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

Natural Language:What is it?

- Answers from linguistics: the scientific study of human language

  Natural Language (NL) vs. Artificial Language
- Genetic basis of human language
- Mysteriously distinct from other species (human language is unique to humans)
- NL is complex, displays recursive structure

Natural Language: What is it?

- Learning of language is an inherent part of NL
- Language has idiosyncratic rules and a complex mapping to thought

For more read The Great Eskimo Vocabulary Hoax by Geoffrey Pullum

Language has structure

- What he did was climb a tree
- What he ran was to the store
- Drink your beer and go home!
- What are drinking and go home?
- Linus lost his security blanket
- Lost Linus blanket security his

Language is recursive

- This is the house
- This is the house that Jack built
- This is the grain that lay in the house that Jack built
- This is the rat that ate the grain that lay in the house that Jack built
- This is the cat that killed the rat that ate the grain that lay in the house that Jack built
- This is the dog that chased the cat that killed the rat that ate the grain that lay in the house that Jack built

Language is recursive

- Finite resources
- Infinite set of utterances
- Recursion

Facets of Language Structure

- Phonetics acoustic and perceptual elements
- Phonology inventory of basic sounds (phonemes) and basic rules for combination, e.g. vowel harmony
- 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
- Semantics how do word meanings recursively compose to form sentence meanings (from syntax to logical formulas)
- 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

Terminology: Parts of Speech

- Nouns: John, cow, can, tomorrow
- Pronouns: he, she, it, who
- Verbs: run, chase, teach
- Auxiliary verbs: be, can, will, might
- Modal verbs: can, might
- Determiners: the, a, each, two or more
- Prepositions: in, at, under

More parts of speech

- Adjectives: blue, former
- Adverbs: quickly, certainly
- Coordinating conjunctions: and, but, or
- Complementizers: that, whether, if
- Possessives: ‘s (Kim ‘s), whose
- Interjections: Hey!

Grammatical Relations

- Subject-Verb-Object

Kim likes olives

- Subject-Object-Verb

Kim-ka olivu-lul cohanta

Kim-Nom olives-Acc like-Present-Decl

- Modifiers: Kim likes olives on Tuesdays
- Optional arguments: Kim donated olives vs. Kim went to the store

Inflections

- Prefix: un-happy
- Suffix: olive-s
- Different types of prefix or suffix information:
  - Plurals: olive-s
  - Past tense: smash-ed
  - …

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 more 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

Ambiguity: a key problem

- 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?
- British Left Waffles on Falkland Islands
  - Is it British/Noun Left/Verb or British Left/Noun Phrase 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 $[_{OBJ}$ in Satellite]" hidden segmentation
- G.I.'s Deployed in Iraq Desert With Lots of American Stuff (New York Times, Aug 13, 2005)
  - the noun desert, not the verb desert

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