CS442/542b: Artificial Intelligence II
Prof. Olga Veksler

Lecture 8
NLP: Introduction

Many slides from: M. Hearst (Berkeley), D. Klein (Berkeley), C. Manning (Stanford), L. Lee (Cornell), R. Barzilay (MIT), L. Venkata Subramaniam, Leila Kosseim (Concordia)

Outline

- Introduction to Natural Language Processing (NLP)
  - What is NLP
  - Applications of NLP
  - Why NLP is hard
  - Brief history of NLP
- Linguistic Essentials

Natural Language Processing

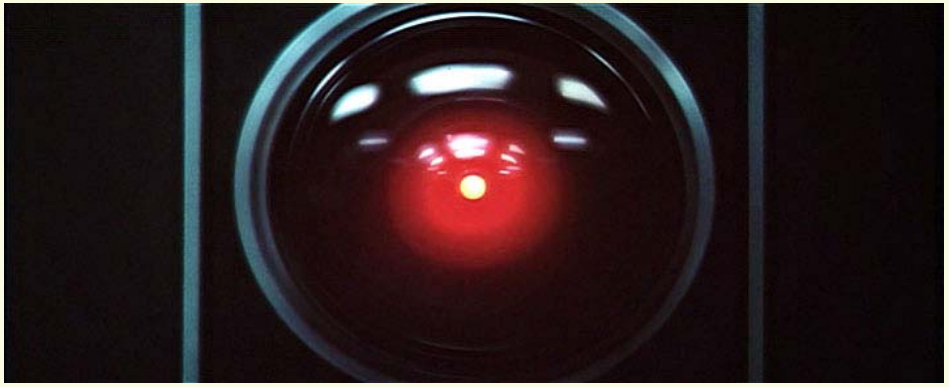
- Computers would be a lot more useful if they could handle our email, do our library research, talk to us, etc …
- But computers are fazed by natural human language
  - Or at least their programmers are, most avoid the language problem by using mice, menus, drop boxes
- How can we tell computers about language?
  - or help them learn it as kids do?
- Can machines understand human language?
  - Define ‘understand’
  - Understanding is the ultimate goal. However, one doesn’t need to fully understand to be useful.
- NLP is also known as Computational Linguistics (CL), Human Language Technology (HLT), Natural Language Engineering (NLE)

A few applications of NLP

- Modest
  - Spelling correction
  - text categorization
- Ambitious:
  - Better search engines
  - Information extraction
  - Speech recognition
  - Speech synthesis
  - Question answering
  - Machine translation
  - Language Teaching/Learning

We've past the year 2001, but we are not close to realizing the dream (or nightmare …)

Dave Bowman: "Open the pod bay doors, HAL"



HAL 9000: "I'm sorry Dave. I'm afraid I can't do that."

Still can't do that

Brief NLP History

- 1950's, empirical approach:
  - data-driven, co-occurrences in language are important sources of information: " You shall know a word by the company it keeps", J. Firth, 1957
  - First speech systems (Davis et al. Bell labs)
  - Text authorship (Hamilton vs. Madison), solved based on patterns of word occurrences in 1941 by F. Mosteller and F. Williams
  - Machine translation: toy system, basically word-substitution, on machines less powerful than pocket calculators
  - Little understanding of natural language syntax and semantics
  - Problem soon appeared intractable: can't store enough data on computers

Brief NLP History

- 1960's and 1970's
  - Data-driven approach falls out of favor
  - Belief that language should be analyzed at a much deeper level than surface statistics
  - N. Chomsky:
    1. "Colorless green ideas sleep furiously"
    2. "Furiously sleep ideas green colorless"
       - Neither (1) nor (2) will never occur. Yet (1) is grammatical, while (2) is not. Therefore (1) should have higher probability of occurrence than (2)
       - However, since neither (1) nor (2) will ever occur, they will both be assigned the same probability of 0
       - The criticism is that the data driven approach will always lack suffer from the lack of data, and therefore doomed to failure
  - Knowledge-based (rule based) approach becomes dominant, human expert encodes relevant information in computer-usable form
    - Development of linguistic
    - Complex language models, parsing, CF grammars
    - Applications in toy domains

