Parser Generation

- **Main Problem**: given a grammar $G$ how to build a top-down parser or a bottom-up parser for it?
- **parser**: a program that, given a sentence, reconstructs a derivation for that sentence --- if done successfully, it “recognize” the sentence
- all parsers read their input **left-to-right**, but construct parse tree differently.
- **bottom-up parsers**: construct the tree from leaves to root
  - shift-reduce, LR, SLR, LALR, operator precedence
- **top-down parsers**: construct the tree from root to leaves
  - recursive descent, predictive parsing, LL(1)

Bottom-Up Parsing

- Construct parse tree “bottom-up” --- from leaves to the root
- **Bottom-up parsing always constructs right-most derivation**
- Important parsing algorithms: **shift-reduce**, LR parsing
- **LR parser** components: input, stack (strings of grammar symbols and states), driver routine, parsing tables.

LR Parsing

- A sequence of new **state** symbols $s_0, s_1, s_2, \ldots, s_m$ --- each state summarize the information contained in the stack below it.
- **Parsing configurations**: (stack, remaining input) written as $(s_0X_1s_1X_2s_2\ldots X_ms_m, a_1a_{1+1}a_{2+2}\ldots a_n\$$)
  - next “move” is determined by $s_m$ and $a_i$
- **Parsing tables**: $\text{ACTION}[s,a]$ and $\text{GOTO}[s,X]$
  - **Table A** $\text{ACTION}[s,a]$ --- $s$ : state, $a$ : terminal
    - its entries
      - (1) shift $s_k$
      - (2) reduce $A \rightarrow \beta$
      - (3) accept
      - (4) error
  - **Table G** $\text{GOTO}[s,X]$ --- $s$ : state, $X$ : non-terminal
    - its entries are states

Constructing LR Parser

**How to construct the parsing table $\text{ACTION}$ and $\text{GOTO}$?**

- **basic idea**: first construct DFA to recognize handles, then use DFA to construct the parsing tables! different parsing table yield different LR parsers SLR(1), LR(1), or LALR(1)
- **augmented grammar** for context-free grammar $G = (T,N,P,S)$ is defined as $G' = (T,N \cup \{ S' \}, P \cup \{ S' \rightarrow S, S' \}$ ---- adding non-terminal $S'$ and the production $S' \rightarrow S$, and $S'$ is the new start symbol. When $S' \rightarrow S$ is reduced, parser accepts.
- **LR(0) item** for productions of a context-free grammar $G$ ---- is a production with dot at some position in the r.h.s.
  - For $A \rightarrow XYZ$, its items are $A \rightarrow .XYZ$, $A \rightarrow X.YZ$
  - For $A \rightarrow \varepsilon$, its items are just $A \rightarrow \$
LR(0) items and LR(0) DFA

- Informally, item $A \rightarrow X \cdot Y \gamma$ means a string derivable from $X$ has been seen, and one from $Y \gamma$ is expected. LR(0) items are used as state names for LR(0) DFA or LR(0) NFA that recognizes viable prefixes.

- Viable prefixes of a CFG are prefixes of right-sentential forms with no symbols to right of the handle; we can always add terminals on right to form a right-sentential form.

- Two way to construct the LR(0) DFA:
  1. first construct LR(0) NFA and then convert it to a DFA!
  2. construct the LR(0) DFA directly!

- From LR(0) DFA to the Parsing Table

Example: LR(0) Items

CFG Grammar:

- $E \rightarrow E + T | T$
- $T \rightarrow T * F | F$
- $F \rightarrow (E) | id$

Augmented Grammar:

- $E' \rightarrow E$
- $E \rightarrow E + T | T$
- $T \rightarrow T * F | F$
- $F \rightarrow (E) | id$

LR(0) terms:

- $E' \rightarrow \cdot E$
- $T \rightarrow \cdot T$
- $\varepsilon$
- $F$
- $F \rightarrow (E)$
- $E \rightarrow \cdot E$
- $T \rightarrow \cdot T$
- $F \rightarrow \cdot F$
- $id$

From LR(0) NFA to LR(0) DFA

- Construct LR(0) NFA with all LR(0) items of G as states, connect states by moving the dot; final states are those with dots at the end.

