ISE 540 Text Analytics

Mayank Kejriwal
Research Assistant Professor/Research Lead
Department of Industrial and Systems Engineering
Information Sciences Institute
USC Viterbi School of Engineering
kejriwal@isi.edu

Advanced NLP Tasks

Information Extraction (IE)

- Identify phrases in language that refer to specific types of entities and relations in text.
- Named entity recognition is task of identifying names of people, places, organizations, etc. in text.

  people organizations places
  - Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.
- Relation extraction identifies specific relations between entities.
  - Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.

Question Answering

- Directly answer natural language questions based on information presented in a corpora of textual documents (e.g. the web).
  - When was Barack Obama born? (*factoid*)
    - August 4, 1961
  - Who was president when Barack Obama was born?
    - John F. Kennedy
  - How many presidents have there been since Barack Obama was born?
    - 9

Text Summarization

- Produce a short summary of a longer document or article.
  - Article: With a split decision in the final two primaries and a flurry of superdelegate endorsements, Sen. Barack Obama sealed the Democratic presidential nomination last night after a grueling and history-making campaign against Sen. Hillary Rodham Clinton that will make him the first African American to head a major-party ticket. Before a chanting and cheering audience in St. Paul, Minn., the first-term senator from Illinois savored what once seemed an unlikely outcome to the Democratic race with a nod to the marathon that was ending and to what will be another hard-fought battle, against Sen. John McCain, the presumptive Republican nominee….
  - Summary: Senator Barack Obama was declared the presumptive Democratic presidential nominee.

Machine Translation (MT)

- Translate a sentence from one natural language to another.
  - Hasta la vista, bebé $\Rightarrow$

    Until we see each other again, baby.

Ambiguity Resolution is Required for Translation

- Syntactic and semantic ambiguities must be properly resolved for correct translation:
  - “John plays the guitar.” → “John toca la guitarra.”
  - “John plays soccer.” → “John juega el fútbol.”
- An apocryphal story is that an early MT system gave the following results when translating from English to Russian and then back to English:
  - “The spirit is willing but the flesh is weak.” ⇒ “The liquor is good but the meat is spoiled.”
  - “Out of sight, out of mind.” ⇒ “Invisible idiot.”

Resolving Ambiguity

- Choosing the correct interpretation of linguistic utterances requires knowledge of:
  - Syntax
    - An agent is typically the subject of the verb
  - Semantics
    - Michael and Ellen are names of people
    - Austin is the name of a city (and of a person)
    - Toyota is a car company and Prius is a brand of car
  - Pragmatics
  - World knowledge
    - Credit cards require users to pay financial interest
    - Agents must be animate and a hammer is not animate