Brief NLP History

- Drawbacks of knowledge-based (rule-based) approach:
  - Rules are often too strict to characterize people's use of language (people tend to stretch and bend rules in order to meet their communicative needs.)
  - Need expert people to develop rules (knowledge acquisition bottleneck)
- 1980's: the empirical revolution
  - In part motivated by success in speech recognition
    - Based on learning from lots of data
  - Corpus-based (data-driven) methods become central
  - Sophisticated machine learning algorithms are developed to learn from the data
  - Linguistics (the rules) is still used
  - Deep analysis often traded for robust and simple approximations

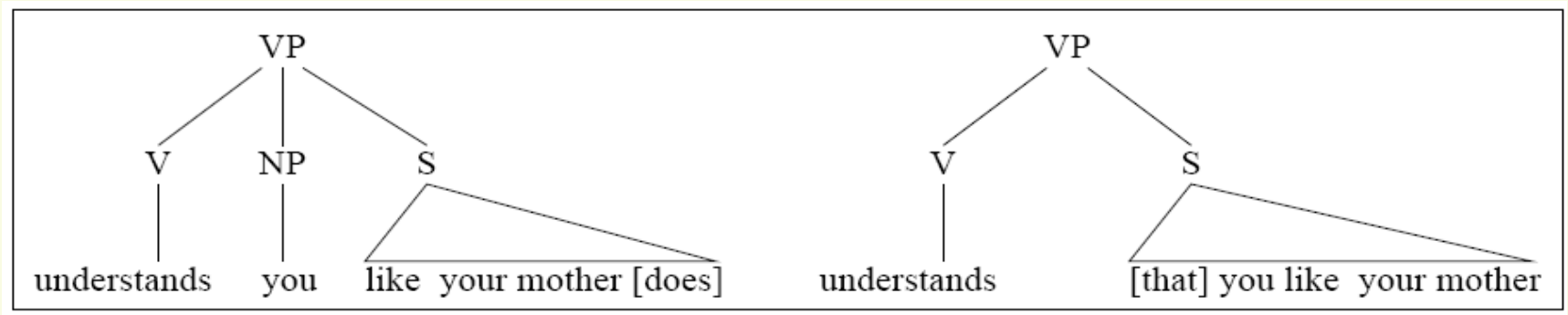
Why is NLP difficult?

- Key problem: language is **ambiguous** at all levels
  - Semantic (word meaning)
  - Syntactic (sentence structure)
  - Acoustic (parsing of speech signal)
- To resolve these ambiguities we often need to use complex knowledge about the world
- Other difficulties
  - Language only reflects the surface of meaning
    - humor, sarcasm, “between the lines” meaning
  - Language presupposes communication between people
    - Persuading, insulting, amusing them
  - Lots of subtleties

Syntactic (Sentence Structure) Ambiguity

"At last, a computer that understands you like your mother"-1985 advertisement from a company claimed to program computer to understand human language

- At least three different interpretations:
  1. The computer understands you as well as your mother understands you
  2. The computer understands that you like your mother
  3. The computer understands you as well as it understands your mother
- Humans would rule out the last two interpretation from their knowledge of the world: we know that the advertisement is trying to convince us of something



different sentence structure leads to different interpretations

Semantic (Word Meaning) Ambiguity

"At last, a computer that understands you like your mother"

- Word "mother" has several meanings:
  - "a female parent"
  - "a cask or vat used in vinegar-making"

Acoustic Ambiguity

"At last, a computer that understands you like your mother"

- For speech recognition:
  - "*a computer that understands you like your mother*"
  - *a computer that understands your lie cured mother*

More Ambiguity

"At last, a computer that understands you like your mother"

- Even if we interpret this as "The computer understands you as well as your mother understands you" does that mean it understands you "well" or "not so well"
  - sarcasm

Another Example Syntactic Ambiguity

- How about simpler sentences?
- Even simple sentences are highly ambiguous
- "*Get the cat with the gloves*"

Headline Ambiguity

- Iraqi Head Seeks Arms
- Ban on Nude Dancing on Governor's Desk
- Juvenile Court to Try Shooting Defendant
- Teacher Strikes Idle Kids
- Kids Make Nutritious Snacks
- British Left Waffles on Falkland Islands
- Red Tape Holds Up New Bridges
- Bush Wins on Budget, but More Lies Ahead
- Hospitals are Sued by 7 Foot Doctors
- Stolen Painting Found by Tree
- Local HS Dropouts Cut in Half

Why is NLP difficult? Language subtleties

- Adjective order and placement
  - A big black dog
  - A big black scary dog
  - A big scary dog
  - A scary big dog
  - ✘ A black big dog
- Antonyms
  - Which sizes go together?
    - Big and little
    - Big and small
    - Large and small
    - ✘ Large and little
- Adjective choice
  - ✘ powerful tea
  - strong tea

Example Problem

- Grammar checker example:

  Which word to use?

