Automated OWL Annotation Assisted by a Large Knowledge Base

Michael Witbrock, Kathy Panton, Stephen L. Reed, Dave Schneider, Bjørn Aldag,
Mike Reimers and Stefano Bertolo

{witbrock, panton, sreed, daves, aldag, mreimers, bertolo}@cyc.com

Abstract. Widespread adoption of the semantic web depends critically on lowering the “barriers to entry” facing document producers. We describe a system that applies automatic partial parsing of web pages into the representations of the large ResearchCyc ontology, combines this with convenient mixed initiative knowledge capture, and produces an OWL annotated document as output. Semantic web publishers can then use this document as a starting point for more elaborate, manual annotation.

Introduction

The rapid adoption of the World Wide Web, in its initial form, was driven in part by the ease with which content could be produced; although specialized tools and techniques quickly evolved, web pages could be produced, reasonably conveniently, by anyone with a text editor and an hour to read a description of the available HTML tags. Semantic markup in languages like OWL has the potential to vastly increase the utility of web content, but describing the logical content of a document is far from straightforward, even without the requirement that that description be done in an XML-based markup language.

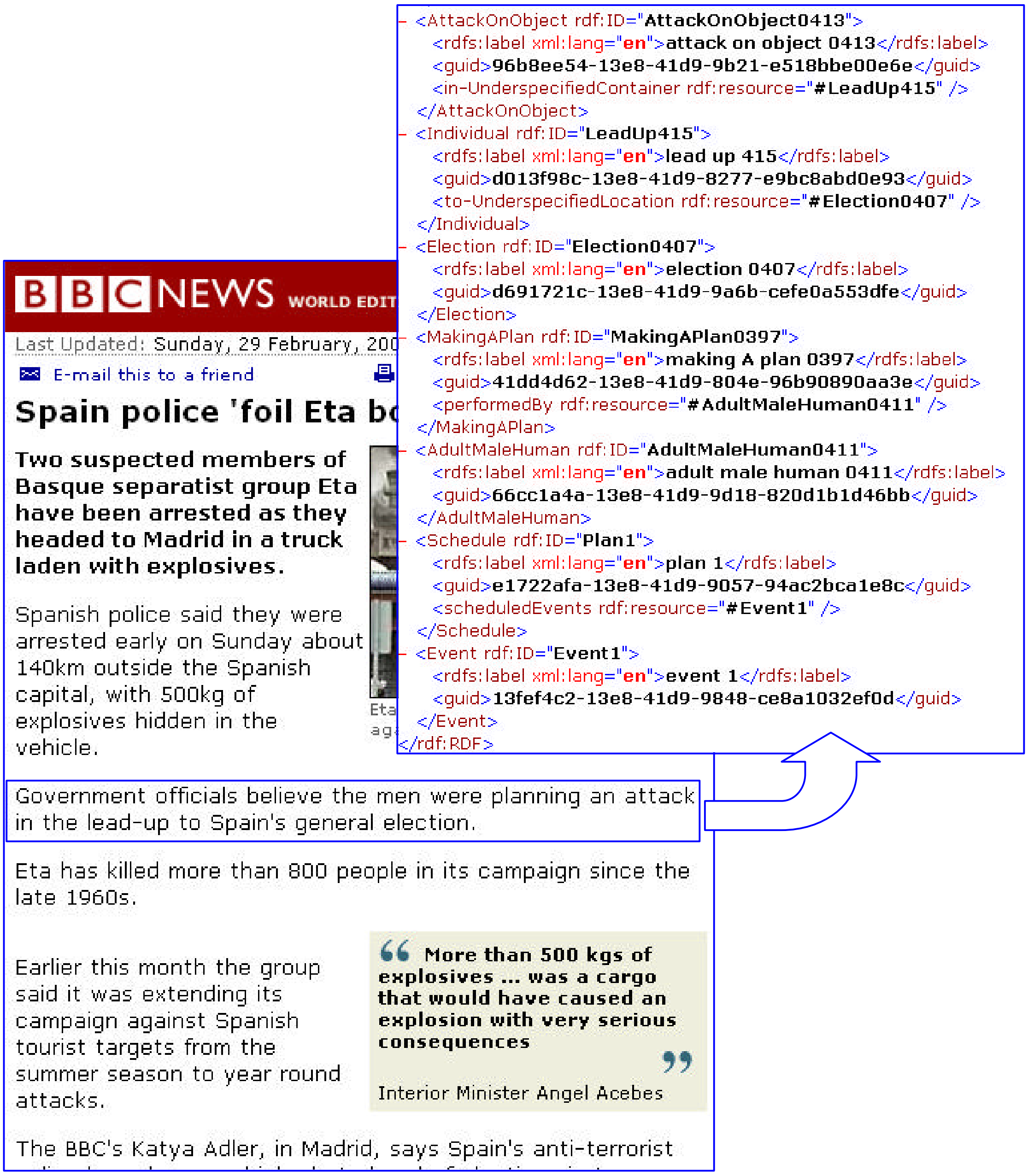
In addition to the simple tools and syntax required for HTML authoring, the ready availability of example pages with mark-up produced by others further flattened the already shallow learning curve for Web authoring. Providing such examples for the semantic web would have similar utility but is not as obviously straightforward. While the syntax of OWL is consistent, the conceptual tag set to be used is highly dependent on the domain of the document, and, even within a domain, is set only by convention. Rather than require prospective authors to identify the appropriate vocabulary, complex XML syntax, and relevant set of example documents before semantic annotation can begin, it seems worthwhile to provide a tool that, while imperfect, can make an initial, automatic pass at annotating a document. From that rough annotation, it should be more straightforward for human content providers to incrementally improve the representation of page content as they increase their understanding of relatively narrow components of the relevant ontology and OWL syntax.

In this paper, a system, based on Cyc, is described that can automatically produce initial OWL annotations of arbitrary text documents. This is done in the vocabulary of the OpenCyc scaffolding ontology, which is freely available[1] and freely usable. The annotation process takes advantage of existing Cyc system components for automated text analysis and guided knowledge entry, as well as newly-created components for interactive disambiguation using natural language and reduction of internal CycL representations to the OWL languages. Interactive components of the process are optional, and annotation can proceed wholly automatically.

Document Analysis

The Cyc OWL annotation system operates in two phases. First, the page is read and as much of the content as possible is represented in the CycL language. Second, the OWL export component of Cyc, developed as part of the DARPA DAML project, is used to generate the appropriate annotation file.

[1] http://www.cyc.com/2004/06/04/cyc



BBC NEWS WORLD EDIT

Last Updated: Sunday, 29 February, 200

E-mail this to a friend

Spain police 'foil Eta bo

Two suspected members of Basque separatist group Eta have been arrested as they headed to Madrid in a truck laden with explosives.

Spanish police said they were arrested early on Sunday about 140km outside the Spanish capital, with 500kg of explosives hidden in the vehicle.

Eta
ag

Government officials believe the men were planning an attack in the lead-up to Spain's general election.

Eta has killed more than 800 people in its campaign since the late 1960s.

Earlier this month the group said it was extending its campaign against Spanish tourist targets from the summer season to year round attacks.

> **More than 500 kgs of explosives ... was a cargo that would have caused an explosion with very serious consequences**
>
> Interior Minister Angel Acebes

The BBC's Katya Adler, in Madrid, says Spain's anti-terrorist

```
- <AttackOnObject rdf:ID="AttackOnObject0413">
    <rdfs:label xml:lang="en">attack on object 0413</rdfs:label>
    <guid>96b8ee54-13e8-41d9-9b21-e518bbe00e6e</guid>
    <in-UnderspecifiedContainer rdf:resource="#LeadUp415" />
  </AttackOnObject>
- <Individual rdf:ID="LeadUp415">
    <rdfs:label xml:lang="en">lead up 415</rdfs:label>
    <guid>d013f98c-13e8-41d9-8277-e9bc8abd0e93</guid>
    <to-UnderspecifiedLocation rdf:resource="#Election0407" />
  </Individual>
- <Election rdf:ID="Election0407">
    <rdfs:label xml:lang="en">election 0407</rdfs:label>
    <guid>d691721c-13e8-41d9-9a6b-cefe0a553dfe</guid>
  </Election>
- <MakingAPlan rdf:ID="MakingAPlan0397">
    <rdfs:label xml:lang="en">making A plan 0397</rdfs:label>
    <guid>41dd4d62-13e8-41d9-804e-96b90890aa3e</guid>
    <performedBy rdf:resource="#AdultMaleHuman0411" />
  </MakingAPlan>
- <AdultMaleHuman rdf:ID="AdultMaleHuman0411">
    <rdfs:label xml:lang="en">adult male human 0411</rdfs:label>
    <guid>66cc1a4a-13e8-41d9-9d18-820d1b1d46bb</guid>
  </AdultMaleHuman>
- <Schedule rdf:ID="Plan1">
    <rdfs:label xml:lang="en">plan 1</rdfs:label>
    <guid>e1722afa-13e8-41d9-9057-94ac2bca1e8c</guid>
    <scheduledEvents rdf:resource="#Event1" />
  </Schedule>
- <Event rdf:ID="Event1">
    <rdfs:label xml:lang="en">event 1</rdfs:label>
    <guid>13fef4c2-13e8-41d9-9848-ce8a1032ef0d</guid>
  </Event>
</rdf:RDF>
```

