TD2: LL(1) Parsing

Top-down Parsing

CMPT 379: Compilers

Instructor: Anoop Sarkar

anoopsarkar.github.io/compilers-class

Parsing - Roadmap

- Parser:
 - decision procedure: builds a parse tree
- Top-down vs. bottom-up
- LL(1) Deterministic Parsing
 - recursive-descent
 - table-driven
- LR(k) Deterministic Parsing
 - LR(0), SLR(1), LR(1), LALR(1)
- Parsing arbitrary CFGs Polynomial time parsing

Top-Down vs. Bottom Up

Grammar: $S \rightarrow AB$ Input String: ccbca

 $A \rightarrow c \mid \epsilon$

 $B \rightarrow cbB \mid ca$

Top-Down/leftmost		Bottom-Up/rightmost		
$S \Rightarrow AB$	$S \rightarrow AB$	ccbca ← Acbca	A→c	
\Rightarrow cB	$A \rightarrow c$	← AcbB	B→ca	
⇒ ccbB	B→cbB	←AB	B→cbB	
⇒ccbca	B→ca		$S \rightarrow AB$	

Leftmost derivation for

$$E \rightarrow E + E$$

$$E \rightarrow E * E$$

$$E \rightarrow (E)$$

$$E \rightarrow -E$$

$$E \rightarrow id$$

$$E \Rightarrow E + E$$

$$\Rightarrow$$
 id + E

$$\Rightarrow$$
 id + E * E

$$\Rightarrow$$
 id + id * E

$$\Rightarrow$$
 id + id * id

$$E \Rightarrow^*_{lm} id + E \setminus^* E$$

Predictive Top-Down Parser

- Knows which production to choose based on single lookahead symbol
- Need LL(1) grammars

First L: reads input Left to right

Second L: produce Leftmost derivation

- 1: one symbol of lookahead

- Cannot have left-recursion
- Must be left-factored (no left-factors)
- Not all grammars can be made LL(1)

LL(1) Parser

- In recursive-descent
 - for each non-terminal and input token, many choices of production to use
 - Backtracking to remove bad choices
- In LL(1)
 - for each non-terminal and each token, only one production

```
S \rightarrow^* \omega A \beta and next input token: t
A \rightarrow \alpha \quad \text{is the only production}
\omega \alpha \beta
```

Left Factoring

Consider this grammar

```
-E \rightarrow T + E \mid T
-T \rightarrow id \mid id * T \mid (E)
```

- Hard to predict because
 - For T two productions start with id
 - For E it is not clear how to predict
- The grammar must not have left-recursion
- The grammar should be left-factored

Left Factoring

In general, for rules

$$A \to \alpha \beta_1 \mid \alpha \beta_2 \mid \ldots \mid \alpha \beta_n \mid \gamma$$

 Left factoring is achieved by the following grammar transformation:

$$A \rightarrow \alpha A' \mid \gamma$$

$$A' \rightarrow \beta_1 \mid \beta_2 \mid \dots \mid \beta_n$$

Left Factoring

Recall the grammar

$$-E \rightarrow T + E \mid T$$

$$-T \rightarrow id \mid id * T \mid (E)$$

Factor out common prefixes for productions

```
-E \rightarrow TX
-X \rightarrow +E \mid \epsilon
-T \rightarrow id Y \mid (E)
-Y \rightarrow *T \mid \epsilon
```

- Can be specified via 2D tables
 - One dimension for current (leftmost) non-terminal to expand
 - One dimension for next token
 - Each table entry contains one production

	Productions		
1	$E \rightarrow T X$		
2	$X \rightarrow \epsilon$		
3	$X \rightarrow + E$		
4	$T \rightarrow (E)$		
5	$T \rightarrow id Y$		
6	$Y \rightarrow *T$		
7	$Y \rightarrow \epsilon$		

	+	*	()	id	\$
E			TX		TX	
X	+ E			3		3
T			(E)		id Y	
Y	3	* T		3		3

- Consider [E, id] entry
 - When current non-terminal is E and the next input is id, use production $E \rightarrow T X$

	Productions		
1	$E \rightarrow T X$		
2	$X \rightarrow \epsilon$		
3	$X \rightarrow + E$		
4	$T \rightarrow (E)$		
5	$T \rightarrow id Y$		
6	$Y \rightarrow *T$		
7	$Y \rightarrow \epsilon$		

	+	*	()	id	\$
E			TX		TX	
X	+ E			3		3
T			(E)		id Y	
Y	3	* T		3		3

- Consider [Y, +] entry
 - When current non-terminal is Y and the next input is +, get rid of Y
 - Y can be followed by + only if $Y \rightarrow \varepsilon$

	Productions		
1	$E \rightarrow T X$		
2	$X \rightarrow \epsilon$		
3	$X \rightarrow + E$		
4	$T \rightarrow (E)$		
5	$T \rightarrow id Y$		
6	$Y \rightarrow *T$		
7	$Y \rightarrow \epsilon$		

	+	*	()	id	\$
E			TX		TX	
X	+ E			ε		3
T			(E)		id Y	
Y	3	* T		ε		3

- Blank entries indicate error situations
- Consider [E, *] entry
 - There is no way to derive a string starting with * from non-terminal E

	Productions		
1	$E \rightarrow T X$		
2	$X \rightarrow \epsilon$		
3	$X \rightarrow + E$		
4	$T \rightarrow (E)$		
5	$T \rightarrow id Y$		
6	$Y \rightarrow *T$		
7	$Y \rightarrow \epsilon$		

	+	*	()	id	\$
E			TX		TX	
X	+ E			3		3
T			(E)		id Y	
Y	3	* T		3		3

Predictive Parsing

- Method similar to recursive descent, except
 - For each non-terminal S
 - We look at the next token a
 - And chose the production shown at entry [S,a]
- We use a stack to keep track of pending nonterminals (frontier of parse tree)
- We reject when we encounter an error state
- We accept when we encounter end-of-input and empty stack

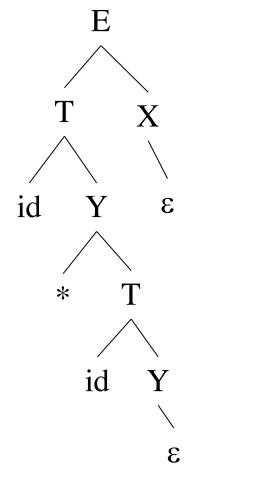
Table-Driven Parsing

```
Stack: to keep track
stack.push($); stack.push($);
                                      of pending non-terminals
a = input.read();
                                      in the derivation
forever do begin
                                      (leaves in parse tree)
  X = stack.peek();
  if X = a and a = $ then return SUCCESS;
  elsif X = a and a != $ then
    stack.pop(X); a = input.read();
  elsif X = a and X \in N and M[X,a] not empty then
    stack.pop(X);
    stack.push(M[X,a]); /* M[X, a] = Y_1...Y_n */
                                  X \longrightarrow Y_1 \dots Y_n
  else ERROR!
end
```

Trace "id*id"

	+	*	()	id	\$
Е			ΤX		ΤX	
X	+ E			3		3
T			(E)		id Y	
Y	3	* T		3		3

Stack	Input	Action
E \$	id*id\$	ΤX
T X \$	id*id\$	id Y
id Y X \$	id*id\$	terminal
Y X \$	*id\$	* T
* T X \$	*id\$	terminal
T X \$	id\$	id Y
id Y X \$	id\$	terminal
Y X \$	\$	ε
X \$	\$	ε
\$	\$	Accept!



When to pick $Y \rightarrow \epsilon$?

	Productions		
1	$E \rightarrow T X$		
2	$X \rightarrow \epsilon$		
3	$X \rightarrow + E$		
4	$T \rightarrow (E)$		
5	$T \rightarrow id Y$		
6	$Y \rightarrow *T$		
7	$Y \rightarrow \epsilon$		

- Choice between $Y \rightarrow *T$ and $Y \rightarrow \varepsilon$
- $FIRST(*T) = \{ * \}$
- For $Y \rightarrow \varepsilon$ we compute FOLLOW(Y)
- FOLLOW(Y) = ?
- FOLLOW(Y) = FOLLOW(T)
- FOLLOW(T) = (FIRST(X) $\{\epsilon\}$) + FOLLOW(E)
- $FOLLOW(T) = \{ +,), \}$
- $FOLLOW(Y) = \{ +, \}, \}$

- Given a grammar produce the predictive parsing table
- We need to to know for all rules A $\rightarrow \alpha \mid \beta$ the lookahead symbol
- Based on the lookahead symbol the table can be used to pick which rule to push onto the stack
- This can be done using two sets: FIRST and FOLLOW

- For Nonterminal A, rule $A \rightarrow \alpha$, and the token t, $M[A, t] = \alpha$ in two cases:
- If $\alpha \Rightarrow * t \beta$
 - $-\alpha$ can derive a t in the first position
 - We say that $t \in First(\alpha)$
- A $\rightarrow \alpha$ and $\alpha \Rightarrow * \varepsilon$ and S $\Rightarrow * \beta$ A t δ
 - Useful if stack has A, input is t and A cannot derive t
 - In this case only option is to get rid of A (by $\alpha \Rightarrow * \varepsilon$)
 - Can work only if t can follow A in at least on derivation
 - We say t ∈ Follow(A)

FIRST and FOLLOW

$$a \in \text{FIRST}(\alpha) \text{ if } \alpha \Rightarrow^* a\beta$$

if $\alpha \Rightarrow^* \epsilon \text{ then } \epsilon \in \text{FIRST}(\alpha)$
 $a \in \text{FOLLOW}(A) \text{ if } S \Rightarrow^* \alpha A a\beta$
 $a \in \text{FOLLOW}(A) \text{ if } S \Rightarrow^* \alpha A \gamma a\beta$
and $\gamma \Rightarrow^* \epsilon$

Conditions for LL(1)

- Necessary conditions:
 - no ambiguity
 - no left recursion
 - Left factored grammar
- A grammar G is LL(1) if whenever $A \rightarrow \alpha \mid \beta$
 - 1. First(α) \cap First(β) = \emptyset
 - 2. $\alpha \Rightarrow^* \epsilon$ implies !($\beta \Rightarrow^* \epsilon$)
 - 3. $\alpha \Rightarrow^* \epsilon \text{ implies First}(\beta) \cap \text{Follow}(A) = \emptyset$

ComputeFirst(α : string of symbols)

```
// assume \alpha = X_1 X_2 X_3 \dots X_n
if X_1 \in T then First[\alpha] := \{X_1\}
else begin
  i:=1; First[\alpha] := ComputeFirst(X_1)\{\varepsilon};
  while X_i \Rightarrow^* \varepsilon do begin
    if i < n then
       First[\alpha] := First[\alpha] \cup ComputeFirst(X_{i+1})\{\epsilon};
    else
      First[\alpha] := First[\alpha] \cup \{\epsilon\};
    i := i + 1;
                           Recursion in computing FIRST
  end
                           causes problems when faced with
end
                           recursive grammar rules
```

ComputeFirst; modified

```
foreach X \in T do First[X] := {X};
foreach p \in P : X \to \varepsilon do First[X] := \{\varepsilon\};
repeat foreach X \in \mathbb{N}, p: X \to Y_1 Y_2 Y_3 ... Y_n do begin
   i:=1;
     while Y_i \Rightarrow^* \varepsilon and i \le n do begin
       First[X] := First[X] \cup First[Y<sub>i</sub>]\{\varepsilon};
       i := i+1;
     end
    if i = n+1 then First[X] := First[X] \cup \{\epsilon\};
until no change in First[X] for any X;
```

ComputeFirst; modified

```
foreach X \in T do First[X] := X;
foreach p \in P : X \to \varepsilon do First[X] := \{\varepsilon\};
repeat foreach X \in \mathbb{N}, p: X \to Y_1 Y_2 Y_3 ... Y_n do begin
   i:=1;
                     Non-recursive FIRST computation
    while Y<sub>i</sub>⇒* { works with left-recursive grammars.
      First[X] := F Computes a fixed point for FIRST[X]
      i := i+1; for all non-terminals X in the grammar.
                     But this algorithm is very inefficient.
    end
   if i = n+1 then First[X] := First[X] \cup \{\epsilon\};
until no change in First[X] for any X;
```

First Sets

First(+) = {+} First(E) = ?

First(*) = {*} First(T)
$$\subseteq$$
 First(E)

First('(') = {'(')} First(T) = {id, '(')}

First(')') = {')'} First(E) = {id, '(')}

First(id) = {id} First(X) = {+, ε }

First(Y) = {*, ε }

	Productions				
1	$E \rightarrow T X$				
2	$X \rightarrow \epsilon$				
3	$X \rightarrow + E$				
4	$T \rightarrow (E)$				
5	$T \rightarrow id Y$				
6	$Y \rightarrow *T$				
7	$Y \rightarrow \epsilon$				

Follow Sets

- Algorithm sketch
 - 1. Add \$ to Follow(S)
 - 2. For each production $A \longrightarrow \alpha X \beta$
 - Add First(β) { ϵ } to Follow(X)
 - 3. For each $A \rightarrow \alpha X \beta$ where $\epsilon \in First(\beta)$
 - Add Follow(A) to Follow(X)
 - Repeat steps 2-3 until no follow set grows

