Using Deduction to Choreograph Multiple Data Sources

Richard Waldinger and **Peter Jarvis**
Artificial Intelligence Center, SRI International
Menlo Park, California, US
{waldinger,jarvis}@ai.sri.com
http://www.ai.sri.com/

Jennifer Dungan
Ecosystem Science and Technology Branch
NASA Ames Research Center
Moffett Field, California, US
Jennifer.L.Dungan@nasa.gov
http://geo.arc.nasa.gov/~jennifer/dungan.html

Abstract

Automatic theorem proving is employed to coordinate multiple data and knowledge sources. Sources are related to a central axiomatic theory so that their interaction can be inferred. The method is applied to human language question answering in geography and earth science.

The Problem

More and more information is available online, especially on the Web, but it is not always easy to find the appropriate information source or to combine data from different sources. Web browsers employ key words but are tripped up by syntactic coincidences. Web sites have interfaces that are not always transparent to the human user, and do not facilitate access by machine. Different sources adapt different representational schemes and notational conventions.

The Approach

In the system GeoLogica, an automated deduction system operating on an axiomatic *application domain theory* is used to solve these problems. The capabilities of the knowledge sources of interest are specified by axioms, logical sentences in the language of the theory. These axioms will serve as an advertisement for the knowledge sources—they say what questions each source is competent to answer. Other axioms in the theory define the concepts that are used in these specifications, and express relationships among these concepts.

Questions are expressed in a subset of English. The question is parsed and translated into a logical form by Gemini, a broad-coverage English parser. The logical form is phrased as a conjecture in the language of the application domain theory and submitted to the automated deduction system SNARK, a general-purpose first-order-logic theorem prover with special capabilities for temporal (Allen and Ferguson 1994) and spatial (Cohn et al. 1997) reasoning. SNARK attempts to determine if the conjecture follows from the axioms in the application domain theory.

SNARK does not merely attempt to prove the truth of the conjecture; it also has capabilities for *extracting answers* from proofs. If the conjecture posits the existence of an entity that satisfies certain conditions, SNARK will be able to find such an entity from its proof.

Certain symbols in the application domain theory stand for external knowledge sources. When axioms containing those symbols are employed in the search for the proof, the corresponding source is invoked, via a *procedural-attachment* mechanism. This causes SNARK to behave as if some of the knowledge possessed by the source were present in the application domain theory, when it is appropriate to the question at hand. We use the term *agent* for a data or program knowledge source that is invoked by means of this mechanism.

The combination of deductive reasoning, answer extraction from proofs, and procedural attachment allow a theorem prover to be used to coordinate multiple information sources to cooperate in solving a common problem.

Geospatial theory and Knowledge Sources

While the above approach is perfectly independent of the choice of application domain, we have primarily experimented with its application to geography and earth sciences. We are developing a geospatial theory to serve as the application domain theory for GeoLogica. Some of the knowl-

edge sources that have been attached procedurally to the geospatial theory include the following:

The Alexandria Digital Library Gazetteer (Hill, Frew, and Zheng 1999). A dictionary of about six million place names, the ADL Gazetteer provides for each one a geographical feature type, a latitude and longitude or bounding box, a list of variant names and alternative spellings, and a list of regions that contain the place. Given a name and a type, it can search for all places of that type with that name. It can restrict the search to places within a given bounding box. If the name is not provided, it can search for all places of the appropriate type (e.g., all forests within the bounding box of Oregon.)

The CIA World Factbook (CIA 2002). An almanac of the world's countries, the Factbook contains geographic, economic, social, and military information about each. Its information is complementary with that of the ADL Gazetteer. Unlike the Gazetteer, it contains no latitudes or longitudes or other information for cities or other geographical types, but for each country it gives its area, its extremes of elevation, its bordering countries, its principal subdivisions, its exports, a map, and other such information.

Agent Semantic Communication Service (Pease, Li, and Barbee 2002). ASCS is a search engine, developed by Teknowledge, that accesses, indexes, and extracts information from all Web pages that are annotated with the DARPA Agent Markup Language (DAML; `www.daml.org`). Although we have extracted some information from the Factbook directly, most of that information comes to us through ASCS, because a DAML-annotated version of the Factbook has been produced.

