CSI 5180: Topics in AI: Natural Language Processing, A Statistical Approach

Instructor: Diana Inkpen
e-mail: diana@site.uottawa.ca

Preliminaries

Why study Natural Language Processing (NLP)?

- NLP is a very important current area of investigation as it is necessary to many useful applications.
- These applications include: information retrieval, extraction, and filtering; intelligent Web searching; spelling and grammar checking; automatic text summarization; pseudo-understanding and generation of natural language; and multi-lingual systems including machine translation.

Linguistics

- What kind of things people say?
  - how do people acquire, produce, and understand language
- What do these things say/ask/request about the world?
  - how to connect utterances to the world

NLP and related terms

- Natural language processing (NLP) = manipulation, processing, "understanding" of natural language text or utterances. Not necessarily full-blown AI or "language understanding the way people do it".
- Language engineering = Building systems that apply the techniques of NLP; has an emphasis on the creation of large systems, software engineering
- Computational linguistics (CL) = Research side of NLP, including relevant parts of AI, linguistics, and cognitive science.

Why study NLP Statistically?

- Up until the late 1980's, NLP was mainly investigated using a ***rule-based*** approach.
- However, rules appear too strict to characterize people's use of language.
- This is because people tend to stretch and bend rules in order to meet their communicative needs.
- Methods for making the modeling of language more accurate are needed and ***statistical methods*** appear to provide the necessary flexibility.

Subdivisions of NLP

- ***Parts of Speech and Morphology*** (words, their syntactic function in sentences, and the various forms they can take).
- ***Phrase Structure and Syntax*** (regularities and constraints of word order and phrase structure).
- ***Semantics*** (the study of the meaning of words (*lexical semantics*) and of how word meanings are combined into the meaning of sentences, etc.)
- ***Pragmatics*** (the study of how knowledge about the world and language conventions interact with literal meaning).

Topics Covered in this course

- ***Studying Words:***
  - Collocations
  - N-gram Models
  - Word Sense Disambiguation
  - Lexical Acquisition
- ***Studying Grammars:***
  - Markov Models
  - Part-of-Speech Tagging
  - Probabilistic Grammars
  - Parsing

Applications: Information Retrieval, Text Categorization, Statistical Alignment and Machine Translation

Tools and Resources Used

- ***Probability/Statistical Theory***: Statistical Distributions, Bayesian Decision Theory.
- ***Linguistics Knowledge***: Morphology, Syntax, Semantics and Pragmatics.
- ***Corpora:*** Bodies of marked or unmarked text to which statistical methods and current linguistic knowledge can be applied in order to discover novel linguistic theories or interesting and useful knowledge organization.

Course Requirements

- 2 written and programming assignments (20% each)
- An in-class presentation of a current research paper (15%)
- Class participation (5%)
- A Final Project (40%)

Textbook and other useful information

- ***Foundations of Statistical Natural Language Processing***, by Chris Manning and Hinrich Schütze, MIT Press, 1999.
- Current literature available from the Web will be used for class presentations.
- ***Course Website***: http://www.site.uottawa.ca/~diana/csi5180/
- Check the class website for a companion website for the textbook and other statistical NLP resources.