Last Class: Parsing Intro

1. Grammars and parsing
2. Top-down and bottom-up parsing

Today: Chart Parsing

1. Chart parsers
2. Bottom-up chart parsing
3. The Earley Algorithm

Chart Parsers

chart: data structure that stores partial results of the parsing process in such a way that they can be reused. The chart for an n-word sentence consists of:

- $n + 1$ **vertices**
- a number of **edges** that connect vertices

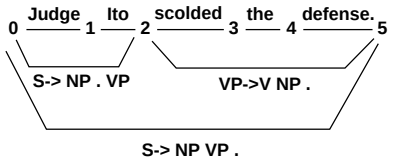


Chart Parsing: The General Idea

The process of parsing an n-word sentence consists of forming a chart with $n + 1$ vertices and adding edges to the chart one at a time.

- Goal: To produce a complete edge that spans from vertex 0 to n and is of category S.
- There is no backtracking.
- Everything that is put in the chart stays there.
- Chart contains all information needed to create parse tree.

Bottom-UP Chart Parsing Algorithm

Do until there is no input left:

1. If the agenda is empty, get next word from the input, look up word categories, add to agenda (as constituent spanning two postions).
2. Select a constituent from the agenda: constituent C from p_1 to p_2.
3. Insert C into the chart from position p_1 to p_2.
4. For each rule in the grammar of form $X \rightarrow C\ X_1 \ldots X_n$, add an active edge of form $X \rightarrow C \circ X_1 \ldots X_n$ from p_1 to p_2.

5. Extend existing edges that are looking for a C.

   (a) For any active edge of form $X \rightarrow X_1 \ldots \circ C X_n$ from p_0 to p_1, add a new active edge $X \rightarrow X_1 \ldots C \circ X_n$ from p_0 to p_2.

   (b) For any active edge of form $X \rightarrow X_1 \ldots X_n \circ C$ from p_0 to p_1, add a new (completed) constituent of type X from p_0 to p_2 to the agenda.

Slide CS474–5

Grammar and Lexicon

Grammar:

1. S → NP VP
2. NP → ART N
3. NP → ART ADJ N
4. VP → V NP

Lexicon:

the: ART
old: ADJ, N
man: N, V
boat: N

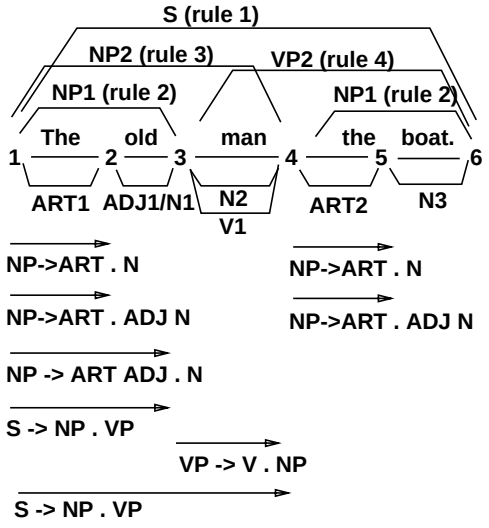
Sentence: $_1$ The $_2$ old $_3$ man $_4$ the $_5$ boat $_6$

Slide CS474–6

Example

[See .ppt slides]

Slide CS474–7



Slide CS474–8

Efficient Parsing

n = sentence length
Time complexity for naive algorithm: exponential in n
Time complexity for bottom-up chart parser: $\bigcirc(n^3)$

Options for improving efficiency:

1. Don't do twice what you can do once.
2. Don't represent distinctions that you don't need.

   Fall leaves fall and spring leaves spring.
3. Don't do once what you can avoid altogether.

   The can holds the water. ("can": AUX, V, N)

Slide CS474–9

Earley Algorithm: Top-Down Chart Parser

For all S rules of the form $S \rightarrow X_1 \ldots X_k$, add a (top-down) edge from 1 to 1 labeled: $S \rightarrow \circ X_1 \ldots X_k$.

Do until there is no input left:

1. If the agenda is empty, look up word categories for next word, add to agenda.
2. Select a constituent from the agenda: constituent C from p_1 to p_2.
3. Using the (bottom-up) edge extension algorithm, combine C with every active edge on the chart (adding C to chart as well). Add any new constituents to the agenda.
4. For any active edges created in Step 3, add them to the chart using the top-down edge introduction algorithm.

Slide CS474–10

Top-down edge introduction.

To add an edge $S \rightarrow C_1 \ldots \circ C_i \ldots C_n$ ending at position j:

For each rule in the grammar of form $C_i \rightarrow X_1 \ldots X_k$,

recursively add the new edge $C_i \rightarrow \circ X_1 \ldots X_k$ from j to j.

Slide CS474–11

Grammar and Lexicon

Grammar	Lexicon
1. S → NP VP	the: ART
2. NP → ART ADJ N	large: ADJ
3. NP → ART N	can: N, AUX, V
4. NP → ADJ N	hold: N, V
5. VP → AUX VP	water: N, V
6. VP → V NP	

Sentence: $_1$ The $_2$ large $_3$ can $_4$ can $_5$ hold $_6$ water $_7$

Slide CS474–12