An ontology-driven reading agent

Deryle W. Lonsdale, David W. Embley, Stephen W. Liddle

Abstract

Textual data—from manuscripts to publications to website content—contains much of extant human knowledge. Unfortunately, the ability to harvest and effectively use this information beyond simple search/retrieval is greatly hampered by the scale of the "reading" problem: there is too much for any one person to read, and computers are not entirely adept at comprehending all information—explicit and implicit—contained in natural language text. Developing increased capability in this area is the focus of ongoing "machine reading" and "reading the web" research initiatives. Interested parties include businesses, the military, and intelligence-gathering agencies. Our own ongoing work with the Church Family History Department's vast digitized repository has led us to consider increased participation in this area of research.

We propose to unite the efforts of two different BYU research labs to create a sophisticated machine reading system. Each lab has concentrated on specific aspects of machine reading: (1) data extraction, integration and modeling in the BYU Data Extraction Group (DEG) lab, and (2) sophisticated natural language parsing and cognitive modeling in the BYU NL-Soar lab. Both research efforts are mature, having produced many academic and scholarly products; both have also benefited from prior support from on-campus and off-campus funding.

Our project will involve designing, implementing, and evaluating a new system, OntoSoar, which integrates our OntoES system with our Soar-based natural language processing systems. Soar is an agent-based cognitive modeling system that has served as an integration platform for several complex multi-task implementations. Our new reading agent will be capable of extracting low-level information in its first-pass treatment of a text; it will then perform a careful re-reading of the text to find more subtle conceptual relationships. OntoSoar will then compare extracted content from both processes and merge or supplement its growing knowledge base accordingly. We will evaluate the system against current research datasets.

1 Introduction

A vast amount of human knowledge is available in written form: in hand-written manuscripts, published journals and books, and via the internet. Coping with large-scale efforts to understand and use textual knowledge is an ongoing challenge for businesses, governments, and researchers. Using computers to help in this task has its limits: fully processing natural language text to extract all recorded information—explicit and implicit—is still an open research question. Search engines can instantly locate documents based on user-specified words and phrases, but no deeper processing is typically carried out on the textual content.

For example, the Church has been digitizing family history books and other publications for decades, making copies via microfilm, microfiche, and digital imagery. Yet only a modest amount of work has been done to systematically harvest this information from these resources; extraction has been largely a matter of individuals who research and hand-record snippets of data that is of interest to them.

Ideally, software could be developed to carry out this work. Presumably it could process text of any type to understand what it's "about", and update stored knowledge representations accordingly. Such software could be particularly adept at recognizing what it already knows and what it doesn't, and would be able to focus on novel information. Work indeed is being done in this area, but much work remains.

A few research efforts have attempted to build systematic large-scale text reading applications. At Microsoft Research a group developed a comprehensive semantic knowledge network called

MindNet by parsing information sources such as dictionary entries from the *Longman's Dictionary of Contemporary English* and from Encarta encyclopedia articles.[1]

A team at Carnegie Mellon University has created an agent-based system called Never Ending Language Learner (NELL), a Read the Web research project[2]. The system, which has been operational for about a year now, augments its large database of acquired facts by reading anything it can find on the web. An impressive real-time sampling of the facts it is acquiring is available for introspection at its website. It doesn't do a full parse, however; the extraction of information is shallow and linguistically naïve.

DARPA, the federal military/intelligence funding agency for advanced research projects, is interested in computerized reading capability. They have sponsored one program scheduled to end in December 2012 called Machine Reading[3] to "develop learning systems that can "read" natural text and insert it into AI knowledge bases". DARPA has also recently initiated another project called Deep Exploration and Filtering of Text (DEFT) aimed to "address remaining capability gaps related to inference, causal relationships, and anomaly detection".[4]

Our proposed ORCA MEG project is to build a reading agent that uses ontologies as one type of knowledge source. Developing this kind of computer system from the ground up would be well beyond the scope of an ORCA MEG grant. However, we believe we can achieve this goal because our project would involve combining the efforts of two of our on-campus research groups and the systems we have already developed independently in those groups. Support for developing these systems has involved prior funding from ORCA and from other organizations both within and outside of BYU.

In this project profile we summarize our prior work in each of these areas and then describe our vision for how they can be integrated in a coherent meta-project here at BYU.

