Natural Language Processing

George Konidaris

gek@cs.brown.edu

Fall 2018
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
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)
Natural Language Processing

“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)

(examples via Ernest Davis, NYU)
Component Problems

perception

“the cat sat on the mat”

syntactic analysis

semantic analysis

SatOn(x = Cat, y = Mat)

disambiguation

Cat?

Mat?

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
Speech Recognition Using HMMs

Must store:
- \( P(O \mid S) \)
- \( P(S_{t+1} \mid S_t) \)

prob. of observed audio given phoneme
prob. of one phoneme following another

transition model
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

“the cat sat on the mat”

Perception

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.*

```
S
  /   \\ 
NP   VP
  /     \\ 
Article Noun VP Prep NP
  /     /     \\ 
the cat sat on the mat
```
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.

\[
S \rightarrow AB \\
A \rightarrow AA \\
A \rightarrow a \\
B \rightarrow BBB \\
B \rightarrow b
\]

\[
ab \\
aaaaaaab \\
abbb \\
aabbbbbb
\]
Context-Free Grammars

Rules must be of the form:

\[ A \rightarrow B \]

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

Why is this called \textit{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.*
## Lexicon

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

\[
\mathcal{E}_0 : \\
S \rightarrow NP \ VP \ [0.90] \ I + \ feel \ a \ breeze \\
\quad | \ S \ Conj \ S \ [0.10] \ I \ feel \ a \ breeze + \ and + \ It \ stinks \\
NP \rightarrow \ Pronoun \ [0.30] \ I \\
\quad | \ Name \ [0.10] \ John \\
\quad | \ Noun \ [0.10] \ pits \\
\quad | \ Article \ Noun \ [0.25] \ the + \ wumpus \\
\quad | \ Article \ Adjs \ Noun \ [0.05] \ the + \ smelly \ dead + \ wumpus \\
\quad | \ Digit \ Digit \ [0.05] \ 3 \ 4 \\
\quad | \ NP \ PP \ [0.10] \ the \ wumpus + \ in \ 1 \ 3 \\
\quad | \ NP \ RelClause \ [0.05] \ the \ wumpus + \ that \ is \ smelly \\
VP \rightarrow \ Verb \ [0.40] \ stinks \\
\quad | \ VP \ NP \ [0.35] \ feel + \ a \ breeze \\
\quad | \ VP \ Adjective \ [0.05] \ smells + \ dead \\
\quad | \ VP \ PP \ [0.10] \ is + \ in \ 1 \ 3 \\
\quad | \ VP \ Adverb \ [0.10] \ go + \ ahead \\
Adjs \rightarrow \ Adjective \ [0.80] \ smelly \\
\quad | \ Adjective \ Adjs \ [0.20] \ smelly + \ dead \\
PP \rightarrow \ Prep \ NP \ [1.00] \ to + \ the \ east \\
RelClause \rightarrow \ RelPro \ VP \ [1.00] \ that + \ is \ smelly
\]
The cat sat on the mat.
Component Problems

perception

“the cat sat on the mat”
syntactic analysis

semantic analysis
SatOn(x = Cat, y = Mat)

Cat?

Mat?

disambiguation

incorporation
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"

\[
\text{SatOn}(x = \text{Cat}, y = \text{Mat}) \\
\text{SatOn}(\text{cat3}, \text{mat16})
\]
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

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

variables

probability depends on variable bindings

ate bandanna vs. ate banana
Semantic Analysis

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

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

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

Sentence to add to KB:

\[ S(\text{Loves}(\text{John}, \text{Mary})) \]

NP(John) \rightarrow Name(John) \rightarrow John

Verb(\lambda y, \lambda x \text{Loves}(x, y)) \rightarrow \lambda -expression

NP(Mary) \rightarrow Name(Mary) \rightarrow Mary

VP(\lambda x \text{Loves}(x, \text{Mary})) \rightarrow symbols in KB

Name(John) \rightarrow John

Name(Mary) \rightarrow Mary

\text{loves}
Machine Translation

Major goal of NLP research for decades.

Document in Russian → Document in English
Competing Approaches

- Document in Russian
- Document in English

Formal Language
Competing Approaches

Document in Russian ➔ Document in English
Google Translate

100 languages, 200 million people daily