature of the data. Only then are they able to attempt to answer the questions that have been posed to them. The goal of our work is to automate a large portion of this process.

EE&E Data

The E3 KB would provide data in four categories:

(a)'macro', or 'extensive' numbers, characterizing the stocks and flows of resources in systems of interest, *e.g.*,

the amount of cropland in the US, or the number of Gigatons of CO2 being pumped into the atmosphere per year.

(b)'micro', or 'intensive' numbers, characterizing technologies or materials out of which systems are built, *e.g.*, the conversion efficiencies of aeroderivative turbines, the energy per unit weight of dry wood, or the cost per installed Watt of a coal-fired plant.

(c)'relational' information, including links between objects in the KB, which indicate the relationships which exist among these objects, and equations giving the Apprentice an understanding of important EE&E concepts that both it and the user can call upon in performing analyses.

(d)'support' information, including help information on the meaning and use of the symbols and functions in the system, notes on the proper interpretation of the numbers in the KB, the source from which they were drawn, and pointers to useful references and authorities on a specified topic.

The first step in our research is to develop an ontology (Gruber 1993), a set of interrelated categories, to organize this material. To make our work compatible with related efforts, we are utilizing HPKB-UL, the Upper Level Ontology being used in Defense Advanced Research Projects Agency's (DARPA) High Performance Knowledge Base (HPKB) project[1]. The HPKB-UL is the upper-level ontology of the Cyc KB (Lenat & Guha 1989) augmented by some links to the Sensus ontology (Knight & Luk 1994). HPKB-UL provides a taxonomy of about 3000 terms and relations for general terms such as tangible-object, action, and transportation. It also defines some relations between them, such as the starting time of an event, the relationship between an object and its parts, and the borders of a geographic region. The X3T2 working group of ANSI has adopted HPKB-UL as the current draft for a standard upper ontology. Many of the terms necessary for our ontology, for example, geographical regions, countries, are already defined in the HPKB-UL. The HPKB-UL, however, does not include many specialized terms necessary for our work, for example, green house gas emissions and air pollution rates. The categories that we define will be linked to the concepts in the HPKB-UL making it easy for us to disseminate our work.

The Knowledge-Base

Most of this information is to be coded in a frame-based language. We will utilize the frame language defined in the Open Knowledge Base Connectivity knowledge model (Chaudhri *et al.* 1998). OKBC is an application programming interface for accessing frame representation systems (FRSs). The OKBC knowledge model defines a frame-based language that is supported by a large class of FRSs. This language will be supplemented with Knowledge Interchange Format (KIF) (Genesereth & Fikes 1992) for any information that cannot be directly captured in the OKBC knowledge model. KIF is a highly expressive language for the interchange of knowledge among disparate programs. We anticipate that as we formalize the domain knowledge for the E3 domain, we will have to make any necessary extensions to the OKBC knowledge model and investigate what combination of the OKBC knowledge model, KIF and domain-specific extension will be most suitable.

The Implementation

We have constructed an embryonic E3 prototype in Mathematica, to sharpen our understanding of where the possibilities and challenges lie in building the system. Mathematica has proven to be a very good environment in which to perform the development thus far. The capabilities of a computer-algebra system (CAS) like Mathematica are useful in a number of respects, in organizing and manipulating the heterogeneous kind of information EE&E work entails. Mathematica is a modern, widely used and well supported CAS, with excellent capabilities for not only symbolic algebra (important for treatment of real problems of a quantitative nature), but also for graphics (plotting data, drawings, color etc.) and for the numerical solution of algebraic and differential equations (which will lend itself well to incorporating facilities for modeling). Mathematica contains a sophisticated programming language that in a unique but systematic fashion incorporates pattern matching facilities (similar to those of Prolog) and facilities allowing ready implementation of object oriented capabilities (similar to those of Smalltalk and C++).

All of these capabilities are important in the construction of the prototype Apprentice being described here. Therefore, we are basing our architecture around the use of Mathematica, but incorporate components written in other languages that are either already available or that we will write during the course of this project. Mathematica has facilities for interacting with outside software, so that it can call other software, and programs can be written that call Mathematica and then utilize the results of a computation performed within Mathematica.

For the construction of the Knowledge-Base we are using the Generic Knowledge-Base Browser and Editor (GKB-Editor) (Karp, Chaudhri, & Paley 1998) developed at SRI. The editor is designed especially to provide a graphical interface to facilitate the development of complex frame-based knowledge bases. Its facilites include ones which graphically represent the structure of the knowledge base as an aid in comprehension, as well as graphical facilities for modifying, deleting, replacing, and editing frames and slots. The GKB-Editor is designed to be portable across multiple frame representation systems (FRS) by utilizing the OKBC protocol which is a set of functions that form a generic application-program interface to the FRS. We will use the OKBC protocol in our implementation.

