Natural language

- "Natural" language

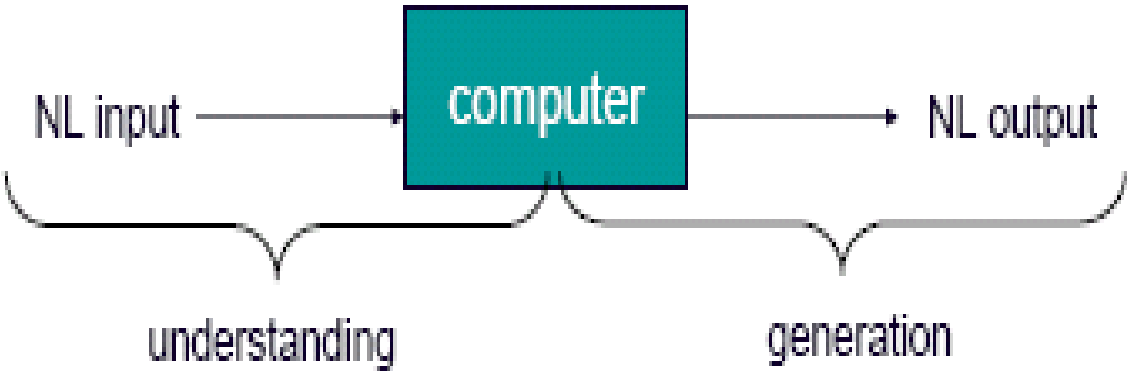
– Languages that people use to communicate with one another e.g. English, Japanese, Swahili, as opposed to artificial languages, like C++, Java, etc.

- Engineering Goal:
  - *Design, implement, and test systems that process natural languages for practical applications*

Computers Lack Knowledge!

- Computers “see” text in English the same you see
- People have no trouble understanding language
  - Common sense knowledge
  - Reasoning capacity
  - Experience
- Computers have
  - No common sense knowledge
  - No reasoning capacity

Language processing



Language technology

- Applications that deal with natural language

- can offer insights into language, a challenging & interesting topic
- can help navigate the web, since language is the medium of the web
- can help in communication
  - With computers (ASR, TTS)
  - With other humans (MT)

Language Technology – Natural Language Processing

- Huge amounts of data
  - Internet = at least 20 billions pages
  - Intranet
- Applications for processing large amounts of texts

  require NLP expertise

- Classify text into categories
- Index and search large texts
- Automatic translation
- Speech understanding
  - Understand phone conversations
- Information extraction
  - Extract useful information from resumes
- Automatic summarization
  - Condense 1 book into 1 page
- Question answering
- Knowledge acquisition
- Text generations / dialogues
- Machine translation

Linguistics Levels of Analysis

Basic NLP Task and Stages of Processing

- Phonology and phonetics: Process Sound
- Morphology: Process word forms
- Lexical: Process words and their properties
- Syntactic: Process structure
- Semantic: Process Meaning
- Pragmatics: Process intention
- Discourse: Process connections amongst sentences

Ambiguity of language

- Ambiguity at different levels of language:

1) Phonetic
   - [raIt] = *write*, *right*, *rite*
2) Lexical
   - *can* = noun, verb
3) Structural
   - *I saw the man with the telescope*
4) Semantic
   - *dish* = physical plate, menu item

→ All of these make NLP difficult

Issues in Syntax

1. Identify the part of speech (POS)

"*the dog ate my bread*"

Dog = noun ; ate = verb ; bread = noun

2. Identify collocations

mother in law, hot dog

Issues in Syntax contd…

3. Shallow parsing:

"the dog chased the bear"

"the dog" "chased the bear"

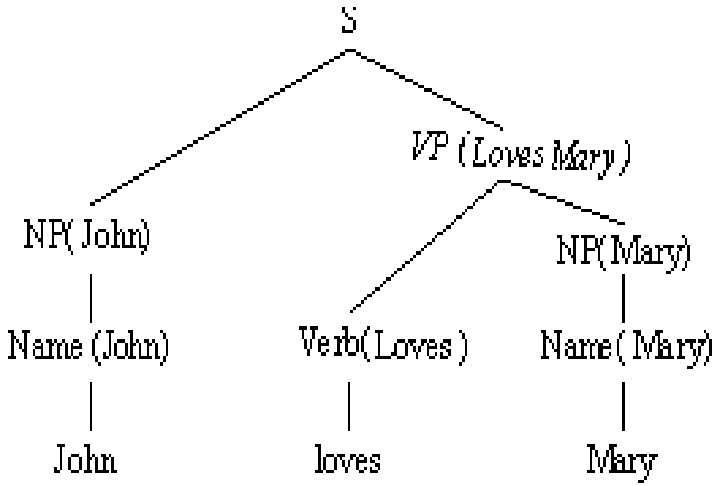
subject - predicate

Identify basic structures

NP-[the dog] VP-[chased the bear]

Issues in Syntax

4. Full parsing: John loves Mary



Help figuring out (automatically) questions like: Who did what and when?

More Issues in Syntax

5. Anaphora Resolution:

"*The dog entered my room. It scared me*"

6. Preposition Attachment

"I saw the man in the park with a telescope"

Collocation

- A COLLOCATION is an expression consisting of two or more words that correspond to some conventional way of saying things.
- The words together can mean more than their sum of parts:

The Times of India, disk drive, hot dog, mother in law

- Examples of collocations
  - noun phrases like *strong tea* and *weapons of mass destruction*
  - phrasal verbs like *to make up*, and other phrases like *the rich and powerful*.
- Valid or invalid?
  - *a stiff breeze* but not *a stiff wind* (while either *a strong breeze* or *a strong wind* is okay).
  - *broad daylight* (but not *bright daylight* or *narrow darkness*).

Compositional versus non-compositional collocates

A phrase is compositional if the meaning can be predicted from the meaning of the parts.

- E.g. new companies

A phrase is non-compositional if the meaning cannot be predicted from the meaning of the parts

- E.g. hot dog

Issues in Semantics

- Understand language! How?
- *"plant" = industrial plant*
- *"plant" = living organism*
- Words are ambiguous
- Importance of semantics?
  - Machine Translation: wrong translations
  - Information Retrieval: wrong information
  - Anaphora Resolution: wrong referents

Why Semantics?

- The sea is at the home for billions factories and animals
- The sea is home to million of plants and animals

Issues in Semantics

- How to learn the meaning of words?
- From dictionaries:

plant, works, industrial plant -- (buildings for carrying on industrial labor; "they built a large plant to manufacture automobiles")

plant, flora, plant life -- (a living organism lacking the power of locomotion)

They are producing about 1,000 automobiles in the new plant

The sea flora consists in 1,000 different plant species

The plant was close to the farm of animals.

Issues in Machine Translations

- Text to Text Machine Translations
- Speech to Speech Machine Translations

- Most of the work has addressed pairs of widely spread languages like
- English- Hindi, Oriya, Marathi, Bengali, Urdu, Tamil
- English – French, Chinese
- Bangla, Marathi, Punjabi, Tamil, Telugu, Kannada, Urdu – Hindi
- Tamil-Telugu
- Malayalam – Tamil

Issues in Machine Translations

- How to translate text?
  - Learn from previously translated data
- → Need parallel corpora
- English – Hindi, French-English, Chinese-English
- Reasonable translations
- Chinese-Hindi – no such tools available today!