Communication

“Classical” view (pre-1953):
  language consists of sentences that are true/false (cf. logic)

“Modern” view (post-1953):
  language is a form of action

Wittgenstein (1953) Philosophical Investigations
Austin (1962) How to Do Things with Words
Searle (1969) Speech Acts

Why?
To change the actions of other agents

Need a deeper understanding of language

Grammar

Address Ch 22 data sparsity through generalization (categories)

Vervet monkeys, antelopes etc. use isolated symbols for sentences
⇒ restricted set of communicable propositions, no generative capacity
(Chomsky (1957): Syntactic Structures)

Grammar specifies the compositional structure of complex messages
e.g., speech (linear), text (linear), music (two-dimensional)

A formal language is a set of strings of terminal symbols

Each string in the language can be analyzed/generated by the grammar

The grammar is a set of rewrite rules, e.g.,

\[ S \rightarrow NP \ VP \]
\[ Article \rightarrow \text{the} \mid a \mid an \mid \ldots \]

Here \( S \) is the sentence symbol, \( NP \) and \( VP \) are nonterminals
Grammar types

Regular: nonterminal → terminal[nonterminal]

\[ S \rightarrow aS \]
\[ S \rightarrow \Lambda \]

Context-free: nonterminal → anything

\[ S \rightarrow aSB \]

Context-sensitive: more nonterminals on right-hand side

\[ ASB \rightarrow AAaBB \]

Recursively enumerable: no constraints

Related to Post systems and Kleene systems of rewrite rules

Natural languages probably context-free, parsable in real time!

Wumpus lexicon

Noun → stench | breeze | glitter | nothing
      | wumpus | pit | pits | gold | east | ...
Verb → is | see | smell | shoot | feel | stinks
      | go | grab | carry | kill | turn | ...
Adjective → right | left | east | south | back | smelly | ...
Adverb → here | there | nearby | ahead
      | right | left | east | south | back | ...
Pronoun → me | you | I | it | ...
Name → John | Mary | Boston | UCB | PAJC | ...
Article → the | a | an | ...
Preposition → to | in | on | near | ...
Conjunction → and | or | but | ...
Digit → 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9

Divided into closed and open classes

Wumpus grammar

\[ S \rightarrow NP \ VP \]
\[ S \rightarrow NP \ Conjunction \ S \]
\[ NP \rightarrow Pronoun \]
\[ NP \rightarrow Noun \]
\[ NP \rightarrow Article \ Noun \]
\[ NP \rightarrow Digit \ Digit \]
\[ NP \rightarrow PP \]
\[ NP \rightarrow RelClause \]
\[ VP \rightarrow Verb \]
\[ VP \rightarrow VP \ NP \]
\[ VP \rightarrow VP \ Adjective \]
\[ VP \rightarrow VP \ PP \]
\[ VP \rightarrow VP \ Adverb \]
\[ PP \rightarrow Preposition \ NP \]
\[ RelClause \rightarrow that \ VP \]

I feel a breeze
I feel a breeze + and + I smell a wumpus
I + feel a breeze
+ and + I smell a wumpus

Divided into closed and open classes

Related to Post systems and Kleene systems of rewrite rules

Natural languages probably context-free, parsable in real time!
Grammaticality judgements

Formal language $L_1$ may differ from natural language $L_2$

Adjusting $L_1$ to agree with $L_2$ is a learning problem!

* the gold grab the wumpus
* I smell the wumpus the gold
  I give the wumpus the gold
* I donate the wumpus the gold

Intersubjective agreement somewhat reliable, independent of semantics!
Real grammars 10–500 pages, insufficient even for “proper” English

Probabilistic CFGs

$Noun \rightarrow stench(.05) | breeze(.10)\ldots$

$S \rightarrow NP \ VP \ (0.9) \ I + feel \ a \ breeze$
| $S \ Conjunction \ S \ (0.1) \ I \ feel \ a \ breeze \ + \ and + I \ smell \ a \ wumpus$

- Sum of the probabilities for each category is 1

Parse trees

Exhibit the grammatical structure of a sentence

I shoot the wumpus

Pronoun | Verb | Article | Noun
---|---|---|---
I | shoot | the | wumpus
Parse trees

Exhibit the grammatical structure of a sentence

```
NP  VP  NP
Pronoun  Verb  Article  Noun
I  shoot  the  wumpus
```

Parse trees - probabilistic

Exhibit the grammatical structure of a sentence

```
S
  VP
   NP  VP  NP
    Pronoun  Verb  Article  Noun
     I  shoot  the  wumpus
```