In this environment, the GKB-Editor is utilized as

[1]See `http://www.teknowledge.com/`

a front end to OCELOT (an OKBC compliant FRS). Finally the ORACLE data base management system is used as the back end. The frame based language is translated into a standard relational data base and various operations on the frame base language are translated into SQL commands. The end result is that all of the data base technology from the ORACLE back end is inherited by the knowledge based system (Karp & Paley 1995).

Some Simple Illustrations

As mentioned earlier, we have developed an embryonic E3 within Mathematica. Below we give a series of illustrations of accessing and combining information in the KB as already implemented. Data is represented as properties on objects. Regions of the world (countries or states such as US, and NJ) have properties such as energy or area and specializations of those properties such as year, or type (e.g. fossil, coal, cropland).

```
In[1]:= energy[US,fossil,coal,1988]
Out[1]= 19.93 EJ,

In[2]:= area[US,cropland,1988]
                6   2
Out[2]= 1.9 10  km

In[4]:=
   energy[NJ+NY,1988]/population[NJ+NY,1988]
                   -7
        2.42483 10   EJ
Out[4]= --------------- .
                cap

In[5]:= energy[US+Canada,fossil+nuclear,1988]
energy[Canada, fossil, 1988] not in KB.
energy[Canada, nuclear, 1988] not in KB.
Out[5]= 81.59 EJ +
        energy[Canada, fossil, 1988] +
        energy[Canada, nuclear, 1988]

In[6]:= ?EJ ; ?cap
An EJ is a unit of energy.
Use SI[] to evaluate.
cap (capita) is the base unit of
     population, meaning one head.

In[8]:= SI[EJ]
               18
Out[8]= 1. 10   Joule

In[9]:= convert[%4,kWh/cap]
         67356.5 kWh
Out[9]= -----------
               cap

In[10]:= SI[boardfoot]
                                3
Out[10]= 0.00235974 meter

In[11]:= SI[200 MW * 1 yr * .05 $US/kWh * .75]
                 7
Out[11]= 6.57 10  $US

In[12]:= ICC[BIG$ISTIG1]
         770 $US
Out[12]= -------
           kW

In[13]:= notes[BIG$ISTIG1]
Out[13]= noteA: Numbers for a biomass-fueled
>       BIG/ISTIG system, with a
>       fixed-bed Lurgi gasifier without chemical
>       hot-gas cleanup, and  a 110
>       MW ISTIG unit.

In[14]:= energyDensity[poplar]
         19.4 MJ
Out[14]= -------
           kg
```

Here, the user's query and E3's response are on lines labeled with the standard Mathematica designation In[i] and Out[i], respectively (jointly referred to as [i] below).

In queries [1,2] are simple requests for 2 macro-numbers. To make matters as simple as possible for the user, E3 is fairly relaxed about the order of arguments, so that, for example, query [2] could also have been posed as area[1988,cropland,US]. In [4] is shown a query somewhat more involved than [1,2] in 2 ways: (a)one can combine the numbers for different regions such as New Jersey and New York by the summation designation in the argument, and (b)the numbers drawn from the KB can be combined in arbitrary algebraic combinations, such as the ratio energy[.]/population[.] shown there.

A more elaborate example of the summation syntax noted in (a) above is given in query [5], meaning 'the sum of the fossil fuel plus nuclear energy used by the sum of the US and Canada in 1988.' Here, E3 finds that the KB knows the amount of fossil+nuclear energy the US consumed in 1988, but for Canada, it presently only knows the total amount of energy consumed. It therefore fails gracefully, giving the messages 'energy[.] not in KB' for the items not known, and providing a symbolic result in the answer-line Out[5]. The given result is therefore still correct, though partly symbolic, and could be used in subsequent manipulations, or further evaluated by hand. This is one virtue of using a CAS like Mathematica over other computer languages or database query languages.

A first kind of support information is requested in [6], the meaning of the symbols EJ and cap which were returned earlier. The ?⟨symbol⟩ notation is part of the help machinery which Mathematica itself provides.

From [6] one learns that an EJ is an energy unit, whose value one can elicit as in [8] by calling the func-

tion SI[], which converts its argument to SI units ('Standard International'/MKS units, extended to units not encountered in physics, such as population, money, etc). In [9] a bit more general function convert[] is called, which uses SI[] to convert the result Out[4] (Mathematica shorthand %4) to kWh/capita.

These conversion functions should be a very useful, albeit simple, facility to have at one's fingertips, while absorbing the heterogeneous literature accessed in EE&E work, and where one regularly has to use and combine numbers in various, sometimes exotic units. Two typical examples are given in [10,11]. Query [11] is a one-line calculation answering the question(Goldston 1997) "What would be the annual revenue of the envisioned ITER prototype fusion reactor if it produces 200 MW electric with a duty factor of 75%, and electricity sells at 5 cents/kWh?"

In [12], a first query for micro-numbers is made, the installed capital cost (ICC) for a particular Biomass-fueled Integrated Gasification/ Intercooled Steam-Injected Gas Turbine (BIG/ISTIG) system (Williams 1989), designated BIG$ISTIG1. In the EE&E literature, a given technology is often characterized by a table or two of its salient features, eg, its power output, conversion efficiencies, the costs of its components, and the busbar cost of electricity (COE) it produces (see for example (Williams 1989)). Such information will be incorporated into the KB with much the same organization as used for the macro-numbers. Similarly for micro-numbers specifying the characteristics of materials, illustrated in [14]. The mature KB would contain the information in such tables for many technologies and materials, and be able to access and make use of them.

Needed Features of the E3 Prototype

Our current work is focused upon extending the features of this embryonic E3 so that it can fully utilize the information encoded in the knowledge-base and also to incorporate a number of features designed to make it into a question answering system suitable for a wide variety of users.

Query Language

For E3, one needs a user/Apprentice communication language which is natural enough that the user can get started quickly, having read perhaps a few introductory pages and seen a few examples of E3's operation. The user will know a modest amount about algebraic notation, but should not be required to have much expertise in either mathematics, computer programming, or Mathematica itself. We believe that the computer-algebraic form of the Mathematica expressions in terms of which E3 currently operates, supplemented by online help-facilities (see below), should permit this envisioned population of users to quickly begin making use of the Apprentice.

However, eventually the initial 'kernel' system will be overlaid with a natural-language interface, to allow questions to be posed in a subset of English.

Help Facilities

E3 will have online system-help capabilities, some simple illustrations of which have just been given.

High Level Queries

Automating Economic Analysis

A common task in EE&E work is assessing the economic characteristics of a given system, once the numbers specifying the system have been given. For example, the COE for BIG$ISTIG1 discussed in Sec. is computed from other information about BIG$ISTIG1, such as its ICC, its assumed capacity factor and lifetime, and from more general economic relations and conventions for assessing the cost of energy technologies, such as described in the EPRI *Technical Assessment Guide* (TAG). The Apprentice KB would include a knowledge of these relations, and the mill portion of the Apprentice would use these to compute such derived quantities, asking the user to make further choices when insufficient information has been given:

```
In[19]:= COE[BIG$ISTIG1]
 E3: Specify discount rate (in percent):
 >> 6.1
 E3: Specify corporate tax rate (in percent):
 >> 38
 E3: Specify biomass[price] (in $US/GJ):
 >> biomass[price]
          $US (0.0179 + 0.00855 biomass[price])
Out[19]= ------------------------------------
                          kWh
```

Note that the user here leaves biomass[price] symbolic, so that the resultant expression can then be submitted to Mathematica's plotting routines, or further manipulated for use in subsequent tasks. In addition to eliminating tedium and potential for error, use of such routines could impose a more uniform framework on economic analyses.

Rather than simply writing a separate algorithmic function for each analytic task desired, the Apprentice could be given a much more general and flexible capability for setting up, analyzing and solving problems of a mathematical nature. In the example just given, the object COE would have a 'givenby' slot, indicating the general economic equation PVC = PVR (Present Value of Costs = Present Value of Revenues) by which the COE is usually computed. E3 would then look at givenby slots of the variables in the expressions for PVC and PVR, and so on, iteratively, generating the system of relations to be solved in order to compute the COE for the system under consideration. Implementing this more general approach would provide an opportunity to study reasoning about and solving mathematical problems, including heuristic solution techniques for solving irregular systems of equations such as problem

decomposition and problem transformation, as well as methods of mathematical approximation (Mynick 1987; 1989). The algebraic capability provided by Mathematica will be of great importance here. The inferential capabilities to be discussed in the following subsection should also enhance the versatility of this economic facility.

Automating Higher–Level Analysis

We are extending the core E3 discussed so far with facilities for solving high-level queries, comparable to those addressed in a typical paper or section of a paper in this area. One example is to construct a system for making windpower a baseload (versus peak-hour) source of electrical power, using technologies presently available or soon to be available. Another is to project an energy mix for the US/World in 2050 which best satisfies the following objectives: minimize energy cost, minimize greenhouse gas emissions, be environmentally benign, and be politically feasible. Processing such a high-level query would result in the execution of a series of basic Mathematica commands with the output of each command appropriately gathered and processed.

Some of the actions that we need to represent are similar to those considered in the planning literature. For example, there may be actions needed to construct a windpower plant in a particular location. On the other hand many of the actions (e.g. applying a particular analytic tool to some data) are quite a bit different than those generally considered in A.I. work on reasoning about actions and planning. They are closest to perceptual actions. For example, as discussed earlier, routines will be available to calculate the cost of electricity given the use of a particular technology under particular conditions.

