Overview

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

[a] 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: $< S > \rightarrow < NP >< VP >$

Languages: Generative Capacity

Chomsky's four classes of grammatical formalisms:

- Regular grammar:
  $< NonTerm > \rightarrow Term \ < Nonterm >$ (Equivalent to finite state machines.)
- Context-free grammar: $< NonTerm > \rightarrow ...$ (Equivalent to push-down automata.)
- Context-sensitive grammar: symbols on the left hand side $\leq$ symbols on the right hand side.
- Recursively enumerable: both sides of the rewrite rule can have any number of terminal/nonterminal. (Equivalent to Turing machines.)

Steps in Communication

- Intention: thought
- Generation: form sentence to utter
- Synthesis: utter the sentence
- Perception: hear the utterance
- Analysis:
  - Parse
  - Semantic interpretation: infer meaning of the parse tree
  - Pragmatic interpretation: infer meaning in reference to the current context.
- Disambiguation (this or that) and Incorporation (believe it or not)

Parsing: Grammar

<SYMBOL>: nonterminal
WORD: terminal.

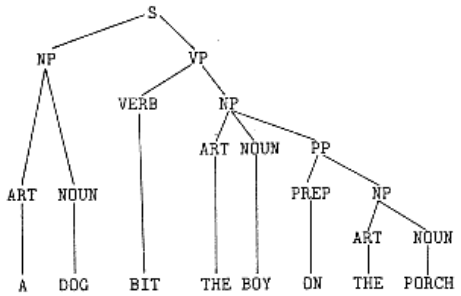
```
<S>      -->   <NP>  <VP>
<NP>     -->   <ART>  <ADJ>  <NOUN>
<NP>     -->   <ART>  <NOUN>
<NP>     -->   <ART>  <NOUN>  <PP>
<VP>     -->   <VERB>  <NP>
<VP>     -->   <VERB>  <NP>  <PP>
<PP>     -->   <PREP>  <NP>

<ART>    -->    A | AN | THE
<NOUN>   -->    BOY | DOG | LEG | PORCH
<ADJ>    -->    BIG
<VERB>   -->    BIT
<PREP>   -->    ON
```


Language Generation

- Start with the sentence symbol, <S>.
- Repeat until no nonterminal symbols remain: (1) Choose a nonterminal symbol in the current string; (2) Choose a production that begins with that nonterminal; (3) Replace the nonterminal by the right-hand side of the production.

```
< S>
< NP>  < VP>
< ART>  < NOUN>  < VP>
THE < NOUN>  < VP>
THE DOG < VP>
THE DOG < VERB>  < NP>
THE DOG < VERB>  < ART>  < NOUN>
THE DOG < VERB>  THE < NOUN>
THE DOG BIT THE < NOUN>
THE DOG BIT THE BOY
```

Parsing

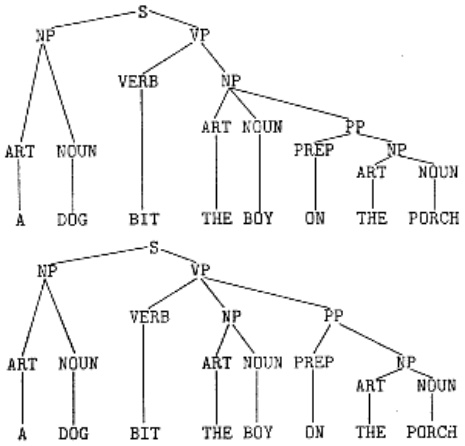


Inverse of generation.

Parsing Techniques

- Top-down: Start with the symbol $< S >$ and hope to produce the string. (Very inefficient.)
- Bottom-up: Reduce phrases using production rules.
- Chart parser: Eliminates the redundant work in rephrasing by saving partial results in a chart.
- Augmented transition networks: (1) arbitrary tests added to arcs; (2) structure-building actions added to arcs (save state, etc.); (3) phrase names on arcs (can name subroutines). (Equivalent to a Turing machine.)

Problem: Ambiguity



- Different parse trees can be generated depending on the order you picked the rewrite rules.
- Lexical ambiguity and grammatical ambiguity.

Foreign Languages

Grammars are quite different from English:

- Word ordering
- Number and gender agreement
- Tense
- Familiar, formal, honorific forms

Number of languages: several thousand (Native American: $>$ 1,000; Africa $\sim$ 1,000; New Guinea $\sim$ 700; India $>$ 150; Russia ~ 100, etc.

Semantics

- Selecting correct word sense meanings.
- Removing ambiguity: choosing interpretations that “make sense” when many interpretations are syntactically possible.

  - John saw my dog driving to work this morning.
- Resolving pronoun references.
  - Bill wanted John’s bike. He stole it.
- Resolving other references.
  - ... a ladder ... A man is 10 ft from the top.
  - A bridge is supported at each end.

Case Theory

Fillmore [Charles Fillmore, The case for case, 1968.]

- Theory of deep case structures: elements of a sentence are related to the verb by deep case. [a]
- Cases used in English are:

Formal Name	Description:	Example:	Example Use:
Nominative	subject	he	He hit the ball.
Objective	direct object	him	John hit him.
Dative	indirect object	him	I gave him a book.
Genitive	possessive	his	He lost his keys.

- Although English does not make the cases obvious, Fillmore argued that the cases are still present in English.

[a] [case: an inflectional form of a noun, pronoun, or adjective indicating its grammatical relation to other words; such a relation whether indicated by inflection or not. – Webster’s 9th New Collegiate Dictionary] relationships.

Cases: Example

- Mother baked for three hours.
- The pie baked for three hours.
- **Mother and the pie baked for three hours.** ⟵ even though the syntax is correct, the sentence is anomalous when the deep case is considered.