1.1 DEG and OntoES

For several years the BYU Data Extraction Group (DEG) has been carrying out research on extracting and structuring data from unstructured and semi-structured electronic documents, such as those found on the web and in many different data warehouses.[5] Our central knowledge source is an ontology: a machine-readable, mathematically specified conceptualization based on a collection of facts.

For example, we have shown [2] how our approach extracts information from English, French, and Korean obituaries. Our process involves the following steps:

1. Development of a narrow-domain ontology to model the concepts associated with an obituary.

2. Parsing the ontology to create a database schema and rules for matching textual information.

3. Processing web-based obituaries and recognizing/removing extralinguistic information (markup, record boundaries, etc.).

4. Marshalling recognizers that use the matching rules to extract textual information corresponding to the ontology's concepts.

5. Populating a relational database with the extracted results so that they can be queried via standard access methods.

[1] See http://research.microsoft.com/en-us/projects/mindnet/default.aspx.

[2] See http://rtw.ml.cmu.edu/rtw/.

[3] For more details see http://www.darpa.mil/Our_Work/I2O/Programs/Machine_Reading.aspx.

[4] See http://www.darpa.mil/Our_Work/I2O/Programs/Deep_Exploration_and_fFiltering_of_Text_(DEFT).aspx.

[5] See our website at http://deg.byu.edu.

This work was principally funded by the National Science Foundation (NSF) in a 3-year project.[6] We have also since extended our extraction work in three other ways:

- Table interpretation: Our system can analyze the structure of tables, infer relations among the information contained in their cells, and generate ontologies reflecting the tables' content. This work was also funded by the NSF[7]. See also [10, 12];

- Workbench interface: We have developed a comprehensive user interface that combines several of the capabilities of our system (text preprocessing, ontology creation, ontology editing, table analysis, corpus annotation, etc.) [3, 4]. It has recently been used by Church service missionaries with the Family History department in a prototype project to annotate and extract content from digitized genealogical documents.

- Multilingual extraction: Our system is now able to perform crosslinguistic information extraction: an English-speaking user can search for information from Korean Web pages, for example, and have the pertinent results returned in English. This functionality is possible via multilingual mappings with our ontologies. An ORCA MEG year 2010 grant provided the support necessary to achieve this functionality [1, 5].

Our system, called the Ontology-based Extraction System (OntoES), is very successful at extraction of low-level items from text—such as dates and names—and at identifying the relationships between them. It is even adept at implementing inferencing rules that can draw further conclusions from the data, such as the fact that if two people have the same parents, they are siblings. However, the system does not perform a deep level of linguistic processing of the information, so many high-level items of information are lost, particularly those not included in an ontology. For example, it is not currently able to reliably resolve the referent of a pronoun: in a sentence like "Susan was born in Toledo and married Fred; she died in Tuscon." it cannot figure out that the referent of "she" is "Susan".

Our proposed solution is to integrate OntoES with a cognitive modeling system that is specialized to handle the complexity and nuances of natural language. We next describe the proposed system, and why a cognitive modeling architecture is needed.

1.2 Cognitive modeling of language use

Cognitive modeling systems are computer systems that are designed to carry out problem solving in a manner not followed by most computer applications: they operate in a self-directed, goal-oriented, interruptible fashion. Whether performing highly routine tasks or extremely difficult, open-ended problems they implement human problem solving methods. This includes inferencing, deduction, learning, and drawing on several types of memory mechanisms: semantic memory, episodic memory, and procedural memory. Also of interest is how the system represents different kinds of knowledge, and how that knowledge is brought to bear in task performance.

Some cognitive modeling systems are agent-based, which means that they are designed to interface seamlessly with an external environment via perceptual mechanisms. An agent-based system is also capable of situating itself and its cognition in its environment, and hence is capable of differentiating what it knows and what it doesn't. Thus it is able to solve novel problems, to learn from experience, and to attend proactively to new information.

One widely used agent-based general cognitive architecture is called Soar. As a programming architecture, it serves as an ideal foundation for implementing software systems. As a general

[6] NSF CISE-IDM-IIS-0083127, code-named TIDIE

[7] NSF CISE-IDM-IIS-0414644, code-named TANGO

cognitive architecture it is designed to carry out problem solving in the same manner that humans do, instead of how computer systems typically operate. As an agent-based system it is capable of controlling its own task performance and learning from its experience. Soar also exhibits other aspects of cognition-based processing: long-term declarative memory, working memory, chunking, reinforcement learning, and episodic learning. Current research even addresses modeling sentiment, emotion, attention, and other phenomena.