Figure 1: The Cyc Document Annotator assists organizations and individuals interested in adapting their document production processes to the Semantic Web. By providing an approximate OWL annotation of an existing document, the system simplifies the initial learning curve, allowing editing to improve the annotation to replace the complex task of manually annotating a document from scratch. Interoperability is supported by annotation using the more than 60,000 freely usable terms in the OpenCyc scaffolding ontology.

The OWL export component of the system is described in more detail later, but the core of the annotation system depends on Cyc’s imperfect but growing ability to interpret free text into a detailed logical representation in CycL. This is provided by combined application of Cyc’s natural language processing subsystem, disambiguation dialogue, and the Factivore, a highly usable knowledge-driven knowledge acquisition interface.

Parsing into the CycL Logical Language

CycL is a fully higher order and modal knowledge representation formalism[2], which makes it suitable for representing a wide range of natural language constructions. Cyc also allows the partition of knowledge into separate ‘microtheories' arranged in a subsumption hierarchy which enables the consistent management of contradictory information and the representation of context (e.g. statement of background assumptions). The strategy followed by our annotation systems is to parse input documents, rendering as much as currently possible into a CycL representation, to provide users with the opportunity, but not the necessity, to interactively disambiguate and elaborate the CycL representation, and then to project the resulting assertions onto the subset of representations allowed by the OWL language, yielding an XML annotation file.

Extracting the Text Content of target web pages

We use two packages from the Apache Project (CyberNeko,[3] and Xerces[4]) to convert an HTML document into a Document Object Model (DOM) as a Java Object. The application traverses the DOM tree, extracting the web page title, meta-description, and text leaf nodes. This will provide us with the ability, in future versions of the annotator, to tailor its focus onto salient content and cause it to ignore distractions (e.g. sidebars, menu items, advertisements, navigation links, and so forth often found with news articles). This will be a substantial improvement over simple web page text extractors, which apply the simple algorithm of stripping out HTML tags, thereby omitting most cues to salience and noise.

Chunking Input into Sentences, Phrases and Words

The second stage of the parsing pipeline populates a “TextDocument” object with sentences, phrases and words obtained from the web page’s DOM. Currently, we use the LINGUA sentence splitting module[5] to extract whole sentences from text strings, and the remaining text fragments are then organized as phases and words. All our web page annotation experiments to date have been conducted on English language documents, but, since the character set used for parsing is UTF-8, it should in principle be straightforward to apply this step of processing to other languages. Full processing of other languages will depend on extending the Cyc Lexicon beyond its rudimentary coverage outside English, and extending the segmentation and syntactic parsing infrastructure to handle a wider range of syntactic phenomena.

Natural Language Knowledge and English Parsing

Natural language processing in Cyc is supported by the Cyc Lexicon, an increasingly comprehensive collection of syntactic and semantic knowledge about English, and a framework in which knowledge about other languages can be embedded. The table below gives some indication of the current coverage.

	Noun	Verb	Adjective
CycL terms representing Lexemes	15450	4454	4716
Denotations	14442	1838	1640
Semantic Translation Patterns	464	3178	1787

CycL terms representing lexemes include `Burger-TheWord` and `Of-TheWord,` representing the English words “burger” and “of”, respectively; denotations connect word senses to KB concepts. For example,

```
(denotation Burger-TheWord CountNoun 0 HamburgerSandwich)
```

means that “burger”, when used in its first CountNoun sense, refers to a hamburger sandwich;

[2] The Cyc inference engine however currently only supports the first order fragment and some of the second order and modal extensions.

