

CMPT 413

Computational Linguistics

Anoop Sarkar

<http://www.cs.sfu.ca/~anoop>

1/14/08

1

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

1/14/08

2

NLP: Lots of Applications

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

1/14/08

3

Natural Language: What is it?

- Answers from linguistics
Natural Language (NL) vs. Artificial Language
- NL is complex, displays recursive structure
- Learning of language is an inherent part of NL
- Language has idiosyncratic rules and a complex mapping to thought

1/14/08

4

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

1/14/08

5

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

1/14/08

6

Language is recursive

- Finite resources
- Infinite (or exponentially many) utterances
- Recursion

1/14/08

7

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!*

1/14/08

8

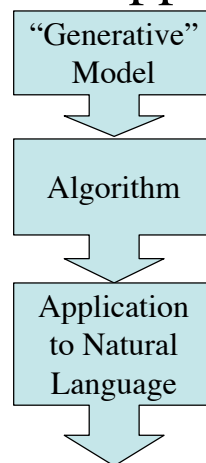
Terminology: Grammar

- 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

1/14/08

9

General Approach



Phonology / Morphology / Syntax / Semantics / Pragmatics

1/14/08

10

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

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

Grammatical Relations

- Subject-Verb-Object

Kim eats olives

- Subject-Object-Verb

김-이 올리브-을 먹-어요

Kim-Nom olives-Acc eat-Present_Decl

- Modifiers: Kim eats olives on Tuesdays

- Optional arguments:

– Kim donated **money** vs. Kim went **to the store**

1/14/08

13

Inflections

- Prefix: **un**-happy

- Suffix: olive-**s**

- Different types of prefix or suffix information:

– Plurals: olive-**s**

– Past tense: smash-**ed**

– ...

1/14/08

14

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
- **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

1/14/08

15

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?

1/14/08

16

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

1/14/08

17

Ambiguity (cont'd)

- Island Monks Fly in Satellite to Watch Pope Funeral
 - (“Monks in Space” languageblog.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 verb **desert**, not the noun **desert**

1/14/08

18

Ambiguity (cont'd)

- We saw her duck (Zwicky & Sadock)
 - “saw [_{NP} her duck]” vs. “saw [_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**
- Squad Helps Dog Bite Victim, and Other Flubs from the Nation's Press (book title, 1980)

1/14/08

19

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**

1/14/08

20