Each interior node is labeled with its probability.
The probability of the tree as a whole is $0.9 \times 0.25 \times 0.05 \times 0.15 \times 0.4 \times 0.1$

```
S
  0.90
   NP
    0.25
     Article  Noun
      0.05  0.15
          Every  wumpus
        0.40
           Verb
            0.10
               smells
```
Syntax in NLP

Most view syntactic structure as an essential step towards meaning;

"Mary hit John" ≠ "John hit Mary"

"And since I was not informed—as a matter of fact, since I did not know that there were excess funds until we, ourselves, in that checkup after the whole thing blew up, and that was, if you'll remember, that was the incident in which the attorney general came to me and told me that he had seen a memo that indicated that there were no more funds."

Context-free parsing

Bottom-up parsing works by replacing any substring that matches RHS of a rule with the rule’s LHS

Efficient algorithms (e.g., chart parsing (Ch. 23.2) - normal forms, dynamic programming again!)

Learning probabilities for PCFGs - treebanks

Logical grammars

BNF notation for grammars too restrictive:
- difficult to add “side conditions” (number agreement, etc.)
- difficult to connect syntax to semantics

Idea: express grammar rules as logic

\[
X \rightarrow YZ \quad \text{becomes} \quad Y(s_1) \land Z(s_2) \Rightarrow X(Append(s_1, s_2))
\]

\[
X \rightarrow \text{word} \quad \text{becomes} \quad X(\text{"word"})
\]

\[
X \rightarrow Y \mid Z \quad \text{becomes} \quad Y(s) \Rightarrow X(s) \quad Z(s) \Rightarrow X(s)
\]

Here, $X(s)$ means that string $s$ can be interpreted as an $X$. 
Now it’s easy to augment the rules

\[ NP(s_1) \land \text{Number}(s_1, n) \land VP(s_2) \land \text{Number}(s_2, n) \implies S(\text{Append}(s_1, s_2)) \]

Parsing is reduced to logical inference:

\[ \text{Ask}(KB, S(\text{"I" "am" "a" "wumpus"})) \]

(Can add extra arguments to return the parse structure, semantics)

Generation simply requires a query with uninstantiated variables:

\[ \text{Ask}(KB, S(x)) \]

Lexicalized PCFGs

- VP(v) -> Verb(v) NP(n) [P1(v,n)]
- VP(v) -> Verb(v) [P2(v)]
- ...
- Noun(banana) -> banana [pn]

Real human languages provide many problems for NLP:

- ambiguity
- anaphora
- indexicality
- vagueness
- noncompositionality
- discourse structure
- metonymy
- metaphor

Squad helps dog bite victim
Squad helps dog bite victim
Helicopter powered by human flies

I ate spaghetti with meatballs

Squad helps dog bite victim
Helicopter powered by human flies
American pushes bottle up Germans

I ate spaghetti with meatballs salad
Ambiguity

Squad helps dog bite victim
Helicopter powered by human flies
American pushes bottle up Germans
I ate spaghetti with meatballs
    salad
    abandon

Ambiguity can be lexical (polysemy), syntactic, semantic, referential

Chapter 23.1-23.3
Indexicality

Indexical sentences refer to utterance situation (place, time, S/H, etc.)

I am over here

Why did you do that?

Anaphora

Using pronouns to refer back to entities already introduced in the text

After Mary proposed to John, they found a preacher and got married.

For the honeymoon, they went to Hawaii

Mary saw a ring through the window and asked John for it
Anaphora

Using pronouns to refer back to entities already introduced in the text
After Mary proposed to John, they found a preacher and got married.
For the honeymoon, they went to Hawaii
Mary saw a ring through the window and asked John for it
Mary threw a rock at the window and broke it

Metonymy

Using one noun phrase to stand for another
I’ve read Shakespeare
Chrysler announced record profits
The ham sandwich on Table 4 wants another beer

Metaphor

"Non-literal" usage of words and phrases, often systematic:
I’ve tried killing the process but it won’t die. Its parent keeps it alive.

Noncompositionality

basketball shoes
Noncompositionality

basketball shoes
baby shoes

Noncompositionality

basketball shoes
baby shoes
alligator shoes

Noncompositionality

basketball shoes
baby shoes
alligator shoes
designer shoes
designer shoes
brake shoes
Noncompositionality

basketball shoes  baby shoes  alligator shoes  designer shoes  brake shoes  red book
red pen
red hair