Three Models for the Description of Language

written by Noam Chomsky

Matt Willsey

6.441: Transmission of Information

May 12, 2006
Background

- **Goal:** Develop simple structural model for the English language
  - Provides a useful method for generating/classifying grammatical English sentences
  - “Reveals” underlying structure of the English Language and how ideas are communicated

- **Three Models**
  - Finite-State Markov Processes
  - \([\Sigma, F]\) Grammars
  - Transformational Grammars

- **Disclaimer:** Sweeping details under the rug
Finite-State Markov Processes

Example

\[ a_{0,1,5} \quad a_{1,2,3} \quad \cdots \]
Finite-State Markov Processes

• Formal System Description
  – finite number of states: \( S_0, S_1, \ldots, S_q \)
  – set of transition symbols:
    \[
    A = \left\{ a_{ijk} \mid 0 \leq i, j \leq q; 1 \leq k \leq N_{ij} \forall i, j \right\}
    \]
  – pairs of connected states: \( C = \{(S_i, S_j)\} \)

• English cannot be modeled with a finite-state Markov process.
• Fails with dependencies: if-then, either-or
  \( a^a^\ldots^b^b \)
\[
\Sigma, F \] Grammars are MORE powerful than finite-state Markov models because they can describe \( a^a \ldots ^b^b \).

\[ Z \rightarrow a^a Z^b \]
Transformation Grammar

- Still $[\Sigma, F]$ grammars are too limited for English
- How do we generate “have taken” from the string “Verb”?
- Define the rules:
  - $\text{Verb} \rightarrow \text{Auxiliary} ^ \wedge \text{V}$
  - $\text{Auxiliary} \rightarrow \text{have} ^ \wedge \text{en}$
  - $\text{V} \rightarrow \text{take}$
- Output is the kernel string “have ^ en ^ take”
- Now define the transformation: $\text{Af} ^ \wedge \text{v} \rightarrow \text{v} ^ \wedge \text{Af} ^ \#$
- Applying the transformation to the kernel string yields “have ^ take ^ en”
What Did We Do?

• Developed a simple model to generate and classify grammatical sentences
  – “Chung asked a hard question.” => Grammatical
  – “Chung a asked question hard.” => Not Grammatical

• Developed a clue how ideas in the brain are converted to sentences, for example, “they are flying planes.”
What else?

- To compress English strings, Lempel-Ziv assumes English is derived from a Markovian source.
- Chomsky proved that English cannot be modeled by a finite-state Markov model because of dependencies.
  
  “If either if the candle falls, then the floor catches on fire, or the toaster smokes, then the fire alarm will sound.”

- English sentences used in practice still can be modeled by a finite Markov source when using Lempel-Ziv.
Questions