1. for each item $A \rightarrow \alpha X \beta$, $A \rightarrow \alpha X \cdot \beta$

2. for each pair $A \rightarrow \alpha X \beta$, $B \rightarrow \cdot \gamma$ (expect to see a string derivable from $\gamma$)

- Convert NFA to DFA using subset construction algorithm.

- The states of the resulting LR(0) DFA --- $C = \{I_1, I_2, \ldots, I_n\}$ are called canonical LR(0) collection for grammar $G'$

- Disadvantage: the NFA is often huge, and converting from NFA to DFA is tedious and time-consuming.

Building LR(0) DFA Directly

- Instead of building DFA from NFA, we can build the LR(0) DFA directly.

- Given a set of LR(0) items $I$, $\text{CLOSURE}(I)$ is defined as repeat for each item $A \rightarrow \alpha X \beta$ in $I$ and each production $B \rightarrow \gamma$ add $B \rightarrow \cdot \gamma$ to $I$, if it's not in $I$ until I does not change

- $\text{GOTO}(I,X)$ is defined as $\text{CLOSURE}(\{all \ items \ A \rightarrow \alpha X \beta \ for \ each \ A \rightarrow \alpha X \cdot \beta \ in \ I\})$

- Canonical LR(0) collection is computed by the following procedure:

  - $I_0 = \text{CLOSURE}\{S' \rightarrow \cdot S\}$ and $C = \{I_2\}$ repeat for each $I \in C$ and grammar symbol $X$

    - $T = \text{GOTO}(I,X)$; if $T = \emptyset$ and $T \notin C$ then $C = C \cup \{ T \}$; until $C$ does not change

    - Resulting LR(0) DFA: $C$ is the set of states; $\text{GOTO}$ is the transition table
Constructing SLR(1) Parsing Table

- From the LR(0) DFA, we can construct the parsing table. The parser based on SLR(1) parsing table is called SLR(1) parser. The SLR(1) grammars are those whose SLR(1) parsing table does not contain any conflicts.

- Algorithm — use $C = \{I_0, \ldots, I_n\}$, GOTO, FOLLOW:

  1. If $A \rightarrow \alpha \cdot$ is in $I_i$ and GOTO($I_i, \alpha$) = $I_j$ where $\alpha$ is a terminal, set $\text{ACTION}[I_i, \alpha]$ to "shift j".
  2. If $A \rightarrow \alpha \cdot$ is in $I_i$, set $\text{ACTION}[I_i, \alpha]$ to "reduce $A \rightarrow \alpha \cdot$" for all terminal $\alpha$ in FOLLOW($A$).
  3. If $S' \rightarrow S.$ is in $I_i$, set $\text{ACTION}[I_i, \emptyset]$ to "accept"
  4. If GOTO($I_i, A$) = $I_j$, set GOTO[$I_i, A$] = $I_j$
  5. set all other entries to "error"
  6. set initial state to be $I_i$ with $S' \rightarrow S.$

Limitation of SLR(1) Parser

- Unfortunately, many unambiguous grammars are not SLR(1) grammars

LR(1) Parsing

- Conflict arises because LR(0) states do not encode enough left context — the previous example, reduction $R \rightarrow L$ is wrong upon input $=$ because "$R = \ldots$" never appears in right-sentential form.

- Solution: split LR(0) states by adding terminals to states, for example, $[A \rightarrow \alpha., a]$ results in reduction only if next symbol is $a$.