[3] http://www.apache.org/~andyc/neko/doc/html/

[4] http://xml.apache.org/xerces-j/

[5] http://people.brandeis.edu/~matthewg/cpan-lingua.html

```
  (verbSemTrans Venerate-TheWord 0 TransitiveNPCompFrame
         (feelsTowardsObject :SUBJECT :OBJECT Reverence highAmountOf)),
```

means that the word “venerate”, when used as the verb in a transitive verb frame taking an NP complement, should be understood in the Cyc logical language, CycL, as meaning that the agent denoted by the subject of the sentence feels a high degree of reverence towards the thing denoted by the object of the sentence. Similarly,

```
(nounSemTrans Bride-TheWord 0 GenitiveFrame
       (and
           (isa :NOUN FemaleHuman)
           (isa ?W WeddingEvent-Entire)
           (eventHonors ?W :NOUN)
           (eventHonors ?W :POSSESSOR)))
```

tells Cyc that, for example, “Frankenstein’s Bride” or “the bride of Frankenstein” should be interpreted as meaning that the bride is a female person, and that some wedding happened that honored both the bride and Frankenstein.

The third stage of the document annotation pipeline iterates over the sentences and phrases in the TextDocument object. Phrases are treated as whole sentences on the first pass. Each sentence is parsed by Cyc’s natural language parsing system, resulting in a list of CycL logical sentences. If the list is empty, then Cyc could not determine a semantic interpretation that covered the entire sentence, and if more than one CycL sentence is returned, then Cyc found one or more ambiguous concepts in the input natural language sentence. Typical performance for a parsing run on a news article is:

Total number of phrase parses attempted	210
Number of phrases for which a CycL translation was found	79
Average time to translate	5 seconds

On the second pass over the TextDocument object, Cyc’s word denotation parser processes the uninterpreted sentences, returning Cyc terms for lexically mapped words and phrases.

Parsing into Semantic Representations

Although a great deal of progress has been made over the past decade in the development of efficient syntactic parsers for natural languages, semantic parsers, which attempt to reach a detailed understanding of the NL input, have been less well studied and less successful. This may be due in part to the lack of a suitable target representation, for which the existence of PropBank[6] [Gildea and Palmer 2002], and, more recently, the availably of OpenCyc and ResearchCyc[7] may offer some relief. The lack may also be due to the difficulty of the process, since unlike syntactic parsing, semantic interpretation depends critically on solutions to difficult linguistic problems, including anaphor resolution, disambiguation, interpretation of metaphors, preposition interpretation, and quantification. It is therefore worth spending a little time to explain the progress we have made during our research and how we have deployed it within this application.

Suppose one is faced with a sentence like “Bill Clinton bought a house in New York”. The first step in interpretation is to perform a syntactic parse targeting the TreeBank tag set. For this prototype we made use of the parser developed by Eugene Charniak at Brown University [Charniak 2000][8]. This parser yields:

```
  [S [NP [NNP “Bill”] [NNP “Clinton”]]
       [VP [VBD “bought”]
              [NP [NP [DT “a”] [NN “house”]]
                     [PP [IN “in”]
                            [NP [NNP “New”] [NNP “York”]]]]]
```

[6] http://www.cis.upenn.edu/~ace/

[7] Open Cyc is a completely unrestricted subset of the Cyc KB and inference system, and includes a scaffolding taxonomy of approximately 60,000 terms that ensure interoperability with other Cyc KB versions. Research Cyc includes all of OpenCyc together with a large number of assertions and rules concerning the scaffolding terms; this high utility version of Cyc is currently in beta and will be available under a research purposes license.

[8] The system, however, is not dependent on the use of this parser; in a current research project our team is collaborating with Stanford University in an effort to achieve semantic parses of English and Chinese using the Stanford Parser (Klein and Manning 2003). We are also exploring the use of the CMU Link parser [Sleator and Temperley 1993].

From this parse, the system identifies the main verb, “bought” in this case, and finds its denotation in the KB (#$Buying) and the appropriate semantic translation pattern (SemTrans):

```
(and (isa :ACTION Buying)
        (buyer :ACTION :SUBJECT)
        (objectPaidFor :ACTION :OBJECT))
```

This is used, in turn to understand the argument structure of the syntactic parse. The syntactic subject,

```
 [NP [NNP “Bill”] [NNP “Clinton”]],
```

and the syntactic object,

```
[NP [NP [DT “a”] [NN “house”]]
                 [PP [IN “in”]
                        [NP [NNP “New”] [NNP “York”]]]
```

are isolated for the purposes of completing the retrieved SemTrans, and interpreted using the Cycorp-developed recursive noun phrase parser, for the base NPs (“Bill Clinton”, “house”, “New York”[9] in this case[10]), combined and compositional parsing of modifiers (“in New York”, in this case), producing the CycL interpretations #$BillClinton and

```
(and
     (isa ?HOUSE House-Modern)
     (in-Underspecified ?HOUSE  NewYork-State)).
```