  <principal> <principle>

- Solution: look at which words surround each use:
  - **I am in my third year as the principal of Anamosa High School.**
  - **School-principal transfers caused some upset.**
  - **This is a simple formulation of the quantum mechanical uncertainty principle.**

Using Very, Very Large Corpora

- Keep track of which words are the neighbors of each spelling in well-edited text, e.g.:
  - Principal: “high school”
  - Principle: “rule”
- At grammar-check time, choose the spelling best predicted by the surrounding words.
- Surprising results:
  - Log-linear improvement even to a billion words!
  - Getting more data is better than fine-tuning algorithms!

The Effects of LARGE Datasets

- From Banko & Brill '01

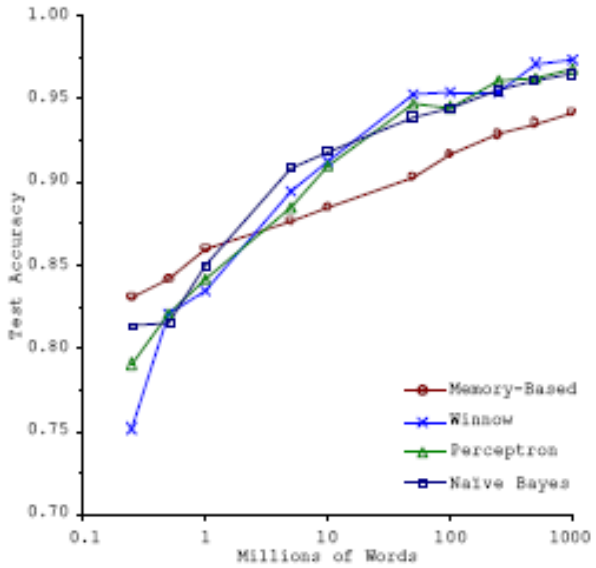


Figure 1. Learning Curves for Confusion Set Disambiguation

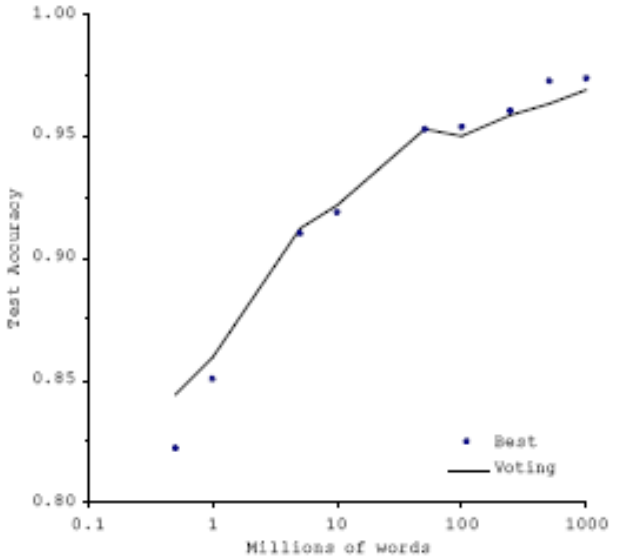


Figure 3. Voting Among Classifiers

Linguistic Essentials

- Parts of speech
- Morphology
- Syntax
- Semantics
- Pragmatics

Linguistic Essentials

- Parts of Speech
  - 3 most important:
    - Noun (objects like ”cat”)
    - verb (action like “go”)
    - Adjective (noun property, like ”black”)
  - Other parts
    - pronoun (refer to person or thing, “he”, “she”,etc)
    - Adverbs (modify verbs, like “often”)
    - Preposition (express spatial relationship,”in”, “over”)
    - Particle (bond with verbs,”gave **in**”)
    - determiners (a, the, this, that)
    - Conjunctions(“and”, “or”)
    - Subordinating conjunction (“that”, “if”, “before”)

Linguistic Essentials

- Morphology:
  - what words (or subwords) are we dealing with?
  - structures and patterns in words
  - analyzes how words are formed from minimal units of meaning, or morphemes, e.g., dogs= dog+s.
  - Example:
    - Input: The fearsome cats attacked the foolish dog
    - Output: The fear-some cat-s attack-ed the fool-ish dog

Why not just Use a Dictionary?

