AeroDAML: Applying Information Extraction to Generate DAML Annotations from Web Pages

Paul Kogut
Lockheed Martin M&DS
PO Box 8048
Philadelphia, PA 19101 USA
610-354-3524
paul.a.kogut@lmco.com

William Holmes
Lockheed Martin M&DS
PO Box 8048
Philadelphia, PA 19101 USA
610-354-6774
william.s.holmes.iii@lmco.com

ABSTRACT
The DARPA Agent Markup Language (DAML) is an emerging knowledge representation for the Semantic Web. DAML can encode the semantics of a document for use by agents on the web. However, DAML annotation of documents and web pages is a tedious and time consuming task. AeroDAML is a knowledge markup tool that applies natural language information extraction techniques to automatically generate DAML annotations from web pages. AeroDAML links most proper nouns and common relationships with classes and properties in DAML ontologies. This paper discusses the design of AeroDAML including linguistic and practical issues related to semantic annotation.

Keywords
Natural language processing, semantic markup, information extraction, software agents

INTRODUCTION

One major motivation for semantic markup is that natural language processing (NLP) is so difficult. An agent that is trying to interpret a page on the web must deal with the full complexity of natural language including syntactic and semantic ambiguities, obscure domain specific terminology and non-literal language like metonymy and metaphors. The extreme complexity of NLP led to major efforts like the DARPA Agent Markup Language (DAML) program which is trying to simplify the agent's interpretation task by allowing the author of a document to add explicit semantic information [1][2]. DAML+OIL is an emerging knowledge representation language that is being developed for the Semantic Web. The DAML+OIL language supports the definition of machine-readable ontologies and the linking of terms in documents to ontologies. DAML+OIL is an extension of RDF and RDF Schema that is able to express a richer variety of constraints and support tractable reasoning.

However, a semantic markup language approach like DAML+OIL has significant drawbacks. DAML annotation of web pages and other documents is a tedious and time consuming task. This is a major problem for people who publish web pages and documents on a regular basis or have a large quantity of legacy documents that they need to annotate. Also, novice and casual users would like to have a simple way to do basic annotation without having to spend time learning about ontologies. In this paper we propose the use of NLP to help reduce the effort and knowledge required from the semantic annotator. Why use NLP to help overcome a problem with the semantic annotation approach that was motivated by the weaknesses of NLP? We believe that moving NLP from the document consumer side to the document producer side will facilitate human intervention to compensate for the weaknesses of NLP. Similar semi-automatic approaches have worked well in applications like machine translation where NLP techniques generate a draft translation that is corrected by a human thus reducing the overall cost of accurate translation.

We have developed a prototype tool called AeroDAML to experiment with the producer side application of NLP. AeroDAML is a knowledge markup tool that applies natural language information extraction techniques to automatically generate DAML annotations from web pages. AeroDAML links most proper nouns and common relationships with classes and properties in DAML ontologies. This paper discusses the design of AeroDAML including linguistic and practical issues related to semantic annotation.

ANNOTATION GENERATION TOOLS

There are at least four basic types of tools for DAML annotation:

- Semi-automatic
- Automatic – default ontologies
- Automatic – customized ontologies
- Hybrid

The semi-automatic tools support assignment of words in a document to DAML classes and properties based on human

judgement. The actual linking is done with some form of drag and drop interface. Ontomat [3] is an example of a semi-automatic annotation tool.

The automatic tools apply NLP techniques to assign words to classes and properties. These tools may work with some predetermined set of default domain independent ontologies (e.g. OpenCyc, IEEE Standard Upper Ontology) or domain dependent ontologies (e.g., Universal Standard Products and Services Classification Code (UNSPSC), Unified Medical Language System (UMLS)). An automated tool may also incorporate more sophisticated support for the development of a domain or application-specific customized ontology and the creation of associated extraction rules.

The final tool type listed above is when semi-automatic and automatic approaches are combined into a hybrid tool that supports word assignments by human judgement and NLP techniques.

We expect that non-technical people will prefer automatic tools. We expect that technical people who occasionally do annotation will prefer semi-automatic tools. Finally, we anticipate that professional publishers of documents and reports will prefer a hybrid approach with customized ontology support.

ANNOTATION USAGE SCENARIOS

There are many potential uses for annotation on the semantic web including workflow, image retrieval, database mediation and device interoperability. In this paper we will focus on three scenarios that involve semantically annotated web pages or text documents:

- Information retrieval (IR) – identify and rank relevant pages or documents
- Simple question answering (Q&A) – e.g., *Who is the governor of Alaska?*
- Complex question answering – e.g., *What is the current situation in Algeria?*

All three scenarios may involve some degree of reasoning and inference. Complex Q&A will often involve significant reasoning and summarization capabilities.

THE AERODAML APPROACH

AeroDAML is an automatic DAML annotation tool. The web-enabled version of AeroDAML [4] supports annotation with a default generic ontology of commonly found word classes and relationships. The user simply enters a URI and AeroDAML returns the DAML annotation for the specified web page.