Substituting these into the SemTrans, and replacing the remaining role key ‘:ACTION’ with an existentially qualified variable, yields the final CycL interpretation:

```
(thereExists ?ACTION
  (thereExists ?HOUSE
    (and (isa ?ACTION Buying)
        (buyer ?ACTION BillClinton)
        (objectPaidFor ?ACTION ?HOUSE)
        (isa ?HOUSE House-Modern)
        (in-Underspecified ?HOUSE NewYork-State))))
```

The rendering of the prepositional phrase as “in-Underspecified” represents a residual ambiguity which future versions of the system will attempt to resolve using background knowledge and discourse context[11]. The current system typically produces translations that render much of the sense of input sentences, but that omit some of the information they contain.

User Interaction in Annotating Partially Translated Documents.

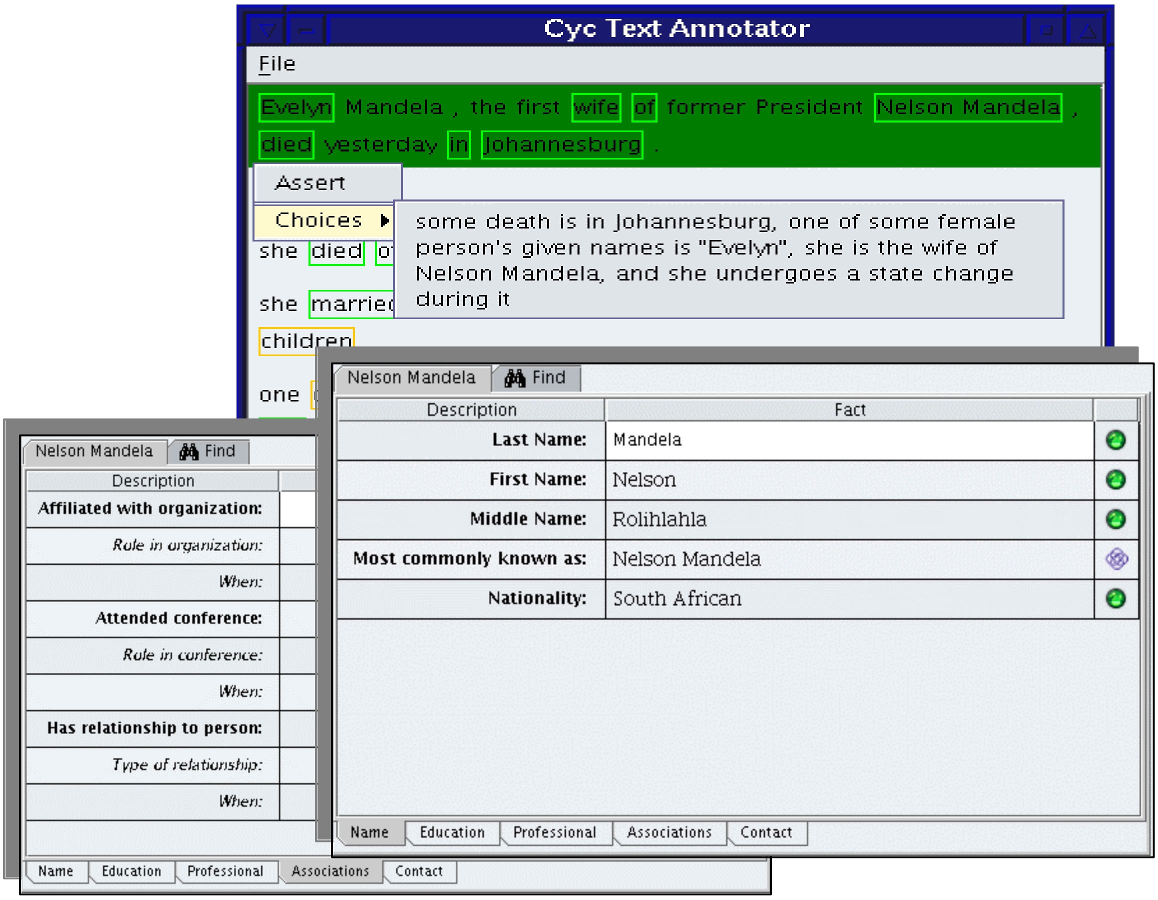
To help ameliorate some of the imperfections in the semantic translation process, the system provides the opportunity, but not the necessity, for users to interact with the current interpretation of a document, resolving ambiguities and adding additional information. Analyzed documents can be displayed in an interface that maintains correspondences between the text of the original document and the current logical interpretation. Fully interpreted terms in the document are highlighted in green; clicking on them takes the user to an appropriate “Factivore” knowledge acquisition form, allowing rapid knowledge entry in natural language. While some of the most commonly used forms have had their representation in the KB hand-crafted by knowledge engineers, the vast majority of forms are produced automatically by the system, using background knowledge and inductive inference over known cases. In experiments performed in the course

[9] Another possible interpretation is New York City. For this example, we assume a user has correctly disambiguated.

