NLP Spring 2005
Cal Watson
Overview

- Data and the current WWW
- Data and The Semantic Web
- Gap between the current WWW and The Semantic Web
- NLP techniques needed to close the gap
- Current state and future work
The World Wide Web

- Data is stored in tagged HTML documents called Web pages
- Data is easily read by humans through the use of a browser
- Since HTML is an open standard, Browsers for the most part will render data tagged with HTML in the same way.
Advantages of HTML

- Easy to publish data in HTML
- Many automated tools
- HTML is an open standard so it is easy to write tools to parse and display HTML
- Great for presenting data in a human readable format
HTML and Knowledge Representation

- Consider the following two paragraphs:
  - `<p>The Twister 6 is a large (12 ft. x 10 ft. 4 in.) hexagonal dome style tent with a full coverage fly, 3 windows, and 3 noncloseable side wall vents for great ventilation. With exceptional waterproofing on floor and fly this is one great shelter. Continuous sleeves make set up quick and easy. 3 large doors make entry/exit easy for up to 6 people.</p>
  - `<p>TV weatherman Bill Harding is trying to get his tornado-hunter wife, Jo, to sign divorce papers so he can marry his girlfriend Melissa. But Mother Nature, in the form of a series of intense storms sweeping across Oklahoma, has other plans. Soon the three have joined the team of stormchasers as they attempt to insert a revolutionary measuring device into the very heart of several extremely violent tornados.</p>`
To a web browser these describe the same things
In fact web browsers have no clue what they are presenting to the user
Although these previous descriptions contained to keyword matches web browsers only distinguish data contained in different presentation tags
So....how can we make data smart?

- We can make data smart by providing metadata or information about data
- Example in NLP, POS tagging this is a form of infusing data with meta data
- This is the basic idea of the Semantic Web.
- Before continuing quick overview of things I want to focus on.
Overview of the Semantic Web

- Instead storing information in html, we want to store it in a machine readable and “understandable” format
- Data must be machine readable because we want computer programs to be able to process it easily
- We also need to incorporate inherent background knowledge that humans take for granted
- We can define background knowledge in an ontology
Ontologies

- So what exactly is an ontology?
- An ontology is a “specification of a conceptualization”
- In other words an ontology defines entities, relationships, classes, vocabulary, and rules within a specific domain knowledge
- Ontologies can also link to other ontologies to form larger ontologies
So what is an ontology good for?

- Ontologies can be used to make systems more intelligent
- Smarter information retrieval
- Support for faster better parsing
- Information extraction and semantic tagging
- Problems of knowledge management, information retrieval, word sense disambiguation can all be addressed using an ontology
Ontologies (cont)

- So how does one construct an ontology?
- There are many languages for constructing ontologies
- Most of them are defined in an RDF or XML syntax
- Host ontology languages are based on logic
- You must decide how detailed you want to be when developing an ontology
Example Ontology

- This is an example was taken from www.schemaweb.info
- It is an ontology that describes geographical places
- Example of ontology defining geographical entities
- Very verbose syntax example
- C:\Documents and Settings\Owner\Desktop\
So how does Natural language processing fit in?

Natural language processing is vital to the success of the semantic web because it is the method of communication between humans and software agents.

Parsing, knowledge representation, Information extraction, and semantic analysis are used in many semantic web technologies.
Word sense disambiguation

- Ontologies provide a way to add context to information.
- By specifying which an ontology an agent should use we can eliminate any ambiguities between words.
- We can take advantage of this technique in parsing and tagging.
Most information on the current Web is stored in natural language documents marked up with HTML.

To improve this situation we need tools that can parse and structure this info.

With the aid of ontologies we can parse unstructured documents more effectively.

Instead of rigid Context Free Grammars, we can use ontologies to provide a richer lexicon and even thesaurus.
Furthermore we can specify logical rules based on the ontology language to help with the parse.

So instead of trying every computing every possible parse we can apply the rules specified in the ontology to the terms we encounter in unstructured documents.
Knowledge Representation

- Ontologies are an excellent way to represent knowledge
- Ontologies can also be very difficult to construct by hand
- For the semantic web to be successful tools must be created to ease the burden of creating ontologies
- For these tools semantic analysis is extremely important
Inferencing

• For all our structured data to be useful we must give our software agents reasoning abilities
• Because many ontology languages are based on Logic, specifically Description logic it is easy to construct symbolic logical rules based on the semantics of the language
Inferencing (cont)

- For example consider the language construct equivalentClass defined in OWL
- With our agent we can define a rule of the form: $A \text{ equivalentClass } B \Rightarrow A == B$
- This will tell the agent that $A$ can be treated the same as $B$ in the structured document.
- Imagine a search engine that indexed logical assertions rather than keyword occurrences
Resources

- http://www.schemaweb.info/default.aspx
- http://www.w3.org/2004/OWL/#ontologies
- Frequently Asked Questions on Ontology Technology
- Juhnyoung Lee
- IBM T. J. Watson Research Center
- Hawthorne, NY
- jyl@us.ibm.com