- An LR(1) term is in the form of $[A \rightarrow \alpha \cdot \beta, a]$ where $A \rightarrow \alpha \beta$ is a production and $a$ is a terminal or $\emptyset$

- To build LR(1) parsing table — we first build LR(1) DFA — then construct the parsing table using the same SLR(1) algorithm except

  2. only if $[A \rightarrow \alpha., a]$ is in $I_1$, then set $\text{ACTION}[I_1, \alpha]$ to "reduce $A \rightarrow \alpha$"

- To way to build LR(1) DFA — from NFA -> DFA or build DFA directly

Building LR(1) DFA

- Construct LR(1) NFA with all LR(1) items of $G$ as states, connect states by moving the dot; then convert the NFA to DFA.

  1. for each item $[A \rightarrow \alpha \cdot X \beta \cdot, a]$

  2. for each pair $[A \rightarrow \alpha \cdot \beta, a] \beta \rightarrow \gamma$ and $b \in \text{FIRST}(\beta)$,

- Construct the LR(1) DFA directly (see the Dragon book)

- Given a set of LR(1) items $I$, $\text{CLOSURE}(I)$ is now defined as

  repeat

  for each item $[A \rightarrow \alpha \cdot \beta, a]$ in $I$ and each production $B \rightarrow \gamma$ and each terminal $b$ in FIRST($\beta$),

  add $[B \rightarrow \gamma, a]$ to $I$, if it's not in $I$

  until $I$ does not change
Constructing LR(1) Parser

- Canonical LR(1) collection is computed by the following procedure:
  \[ I_0 = \text{CLOSURE}(\{S' \rightarrow .S, \epsilon\}) \text{ and } C = \{I_0\} \]
  repeat
  for each \( I \in C \) and grammar symbol \( X \)
  \[ T = \text{GOTO}(I, X); \text{ if } T \neq \emptyset \text{ and } T \notin C \text{ then } C = C \cup \{T\} \]
  until \( C \) does not change

Resulting LR(1) DFA: \( C \) is the set of states; \( \text{GOTO} \) is the transition table

- From the LR(1) DFA, we can construct the parsing table ---- LR(1) parsing table. The parser based on LR(1) parsing table is called LR(1) parser. The LR(1) grammars are those whose LR(1) parsing table does not contain any conflicts (no duplicate entries).

- Example:
  \[
  S' \rightarrow S \\
  S \rightarrow C C \\
  C \rightarrow c C | d
  \]

LALR(1) Parsing

- Bad News: LR(1) parsing tables are too big; for PASCAL, SLR tables has about 100 states, LR table has about 1000 states.

- LALR (LookAhead-LR) parsing tables have same number of states as SLR, but use lookahead for reductions. The LALR(1) DFA can be constructed from the LR(1) DFA.

- LALR(1) states can be constructed from LR(1) states by merging states with same core, or same LR(0) items, and union their lookahead sets.

  Merging \( I_8: C \rightarrow c, C, c / d \) into a new state \( I_{89}: C \rightarrow c, c / d / \)

  Merging \( I_3: C \rightarrow c, C, c / d \) into a new state \( I_{36}: C \rightarrow c, c / d / \)

Summary: LR Parser

- Relation of three LR parsers: \( LR(1) > LALR(1) > SLR(1) \)

- Most programming language constructs are LALR(1). The LR(1) is unnecessary in practice, but the SLR(1) is not enough.

- YACC is an LALR(1) Parser Generator.

- When parsing ambiguous grammars using LR parsers, the parsing table will contain multiple entries. We can specify the precedence and associativity for terminals and productions to resolve the conflicts. YACC uses this trick.

