CMPT-413: Computational Linguistics

Anoop Sarkar
anoop@cs.sfu.ca
www.sfu.ca/~anoop/courses/CMPT-413-Spring-2003.html

- goals of the course
  - hopefully convince you that understanding language is an interesting problem,
  - give insight into various algorithms including some for machine learning that will be generally useful
  - hands-on experience with non-trivial datasets
  - formal models that will be useful in other fields of inquiry
  - useful background for: building software that deals with language (demand for this is increasing every day); artificial intelligence; cognitive science; machine learning.

- flavor of the course: emphasis on models and algorithms
  - after relevant background is covered the main focus of the course will be on working through examples in class
  - these examples will be prototypical of what you will be tested on in your exams
  - hands-on experience with algorithms in the homeworks
  - by doing the required readings and preparing questions for class you will learn more and also prepare for the exams better (might be obvious but it never hurts to remind you of this)
  - first few minutes of each course will be available for questions on material from the previous class

- tour of the course web page:
  - my office and office hours for this course
  - TA, TA office location and office hours
  - URL of web page and where to check the readings for each week
  - announcements section for important dates

- schedule and outline of topics to be covered:
  go over the syllabus from the web page

- required vs. optional readings:
  - required textbook: Jurafsky & Martin
  - syllabus is listed on web page
  - the required readings are all taken from Jurafsky & Martin
  - several review and optional readings are taken from other sources such as: Sipser, Programming Perl, and Manning & Schütze
  - In particular, the readings from Sipser are included for those who wish to review definitions and results from formal language theory which are used in this course

- how to do the readings (work out examples for each topic): write down at least a couple of questions for me.

- homeworks and exams
  - 3 homeworks before mid-term
  - 2 homeworks after mid-term
  - 1st homework will have no programming due to the fact that the csil labs are scheduled to open in late jan

- `perl`
  - scripting language – different from the compiled sort, quick prototyping or even for building systems
  - useful for text processing due to regexps
  - ideal for linguistic data and popularly used for web-based text processing
  - some people prefer `python`, also a useful language for these kinds of tasks. however, the most powerful regexp implementation is still only part of `perl`
  - csil labs will assume that you are using `perl` on `linux` – you are welcome to install and use `perl` on whatever platform you like but

your code cannot contain system specific commands, i.e. they should run on any system without any changes

- `www.perl.com` – contains the perl development system for many platforms, `perl` is typically installed in most platforms such as `linux`, `*bsd`, MacOS X

Natural Language Processing (NLP)

- NLP is the application of the theory of computational linguistics
- Humans use language to communicate with each other
- Perhaps we can build programs that can be more useful to us by "listening" in . . .
- The AI challenge: language will have to play a large part in any mimic
- Lots of data available: newswire, the web

NLP: what's it good for?

- Many useful applications in text: from spam detection to information extraction from large collections
- Many useful applications in speech: transcription, speaker identification
- NLP provides a challenging testbed for Machine Learning algorithms (high dimensionality, complex classification, sparse data)
- Cognitive Science: sentence processing

NLP on text: some common applications

- Information Retrieval
- Named entity recognition
- Information Extraction

NLP on text: some common applications

- Summarization
- Document Classification (spam detection, search engine IR, . . .)
- Machine Translation

NLP on text: some common applications

- Cross Language Information Retrieval
- Language Understanding (Parsing)
- Language Generation

NLP on text: some common applications

- Question Answering
- Knowledge Acquisition (building a dictionary, thesaurus, . . . automatically)
- Improving Speech Recognition (better language models)

$$\arg\max_{w_i} Pr(w_i \mid w_0, \ldots, w_{i-1})$$

NLP on speech: some common applications

- Speaker identification
- Speech recognition and transcription
- Text to speech synthesis
- Speech to speech translation

NLP on speech: some common applications

- Spelling correction, accent restoration, correcting speech recognized or OCR text
- Dialog Systems (call centres)
- Multi-modal dialog systems (AT&T, smartkom)

NLP on text: some uncommon applications

- Plagarism Detection
- Automatic evaluation of test essays – ETS
- Author identification

  Frederick Mosteller and David Wallace, *Inference and Disputed Authorship: The Federalist Papers*, Addison Wesley, 1964
- NLP for biological sequences (finding genes, predicting protein folding characteristics)