TextPro (Appelt and Martin 1999). An information-extraction engine developed at SRI, TextPro allows us to extract information from unstructured text sources. TextPro preprocesses the sources, extracts relational and temporal information, and enters them into a relational database, which can be consulted while the proof is in progress. Currently TextPro has been applied to data provided by the Center for Nonproliferation Studies, Monterey, CA.

Geographical computation agents. GeoLogica invokes a number of procedures for performing geographical computations, such as finding the distance between two lat/long pairs, finding the lat/long of a place so many miles north of a given lat/long, or finding the scale necessary to display a given region.

Conversion agents. Different sources adopt different notations and conventions. We invoke many agents whose sole purpose is to convert between one notation and another. For instance, there are agents that convert between different notations for latitude and longitude.

Visualization agents. GeoLogica invokes a number of providers of maps and satellite imagery. NIMA's Geospatial Engine and Generic Mapping Tools supply maps for a given region; the agent can select features to highlight or provide points or vectors to be superimposed on the map. Satellite imagery is provided by USGS's LandSat Project and the NASA Goddard Distributed Active Archive. TerraVision (Reddy et al. 1999) presents a flight-simulator-like three-dimensional view of a selected region; the user can then "fly" around the region under interactive control.

Recently we have been incorporating ECHO (`www.echo.eos.nasa.gov/operation.shtml`), the Advanced National Seismic System database at Berkeley (`quake.geo.berkely.edu/anss`), the University of Maryland Global Land-Cover Facility browser (`glcfapp.umiacs.umd.edu/index.shtml`), Modis, the USGS Geographic Names Information Service `geonames.usgs.gov`, and other information sources.

The emphasis of the project has been to make it easy to incorporate new information sources. The sources do not need to know about each other or to be intended to work together. They may have chosen different conventions or representational schemes. For instance, the capital of the Czech Republic may be Prague or Praha. There are many representations for latitudes and longitudes, in terms of numbers or strings. The 37th North latitude can be represented by the signed string "37" or the compass string "37N". We can also use decimal notation, or the notation based on degrees, minutes, and seconds. Different knowledge sources will produce different representation of latitude and longitude as outputs, and expect different representations as inputs. For instance, the Alexandria Digital Library Gazetteer accepts and produces latitudes and longitudes in signed string notation. One agent that computes the distance between latitude/longitude pairs requires latitudes and longitudes in compass notation. The axiom that advertises an agent must specify the notations expected and produced. The geospatial theory, therefore, must discriminate between these notations. Also, some conversion agents will be able to convert from one representation to another. The ADL Gazetteer knows about alternative names for the same place, and serves as a conversion agent between them.

Components of GeoLogica

In this section we will describe two important components of GeoLogica.

Gemini

Questions to GeoLogica are translated into a logical form by Gemini (Dowding et al., 1993), a mature, robust parsing and interpretation system that has been used by several projects at SRI, Stanford, NASA, and elsewhere over the past ten years. Although currently, in GeoLogica, Gemini is only used to parse questions, in the future it may be used to parse information supplied by the user in dialogue and text from other source material. Gemini may also be used to generate text to present and explain answers to questions.

A broad-coverage English grammar and lexicon for GeoLogica was compiled from several earlier projects. The open-ended nature of GeoLogica queries required a much larger vocabulary than previous Gemini projects. More than 50,000 new items were added to the lexicon, including 6000 adjectives and 35,000 nouns from Wordnet, and 400 geographical terms from the Alexandria Digital Library Gazetteer and NASA sources.

Gemini also has a capability for guessing the part of speech of an out-of-vocabulary word and temporarily adding that word to the lexicon. This has proved necessary for dealing with the large number of place names that occur in GeoLogica questions that cannot be cataloged in advance.

SNARK

Theorem provers have traditionally excelled at mathematical reasoning, which requires finding non-obvious proofs over theories defined by relatively small sets of axioms. In contrast, SNARK has been developed for applications in artificial intelligence and software engineering, which requires straightforward reasoning on theories defined by large axiom sets. SNARK (Stickel, Waldinger, and Chaudhri 2000) is a first-order logic theorem prover with resolution (for general deductive reasoning) and paramodulation (for reasoning about equality), implemented in Common Lisp. It has a sort mechanism, which allows all expressions to be categorized according to a hierarchical sort structure. It is particularly well suited for question-answering applications, for several reasons: It has strategic controls that allow us to tailor it to exhibit high performance in selected application domains; it has a mechanism for extracting answers from proofs; it has a procedural attachment mechanism; and it has built-in procedures for reasoning efficiently about space and time. SNARK is used in NASA's system Amphion (Lowry et al. 1994), for automatic software composition, and in the Kestrel Institute's software development environment, SPECWARE (Kestrel Institute 2002), as well as several SRI projects.

A Sample Problem: The Petrified Forest

To illustrate our approach, let us consider a simple problem. We are given the following query:

> Show a petrified forest in Zimbabwe that is north of the capital of Botswana and within 200 miles of Lusaka, Zambia.

This is parsed by Gemini, which produces the following logical form:

```
show(?x) &
patient(?x, ?y) &
petrified-forest(?y) &
in(?y, Zimbabwe) &
north(?z, ?y) &
source(?z, ?u) &
capital-of(?u, Botswana) &
within-distance-of(?y, ?v,
    feature(city, Lusaka, Zambia) &
mile-unit(?v) &
count-of(?v, 200)

answer:  ?x
```

This might be translated more literally as "Find a showing ?x of ?y where ?y is a petrified forest and ?z is a northness of ?y and the object of the northness is ?u, the capital of Botswana, and the distance of the petrified forest ?y from the city of Lusaka, Zambia is ?v, where the unit of ?v is miles and the magnitude of ?v is 200." The geospatial theory has axioms for each of the concepts in this logical form. Rather than reproducing the proof, let us see what agents are invoked to solve the problem. The ADL Gazetteer finds the bounding box of Zimbabwe and then searches within it for a petrified forest. It finds one, the "Makuku Fossil Forest," and produces the lat/long for this forest. The CIA World Factbook

Figure 1: Makuku Fossil Forest

reveals that the capital of Botswana is Gabarone. The ADL Gazetteer finds the bounding box for Botswana and then searches within it for the lat/long of Gabarone. A geographical computation agent is invoked to compare the latitudes and verify that the Makuku Fossil Forest is indeed north of Gabarone. The ADL Gazetteer also finds the bounding box for Zambia and searches within it for a lat/long for the city of Lusaka. A geographical computation agent determines the distance between the Makuku Fossil Forest and Lusaka, 112 miles. This is within the specified distance of 200 miles. The lat/long for the fossil forest is then passed to TerraVision, which displays the region around it (Fig 1.) NIMA or Generic Mapping Tools maps or LandSat images for the region can also be displayed.

Related Work

Conventional mediation techniques (e.g., Gupta et al. 1999) restrict the language in which we can describe the relationship between various sources. By employing a theorem prover to aid in the mediation process, GeoLogica can harness the full power of logic in expressing these relationships.

GeoLogica has roots in early work in deductive question answering and program synthesis (e.g., Green 1969, Manna and Waldinger 1980). Deductive program synthesis techniques were used for software composition and applied to data analysis for planetary astronomy in NASA's Amphion project (Lowry et al. 1994).

The approach depends on the development of an appropriate application domain theory, including ontology and axioms. A large axiomatic knowledge base has been under development for many years by Cycorp (Lenat and Guha 1994). Teknowledge has also been developing a public ontology and axiomatic theory for general world knowledge. Fonseca et al. have been developing an ontology specifically for geographical application. There is a group led by Hobbs to develop a spatial ontology in DAML (Hobbs et al. 2003). The Sweet Ontology (Raskin 2003) is specifically for the earth sciences.

Both Infomaster (Genesereth, Keller, and Duschka 1997) and Ariadne (Knoblock and Minton 1998) use deduction to coordinate multiple agents, for applications such as searching through classified ads or making travel arrangements.

Future Research

GeoLogica is very much work in progress. Our first order of business is to enrich our geospatial theory and incorporate a larger set of data sources. We expect to have GeoLogica perform computations on data that it finds, and perhaps produce tables and other visual displays of its results.

Up to now, we have been developing GeoLogica's ability to answer single, isolated questions. In the next phase of our research we shall develop the capability to engage in a dialogue with the user, who would be able to establish a context, provide background information, and ask for modifications of previous questions or elaborations on their answers.

We shall also extend GeoLogica's explanation capability; the user will be able to find out the sources for GeoLogica's answers, and some of the reasoning behind them.

Often there are multiple sources for information, but some sources may be more reliable or more efficient than others. Also, sources we come to rely on may be periodically unavailable. In the future, GeoLogica will be able to juggle alternative sources of information more strategically.