Case Relations

Agent	instigator of the event
Counter-agent	resistance against action
Object/Theme/Patient	entity that moves or changes
Result	entity that comes into existence
Instrument	physical cause of event
Source	place from which something moves
Goal	place to which something moves
Experiencer	entity which receives effects
Locus	place where event occurs
Modality	tense etc. of verb

Case Relation Example

John broke the window with the hammer.
John broke the window.
The hammer broke the window.
The window broke.

```
Break1:
   Tok:         Break
   Modality:    Tense past, Voice active, ...
   Agent:       John
   Object:      Window
   Instrument:  Hammer
```

Such a case structure can also be represented as a semantic network:

Case Frames

A case frame is a lexical definition for a word sense that tells how other phrases are related to it.

- Selection restriction: restrictions on the possible slot fillers for the frame.
- Semantic structure: how to build an output structure to represent the meaning.
- Related phrases: how certain phrases, e.g. prepositional phrases, may fit into the meaning.

```
HIT-1   < subj>   +animate
        < obj>    +concrete
        [with < inst>   +concrete,-animate]

-->   STRIKE   AGENT      < subj>
               THEME      < obj>
               INSTRUMENT < inst>
```


Disambiguation Using Case Frames

- The **selection restrictions** of the case frame can be matched against the semantic markers of different sense meanings of words to choose combinations that make sense.
  - John hit the ball with a bat.
- The selection restrictions of `HIT-1` require the marker +concrete for the object; this is true of the “spherical object” meaning of ball, but not of the “dance” meaning. The correct sense of bat can be found in a similar way.

Conceptual Dependency

Proposed by Roger Schank with stated goals of:

- allowing easier inference from sentences,
- provide a representation independent of the original words, and
- sentences with the same meaning should result in the same representation.

Some material here are drawn from http://www.cs.cf.ac.uk/Dave/AI2/node69.html.

Conceptual Dependency: Outline

- Definition of verbs in terms of underlying concepts called **primitive acts** (14 or so).
- Deep case relations between primitive acts and objects.
- Semantic network representation of these structures.

Primitive Acts

TRANS	Transfer (e.g. give)
INGEST	Ingest into the body (e.g. eat)
EXPEL	Expel from the body
ATRANS	Abstract transfer (e.g. ownership)
MOVE	Move a body part
GRASP	Grasp an object
PROPEL	Propel (e.g. throw) an object
PTRANS	Physical transfer
MTRANS	Mental transfer
MBUILD	Construct new info from old, (e.g. decide)
CONC	Conceptualize
ATTEND	Direct sensory organ to observe

Primitive Conceptual Categories

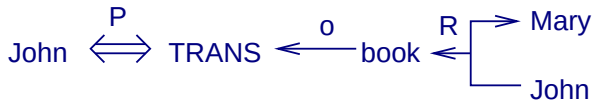
Building blocks for allowable dependencies:

- PP: Real world objects.
- ACT: Real world actions.
- PA: Attributes of objects.
- AA: Attributes of actions.
- T: Time.
- LOC: Location.

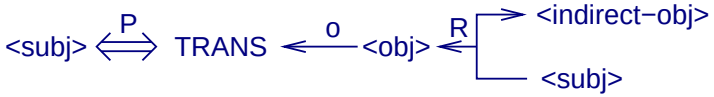
Representing Dependencies

- $\longrightarrow$: dependency (direction of arrow is the direction of dependency).
- o: object
- R: recipient-donor
- I: instrument
- D: destination
- $\Leftrightarrow$: two-way link between actor (PP) and action (ACT).
- Actions can be modified with temporal and other attributes.

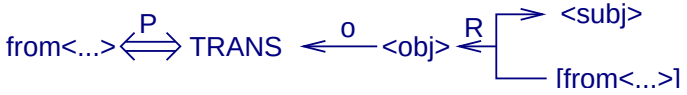
Conceptual Dependency: Example



Definition of "Give"



Definition of "Take"



- John gave a book to Mary.

“Mary received a book from John” would ultimately have the same representation.

Conceptual Dependency: Evaluation

Contributions

- An internal representation that is language-free, using primitive ACTs instead.
- A small number of primitive ACTs (around 14) rather than thousands.
- Different ways of saying the same thing can map to the same internal representation.
- Inference is simpler.
- Many inference rules are already represented in CD.
- The holes in the initial structure help to focus on the points that has yet to be established.

Conceptual Dependency: Evaluation

Limitations

- Very hard to find the right set of primitives.
- The expansion of some verbs into primitive ACT structures can be complex.
- Storage issues.
- Graph matching is NP-hard.

Application by Schank et al.

- Paraphrase:
  - Input: John strangled Mary.
  - Response: John grasped Mary’s neck and she died because she could not ingest air.
- Inferring facts and motivations:
  - Input: Mary was glad that John hit Bill.
  - Response: What did Bill do to make Mary angry?
- etc.

MARGIE, SAM, PAM, etc.

Semantics in General

- Sentences contain the bare minimum.
- Understanding natural language requires filling in a lot of information (based on world knowledge and contextual cues).

Natural Language Applications

- ELIZA: simple pattern matching conversational agent.
- LUNAR: NASA lunar rock database interface.
- PLANES: US Airforce aircraft database interface.
- LADDER: Navy database of ship positions, capabilities, status.
- Babel Fish: machine translation. Try English $\longrightarrow$ German $\longrightarrow$ English.

Key Points

- Why is natural language processing difficult?
- What are the kinds of ambiguities in natural language?
- What are the issues with syntactic and semantic processing?
- How does Case Frames and Conceptual Depencency provide semantic representations? What are their strengths and limitations?