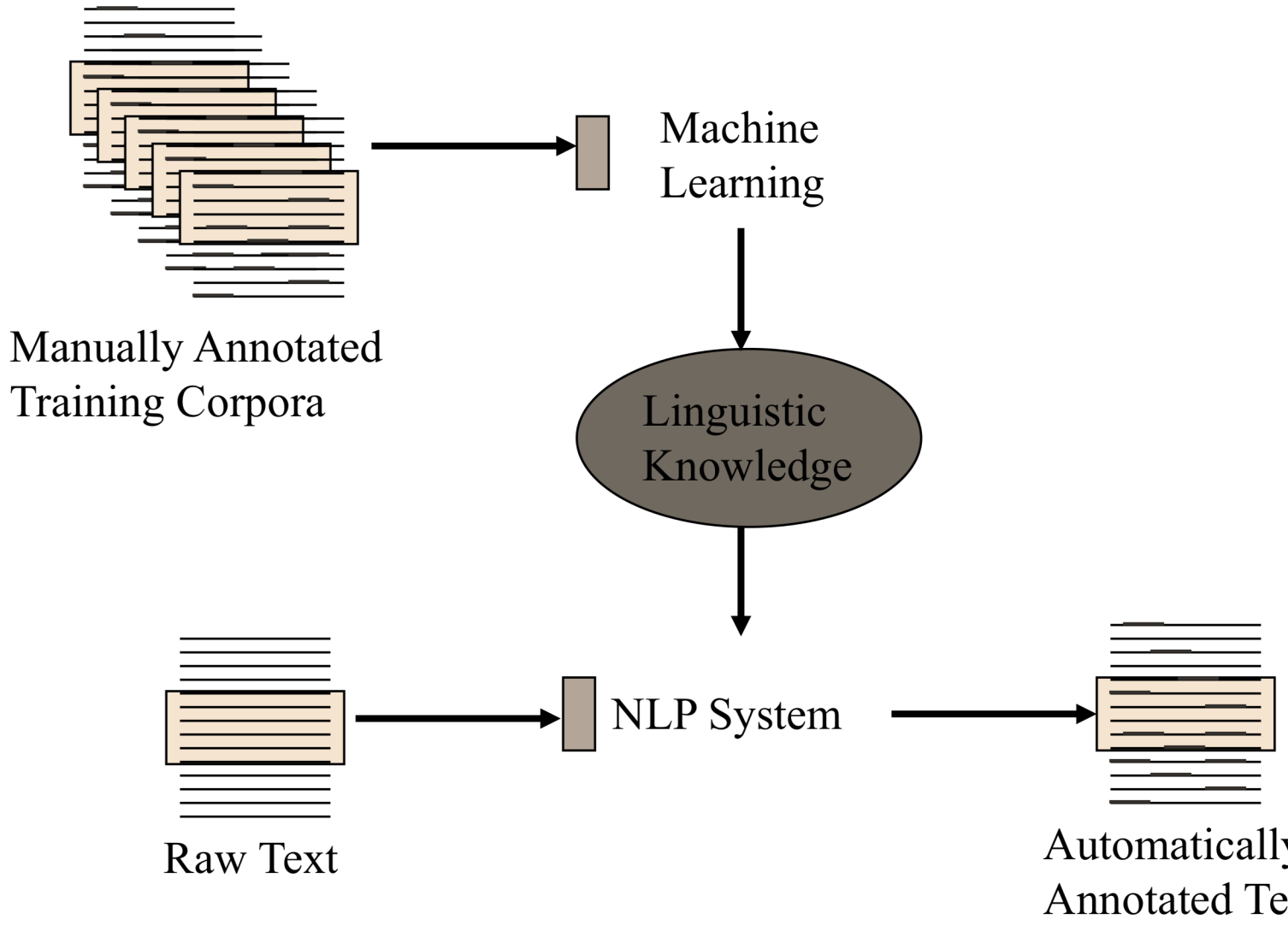
Manual Knowledge Acquisition

- Traditional, “rationalist,” approaches to language processing require human specialists to specify and formalize the required knowledge.
- Manual knowledge engineering, is difficult, time-consuming, and error prone.
- “Rules” in language have numerous exceptions and irregularities.
  - “All grammars leak.”: Edward Sapir (1921)
- Manually developed systems were expensive to develop and their abilities were limited and “brittle” (not robust).

Automatic Learning Approach

- Use machine learning methods to automatically acquire the required knowledge from appropriately annotated text corpora.
- Variously referred to as the “corpus based,” “statistical,” or “empirical” approach.
- Statistical learning methods were first applied to speech recognition in the late 1970’s and became the dominant approach in the 1980’s.
- During the 1990’s, the statistical training approach expanded and came to dominate almost all areas of NLP.

Learning Approach



Advantages of the Learning Approach

- Large amounts of electronic text are now available.
- Annotating corpora is easier and requires less expertise than manual knowledge engineering.
- Learning algorithms have progressed to be able to handle large amounts of data and produce accurate probabilistic knowledge.
- The probabilistic knowledge acquired allows robust processing that handles linguistic regularities as well as exceptions.

