Welcome to
Natural Language Processing
Setting Goals
CSCI-GA.2590 – Lecture 1

Ralph Grishman
Centrality of Natural Language

- a primary (and natural) mode of human communication
- representation for most recorded human knowledge
- a very rich and flexible representation (when compared to most formal representations)
Applications of NLP

- machine translation
- dictation
- document retrieval
- question answering
- information extraction
- personal assistant
- chatbot
Machine Translation

• long history of development
  – Warren Weaver’s 1949 memo: similarity to cryptography
  – early driver of NLP research
  – difficulty initially underestimated
  – benefited from machine learning methods
    • statistical MT (1990)
    • deep learning (2015)
  – quality depends on similarity of language structures (Chinese to English much harder than French to English)
Dictation

• Effective dictation systems rely on strong expectations about what will be said next:

  a language model

= P(current word | preceding words)
Document Retrieval (Web Search)

• For English can be based on word matching
• For languages with richer morphology, *morphological analysis* is required
Question Answering

• Step beyond passage retrieval: providing a direct answer from a text corpus or the Web
• Much harder than document retrieval because facts can be expressed many different ways
• Opportunistic approach
  – Return direct response if one can be found
  – Else return passage
• IBM Watson
Information Extraction

• convert unstructured text to structured data for further processing, capturing selected information

examples:

• name tagger / wikifier: **DBpedia/Spotlight**
• Relation/event classifier:
  – **OpenCalais**
  – **EuropeanMediaMonitor**
Personal Assistant
Relation to Other Fields
Linguistics

• goal of linguistics is to describe language
  – provide simple models which can predict language behavior
  – understand what is *universal* about language
  – through these formal models, understand how language can be *acquired*
Linguistics

• formal models from linguistics have been of value in NLP, but its goals are not the same as NLP:
• a single counterexample can invalidate a model as a linguistic theory, but would not significantly lessen its value for NLP
• NLP must address all phenomena which arise in an application, while linguistics may focus on select phenomena which give insight into the language faculty
Classical [Symbolic] AI

- classical 'symbolic' AI is concerned primarily with generic problem solving strategies and suitable knowledge representations
- there is an inherent link between AI and NLP: some NLP problems require the sort of deep reasoning addressed by these AI methods
- but NLP (and AI) has found increasing success through avoiding deep reasoning and turning instead to
Machine Learning

- early NLP systems (before 1990) were purely symbolic and handcrafted
- statistical methods and models have become more widely used in NLP since the mid 1990’s
- easily trainable and easily computable models have for some NLP tasks proven much more effective than more complex hand-crafted models
- furthermore, they have become more attractive now that lots of training data is available (on the web)
- The past few years have seen the rapid growth of neural network (DeepLearning) models for NLP, achieving better performance than earlier models (such as log-linear models)
Course Goals and Structure

• Many of these applications have a common need to convert the unstructured text or speech input to a structured form which reflects the meaning of the input.

• This is done in a series of processing steps:
  – morphological analysis
  – part of speech tagging
  – parsing
  – semantic analysis
  – reference resolution
• We will look at these stages in turn
• Most of this will be application independent, but when we need a specific application, we will focus on information extraction (an NYU specialty)
Administrative

- assignments
  - 8 or 9
  - a mix of paper-and-pencil and programming
- term project
  - may do individually or in a group
- textbook
  - Jurafsky and Martin: second and draft 3rd editions
- TAs
- final exam
- grading:
  - 40% assignments
  - 30% term project
  - 30% final