[10] In addition to being able to map single and multi-word tokens into CycL terms – e.g. "Bill Clinton" to #$BillClinton – the NP parser can interpret a wide variety of compound NPs, e.g. "Bronze age farmers" are farmers that were active during the Bronze age and "black leather jackets" are jackets made of leather and black in color.

[11] To the predicate #$objectFoundInLocation, in this case.

of entering knowledge about terrorists and their activities, lightly trained domain experts have achieved knowledge entry rates exceeding thirty facts[12] per hour using this interactive interface.
The other interactions available to users are selection from amongst interpretation alternatives (via menus rendered by the Natural Language Generation system) for terms highlighted in orange, and obtaining a complete English paraphrase of the current logical interpretation of a sentence, before it is asserted.



Evelyn Mandela , the first wife of former President Nelson Mandela , died yesterday in Johannesburg .

she died o

she marrie

children

one

some death is in Johannesburg, one of some female person's given names is "Evelyn", she is the wife of Nelson Mandela, and she undergoes a state change during it

Description	Fact
Last Name:	Mandela
First Name:	Nelson
Middle Name:	Rolihlahla
Most commonly known as:	Nelson Mandela
Nationality:	South African

Description
Affiliated with organization:
Role in organization:
When:
Attended conference:
Role in conference:
When:
Has relationship to person:
Type of relationship:
When:

Fig. 2: After the system has analyzed a document, it can be made available to the user for further annotation. Terms recognized within sentences are marked in green, if fully interpreted, and orange, if ambiguous to the system. Users can chose to resolve ambiguities in pull down menus, forcing reinterpretation of the affected sentence, or can leave the ambiguity intact. The current interpretation can be disclosed to the user by automatically paraphrasing it back into English, as shown in the pop up. More information can be provided about terms in the document, at the users whim, by accessing "Factivore" knowledge entry forms, which provide a rapid, NL mechanism for assertion into the knowledge base.

Asserting CycL Sentences into a Unique Cyc Microtheory

The fourth stage of the parsing pipeline asserts the CycL sentences and Cyc terms into a unique Cyc microtheory (context) within the knowledge base. The microtheory represents the propositional content of the target web page, and it is placed within the Cyc microtheory inheritance lattice so that commonsense assumptions about the target web page document are made explicit within Cyc. For example, a current

