

# For Friday

- Finish chapter 23
- Homework
  - Chapter 23, exercises 8 and 9
  - For 8, you need to do 4 different rounds of translation: 2 with the first language, then start over with the original 5 sentences and do two rounds with the second language. Be sure to pick dissimilar languages (Spanish and Italian are not a good pair).

# Program 5

- Any questions?

# Semantics

- Need a semantic representation
- Need a way to translate a sentence into that representation.
- Issues:
  - Knowledge representation still a somewhat open question
  - Composition  
“He kicked the bucket.”
  - Effect of syntax on semantics

# Dealing with Ambiguity

- Types:
  - Lexical
  - Syntactic ambiguity
  - Modifier meanings
  - Figures of speech
    - Metonymy
    - Metaphor

# Resolving Ambiguity

- Use what you know about the world, the current situation, and language to determine the most likely parse, using techniques for uncertain reasoning.

# Discourse

- More text = more issues
- Reference resolution
- Ellipsis
- Coherence/focus

# Survey of Some Natural Language Processing Research

# Speech Recognition

- Two major approaches
  - Neural Networks
  - Hidden Markov Models
    - A statistical technique
    - Tries to determine the probability of a certain string of words producing a certain string of sounds
    - Choose the most probable string of words
- Both approaches are “learning” approaches

# Syntax

- Both hand-constructed approaches and data-driven or learning approaches
- Multiple levels of processing and goals of processing
- Most active area of work in NLP (maybe the easiest because we understand syntax much better than we understand semantics and pragmatics)

# POS Tagging

- Statistical approaches--based on probability of sequences of tags and of words having particular tags
- Symbolic learning approaches
  - One of these: transformation-based learning developed by Eric Brill is perhaps the best known tagger
- Approaches data-driven

# Developing Parsers

- Hand-crafted grammars
- Usually some variation on CFG
- Definite Clause Grammars (DCG)
  - A variation on CFGs that allow extensions like agreement checking
  - Built-in handling of these in most Prologs
- Hand-crafted grammars follow the different types of grammars popular in linguistics
- Since linguistics hasn't produced a perfect grammar, we can't code one

# Efficient Parsing

- Top down and bottom up both have issues
- Also common is chart parsing
  - Basic idea is we're going to locate and store info about every string that matches a grammar rule
- One area of research is producing more efficient parsing

# Data-Driven Parsing

- PCFG - Probabilistic Context Free Grammars
- Constructed from data
- Parse by determining all parses (or many parses) and selecting the most probable
- Fairly successful, but requires a LOT of work to create the data

# Applying Learning to Parsing

- Basic problem is the lack of negative examples
- Also, mapping complete string to parse seems not the right approach
- Look at the operations of the parse and learn rules for the operations, not for the complete parse at once

# Syntax Demos

- [http://nlp.cs.berkeley.edu/Main.html#research\\_overview](http://nlp.cs.berkeley.edu/Main.html#research_overview)
- <http://www2.lingsoft.fi/cgi-bin/engcg>

# Language Identification

- <http://rali.iro.umontreal.ca/>

# Semantics

- Most work probably hand-constructed systems
- Some more interested in developing the semantics than the mappings
- Basic question: what constitutes a semantic representation?
- Answer may depend on application???

# Possible Semantic Representations

- Logical representation
- Database query
- Case grammar

# Distinguishing Word Senses

- Use context to determine which sense of a word is meant
- Probabilistic approaches
- Rules
- Issues
  - Obtaining sense-tagged corpora
  - What senses do we want to distinguish?

# Semantic Demos

- <http://www.cs.utexas.edu/users/ml/geo.html>
- <http://www.cs.utexas.edu/users/ml/rest.html>
- <http://www.ling.gu.se/~lager/Mutbl/demo.html>

# Information Retrieval

- Take a query and a set of documents.
- Select the subset of documents (or parts of documents) that match the query
- Statistical approaches
  - Look at things like word frequency
- More knowledge based approaches interesting, but maybe not helpful

# Information Extraction

- From a set of documents, extract “interesting” pieces of data
- Hand-built systems
- Learning pieces of the system
- Learning the entire task (for certain versions of the task)
- Wrapper Induction

# Question Answering

- Given a question and a set of documents (possibly the web), find a small portion of text that answers the question.
- Some work on putting answers together from multiple sources.

# QA Demos

- [http://l2r.cs.uiuc.edu/~cogcomp/qa\\_news\\_demo.php](http://l2r.cs.uiuc.edu/~cogcomp/qa_news_demo.php)
- <http://demos.inf.ed.ac.uk:8080/qualim/>

# Text Mining

- Outgrowth of data mining.
- Trying to find “interesting” new facts from texts.
- One approach is to mine databases created using information extraction.

# Pragmatics

- Distinctions between pragmatics and semantics get blurred in practical systems
- To be a practically useful system, some aspects of pragmatics must be dealt with, but we don't often see people making a strong distinction between semantics and pragmatics these days.
- Instead, we often distinguish between **sentence** processing and **discourse** processing

# What Kinds of Discourse Processing Are There?

- Anaphora Resolution
  - Pronouns
  - Definite noun phrases
- Handling ellipsis
- Topic
- Discourse segmentation
- Discourse tagging (understanding what conversational “moves” are made by each utterance)

# Approaches to Discourse

- Hand-built systems that work with semantic representations
- Hand-built systems that work with text (or recognized speech) or parsed text
- Learning systems that work with text (or recognized speech) or parsed text

# Issues

- Agreement on representation
- Annotating corpora
- How much do we use the modular model of processing?

# Summarization

- Short summaries of a single text or summaries of multiple texts.
- Approaches:
  - Select sentences
  - Create new sentences (much harder)
  - Learning has been used some but not extensively

# Machine Translation

- Best systems must use all levels of NLP
- Semantics must deal with the overlapping senses of different languages
- Both understanding and generation
- Advantage in learning: bilingual corpora exist--but we often want some tagging of intermediate relationships
- Additional issue: alignment of corpora

# Approaches to MT

- Lots of hand-built systems
- Some learning used
- Probably most use a fair bit of syntactic and semantic analysis
- Some operate fairly directly between texts

# Generation

- Producing a syntactically “good” sentence
- Interesting issues are largely in choices
  - What vocabulary to use
  - What level of detail is appropriate
  - Determining how much information to include