The Soar platform is used by many renowned researchers and is actively maintained: the latest version of the software was released earlier this year. Continual funding for development of the Soar architecture has been provided by several U.S. government agencies including the Office for Naval Research, the Air Force Office for Special Research, and DARPA. Several publications discuss aspects of Soar theory, implementation, and modeling [7, 11]. There is also a comprehensive list of Soar-related publications, a wiki, and a FAQ page at the Soar project website[8].

1.2.1 Soar-based language processing

Natural language Soar (NL-Soar) is a natural language processing system built upon the Soar cognitive architecture. The original version of the system was designed to syntactically parse sentences in order to model human parsing difficulties based on cognitive-architectural constraints [8]. The system was subsequently extended to handle semantic processing, natural language generation, and discourse processing among several other task-related capabilities. In all of this work a theory of syntax was assumed that reflected current thinking around the late 1980's and early 1990's, and was built upon the basic cognitive modeling framework available at the time.

In its incremental parsing mode, sentences are input into the agent one word at a time. A lexical access operator is initiated for each word in turn; it retrieves from several knowledge sources various kinds of lexical, orthographic, morphological, syntactic, and semantic information for that word. Each lexical item is associated with a zero-level node which is then projected to bar-level and phrasal-level nodes. When features license combining words into phrases, corresponding syntactic constituents are built. The resulting parse tree is further converted to a semantic representation, which can be used in subsequent processing. Numerous academic products have resulted from NL-Soar work at BYU.[9]

Link Grammar Soar (LG-Soar) is another Soar-based language processor that was developed specially for information extraction. It is different than NL-Soar because it uses a more robust parser for the input sentences; this parser is not constructed on linguistic theory but entirely on relationships between words. The parse results are loaded into Soar working memory elements, and Soar processing builds up a semantic/pragmatic representation based on Discourse Representation Theory [6]. LG-Soar is more versatile than NL-Soar in handing non-standard language: spelling errors, complex sentences, abbreviations, sentence fragments, etc. Work at BYU has shown LG-Soar's success in processing family history documents, clinical trial reports, and even human/robot interaction. Several academic products have been based on LG-Soar work at BYU. LG-Soar currently serves as a language platform for the ongoing DARPA BOLT Activity E effort in grounded acquisition of natural language by robots.

XNL-Soar is a recent BYU development: the re-write of NL-Soar using more up-to-date linguistic theory and the newest updates to the basic Soar architecture. It performs a full, incremental syntactic parse and creates corresponding sematic representations. The system is able to handle many more syntactic structures than its predecessor NL-Soar. Part of our MEG activity will be to complete the semantic processing, and add pragmatic processing to XNL-Soar.

[8] http://sitemaker.umich.edu/soar

[9] See our website at http://nlsoar.byu.edu.

In fact, CMU NELL team members have inquired about the availability of Soar-based language processing agents for this task; unfortunately we had to decline participation at the time because our agent was not completely ready. One purpose of this MEG proposal is to bring our agent to completion to the extent that we have this capability.

Other Soar-based language work has involved using DAML/OIL ontologies as a declarative semantic memory knowledge source for a Soar system [9], though that work has not yet been implemented in any Soar natural-language system.

2 Methods and Research Plan

For our MEG project we intend to integrate our two separately developed systems to create a text reading system. The system we propose to develop will have several unique characteristics; it will:

- integrate bottom-up (data-driven) and top-down (expectation-driven) processing
- execute a first-read extraction of low-level information as currently performed by OntoES, creating an OSM-X output ontology
- execute a careful re-read, creating a full syntactic/semantic parse via LG-Soar and later XNL-Soar, creating a semantic model
- compare the contents of the ontology with those of the semantic model
- identify overlap, non-overlap of the concepts in these two knowledge structures
- extend the ontology with novel concepts it has discovered
- extend semantic memory as necessary with novel information it has discovered

This integrated system will exhibit several interesting characteristics; it will be:

- a hybrid system: combining the strengths of both bottom-up and top-down processing
- linguistically sophisticated: capable of generating a full linguistic parse (syntax, semantics, pragmatics)
- a learning system: capable of chunking up and recognitionally executing tasks it has encountered before
- agent-based: capable of determining whether information learned is something is does(n't) know already
- a modeling platform: capable of simulating/modeling human performance of tasks and of integration with other tasks

Our integration of the OntoES extraction system and Soar-based cognitive modeling agents involves several steps with identifiable outcomes at each stage.

We do not anticipate the need to recruit students at project onset; our research groups are already staffed with trained and enthusiastic undergraduates and graduates. We may need to replace students who graduate and leave during the course of the project. During the Winter semester, after the announcement of MEG grant recipients, we plan to mentor each student in carrying out the following activities:

1. Install each system (LG-Soar and OntoES) in a versioning system to track changes to the code base.

2. Integrate LG-Soar and OntoES: Design and implement algorithms for passing information between both components. This will result in a prototype system, OntoSoar 1.0. This system will be capable of extracting information in a two-pass reading of text: (1) OntoES-level extraction of objects and relationships, and (2) LG-Soar-level relationships. The system will then be able to perform basic analysis to compare and contrast the information from each pass. It will then make basic updates to its knowledge structures, based on the information learned. We will evaluate the performance of OntoSoar 1.0 by extracting information from family history data.

3. Integrate semantics and discourse processing into XNL-Soar: This is ongoing work that is partially completed. It can be done entirely in parallel with the OntoSoar 1.0 system integration.

4. Integrate XNL-Soar and OntoSoar: This will result in the OntoSoar 2.0 system. It will be capable of performing incremental parsing and extracting much more information (anaphor resolution, etc.) than version 1.0. We will evaluate OntoSoar 2.0 on the family history data (to compare its performance with that of OntoSoar 1.0). We will also evaluate version 2.0 with other type of text to be determined; it will likely consist of information from the Web, perhaps reading task comparative evaluation datasets.

The end result of our work will situate us in the space of state-of-the-art research in machine reading and agent-based information gathering. It will allow us to participate in the Family History extraction research mentioned above, as well as other possible text genres: biomedical research literature, patents, newswire, web pages, social media, journals and blogsites, for example. We will also have a modeling and simulation platform which can serve to investigate human reading, webpage scanning behavior, information gisting, and reading/task integrations. This will foster further on-campus and outside collaborations. Since our Soar-based agents currently support parsing in English, French, and Japanese, whereas OntoES has proven effective for English, Japanese, Korean, and French information extraction we anticipate being able to explore multilingual and crosslingustic aspects of text reading in our agent.

3 Anticipated mentoring outcomes

We currently envision that at least two master's theses will result from this work, and probably one honor's thesis. An MA thesis will focus on the integration of the two engines, and an MS thesis will describe efforts to preprocess texts for treatment by the system (i.e. record and sentence boundary detection, layout information, named entity recognition, etc.).

At least four peer-reviewed publications will result from this work. They will be co-authored with students and faculty. We will be targeting venues that have so far published our work: Springer LNCS, DKE, and conference proceedings (ER, WWW, NLDB, (NA)ACL, CogSci). As is common in our field, the publications will be associated with presentations at a world-class venue where we will present our results to an international audience of our peers.

We also anticipate continuing our practice of hosting visits by world-class researchers who are working in this area, as described above. This will afford local faculty and students networking opportunities, and exposure to our work to outside researchers.

Our labs involve both graduate and undergraduate students who interact regularly in planning, implementation, and evaluation of our collective and individual progress. We build strong peer relationships through daily interaction with our students. Our weekly lab meetings will continue, as will our weekly one-on-one meetings with the students on the project. We also hold occasional social gatherings for relaxed interaction with students' family members. Our work provides a wide range of mentoring situations, from high-pressure deadlines to team software development to public presentation rehearsals. We begin our research meetings with prayer, highlighting the special nature of the BYU environment. We strive to help our students learn how to balance the competing demands of their work, family, church, and community obligations.

4 Mentoring qualifications and experience

We have enjoyed a wonderful cross-disciplinary mentoring environment in our prior work; this MEG project would continute to foster this collaboration. Our students find themselves exposed to several intellectual disciplines and skill sets, and this project will bring together even more fields of endeavor. It will also prepare our students for participation in state-of-the-art activities that will help them in career or graduate studies placement.

The three co-applicants have worked together for over a decade on the research questions described in this proposal. We have been meeting weekly in a research group setting, reviewing literature together, and co-publishing and co-presenting on our information extraction work.

