Introduction to Natural Language Processing

Introduction

Today's Outline

Course Overview

Administration

Natural Language Processing

- The field of *Natural Language Processing (NLP)*, or *Computational Linguistics (CL)*, or *Human Language Technology (HLT)*, is primarily concerned with the creation of computer programs that perform useful and interesting tasks with human languages (e.g., understanding, generation, learning).
- It is secondarily concerned with using computational metaphors to help us come to a better understanding of human language.
- The foundations of the field are in computer science (e.g., AI, theory), linguistics, mathematics and statistics, electrical engineering, and psychology.
- Studying NLP involves studying natural languages, formal representations, and algorithms for their manipulation.

Major Topics of this Course

Knowledge of Language

- words (meaningful components of words)
- syntax (structure of a sentence)
- semantics (explicit meaning of a sentence)
- pragmatics/discourse (implicit/contextual meaning)

Methodologies and Tools

- knowledge-based and statistical
- state machines, rule systems, grammars, logic, search, probability, automata, machine learning, and more

Applications

- from hyphenators to intelligent agents

Current Research

Knowledge of Language

- Example dialogue from *2001: A Space Odyssey*

```
Dave: Open the pod bay doors, HAL.
HAL: I'm sorry Dave, I'm afraid I can't do that
```

- To participate in such a conversation, HAL needs knowledge about many levels of language
  - words: producing contractions, plurals
  - syntax: questions versus statements, word order and grouping
  - semantics: meaning of words in isolation and compositionally
  - pragmatics: politeness and indirectness
  - discourse: between utterance references

Applications

What makes a computer application a language processing application?

- language processing applications require the use of knowledge of language

Little Applications

Little applications typically make use of only a small amount of a single kind of knowledge of language. Many of these types of applications are currently in use, but they are often nearly invisible.

- line breakers
- hyphenators
- spelling correctors
- OCR software
- grammar and style checkers

Big Applications

Big applications often make use of large amounts and varied kinds of knowledge of language.

- information retrieval and extraction
- question answering
- dialogue management
- text summarization
- machine translation
- HAL

Some Current Application Scenarios

Tutor students in areas such as physics

Access the web over the telephone

Speak to your appliances

Grade essays

Generate running commentary for robotic soccer

Generate weather reports in multiple languages

Demos

Dialogue Systems

- Why2-Atlas/ITSPOKE (Pitt; email to be a subject!)
- TOOT / NJFUN (Litman et al. at AT&T)
- JUPITER (1-888-573-TALK)

Question Answering

- AnswerBus (www.answerbus.com/news)
- Ask Jeeves (www.ask.com)
- Opinion Highlighter (Pitt)

Summarization

- Newsblaster (www.cs.columbia.edu/nlp/newsblaster)
- NewsInEssence (www.newsinessence.com/nie.cgi)

Machine Translation

- Babelfish (babelfish.altavista.com)

Speech Synthesis

- Festival Text-To-Speech (festvox.org/voicedemos.html)

Administration

Professor

Teaching Assitant

Students

Textbook

Web page

Requirements

Who should be here

Dr. Diane Litman

Affiliations

- Associate Professor, Computer Science Department
- Research Scientist, LRDC
- Faculty, Intelligent Systems Program
- Member, CIRCLE

Contact Information

- 5105 Sennott Square, (412) 624-8838
- 741 LRDC, (412) 624-1261
- litman@cs.pitt.edu
- http://www.cs.pitt.edu/~litman

Office Hours

- M 12:15-2:15 (5105 Sennott Square)
- Tu 10:00-12:00 (741 LRDC)

Litman, cont.

Education

- 1986: PhD, Computer Science, University of Rochester
- Thesis: Plan Recognition and Discourse Analysis: An Integrated Approach for Understanding Dialogues

Employment

- 2001-present: University of Pittsburgh
- 1985-2001: Technical Staff, AI Principles Research Department., AT&T Labs - Research (formerly AT&T Bell Laboratories)
- 1990-1992: Assistant Professor, Computer Science, Columbia University

2000-2003: Chair, North American Chapter of the Association for Computational Linguistics (NAACL)

Litman, cont.

Research

- Natural Language Processing
  - Spoken Dialogue for Intelligent Tutoring Systems (with Prof. Vanlehn and LRDC)
  - Typicality and Natural Language Learning (with M. Rotaru)
  - Multiple Perspective Question Answering (with Prof. Wiebe, T. Wilson)
  - Predicting Medical Reasoning Codings in Pathology Protocols (with Prof. Crowley, L. Ma)
  - other discourse, dialogue, pragmatics, spoken language, learning
- Natural Language Processing Laboratory
  - http://www.cs.pitt.edu/~litman/nlplab.html
- Other Artificial Intelligence
  - knowledge representation, plan recognition, user modeling

Teaching Assistant

Mihai Rotaru

- Doctoral Student, Computer Science Department

Contact Information

- 5420 Sennott Square, (412) 624-8462
- mrotaru@cs.pitt.edu
- http://www.cs.pitt.edu/~mrotaru

Office Hours

- MW 9.00 - 11.00
- Th 11.00-12.00

Students

Please introduce yourself

The Text

Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, by Daniel Jurafsky and James H. Martin; Prentice-Hall, 2000

- http://www.cs.colorado.edu/∼martin/slp.html
- should be available from the bookstore
- also available from online providers

Course emphasis will be from a "computer science" perspective

Text will be supplemented with a selection of current research papers

Course Web Page

URL

- www.cs.pitt.edu/~litman/courses/cs2731/cs2731.html

Syllabus

- topics
- readings
- assignments
- lecture notes
- announcements
- NOTE: involves viewing/printing postscript, pdf, ppt, etc.

Requirements

Readings (before class!)

Homeworks

- problem sets
- writing programs
- using programs

Exams

- midterm
- final

Project

Leading and participating in research paper discussions

Prerequisites

An interest in language, and . . .

- ability to write and use programs
- background in computer science, linguistics, or speech
- some basic exposure to logic, probability, algorithm analysis, and/or programming language theory would be helpful

For Next Time

Read Chapter 1

- available online

Get a CS account

- http://www.cs.pitt.edu/userdb/acct_app_page.html

Send me email for a class mailing list

Assignment

- try at least one listed demo from each category, and be prepared to report back on your experiences

Look at Reading List

Survey

Name:

Email:

Department:

Year:

Undergraduate Major:

Relevant Courses:

Relevant Research Experience:

Methodologies (state machines, rule systems, grammars, logic, search, probability, automata, machine learning, etc.):

Programming Languages:

Operating Systems:

Goals: