Natural Language Processing

George Konidaris
gdk@cs.brown.edu

Fall 2017
Natural Language Processing

Understanding spoken/written sentences in a natural language.

**Major area of research in AI.**

Why?

- Humans use language to communicate.
- Most natural interface.
- Huge amounts of NLP “knowledge” around.
  - E.g., books.
- Generative power

- Key to intelligence?
  - Hints as to underlying mechanism
  - Key indicator of intelligence
Natural Language Processing

It is also incredibly hard. **Why?**

I saw a bat.

Lucy owns a parrot that is larger than a cat.

John kissed his wife, and so did Sam.

Mary invited Sue for a visit, but she told her she had to go to work.

I went to the hospital, and they told me to go home and rest.

The price of tomatoes in Des Moines has gone through the roof.

Mozart was born in Salzburg and Beethoven, in Bonn.

(examples via Ernest Davis, NYU)
“If you are a fan of the justices who fought throughout the Rehnquist years to pull the Supreme Court to the right, Alito is a home run - a strong and consistent conservative with the skill to craft opinions that make radical results appear inevitable and the ability to build trusting professional relationships across ideological lines.” (TNR, Nov. 2005)

“The juiciest prize is to become the face of a luxury brand such as Dior or Burberry. To have any chance, a model must first have magazine shoots under her designer belt. This fact allows fashion magazines to pay peanuts, even for a cover-shoot.” (Economist, Feb. 2012)
Component Problems

perception

“the cat sat on the mat”

syntactic analysis

SatOn(x = Cat, y = Mat)

semantic analysis

Cat?

Mat?

disambiguation

incorporation

SatOn(cat3, mat16)
Perception

“The cat sat on the mat.”
Major Challenges

Speaker accent, volume, tone.

No pauses - word boundaries?
Noise.
Variation.
Speech Recognition

![Graph showing speech signal with time and pressure axes and labeled segments: th, ah, ca, t.](image)
Speech Recognition Using HMMs

Must store:

- \( P(O_t | S_t) \)
- \( P(S_{t+1} | S_t) \)

prob. of observed audio given phoneme

prob. of one phoneme following another
Issues

Phoneme sequence not Markov
  • Must introduce memory for context
  • k-Markov Models

People speak faster or slower
  • “Window” does not have fixed length
  • Dynamic Time Warping

Quite a simplistic model for a complex phenomenon.

Nevertheless, speech recognition tech based on HMMs commercially-viable mid-1990s.
Speech Recognition with Deep Nets

Mid-to-late 2000s: replace HMM with Deep Net.

ah ca ... th
0.1 0.3 0.1
Speech Recognition with Deep Nets

How to deal with dependency on prior states and observations?

Recurrent nets: form of memory.
Component Problems

perception

“the cat sat on the mat”

syntactic analysis

SatOn(x = Cat, y = Mat)

semantic analysis

disambiguation

Cat?

Mat?

incorporation

SatOn(cat3, mat16)
Syntactic Analysis

**Syntax**: characteristic of language.

- Structure.
- Composition.

But observed in linear sequence.
Syntactic Analysis

How to describe this structure?

Formal grammar.
- Set of rules for generating sentences.
- Varying power:
  - Recursively enumerable (equiv. Turing Machines)
  - Context-Sensitive
  - Context-Free
  - Regular

Each uses a set of rewrite rules to generate syntactically correct sentences.

*Colorless green ideas sleep furiously.*
Formal Grammars

Two types of symbols:
• Terminals (stop and output this)
• Non-terminals (one is a start symbol)

Production (rewrite) rules that modify a string of symbols by matching expression on left, and replacing it with one on right.

\[
\begin{align*}
S & \rightarrow AB & ab \\
A & \rightarrow AA & aaaaaaab \\
A & \rightarrow a & abbb \\
B & \rightarrow BBB & aabbbb \\
B & \rightarrow b \\
\end{align*}
\]
Rules must be of the form:

\[ A \rightarrow B \]

where \( A \) is a **single** non-terminal and \( B \) is any sequence of terminals and non-terminal.

Why is this called *context-free*?
Probabilistic CFGs

Attach a probability to each rewrite rule:

\[ A \rightarrow B[0.3] \]
\[ A \rightarrow AA[0.6] \]
\[ A \rightarrow a[0.1] \]

Probabilities for the same left symbol sum to 1.

Why do this?

More vs. less likely sentences.

*Probability distribution over valid sentences.*
<table>
<thead>
<tr>
<th>Part of Speech</th>
<th>Example Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>stench 0.05</td>
</tr>
<tr>
<td></td>
<td>breeze 0.10</td>
</tr>
<tr>
<td></td>
<td>wumpus 0.15</td>
</tr>
<tr>
<td></td>
<td>pits 0.05</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
<tr>
<td>Verb</td>
<td>is 0.10</td>
</tr>
<tr>
<td></td>
<td>feel 0.10</td>
</tr>
<tr>
<td></td>
<td>smells 0.10</td>
</tr>
<tr>
<td></td>
<td>stinks 0.05</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
<tr>
<td>Adjective</td>
<td>right 0.10</td>
</tr>
<tr>
<td></td>
<td>dead 0.05</td>
</tr>
<tr>
<td></td>
<td>smelly 0.02</td>
</tr>
<tr>
<td></td>
<td>breezy 0.02</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
<tr>
<td>Adverb</td>
<td>here 0.05</td>
</tr>
<tr>
<td></td>
<td>ahead 0.05</td>
</tr>
<tr>
<td></td>
<td>nearby 0.02</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
<tr>
<td>Pronoun</td>
<td>me 0.10</td>
</tr>
<tr>
<td></td>
<td>you 0.03</td>
</tr>
<tr>
<td></td>
<td>I 0.10</td>
</tr>
<tr>
<td></td>
<td>it 0.10</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
<tr>
<td>RelPro</td>
<td>that 0.40</td>
</tr>
<tr>
<td></td>
<td>which 0.15</td>
</tr>
<tr>
<td></td>
<td>who 0.20</td>
</tr>
<tr>
<td></td>
<td>whom 0.02</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
<tr>
<td>Name</td>
<td>John 0.01</td>
</tr>
<tr>
<td></td>
<td>Mary 0.01</td>
</tr>
<tr>
<td></td>
<td>Boston 0.01</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
<tr>
<td>Article</td>
<td>the 0.40</td>
</tr>
<tr>
<td></td>
<td>a 0.30</td>
</tr>
<tr>
<td></td>
<td>an 0.10</td>
</tr>
<tr>
<td></td>
<td>every 0.05</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
<tr>
<td>Prep</td>
<td>to 0.20</td>
</tr>
<tr>
<td></td>
<td>in 0.10</td>
</tr>
<tr>
<td></td>
<td>on 0.05</td>
</tr>
<tr>
<td></td>
<td>near 0.10</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
<tr>
<td>Conj</td>
<td>and 0.50</td>
</tr>
<tr>
<td></td>
<td>or 0.10</td>
</tr>
<tr>
<td></td>
<td>but 0.20</td>
</tr>
<tr>
<td></td>
<td>yet 0.02</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
<tr>
<td>Digit</td>
<td>0 0.20</td>
</tr>
<tr>
<td></td>
<td>1 0.20</td>
</tr>
<tr>
<td></td>
<td>2 0.20</td>
</tr>
<tr>
<td></td>
<td>3 0.20</td>
</tr>
<tr>
<td></td>
<td>4 0.20</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>
**Grammar**

$$E_0 : \quad S \to NP \ VP \quad [0.90] \quad \text{I + feel a breeze}$$

$$\quad \quad \quad \quad | \quad S \text{ Conj } S \quad [0.10] \quad \text{I feel a breeze + and + It stinks}$$

$$NP \to \quad \text{Pronoun} \quad [0.30] \quad \text{I}$$

$$\quad | \quad \text{Name} \quad [0.10] \quad \text{John}$$

$$\quad | \quad \text{Noun} \quad [0.10] \quad \text{pits}$$

$$\quad | \quad \text{Article Noun} \quad [0.25] \quad \text{the + wumpus}$$

$$\quad | \quad \text{Article Adj Noun} \quad [0.05] \quad \text{the + smelly dead + wumpus}$$

$$\quad | \quad \text{Digit Digit} \quad [0.05] \quad 3 4$$

$$\quad | \quad NP \ PP \quad [0.10] \quad \text{the wumpus + in 1 3}$$

$$\quad | \quad NP \ RelClause \quad [0.05] \quad \text{the wumpus + that is smelly}$$

$$VP \to \quad \text{Verb} \quad [0.40] \quad \text{stinks}$$

$$\quad | \quad VP \ NP \quad [0.35] \quad \text{feel + a breeze}$$

$$\quad | \quad VP \ Adjective \quad [0.05] \quad \text{smells + dead}$$

$$\quad | \quad VP \ PP \quad [0.10] \quad \text{is + in 1 3}$$

$$\quad | \quad VP \ Adverb \quad [0.10] \quad \text{go + ahead}$$

$$Adjs \to \quad \text{Adjective} \quad [0.80] \quad \text{smelly}$$

$$\quad | \quad \text{Adjective Adj} \quad [0.20] \quad \text{smelly + dead}$$

$$PP \to \quad \text{Prep NP} \quad [1.00] \quad \text{to + the east}$$

$$RelClause \to \quad \text{RelPro VP} \quad [1.00] \quad \text{that + is smelly}$$
The cat sat on the mat.
Component Problems

Perception: “the cat sat on the mat”

Semantic analysis: SatOn(x = Cat, y = Mat)

Disambiguation: Cat? Mat?

Syntactic analysis: SatOn(cat3, mat16)
Semantic Analysis

Semantics: what the sentence actually means, eventually in terms of symbols available to the agent (e.g., a KB).

“the cat sat on the mat”

SatOn(x = Cat, y = Mat)
SatOn(cat3, mat16)
Semantic Analysis

Key idea: *compositional semantics.*

*The semantics of sentences are built out of the semantics of their constituent parts.*

“*The cat sat on the mat.*”

Therefore there is a clear relationship between syntactic analysis and semantic analysis.
Semantic Analysis

Useful step:
• Probability of parse depends on words
• Lexicalized PCFGs

\[ VP(v) \rightarrow \text{Verb}(v)NP(n)[P_1(v, n)] \]

variables

ate bandanna
vs.
ate banana

probability depends on variable bindings
Semantic Analysis

“John loves Mary”
Desired output: Loves(John, Mary)

Semantic parsing:
• Exploit compositionality of parsing to build semantics.

\[
S(\text{pred}(\text{obj})) \rightarrow \text{NP}(\text{obj}) \text{ VP}(\text{pred}) \\
\text{VP}(\text{pred}(\text{obj})) \rightarrow \text{Verb}(\text{pred}) \text{ NP}(\text{obj}) \\
\text{NP}(\text{obj}) \rightarrow \text{Name}(\text{obj}) \\
\text{Name}(\text{John}) \rightarrow \text{John} \\
\text{Name}(\text{Mary}) \rightarrow \text{Mary} \\
\text{Verb}(\lambda y \lambda x \text{ Loves}(x, y)) \rightarrow \text{loves}
\]
Semantic Analysis

S(Loves(John, Mary))

NP(John)  VP(\lambda x \text{ Loves}(x, \text{ Mary}))

Name(John)  Verb(\lambda y, \lambda x \text{ Loves}(x, y))  NP(Mary)

John  loves  Mary

\lambda\text{-expression}
symbols in KB

sentence to add to KB
Machine Translation

Major goal of NLP research for decades.
Competing Approaches

Formal Language

Document in Russian

Document in English
Competing Approaches

Document in Russian

Document in English
Google Translate

100 languages, 200 million people daily