Computational Linguistics (CL)

- A.k.a. Natural Language Processing (NLP), Language Engineering
- Study of making computers use human languages for specific tasks.
  - Conversational Agents
  - Machine Translation
  - Automatic Speech Recognition/Synthesis
  - Automatic Summarization
  - Text Categorization
  - Information Retrieval
  - Text/Dialogue Segmentation
  - Intelligent Tutoring Systems
  - ...
Some CL/NLP Problems

- Speech Recognition:
  - People’s voices and accents vary greatly.
  - Many words/phrases sound similar, e.g., he’s, his; hour, our; striker, strike her.
- Part-of-Speech Tagging:
  - Many words belong to multiple parts of speech.
    - Fruit flies like a banana.
    - Fruit/NN flies/NN like/V a/DT banana/NN ./.
- Syntactic Parsing:
  - Even though grammar rules are finite, the combinatorial possibilities of rules make an extremely complex system.
- Semantic Analysis:
  - What is the best way to represent meaning?
  - How much knowledge is needed for semantic representation?

Some CL/NLP Problems

- Dialogue Agents:
  - How are human dialogues structured?
  - How do we know what to say next?
- Automatic Summarization:
  - How do we find more salient sentences in a text that are representative of the content?
- Text Data Mining:
  - How do we discover new knowledge from large collections of text?

Career Opportunities in CL/NLP

- Grammar Writer
- Knowledge Engineer
- Computational Lexicographer
- Language Engineer
- Academic Positions
Forensic Linguistics

- The practice of finding information from linguistic data useful to law enforcement or for presentation at court.
  - Forensic Phonetics
  - Authorship Analysis
  - Statement Analysis

Forensic Phonetics

- Speaker Identification
  - Who said it?
- Transcription
  - What did the speaker say?
- Authentication of recordings
  - Is the recording genuine or has it been tampered with?
- Language or Accent Identification
  - Where is the speaker from?

Authorship Analysis

- Authorship Attribution
  - Ransom letters
  - Hate mail
  - Any other document
- Similarity Analysis; Stylometry
  - Detecting plagiarism
- Authorship Profiling
  - Determining author’s cohort

Authorship Analysis

- Finding author’s linguistic “fingerprint.”
  - Assumption: Each author has a set of subconscious writing habits that are constant over time.
Authorship Analysis

- Finding author’s linguistic “fingerprint.”
  - Assumption: Each author has a set of subconscious writing habits that are constant over time.

Authorship Profiling

- The Anthrax Scare:
  - Uses “-” in dates
  - Uses block letters
  - Addresses slanted
  - Quotes Russian style, Ns resemble Russian I
  - Fakes being Russian
  - Many clues pointing to the author being an American Scientist.

  [Image of a letter with a URL: http://www.anthraxinvestigation.com/AnthraxFoster.pdf]

Statement Analysis

- Detecting deception
  - Word definitions
    - You know, I’m trying to be as honest as possible.
  - Grammar
    - A man pointed a gun at me and told me to get in the car.
  - Research
    - Things people typically say when they lie.
      - I swear to God. I only had two drinks. There were three men that attacked me.

  [URL: http://www.statementanalysis.com/]

Thank You!