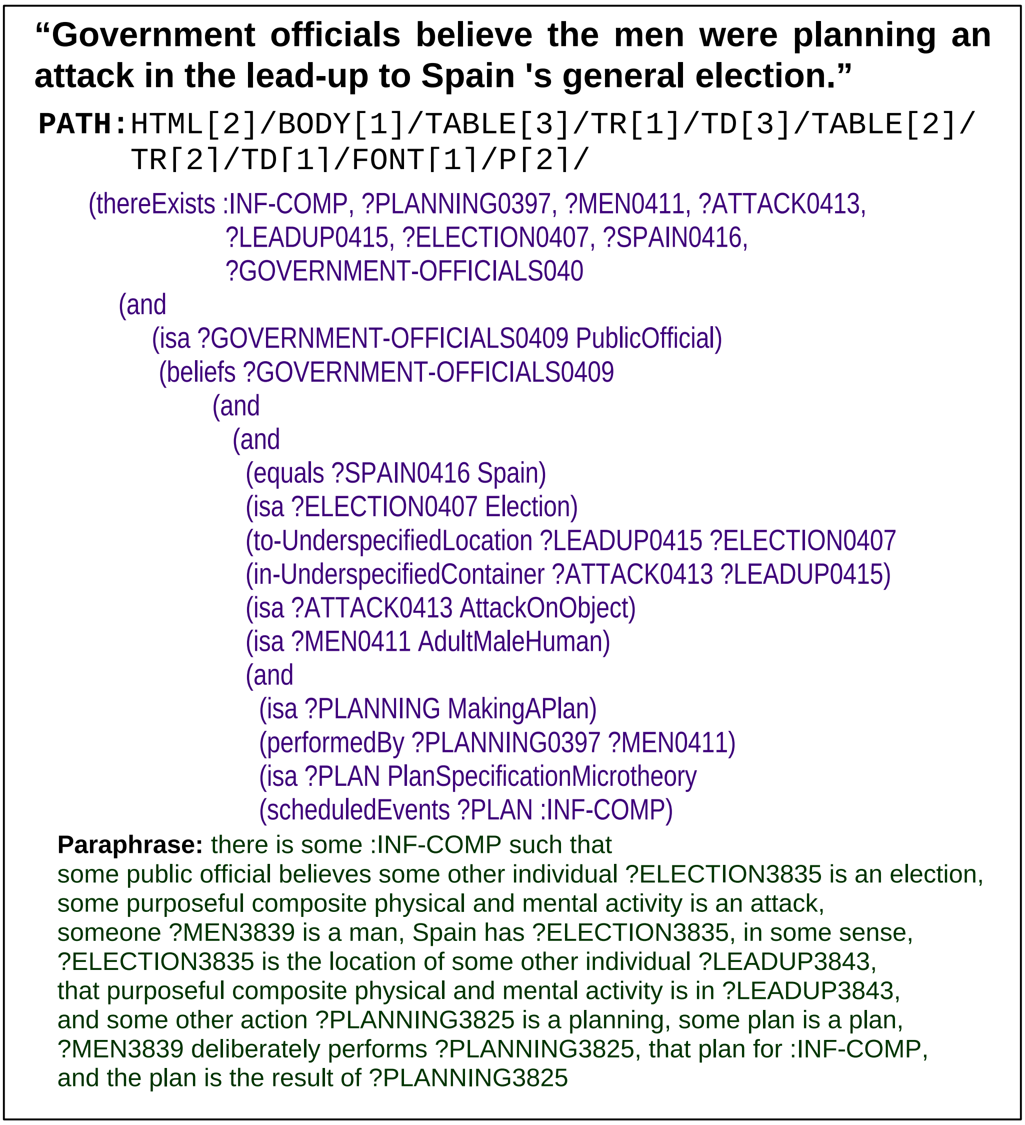
[12] A fact is a single assertion made into the Cyc KB. Facts can express simple concepts (such as "George W. Bush is a person") or more complicated concepts (such as "something is consumed during every eating event").

news article microtheory inherits rules and facts from Cyc's CurrentWorldDataCollectorMt. Existential variables are replaced by concrete terms during the CycL sentence assertion. Below is an assertion as parsed from the text "Bill Clinton bought a house in New York":

```
(thereExists ?ACTION
  (thereExists ?HOUSE
    (and (isa ?ACTION Buying)
         (buyer ?ACTION BillClinton)
         (objectPaidFor ?ACTION ?HOUSE)
         (isa ?HOUSE House-Modern)
         (in-Underspecified ?HOUSE NewYork-State))))
```

Replacing the existentially quantified variables with their skolem equivalents in the formula yields:

```
(and (isa Buying21 Buying)
    (buyer Buying21 BillClinton)
    (objectPaidFor Buying21 House-Modern22)
    (isa House-Modern22 House-Modern)
    (in-Underspecified House-Modern22 NewYork-State))))
```



"Government officials believe the men were planning an attack in the lead-up to Spain 's general election."

```
PATH:HTML[2]/BODY[1]/TABLE[3]/TR[1]/TD[3]/TABLE[2]/
     TR[2]/TD[1]/FONT[1]/P[2]/
```

```
(thereExists :INF-COMP, ?PLANNING0397, ?MEN0411, ?ATTACK0413,
             ?LEADUP0415, ?ELECTION0407, ?SPAIN0416,
             ?GOVERNMENT-OFFICIALS040
  (and
    (isa ?GOVERNMENT-OFFICIALS0409 PublicOfficial)
     (beliefs ?GOVERNMENT-OFFICIALS0409
         (and
           (and
             (equals ?SPAIN0416 Spain)
             (isa ?ELECTION0407 Election)
             (to-UnderspecifiedLocation ?LEADUP0415 ?ELECTION0407
             (in-UnderspecifiedContainer ?ATTACK0413 ?LEADUP0415)
             (isa ?ATTACK0413 AttackOnObject)
             (isa ?MEN0411 AdultMaleHuman)
             (and
               (isa ?PLANNING MakingAPlan)
               (performedBy ?PLANNING0397 ?MEN0411)
               (isa ?PLAN PlanSpecificationMicrotheory
               (scheduledEvents ?PLAN :INF-COMP)
```

