CMPT 413
Computational Linguistics

Anoop Sarkar

`http://www.cs.sfu.ca/~anoop`

Natural Language Processing (NLP)

- NLP is the application of a computational theory of human language
- Language is the predominant repository of human interaction and knowledge
- Goal of NLP: programs that “listen in”
- The AI Challenge: the Turing test
- Lots of speech and text data available

NLP: Lots of Applications

- Doc classification
- Doc clustering
- Spam detection
- Information extraction
- Summarization
- Machine translation
- Cross Language IR
- Multiple language summarization
- Language generation
- Plagarism or author detection

- Error correction, language restoration
- Language teaching
- Question answering
- Knowledge acquisition (dictionaries, thesaurus, semantic lexicons)
- Speech recognition
- Text to Speech
- Speaker Identification
- (multi-modal) Dialog systems
- Deciphering ancient scripts

Language has structure

- Finnish word structure
  - talossansakaanko ‘not in his house either?’
  - kynässänsäkäänkö ‘not in his pen either?’
- English phrase structure
  - It is likely that John went home.
  - That John went home is likely.
  - OK: Where is it likely that John went t?
  - Not OK: *Where is that John went t likely?

Language is recursive

- Combine the following two sentences:
  - The clown watches the ballerina

    NP1 V1 NP2
  - The musician hits the clown

    NP3 V2 NP4
- Many possible combinations of the two sentences:
  - The clown watches the ballerina and the musician hits the clown
- Use a modifier to combine them:
  - The clown who the musician hits watches the ballerina

    NP1/4 NP3 V2 V1 NP2
  - The musician hits the clown who watches the ballerina

    NP3 V2 NP4/1 V1 NP2

Children's comprehension of relative clauses.
De Villiers et al. J. Psych Res 8(5) 2005

Language is recursive

- Finite resources but possibly infinite utterances (via recursion)
- **Sparse** language:
  - a sparse language is a set of strings where the number of strings of length *n* is bounded by a polynomial function of *n*
  - Regular and context-free languages are **dense**

    as shown by Chomsky, Flajolet, Incitti

Language is Parsed

- Google's Computer Might Betters Translation Tool
  - New York Times March 8, 2010
- Number of Lothian patients made ill by drinking rockets
  - Edinburgh Evening News, March 4, 2010
- Violinist linked to JAL crash blossoms
  - *http://languagelog.ldc.upenn.edu/nll/?p=1693*

Language is ambiguous

- Lung cancer in women mushrooms
  - Mushrooms is noun or a verb?
- Teacher Strikes Idle Kids
  - Strikes is a verb or a noun?
- Two sisters reunited after 18 years in checkout counter
  - Is it reunited in checkout counter or 18 years in checkout counter?
- Ban on nude dancing on governor's desk
  - Another case of “if-then-else” ambiguity
- British Left Waffles on Falkland Islands
  - Is it British/Noun Left/Verb or British Left/NP Waffles/Verb?

Ambiguity (cont’d)

- Kids make nutritious snacks
  - make can mean different things, which is it?
- Iraqi Head Seeks Arms
  - Arms can mean different things, which is it?
- Two Soviet Ships Collide, One Dies
  - What does one refer to in this case?
- Chef throws his heart into feeding needy
  - Throws his heart is not decomposed normally in this case: idiom finding

Ambiguity (cont’d)

- Island Monks Fly in Satellite to Watch Pope Funeral

  (“Monks in Space” languagelog.com/archives/002045.html)

  - “fly in” vs. “fly $[_{\text{OBJ}}$ in Satellite]” hidden segmentation
- G.I.'s Deployed in Iraq Desert With Lots of American Stuff (New York Times, Aug 13, 2005)
  - the verb desert, not the noun desert
- McDonald's fries the holy grail for potato farmers
  - *http://languagelog.ldc.upenn.edu/nll/?p=1762*

Ambiguity (cont’d)

- We saw her duck (Zwicky & Sadock)
  - “saw [$_{\text{NP}}$ her duck]” vs. “saw [$_{\text{S}}$ her duck]” duck: Noun/ Verb, her: ambiguous pronoun
- Leahy wants FBI to help corrupt Iraqi police force (CNN, Dec 13, 2006)
  - the adjective corrupt, not the verb corrupt
- Last Alder Hey hospital child remains buried
- Red tape holds up new bridges