The Importance of Probability

- Unlikely interpretations of words can combine to generate spurious ambiguity:
  - “The a are of I” is a valid English noun phrase (Abney, 1996)
    - “a” is an adjective for the letter A
    - “are” is a noun for an area of land (as in hectare)
    - “I” is a noun for the letter I
  - “Time flies like an arrow” has 4 parses, including those meaning:
    - Insects of a variety called “time flies” are fond of a particular arrow.
    - A command to record insects’ speed in the manner that an arrow would.
- Some combinations of words are more likely than others:
  - “vice president Gore” vs. “dice precedent core”
- Statistical methods allow computing the most likely interpretation by combining probabilistic evidence from a variety of uncertain knowledge sources.

Human Language Acquisition

- Human children obviously learn languages from experience.
- However, it is controversial to what extent prior knowledge of “universal grammar” (Chomsky, 1957) facilitates this acquisition process.
- Computational studies of language learning may help us to understand human language learning, and to elucidate to what extent language learning must rely on prior grammatical knowledge due to the “poverty of the stimulus.”
- Existing empirical results indicate that a great deal of linguistic knowledge can be effectively acquired from reasonable amounts of real linguistic data without specific knowledge of a “universal grammar.”

Pipelining Problem

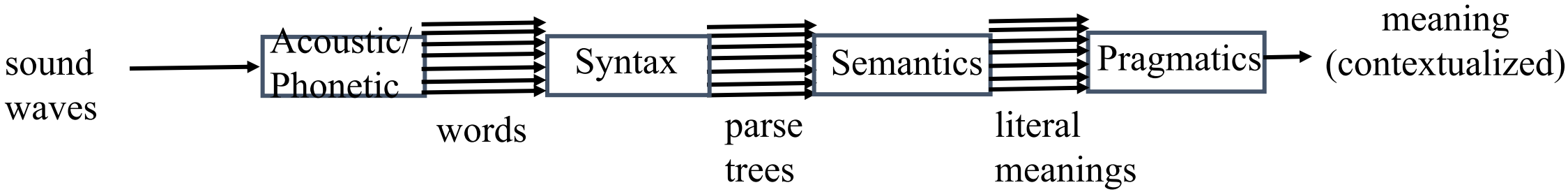
- Assuming separate independent components for speech recognition, syntax, semantics, pragmatics, etc. allows for more convenient modular software development.
- However, frequently constraints from “higher level” processes are needed to disambiguate “lower level” processes.
  - Example of syntactic disambiguation relying on semantic disambiguation:
    - At the zoo, several men were showing a group of students various types of flying animals. Suddenly, one of the students hit the man **with** a **bat**.

Pipelining Problem (cont.)

- If a hard decision is made at each stage, cannot backtrack when a later stage indicates it is incorrect.
  - If attach “with a bat” to the verb “hit” during syntactic analysis, then cannot reattach it to “man” after “bat” is disambiguated during later semantic or pragmatic processing.

Increasing Module Bandwidth

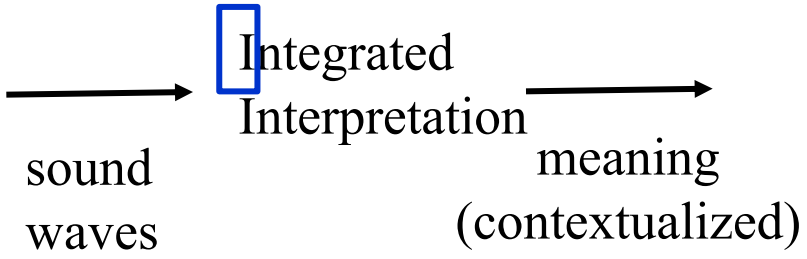
- If each component produces multiple scored interpretations, then later components can rerank these interpretations.



- **Problem:** Number of interpretations grows combinatorially.
- **Solution:** Efficiently encode combinations of interpretations.
  - Word lattices
  - Compact parse forests

Global Integration/ Joint Inference

- Integrated interpretation that combines phonetic/syntactic/semantic/pragmatic constraints.



- Difficult to design and implement.
- Potentially computationally complex.