Paraphrase: there is some :INF-COMP such that
some public official believes some other individual ?ELECTION3835 is an election,
some purposeful composite physical and mental activity is an attack,
someone ?MEN3839 is a man, Spain has ?ELECTION3835, in some sense,
?ELECTION3835 is the location of some other individual ?LEADUP3843,
that purposeful composite physical and mental activity is in ?LEADUP3843,
and some other action ?PLANNING3825 is a planning, some plan is a plan,
?MEN3839 deliberately performs ?PLANNING3825, that plan for :INF-COMP,
and the plan is the result of ?PLANNING3825

Figure 3: The result of translating one sentence of a document into CycL. These translations are often quite complex, and, as in this case, imperfect, but provide a good basis for editing the OWL representation into an accurate reflection of document semantics. The paraphrase is the result of automatic conversion of the CycL translation back into English, and is given as an aid to reading. Paraphrase into English is not present in the Cyc Annotator output.

Exporting CycL into OWL

The fifth and final stage of the web page annotation pipeline exports the document microtheory contents into an OWL XML document. All the built-in OWL Classes and properties have CycL equivalents. Here are sample rules for exporting some CycL predicates that happen to have built-in OWL definitions:

```
  #$disjointWith --> owl:disjointWith
  #$equals --> owl:sameAs
  #$genlPreds --> rdfs:subPropertyOf
  #$genls --> rdfs:subClassOf
  #$isa --> rdf:type
  #$TransitiveBinaryPredicate --> owl:TransitiveProperty
```

The sample CycL formula results in the following OWL RDF triples, with boldface to indicate the transformation of CycL predicates that are defined in Cyc's OWL ontology:

```
  <Buying rdf:ID="Buying21">
      <buyer rdf:resource="#BillClinton">
      <objectPaidFor rdf:resource="#House-Modern22">
  </Buying21>
  <House-Modern>
      <in-Underspecified rdf:resource="#NewYork-State">
  </House-Modern>
```

A portion of the OWL output for a particular news story is included in Figure 1, above. The primary difficulty in the OWL export process was the expressiveness limitation of OWL with respect to CycL. We overcame this by ensuring that the CycL assertions were ground atomic formulae, without functional terms and using only binary predicates. For cases such as rules, where the representation is not amenable to OWL export, we omit them from the OWL markup.

Conclusions and Future Work

The Cyc OWL annotator seeks to lower the barriers to the acceptance and growth of the semantic web by using the Cyc system to produce fully automatic, partial OWL markup for unrestricted text documents. This is done by applying lexical information and background knowledge from the Cyc knowledge base, subsystems for text analysis, optional interactive knowledge acquisition and disambiguation, isolation of incomplete knowledge within a microtheory structure, and down-projection of CycL logical representations into OWL.

One of the central thrusts of our research is improving the process of translation from unrestricted natural language text into full logical representations; over the next year we expect substantial improvements in the quality of English interpretation, and initial results for Chinese interpretation; these improvements should directly improve the resulting OWL annotations.

An independent research direction involves adding the ability for the system to optionally produce OWL extended with RuleML and other proposed extensions to the language of the semantic web, improving the quality of the output produced by down-projection from CycL. These extensions should be straightforward to produce once the relevant standards are adopted.

This work was supported by DARPA's DAML program, and used additional technology supported by ARDA's AQUAINT program and a Phase I SBIR grant.

References

Burns, Kathy J. and Anthony R. Davis. 1999. "Building and Maintaining a Semantically Adequate Lexicon Using CYC" in Evelyne Viegas, *Breadth and Depth of Semantic Lexicons.* Kluwer: Dordrecht.

Charniak, Eugene. 2000. "A Maximum-Entropy-Inspired Parser". *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics* (NAACL'2000), Seattle, Washington.

Gildea, Daniel and Martha Palmer. 2002. "The Necessity of Parsing for Predicate Argument Recognition" In *Proceedings of ACL 2002*, Philadelphia, PA.

Klein, Dan and Christopher D Manning. 2003. "A* Parsing: Fast Exact Viterbi Parse Selection." HLT-NAACL 2003, Edmonton, Canada.

Sleator, Daniel and Davy Temperley. 1993. "Parsing English with a Link Grammar". Third International Workshop on Parsing Technologies, Tilburg, The Netherlands and Durbuy, Belgium.