The theorem prover SNARK has well developed capabilities for reasoning about space and time, which we have not exploited very much yet in GeoLogica. It includes time and date arithmetic and an implementation of the Allen Temporal Interval Calculus (Allen and Ferguson, 1994) for reasoning about time; we plan to use this for detecting and reasoning about motion, environmental change, events, and other objects that have a temporal as well as a spatial dimension. Eventually we should be able to produce animations as well as static visualizations.

Acknowledgments

We would like to thank Douglas E. Appelt, John Fry, Jerry Hobbs, David J. Israel, David Martin, Susanne Riehemann, Mark Stickel, and Mabry Tyson for contributions to the research, its implementation, and its presentation. This work has been supported by NASA, under the Intelligent Systems Program; ARDA, under the Aquaint program; and DARPA, under the DAML program.

References

Allen, J. F., Ferguson, G. 1994. Actions and events in interval temporal logic. *J. Logic and Computation* 4:5.

Appelt, D. E., Martin, D. L. 1999. Named Entity Recognition in Speech: Approach and Results using the TextPro System `www.nist.gov/speech/publications/darpa99/html/ie30/ie30.htm`

Cohn, A. G., Bennett, B., Gooday, J. M., and Gotts, N. 1997. RCC: a Calculus for Region-based Qualitative Spatial Reasoning *GeoInformatica* 1:275–316.

Green, C. C. 1969. Application of Theorem Proving to Problem Solving. *Proc. Int. Joint Conf. on Artificial Intelligence*, 219–239.

Lowry, M., Philpot, A., Pressburger, T., Underwood, I., Waldinger, R., Stickel, M. 1994. Amphion: Automatic Programming for the NAIF Toolkit, *NASA Science Information Systems Newsletter*, Issue 31, 22–25.

Central Intelligence Agency. 2002. The World Factbook 2002. `www.cia.gov/cia/publications/factbook`

Dowding, J., Gawron, J. M., Appelt, D., Bear, J., Cherny, L., Moore, R., Moran, D. 1993. GEMINI: A Natural Language System for Spoken-Language Understanding. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 54–61.

Fonseca, F., Egenhofer, M., Davis, C., Camara, G. 2002. Semantic Granularity in Ontology-Driven Geographic Information Systems *Annals of Mathematics and Artificial Intelligence*, 36(1–2):121–151.

Genesereth, M., Keller, A. M. Duschka, O. M. 1997. Infomaster: an Information Integration System. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 26:539–542.

Gupta, A., Marciano, R., Zaslavsky, I., and Baru, C. 1999. Integrating GIS and Imagery through XML-based Information Mediation. In Agouris, P. and Stefanidis, A. (eds.) *Integrated Spatial Databases: Digital Images and GIS*, Lecture Notes in Computer Science, 1737:211–234.

Hill, L. J., Frew, J., Zheng, Q. 1999. Geographic Names: The Implementation of a Gazetteer in a Georeferenced Digital Library. *D-Lib*.

Hobbs, J. et al. 2003. A DAML Spatial Ontology. `www.daml.org/listarchive/daml-spatial/0002.htm`

Knoblock, C. A., Minton, S. 1998. The Ariadne Approach to Web-based Information Integration. *IEEE Intelligent Systems*, 13(5):17–20.

Lenat, D. B., Guha, R. V. 1994. Enabling agents to work together. *Communications of the ACM*, 37:7.

Kestrel Institute. 2003. Specware. http://www.kestrel.eduy/HTML/prototypes/specware.html

Manna, Z., Waldinger, R. 1980. A Deductive Approach to Program Synthesis. *ACM transactions on programming languages and systems* 2:90–121.

Pease, A., Li, J., Barbee, C. 2002. DAML Agent Semantic Communications Service (ASCS) `oak.teknowledge.com:8080/daml/damlquery.jsp`

Raskin, R. 2003. Semantic Web for Earth and Environmental Terminology (SWEET). *Earth Science Technology Conference*, Adelphi, MD.

Reddy, M., LeClerc, Y.-G., Iverson, L., Bletter, N. 1999. TerraVision II: Visualizing Massive Terrain Databases in VRML. *IEEE Computer Graphics and Applications (Special Issue on VRML)*. 19(2):30–38.

Stickel, M. E., Waldinger, R. J., Chaudhri, V. K. 2000. A Guide to SNARK. SRI International, Menlo Park, CA.