Following is a scenario created in class on March 9, 2006 to illustrate an imagined trace of our target ReadTheWeb system. This primarily focuses on what knowledge the system has and how it uses it to extract information. It doesn't focus on learning methods to acquire that knowledge, though we certainly need to understand that too.

Input Goal to the system will be stated as the problem of populating a particular portion of a given ontology. For example the goal might be given as: find BioDepts, persons who are MemberOf(p,biodept), researchArea(p,r), publications(p,paper), etc.

Assume we start with a lot of facts in the knowledge base (KB), lots of general rule-based knowledge (e.g., coauthors are often in the same dept).

- determine which of these types, relations is most reliably extractable.

Find starting text to process, using one of several strategies:

- 1. Searches web for "biology department"
- 2. crawl and find pages with lots of medline pointers
- ...
- 3. searches for "biology department institution ..."
- Classify returned URLs as bio or not (Richard's project 1)
- Crawl dept and classify each page as bio or not.
- [previously bootstrap trained a la Richard's project 1]
- 4. starting from known names of proteins, conferences, search for papers by poorly-characterized author, on proteins, in relevant conferences.
- 5. search for papers by 'gang of 3' and see if they're in same dept, if so, probably bio.

Finds "Pitts bio dept" page, decides (classifies?) this as bio dept home page
[uses Richard's URL classifier from his first project to classify this as biology page]

Crawls this site [using WIT to crawl, staying with URLs with prefix www.bio.pitt.edu]

Finds page which it classifies as "faculty homepage", extracts [how?] MaryS as the faculty.
[one strategy to extract home page owner: "if first tokens on page are a person name, assume this is the home page of that person." Might need to input this rule manually because no projects are planning to learn rules of this form]

Extract EnglishNames(MaryS, "Mary", "Mary Smith", "Professor Smith")
[co-reference resolution across the homepage to decide which NPs refer to MaryS]

Extracts from this page the belief that AuthorOf(MaryS, pap1). The citation to pap1 on this page also allows extracting AppearsIn(pap1, ISMB2005).

Finds on another page there is a mention of MaryS, and a publication associated with her.
[co-reference across multiple documents decides both are referring to the same MaryS, based in part on knowledge EnglishName(MaryS, “Mary Smith”, “prof smith”). Determined by the co-reference project team]

Continues by reading well-formed sentences on this page, which include a description of the research of MaryS.
[project semanticRoleLabeling: sentence-level analysis to extract that protein ACG is inhibited by transcription factor XYZ]

[project EntityAssociation: finds the protein ACG mentioned on this page, so adds the belief AssociatedWith(MaryS,ACG), and because MaryS is of type person adds ExptertiseOf(MaryS, ACG). Also notices that ACG frequently co-occurs with Mitochondria, so (probably) Expertise(MaryS, Mitochondria).

Expertise(MaryS, yeast), CoOccursWith(yeast, PTP), therefore more likely that Expertise(MaryS, PTP)]

Goes inside the paper to find protein names and relations among these.
[projects relevant here include all those dealing with extraction for full grammatical text: semantic role labeling, relation extraction, co-reference resolution, ...]

Decides “it” in “It is synthesized when gene xyz is expressed.” refers to ACG
[CO-REFERENCE RESOLUTION USED HERE]

[SYNERGY: RELATION EXTRACTION HAS RULE THAT APPLIES TO ABOVE SENTENCE TO EXTRACT PROTEIN_GENE(IT, XYZ). USING CO-REF RESOLUTION, CHANGES THIS TO PROTEIN_GENE(ACG,XYZ)
...

Kb contains knowledge that “advisees and advisors are affiliated usually with the same university” We extract InstitutionOf(JoeJ, UPitt), and AdvisorOf(MaryS, JoeJ), therefore suspect that InstitutionOf(MaryS, UPitt).

==========
To learn the above kind of knowledge

[Projects relationLearning and ActiveLearning will bootstrap, and use limited trainer input, to learn extraction rules.]

Possibly restrict relation learning to rules of the form:
If sentence is of the form .... Then R(NP1,NP2)
Where NP1, NP2 are noun phrases within the same sentence

==========

Needs:

- focus of attention / top level control

Notes:

- many of these ‘facts’ and ‘relations’ were extracted not from sentences, but other HTML structure.

- some facts such as “worksOn(author,protein)” more easily extracted from paper citations than from running text.

- Include in the system lots of hand-written rules of unknown reliability. System can then try to learn the reliability of (and perhaps refinements to) these rules, using redundant sources and active learning.

- Possibly populate the knowledge base (KB) with a starting fact base, then mine that for regularities that can be used by the datamining project

- See Wikipedia set of features used to describe people

- Wikipedia is a good source of full-sentence, grammatical text. So are the faculty home pages at http://www.pitt.edu/~biology/

Synergy

- coref resolution and relation learning (leverage coreference)