Co-proposer Embley has taught and directed research at BYU since 1982. His research interests include database systems and theory, information extraction from data-rich unstructured documents, heterogeneous data integration, revitalization of data from historical documents, model-driven software development, and the semantic web. He is a member of the steering committee for the International Conferences on Conceptual Modeling, and recently received the Peter Chen Award for exceptional contributions to conceptual modeling. He has directed the DEG research team since its founding at BYU in 1998, and has been a principal investigator for the two NSF research grants (NSF-TIDIE and NSF-TANGO 0414644) mentioned above, both of which have been central to the DEG research agenda. Several dozen of Embley's publications in the last decade have been on topics related to DEG research.

Co-proposer Liddle is the chief architect of the OntoES system. He is a professor in the Department of Information Systems and Academic Director of the Center for Entrepreneurship and Technology. Liddle was awarded the Marriott School's Gunnell Professorship. He is a computer scientist with expertise in web and mobile development, software engineering, entrepreneurship, and related topics.

Co-proposer Lonsdale is an associate professor in the Department of Linguistics and English Language. He has developed Soar-based language processing systems since the early 1990's, as one of the original NL-Soar team members at Carnegie Mellon University. Since coming to BYU Lonsdale has continued to be the principal Soar developer of language applications, and BYU is currently the center site for language-related development using Soar. He has also been a member of the DEG lab since its inception.

References

[1] David W. Embley, Stephen W. Liddle, Deryle W. Lonsdale, Joseph Park, Byung-Joo Shin, and Andrew Zitzelberger. Cross-language hybrid keyword and semantic search. In *Proceedings of*

the 31st International Conference on Conceptual Modeling (ER 2012), pages 190–203. Springer, 2012.

[2] D.W. Embley, D.M. Campbell, Y.S. Jiang, S.W. Liddle, D.W. Lonsdale, Y.-K. Ng, and R.D. Smith. Conceptual-model-based data extraction from multiple-record web pages. *Data & Knowledge Engineering*, 31(3):227–251, November 1999.

[3] D.W. Embley, S.W. Liddle, D. Lonsdale, G. Nagy, Y. Tijerino, R. Clawson, J. Crabtree, Y. Ding, P. Jha, Z. Lian, S. Lynn, R.K. Padmanabhan, J. Peters, C. Tao, R. Watts, C. Woodbury, and A. Zitzelberger. A conceptual-model-based computational alembic for a web of knowledge. In *Proceedings of the 27th International Conference on Conceptual Modeling (ER08)*, pages 532–533, Barcelona, Spain, October 2008.

[4] D.W. Embley, S.W. Liddle, and D.W. Lonsdale. Conceptual modeling foundations for a web of knowledge. In D.W. Embley and B. Thalheim, editors, *Handbook of Conceptual Modeling: Theory, Practice, and Research Challenges*, chapter 15, pages 477–516. Springer, Heidelberg, Germany, 2011.

[5] D.W. Embley, S.W. Liddle, D.W. Lonsdale, and Y. Tijerino. Multilingual ontologies for cross-language information extraction and semantic search. In *Proceedings of the 30th International Conference on Conceptual Modeling (ER 2011)*, pages 147–160, Brussels, Belgium, October/November 2011.

[6] Hans Kamp and Uwe Reyle. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic, and Discourse Representation Theory.* Kluwer Academic, 1993.

[7] John E. Laird. *The Soar Cognitive Architecture.* MIT Press, 2012.

[8] Richard Lewis. *An Architecturally-based Theory of Human Sentence Comprehension.* PhD thesis, Carnegie Mellon, 1993.

[9] Sean Lisse. DAML2Soar: A DAML+OIL Ontology to Soar Knowledge Translator. Presented at the 23rd Soar Workshop, Ann Arbor, MI, 2003.

[10] G. Nagy, S. Seth, D.W. Embley, M. Krishnamoorthy, D. Jin, , and S. Machado. Data extraction from web tables: the devil is in the details. In *Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011)*, pages 242–246, Beijing, China, September 2011.

[11] Allen Newell. *Unified Theories of Cognition.* Harvard University Press, Cambridge, MA, 1990.

[12] C. Tao and D.W. Embley. Automatic hidden-web table interpretation, conceptualization, and semantic annotation. *Data & Knowledge Engineering*, 68(7):683–703, July 2009.