The approach that we will first consider is to develop a language, based upon GOLOG (Levesque *et al.* 1997), for writing general high-level plans specifying the pieces of knowledge and tests that need to be performed to answer such questions. The agent programming language GOLOG is based on a set of *complex actions* that include programming constructs such as sequences of actions, conditions, and while loops. These high-level plans expand in context into expressions in the situation calculus (with the details filled in at execution time). A particular execution sequence of a complex action expression will be a sequence of situation calculus primitive actions (Reiter 1991) which can then be executed by the Apprentice.

For example, the following simple GOLOG[2] based routine finds a set of regions satisfying certain conditions (location, average wind speed) that total to an area of at least 2000 acres.

```
while (AREA < 2000)
    (Πe) QUERY(ARIZONA_REGION(e));
        DETERMINEP(e);DETERMINEQ(e);
        if HIGH_WIND(e)
          then AREA = COMP_AREA(e) + AREA;
```

In this fashion, many much more complex procedures would be written. They would both call and be called by other procedures.

To model the effects of actions which involve the application of an analytic tool or a query of the E3 KB, we can draw upon the work on modeling the effects of perceptual actions (Moore 1980; Scherl & Levesque 1993). For example, consider a DETERMINEP action, such that after doing a DETERMINEP, the truth value of P is known. Here P may be defined to be equivalent to a relatively involved formula describing a particular climate type. Or it could be defined to be stating that the cost of using a particular energy technology in a particular region is above a certain value. We introduce the notation **Kwhether**(P, s) as an abbreviation for the formula **Knows**(P, s) $\vee$ **Knows**(P, s). The result of executing DETERMINE$_P$ is **Kwhether**(P, s), which means that the truth value of P is known. In the GOLOG example above, knowledge of P, Q and possibly additional background knowledge must enable the program to know the truth value of HIGH_WIND for the area under consideration.

Summary

The development of E3 would enhance scientific productivity and quality within the area of EE&E and could radically alter the way research and teaching is carried out within this field. Additionally, the project provides an excellent domain for exploring techniques needed in Question Answering Systems.

Acknowledgments

We thank Vinay Chaudhri for a numer of helpful discussions related to the topic of this paper.

References

Chaudhri, V. K.; Farquhar, A.; Fikes, R.; Karp, P. D.; and Rice, J. P. 1998. OKBC: A Foundation for Knowledge Base Interoperability. In *Proceedings of the National Conference on Artificial Intelligence.* To appear.

Genesereth, M. R., and Fikes, R. E. 1992. Knowledge Interchange Format, Version 3.0 Reference Manual. Technical Report Logic-92-1, Computer Science Department, Stanford University, Stanford, CA.

Goldston, R. 1997. private communications.

Gruber, T. 1993. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition* 5(2):199–220.

Karp, P., and Paley, S. 1995. Knowledge representation in the large. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence.*

[2]Here Πe represents nondeterministic choice of an object e. Also, the procedures DETERMINEP(e) and DETERMINEQ(e) result in the agent knowing the truth value of P and Q respectively. They may involve the execution of code written in Mathematica.

Karp, P. D.; Chaudhri, V. K.; and Paley, S. M. 1998. A Collaborative Environment for Authoring Large Knowledge Bases. *Journal of Intelligent Information Systems.* To appear.

Knight, K., and Luk, S. 1994. Building a Large-Scale Knowledge Base for Machine Translation. In *Proceedings of the National Conference on Artificial Intelligence.*

Lenat, D. B., and Guha, R. 1989. *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project.* Reading, Massachusetts: Addison-Wesley.

Levesque, H.; Reiter, R.; Lespérance, Y.; Lin, F.; and Scherl, R. B. 1997. Golog: A logic programming language for dynamic domains. *Journal of Logic Programming.*

Moore, R. 1980. Reasoning about knowledge and action. Technical Note 191, SRI International.

Mynick, H. 1987. Alpha prticle effects as a test domain for pap: A plasma apprentice program. *Physica Scripta* T16:133–142.

Mynick, H. 1989. Toward the automated analysis of plasma physics problems. Technical Report PPPL-2605, Princeton University.

Reiter, R. 1991. The frame problem in the situation calculus: A simple solution (sometimes) and a completeness result for goal regression. In Lifschitz, V., ed., *Artificial Intelligence and Mathematical Theory of Computation: Papers in Honor of John McCarthy.* San Diego, CA: Academic Press. 359–380.

Scherl, R. B., and Levesque, H. J. 1993. The frame problem and knowledge producing actions. In *Proceedings, Eleventh National Conference on Artificial Intelligence*, 689–695.

Williams, R. 1989. Biomass energy strategies for coping with the greenhouse problem. Center for Energy and Environmental Studies, Princeton, New Jersey.