Introduction to Symbolic and Statistical NLP in Scheme

Damir Ćavar
dcavar@unizd.hr

ESSLLI 2006, Malaga

July/August 2006

Grammar

- Replacement rules:
  - Set of symbols
  - Set of terminals
  - Production rules
  - $X \rightarrow ab$
  - $X \rightarrow aXb$

Grammar

- Context-free grammar:
  - The left-hand side of a rule can only consist of one symbol.
  - The right-hand side may consist of any number of symbols and terminals.
  - Replacement of a left-hand side symbol is possible any time, independent of the context.

Grammar

- $G = (\{VP, NP, V, N\}, \{see, John\}, S, P)$
  - $S = VP$
  - $P =$

    $VP \rightarrow V \;\; NP$

    $NP \rightarrow N$

    $N \rightarrow John$

    $V \rightarrow see$

Parsing and Phrase Structure Grammar

- Top-down parsing
  - Replace goal symbol with symbols and symbols with terminals until the terminals match.
- Bottom-up parsing
  - Replace terminals with symbols and symbols with symbols until the goal symbol is reached.

Parsing

- Parsing strategies
  - Top-down parsing
  - Bottom-up parsing
- Processing strategies
  - Breath first
  - Depth first

Parsing Strategy

- Problems observed
  - Reanalysis of already analyzed constituents
  - Search through all grammar rules
- Solution
  - Memorize analyzed constituents
  - Choose appropriate rules

Parsing Strategy

- Solution
  - Chart Parsing
    - Chart as memory
    - Selection of relevant rules from grammar

Chart Parsing

- Chart:
  - Storage for complete and incomplete constituents
  - Edges
    - Dotted rule
    - Index

Chart Parsing

- Chart:
  - Storage for complete and incomplete constituents
  - Edges
    - Dotted rule: VP $\rightarrow$ V $\bullet$ NP
    - Index:
      - Left and right position of the edge span
      - Position of the dot in the RHS

Chart Parsing

- Edges:
  - Dotted rule: VP $\rightarrow$ V $\bullet$ NP
    How much of the input at which position matches which part of the RHS of the rule?
  - Example:
    * Input: *John loves Mary*
    * Edge: (1, 2, 1, V $\rightarrow$ *loves* $\bullet$)

Chart Parsing

- Edges:
  - Inactive edge: (1, 2, 1, V $\rightarrow$ *loves* $\bullet$)
    - Complete constituent
  - Active edge: (1, 2, 1, VP $\rightarrow$ V $\bullet$ NP)
    - Incomplete constituent

Chart Parsing

- Adding edges to chart:
  - Initialization
  - Rule invocation: Matching edges with rules
  - Fundamental rule: Matching active and inactive edges on the chart

Chart Parsing

- Initialization
  - Bottom-up strategy:
  - For every token add an inactive edge to chart:
    - edge(0, 1, 1, N $\rightarrow$ *John* $\bullet$)
    - edge(1, 2, 1, V $\rightarrow$ *kissed* $\bullet$)
    - edge(2, 3, 1, N $\rightarrow$ *Mary* $\bullet$)

Chart Parsing

- Rule Invocation
  - Bottom-up strategy:
  - For every inactive edge on chart:
    - Find rules that have its LHS on their left periphery in RHS
    - Create new edges and add to chart.

Chart Parsing

- Rule Invocation
  - Example:
    - Inactive edge: edge(0, 1, 1, N $\rightarrow$ *John* $\bullet$)
    - Rule: NP $\rightarrow$ N
    - New edge: edge(0, 0, 0, NP $\rightarrow$ $\bullet$ N)

Chart Parsing

- Fundamental Rule
  - Move inactive edge from agenda to chart
  - For inactive edge find edge that expects it
    - edge(0, 1, 1, NP $\rightarrow$ N $\bullet$)
    - edge(0, 0, 0, S $\rightarrow$ $\bullet$ NP VP)
  - Add resulting edge to agenda:
    - edge(0, 1, 1, S $\rightarrow$ NP $\bullet$ VP)

Chart Parsing

- Bottom-up:

```
1: Initialize agenda
2: Repeat until edges in agenda
   Process first edge on agenda
   If edge inactive:
      move inactive edge to chart
      Function RuleInvocation
   Function FundamentalRule
```

- Result:

  If chart contains over-spanning edges, these represent possible parses of the input.