Ambiguity (cont’d)

- Massive fish kill blankets Arkansas River
  - CNN 3 January 2011
- Suspect In Mumbai Attacks A Thorn In U.S.-India Ties
  - NPR 15 November 2010
- Baby Steps to New Life-Forms
  - New York Times 27 May 2010

Ambiguity (cont’d)

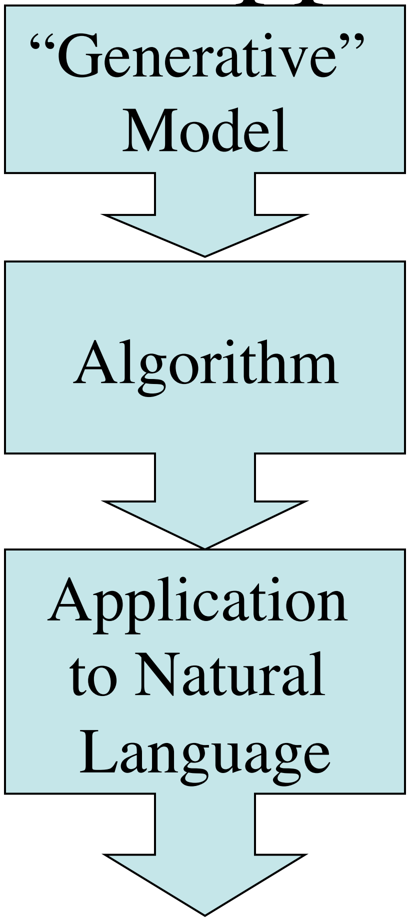
- Ambiguity can occur locally or globally
- Here’s an example of local ambiguity:
  - First black woman elected to Congress
  - First black woman elected to Congress dies
- dies causes a reanalysis of the structure of the sentence
  - before dies we analyze elected as the main verb
  - after we see dies we analyze elected as a sub-clause modifying the word elected

- Phonetics acoustic and perceptual elements
- Phonology inventory of basic sounds (phonemes) and basic rules for combination
  - e.g. vowel harmony. Anupu is pronunciation of Anoop in Classic Period Mayan
- Morphology how morphemes combine to form words, relationship of phonemes to meaning
  - e.g. delight-ed vs. de-light-ed
- Syntax sentence (utterance) formation, word order and the formation of constituents from word groupings
  - e.g. The clown who the musician hits watches the ballerina
- Semantics how do word meanings recursively compose to form sentence meanings (from syntax to logical formulas)
  - e.g. Everyone is not here => what does this mean? Nobody / Not everyone is here.
- Pragmatics meaning that is not part of compositional meaning,
  - e.g. This professor dresses even worse than Anoop!

Terminology: Grammar

- Grammar can be prescriptive or descriptive
- *Descriptive grammar* is a model of the form and meaning of a speaker of a language
- Grammar books for learning a language are *prescriptive grammars*, usually style manuals or rules for how to write clearly
- Except for some NLP apps like grammar checking or teaching, we are usually interested in creating models of language

General Approach



“Generative” Model

Algorithm

Application to Natural Language

Phonology / Morphology / Syntax / Semantics / Pragmatics

Formal Languages and NLP

Formal Language Theory	NLP
Language (possibly infinite)	Text Data, Corpus (finite)
Grammar	Grammar (usually inferred from data, produces infinite set)
Automata	Recognition/Generation algorithms

Some definitions

- Classification: assigning to the input one out of a finite number of classes, e.g.: Document -> spam, formalization -> Noun
- Sequence learning/Tagging: assigning a sequence of classes, e.g.: I/Pron can/Modal open/Verb a/Det can/Noun
- Parsing: assigning a complex structure, e.g.: formalization -> (Noun (Verb (Adj formal) -ize) -ation)
- Grammar development: human driven creation of a model for some linguistic data
- Transduction: transforming one linguistic form to another, e.g. summarization, translation, tokenization
- Tracking/Co-reference: after detecting an entity (say a person) tracking that entity in subsequent text; co-reference of a pronoun to its antecedent; “lexical chains” of similar concept
- Clustering: unsupervised grouping of data using similarity, constructing “phylogenetic” trees