The client server version of AeroDAML supports annotation with customized ontologies. The user enters a file name and AeroDAML returns the DAML annotation for the text document. Eventually, we plan to integrate a semi-automatic tool into AeroDAML to create a hybrid annotation tool.

AeroDAML consists of a commercial information extraction product called AeroText™ and components for DAML generation. AeroText™ is a high performance information extraction system for developing NLP-based content analysis applications. It provides advanced graphical tools in an integrated development environment to simplify the creation and maintenance of application knowledge bases.

AeroText™ has a versatile architecture that is designed to support a variety of text processing tasks. It is composed of the following major components:

Knowledge Base Compiler, which converts linguistic data files into an efficient run-time knowledge base (KB)

Knowledge Base Engine, which applies the knowledge base to input documents

Integrated Development Environment (IDE), which provides a complete environment to build, test, and analyze linguistic knowledge bases.

Common Knowledge Base, which contains domain independent rules for extracting most proper nouns and frequently occurring relations. The elements in this common KB can be quickly composed into domain specific relationship and event extraction rules.

AeroText™ has a Java API that is used to access an internal form of the extraction results. DAML generation components access this internal form, then translate the extraction results into a corresponding RDF triple model that utilizes the DAML+OIL syntax. This is accomplished by referencing a default ontology that directly correlates to the linguistic knowledge base used in the extraction process. In the final step, the RDF model is serialized to produce the resulting DAML annotation.

The lower level of the current default AeroDAML ontology is based on the common knowledge base of AeroText™. The upper level of the default ontology is based on the WordNet noun synset hierarchy [5]. The ontology was modeled as a set of UML class diagrams in Tau UML Suite [6] and ontology engineering tools generated the DAML compliant ontology [7].

The current AeroDAML can generate annotations that consist of words (entities) linked to ontologies as instances of classes and relationships that are linked to ontologies as instances of properties. A list of typical examples is shown below:

- Proper nouns – example: *Japan* instanceOf *nation*
- Common nouns – example: *gun* instanceOf *weapon*
- Co-references – example: *Clinton* equivalent to *Bill Clinton* instanceOf *person*
- Measure – example: *22 inches* instanceOf *measure*
- Money – example: *$200* instanceOf *money*

- Absolute date – example: *December 19, 1997* instanceOf *absolute date*
- Relative date – example: *last month* instanceOf *relative date*
- Temporal co-references – example: *December 19, 1997* is event date then *last month* is *November 1997*
- Organization to location – example: *Tyrolean Airways* to *Austria*
- Person to organization – example: *Bill Gates* to *Microsoft*

Our hypothesis is that the annotation generated from the current AeroDAML will support information retrieval and simple question answering applications because the type of class and property instances illustrated above provide high information content. IR experiments will be conducted to test this hypothesis. Another hypothesis is that automatic annotation capabilities must be extended to support more complex Q&A applications. This would include the extraction and annotation of additional common nouns and frequently occurring verbs.

RELATED WORK

A good example of the consumer side approach to the application of NLP to interpretation of web pages can be found in [8]. This approach requires a combination of NLP and machine learning techniques to deal with full complexity of natural language. In contrast, the producer side approach described in this paper is amenable to leveraging human judgement for cases that cannot be handled automatically.

The effort to develop the WordNet Semantic Concordance [5] applied semi-automatic semantic annotation to a wide variety of nouns, verbs, adjectives and adverbs. We believe this work provides insights into the development of richer automatic annotation generation.

CONCLUSIONS

The feasibility of automatic annotation generation has been demonstrated by the AeroDAML prototype. To support professional document publishers we plan to extend AeroDAML to annotate more words and relationships and link these terms to a more semantically rich ontology. Also we plan to integrate AeroDAML with a semi-automatic annotation capability to form a hybrid tool. Some major open research issues include:

- What words and relationships need to be annotated to support complex question answering applications?
- What form of graphical interface will best support a semi-automated refinement of automatically generated DAML annotation?

ACKNOWLEDGMENTS

This material is based upon work supported by the Air Force Research Laboratory, Contract Number F30602-00-C-0188. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.

REFERENCES

1. T. Berner-Lee, J.Hendler and O. Lassila The Semantic Web. *Scientific American*, May 2001.
2. J.Hendler and D. McGuinness The DARPA Agent Markup Language. *IEEE Intelligent Systems*, 15, No. 6:67-73, 2000.
3. http://ontobroker.semanticweb.org/annotation/ontomat/index.html.
4. http://ubot.lockheedmartin.com/ubot/hotdaml.
5. Christiane Fellbaum (Editor*), Wordnet: An Electronic Lexical Database* MIT Press 1998.
6. Telelogic http://www.telelogic.com.
7. UML Based Ontology Toolset (UBOT) Project http://ubot.lockheedmartin.com/.
8. Mark Craven et al. “Learning to construct knowledge bases from the World Wide Web” *Artificial Intelligence*, 118 (1-2) (2000) pp. 69-113