Overview

- Natural language processing[*]
- Introduction
- Properties
- Syntax
- Semantics: Case; Conceptual dependency

[*] Material mostly drawn from Gordon Novak's AI lectures http://www.cs.utexas.edu/users/novak/.

Natural Language Processing

A form of communication.

- Intentional exchange of information through the **production and perception of signs**.
- **Signs** are drawn from a system of conventional signs.
- Allows the use of information learned or observed by others.
- Natural language is an example.

* Study of signs is a complex discipline in its own right, called **semiotics**.

NLP and AI

NLP is a classical AI problem:

- Minimal input data
- Knowledge based
- Reference to context
- Local ambiguity
- Global constraints (on interpretation)
- Capturing the infinite: finite system for understanding an infinite set of sentences

Areas of NLP

- Text understanding
- Speech recognition
- Language generation (written or speech)
- Machine translation (e.g. Babel Fish at Altavista).

Why Study NL?

- Theoretical: (1) understand how language is structured; (2) understand the mental mechanisms necessary to support language use, e.g., memory.
- Practical: (1) easier human-computer interaction; (2) machine translation (www, globalization, ...); (3) computer-computer interaction (future).

Efficiency of Natural Language

- Serial in nature: limited bandwidth.
- Information theoretic concerns (bits per symbol).
- Only say things that may not be known to the listener.
- Zipf's law: **frequently used words are short!**
  - Example: mom, dad, eat, ...
- Often used long words tend to get abbreviated:
  - Fax, Cell, ASAP, PC, ...

Characteristics of NLP

- Ambiguity: multiple interpretations

  One morning I shot an elephant in my pajamas.

  How he got in my pajamas I'll never know.

- Incompleteness: only a bare outline is given

  I was late for work today. My car wouldn't start. The battery was dead.

Major Challenges

- Lexical ambiguity:
  - The pitcher broke his arm.
  - The pitcher broke.
- Grammatical ambiguity:
  - I saw the man on the hill with the telescope.
- Anaphora: words that refer to others.
  - John loaned bill his bike.
- Semantics: understanding the meaning.
  - Need a vast amount of world knowledge.

Approaches

- Formal approaches: parsing, reasoning, etc.
- Statistical approaches: data mining, text mining, usage of surrounding context (word cooccurrence statistics: **N-grams**).

Speech Act and Understanding

- Speech act: actions that allow **production of language**.
- Examples: query, inform, request, acknowledge, promise, etc.
- Characteristics: informative, declarative, etc.
- Communicating agents' task: to **understand** speech acts.

Fundamentals

- Formal languages: strings of symbols (terminals).
- Grammar (Syntax): finite set of rules that specifies a language (legal ways of ordering the string).
- Semantics: meaning of the string of terminals.
- Pragmatics: the meaning of the string within the **context** it is currently being used (need knowledge of the world and the social context of language).

(part)

Grammar

- Phrase structure: **phrases** are **substrings**, and can come in different categories.
- Phrase categories:
  - Sentence (S)
  - Noun phrase (NP)
  - Verb phrase (VP)
  - Prepositional phrase (PP)
  - ...
- Terminal vs. **nonterminal**: words are terminals (leaves), and symbols S, NP, VP, etc. are nonterminals (internal nodes in a parse tree).
- Rewrite rule: $< S > \rightarrow < NP >< VP >$