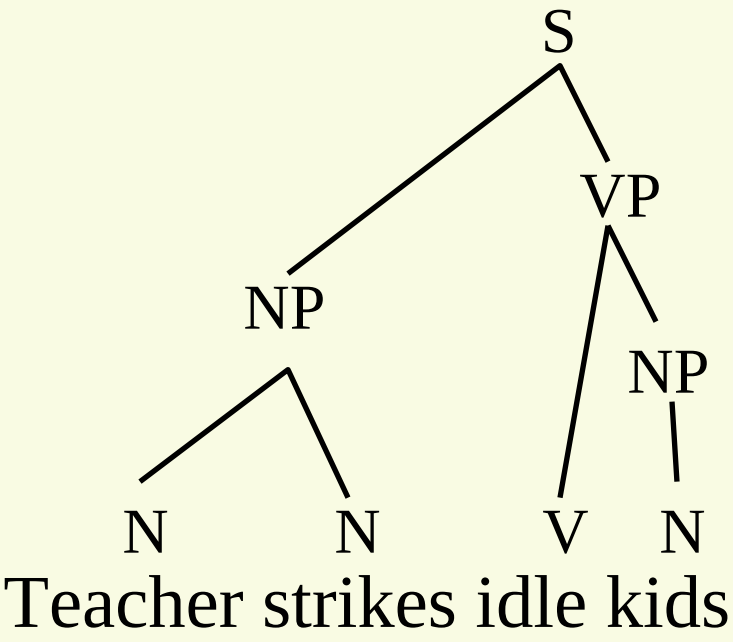
- How many words are there in a language?
  - English: OED 400K entries
  - Turkish: 600×10^6 forms
  - Finnish: 10^7 forms
- New words are being invented all the time
  - e-mail
  - IM

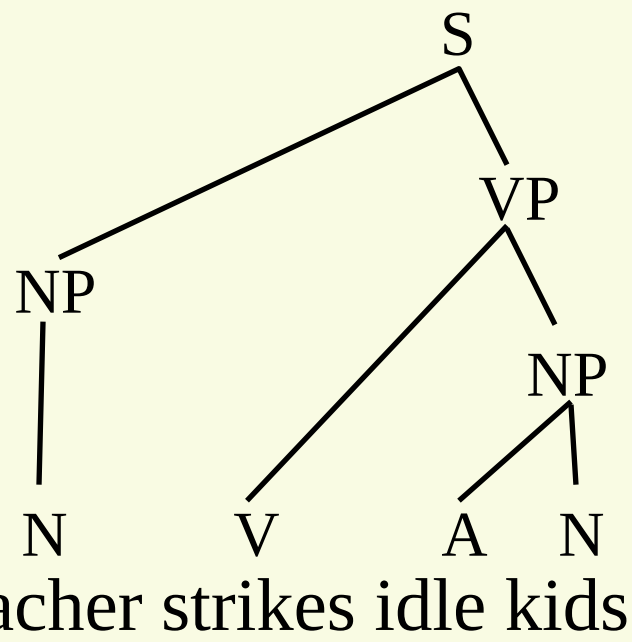
Linguistic Essentials

- Syntax:
  - What phrases are we dealing with? Which words modify one another?
  - Sentences have structures and are made up of constituents.
  - The constituents are phrases.
  - A phrase consists of a head and modifiers.
  - The category of the head determines the category of the phrase
    - e.g., a phrase headed by a noun is a noun phrase

Parsing

- Analyze the structure of a sentence





Linguistic Essentials

- Semantics: the meaning of a word or phrase within a sentence
  - Meaning of words
  - Meaning of sentences
- Pragmatics: structures and patterns in discourses
  - What should you conclude from the fact that I said something? How should you react?
  - Co-reference resolution
    - Jane races Mary on weekends. She often beats her.

Tools and Resources Needed

- Probability/Statistical Theory:
  - Statistical Distributions, Bayesian Decision Theory.
- Linguistics Knowledge:
  - Morphology, Syntax, Semantics, 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 to build applications.