- Other Issues in parser implementation: 1. compact representation of parsing table 2. error recovery and diagnosis.
Top-Down Parsing

- Starting from the start symbol and "guessing" which production to use next step. It often uses next input token to guide "guessing".

example:  
\[ S \rightarrow c \ A \ d \]  
\[ A \rightarrow a b \ | \ a \]

```
input symbols: cad  
we are looking ahead only one at a time!
```

```
“c” matches  
try to decide which rule of A to use here?
```

```
we guessed wrong, backtrack! try another one!
```

```
Top-Down Parsing (cont’d)
```

- Typical implementation is to write a recursive procedure for each non-terminal (according to the r.h.s. of each grammar rule)

```plaintext
Grammar:
E → T E'  
E' → + T E' | ε  
T → F T'  
T' → * F T' | ε  
F → id  
| ( E )
```

```
Algorithm: Recursive Descent
```

- Parsing Algorithm (using 1-symbol lookahead in the input)

1. Given a set of grammar rules for a non-terminal \( A \)

\[ A \rightarrow \alpha_1 \ | \ \alpha_2 \ | \ldots \ | \alpha_n \]

we choose proper alternative by looking at first symbol it derives ---- the next input symbol decides which \( \alpha_i \) we use

2. for \( A \rightarrow \varepsilon \), it is taken when none of the others are selected

- Algorithm: constructing a recursive descent parser for grammar \( G \)

1. transform grammar \( G \) to \( G' \) by removing left-recursions and do the left-factoring.

2. write a (recursive) procedure for each non-terminal in \( G' \)
Left Recursion Elimination

- **Elimination of Left Recursion** (useful for top-down parsing only)

  replace productions of the form

  \[ A \rightarrow A \alpha | \beta \]

  with

  \[ A \rightarrow \beta A' \]

  \[ A' \rightarrow \alpha A' | \epsilon \]

(yields different parse trees but same language)

**Example:**

- \[ E \rightarrow E + T | T \]
- \[ T \rightarrow T * F | F \]
- \[ F \rightarrow (E) | id \]

become

- \[ E \rightarrow T E' \]
- \[ E' \rightarrow + T E' | \epsilon \]
- \[ T \rightarrow F T' \]
- \[ T' \rightarrow * F T' | \epsilon \]
- \[ F \rightarrow (E) | id \]

**Important:** read Appel pp 51 - 53 for details

Left Factoring

- **Some grammars are unsuitable for recursive descent, even if there is no left recursion**

  "dangling-else"

  \[ stmt \rightarrow if \ expr \ then \ stmt | if \ expr \ then \ stmt \ else \ stmt | \ldots.. \]

  input symbol if does not uniquely determine alternative.

  - **Left Factoring** --- factor out the common prefixes (see AHU pp 178)

  change the production

  \[ A \rightarrow x y | x z \]

  to

  \[ A \rightarrow x A' \]

  \[ A' \rightarrow y | z \]

  thus

  \[ stmt \rightarrow if \ expr \ then \ stmt S' \]

  \[ S' \rightarrow else \ stmt | \epsilon \]

Predictive Parsing

- **Predictive parsing** is just table-driven recursive descent; it contains:

  - A **parsing stack** --- contains terminals and non-terminals
  - A **parsing table** : a 2-dimensional table \[ M[A,a] \] where \[ X \] is non-terminal, \[ a \] is terminal, and table entries are grammar productions or error indicators.

  **Algorithm**

  - \$ \ is end-of-file, \$ \ is start symbol
  - \[ push(\$); push(\$); \]
  - \[ while top \neq \$ \ do \{
      a <= \ the \ input \ char
      if top is a terminal or \$ \ then
      \[ if \ top \neq \ a \ then
        pop(); \ advance();
      \] else \ err();
    \] else if \[ M[top,a] \] is \[ X \rightarrow Y_1 Y_2 \ldots Y_k \] then
    \[ pop(); \]
    \[ push(Y_k); \ldots; \ push(Y_1); \]
    \[ else \ err(); \]
  \} )

  **Constructing Predictive Parser**

  - The key is to build the parse table \[ M[A,a] \]

    \[ for \ each \ production \ A \rightarrow \alpha \ do \]

    \[ if \ a \in \ FIRST(\alpha) \ do \]

    \[ add \ A \rightarrow \alpha \ to \ M[A,a] \]

    \[ if \ \epsilon \in \ FIRST(\alpha) \ then \]

    \[ for \ each \ b \in \ FOLLOW(A) \ do \]

    \[ add \ A \rightarrow \alpha \ to \ M[A,b] \]

    \[ rest \ of \ M \ is \ error \]

  - **FIRST(\alpha)** is a set of terminals (plus \$\) that begin strings derived from \alpha, where \alpha is any string of non-terminals and terminals.