Chart Parsing

- Step by step
  - Initialize chart with the next word of the utterance, i. e. create edge with the lexical rule
  - Find rules in the grammar that consume the symbol of the inactive edges on the chart, i. e. extend the chart with edges that have LHS-symbols of inactive edges at the left periphery of their RHS

Chart Parsing

- Step by step
  - Create new edges by combining an active edge with an inactive edge:
    * the end of the one is the beginning of the other
    * the expectation symbol of the active edge corresponds with the LHS of the inactive edge

Chart Parsing

- Motivation
  - Problems with backtracking (our brute-force) parsers:
    * Repetitive parsing of same token(list)s
    * Repetitive parsing of paths that turned out to be unsuccessful
    * Unknown words and partial structures lead to a failure

Chart Parsing

- Motivation
  - Chart parser (e. g. Earley parser):
    - Avoid parsing of same token(list)s by memorization in chart
    - Memorize parses for partial structures
      - If a spanning analysis is impossible, the chart contains the partial analyses

Chart Parsing

- Motivation
  - Chart parser (e. g. Earley parser):
    - Compact representation for ambiguous structures (multiple parses)

Chart Parsing

- Chart
  - Edges
    - Directed graph: start point, end point, analysis
    - Input: *John kissed Mary*
    - Final chart:

(0, 1, N, [*John* •])	(0, 1, NP, [N •])
(1, 2, V, [*kissed* •])	(2, 3, NP, [N •])
(2, 3, N, [*Mary* •])	(1, 3, VP, [V NP •])
(0, 3, S, [NP VP •])	

Chart Parsing

- Strategies
  - Bottom-up
  - Top-down
- Active strategies (Agenda)
  - Depth-first
  - Breath-first

Chart Parsing

- Bottom-up strategy
  - Initialization (scan, tagging)
    - Add edges with lexical rules for each token (incrementally)
  - Rule invocation (prediction)
  - Fundamental rule (completion)

Chart Parsing

- Bottom-up strategy
  - Rule Invocation:
    For every inactive edge on chart:
    * Find rules that have its LHS on their left periphery in RHS.
    * Create new edges and add to chart.

Chart Parsing

- Bottom-up rule invocation example:
  - Inactive edge:
    - edge(0, 1, N $\rightarrow$ *John* $\bullet$)
  - Rule:
    - NP $\rightarrow$ N
  - New edge:
    - edge(0, 0, NP $\rightarrow$ $\bullet$ N)

Chart Parsing

- Fundamental Rule
  - For every active edge find expected inactive edge:
    - edge(0, 1, N $\rightarrow$ *John* $\bullet$)
    - edge(0, 0, NP $\rightarrow$ $\bullet$ N)
  - Merge edges and add resulting edge to chart:
    - edge(0, 1, NP $\rightarrow$ N $\bullet$)

Chart Parsing

- Top-down strategy
  - Initialization
    - Add edges with rules with goal symbol on LHS (incrementally)
  - Rule invocation (prediction)
  - Fundamental rule (completion)

Chart Parsing

- Top-down strategy
  - Rule Invocation:
    For every active edge on chart:
    * Find rules that have its left peripheral symbol from the expected RHS on their LHS. The left peripheral symbol from the expected RHS is the first symbol following the DOT.
    * Create new edges and add to chart.

Chart Parsing

- Top-down rule invocation example:
  - Active edge:
    - edge(0, 0, S $\rightarrow$ $\bullet$ NP VP)
  - Rule:
    - NP $\rightarrow$ N
  - New edge:
    - edge(0, 0, NP $\rightarrow$ $\bullet$ N)

Chart Parsing

- Top-down rule invocation depth-first:
  - Active edge:
    - edge(0, 0, S $\rightarrow \bullet$ NP VP)
  - Rules: NP $\rightarrow$ N; N $\rightarrow$ John
  - New edges:
    - edge(0, 0, NP $\rightarrow \bullet$ N)
    - edge(0, 0, N $\rightarrow \bullet$ *John*)

Chart Parsing

- Top-down after rule invocation and fundamental rule:
  - New edges:
    - edge(0, 1, S $\rightarrow$ NP $\bullet$ VP)
    - edge(0, 1, NP $\rightarrow$ N $\bullet$)
    - edge(0, 1, N $\rightarrow$ *John* $\bullet$)

Chart Parsing

- Top-down rule invocation breadth-first:
  - Active edge:
    - edge(0, 0, S $\rightarrow \bullet$ NP VP)
  - Rules: NP $\rightarrow$ N; VP $\rightarrow$ V NP
  - New edges:
    - edge(0, 0, NP $\rightarrow \bullet$ N)
    - edge(0, 0, VP $\rightarrow \bullet$ V NP)

Chart Parsing

- Fundamental Rule
  - For every active edge find expected inactive edge:
    - edge(0, 0, NP $\rightarrow$ $\bullet$ N)
    - edge(0, 1, N $\rightarrow$ John $\bullet$)
  - Merge edges and add resulting edge to chart:
    - edge(0, 1, NP $\rightarrow$ N $\bullet$)

Chart Parsing

- Fundamental Rule
  - For every active edge find expected inactive edge:
    - edge(0, 0, S $\rightarrow$ $\bullet$ NP VP)
    - edge(0, 1, NP $\rightarrow$ N $\bullet$)
  - Merge edges and add resulting edge to chart:
    - edge(0, 1, S $\rightarrow$ NP $\bullet$ VP)

Chart Parsing

- Rule Invocation
  - Dependent of parsing strategy.
- Fundamental Rule
  - Independent of parsing strategy.

Chart Parsing

- Differences between top-down and bottom-up parsing:
  - TD: Disambiguates by position.
    - *Calls from Alaska are expensive.*
  - BU: Lexically driven.
  - TD: Has to handle recursion.