  - **FOLLOW(A)** is a set of terminals that can follow \[ A \] in a sentential form, where \[ A \] is any non-terminal.
First & Follow

• To compute \textsc{first}(X) for any grammar symbol \(X\):

\[
\textsc{first}(X) =\begin{cases} 
\{X\}, & \text{if } X \text{ is a terminal;} \\
\textsc{first}(X) \cup \{a\}, & \text{if } X \rightarrow a \alpha; \\
\textsc{first}(X) \cup \{\varepsilon\}, & \text{if } X \rightarrow \varepsilon; \text{ and} \\
\textsc{first}(X) \cup \textsc{first}(Y_1Y_2\ldots Y_k), & \text{if } X \rightarrow Y_1Y_2\ldots Y_k.
\end{cases}
\]

repeat until nothing new is added to any \textsc{first}

• \textsc{first}(Y_1Y_2\ldots Y_k) = \textsc{first}(Y_1) - \{\varepsilon\} \cup \textsc{first}(Y_2) - \{\varepsilon\} \cup \textsc{first}(Y_3) - \{\varepsilon\} \cup \ldots \cup \textsc{first}(Y_k) - \{\varepsilon\}.

\[
\cup \{\varepsilon\} \quad \text{if all } \textsc{first}(Y_{i+1}) \text{ contain } \varepsilon.
\]

First & Follow (cont’d)

• To compute \textsc{follow}(X) for any non-terminal \(X\):

\[
\textsc{follow}(S) = \textsc{follow}(S) \cup \{\$\}, \quad \text{if } S \text{ is start symbol;}
\]

\[
\textsc{follow}(B) = \textsc{follow}(B) \cup \textsc{first}(\beta) - \{\varepsilon\}, \quad \text{if } A \rightarrow \alpha B \beta \text{ and } \beta \neq \varepsilon
\]

\[
\textsc{follow}(B) = \textsc{follow}(B) \cup \textsc{follow}(A) \quad \text{if } A \rightarrow \alpha B \text{ or } A \rightarrow \alpha B \beta \text{ and } \varepsilon \in \textsc{first}(\beta)
\]

• Example:

\[
\textsc{follow}(E) = \{\$\}, \quad \text{E is start symbol}
\]

\[
\textsc{follow}(E') = \textsc{follow}(E) - \{\$\}, \quad \text{E} \rightarrow \text{T E'}
\]

(Read Appel pp 47 - 53 for detailed examples)

Summary: LL(1) Grammars

• A grammar is \textsc{ll}(1) if parsing table \textsc{M}[A, a] has no duplicate entries, which is equivalent to specifying that for each production

\[
A \rightarrow \alpha_1 \mid \alpha_2 \mid \ldots \mid \alpha_n
\]

1. All \textsc{first}(\alpha_i) are disjoint.

2. At most one \(\alpha_i\) can derive \(\varepsilon\); in that case, \textsc{follow}(A) must be disjoint from \textsc{first}(\alpha_2) \cup \textsc{first}(\alpha_3) \cup \ldots \cup \textsc{first}(\alpha_n)

• Left-recursion and ambiguity grammar lead to multiple entries in the parsing table. (try the dangling-else example)

• The main difficulty in using (top-down) predicative parsing is in rewriting a grammar into an \textsc{ll}(1) grammar. There is no general rule on how to resolve multiple entries in the parsing table.