ComputeFollow

```
Follow(S) := {$};
repeat
 foreach p \in P do
     case p = A \rightarrow \alpha B\beta begin
       Follow[B] := Follow[B] \cup ComputeFirst(\beta)\{\epsilon};
       if \varepsilon \in First(\beta) then
        Follow[B] := Follow[B] \cup Follow[A];
     end
    case p = A \rightarrow \alpha B
       Follow[B] := Follow[B] \cup Follow[A];
until no change in any Follow[N]
```

Follow Sets. Example

 $Follow(E) \subseteq Follow(X)$

 $Follow(X) \subseteq Follow(E)$

 $First(X)-\{\varepsilon\}\subseteq Follow(T)$

 $Follow(E) \subseteq Follow(T)$

 $Follow(Y) \subseteq Follow(T)$

 $Follow(T) \subseteq Follow(Y)$

 $Follow(E) = \{\$, \}$

 $Follow(X) = \{\$, \}$

Follow(T) = $\{+, \$, \}$

Follow(Y) = $\{+, \$, \}$

Follow('(') = $\{(, id)\}$

Follow(')') = $\{+, \$, \}$

 $Follow(+) = \{(, id)\}$

Follow(*) = $\{(, id)\}$

Follow(id) = $\{*,+,$,)\}$

Productions				
1	$E \rightarrow T X$			
2	$X \rightarrow \epsilon$			
3	$X \rightarrow + E$			
4	$T \rightarrow (E)$			
5	$T \rightarrow id Y$			
6	$Y \rightarrow *T$			
7	$Y \rightarrow \epsilon$			

Building the Parse Table

- Compute First and Follow sets
- For each production A $\rightarrow \alpha$
 - − For each $t \in First(\alpha)$
 - $M[A,t] = \alpha$
 - − If ε ∈ First(α), for each t ∈ Follow(A)
 - $M[A,t] = \alpha$
 - − If ε ∈ First(α) and φ ∈ Follow(α)
 - M[A,\$] = α
 - All undefined entries are errors

First(E) = {id, '(')}
Follow(E) = {\$,)}
First(X) = {+,
$$\epsilon$$
}
Follow(X) = {\$,)}

$First(T) = \{id, '(')\}$
Follow(T) = $\{+, \$, \}$
$First(Y) = {*, \epsilon}$
Follow(Y) = $\{+, \$, \}$

Productions			
1	$E \rightarrow T X$		
2	$X \rightarrow \epsilon$		
3	$X \rightarrow + E$		
4	$T \rightarrow (E)$		
5	$T \rightarrow id Y$		
6	$Y \rightarrow *T$		
7	$Y \rightarrow \epsilon$		

	+	*	()	id	\$
E			TX		TX	
X	+ E			3		3
T			(E)		id Y	
Y	3	* T		3		3

Example First/Follow

$$S \rightarrow AB$$

$$A \rightarrow c \mid \epsilon$$

 $B \rightarrow cbB \mid ca$

$$First(A) = \{c, \epsilon\}$$

 $First(B) = \{c\}$

First(cbB) =

 $First(ca) = \{c\}$

$$First(S) = \{c\}$$

Not an LL(1) grammar

$$Follow(A) = \{c\}$$

 $Follow(A) \cap$

$$First(c) = \{c\}$$

$$Follow(B) = \{\$\}$$

$$Follow(S) = \{\$\}$$

Converting to LL(1)

$$S \rightarrow AB$$

$$A \rightarrow c \mid \epsilon$$

$$B \rightarrow cbB \mid ca$$

Note that grammar

same as:

$$S \rightarrow cAa$$

$$A \rightarrow cB \mid B$$

$$B \rightarrow bcB \mid \epsilon$$

Verifying LL(1) using F/F sets

$$S \rightarrow cAa$$

 $A \rightarrow cB \mid B$
 $B \rightarrow bcB \mid \epsilon$
First(A) = {b, c, \epsilon} Follow(A) = {a}
First(B) = {b, \epsilon} Follow(B) = {a}
First(S) = {c} Follow(S) = {\$}

Building the Parse Table

- Compute First and Follow sets
- For each production A $\rightarrow \alpha$
 - foreach a ∈ First(α) add A \rightarrow α to M[A,a]
 - If $\varepsilon \in First(\alpha)$ add $A \to \alpha$ to M[A,b] for each b in Follow(A)
 - If $\varepsilon \in First(\alpha)$ add $A \rightarrow \alpha$ to M[A,\$] if \$ \in Follow(α)
 - All undefined entries are errors

Productions1 $T \rightarrow F T'$ 2 $T' \rightarrow \varepsilon$ 3 $T' \rightarrow * F T'$ 4 $F \rightarrow id$ 5 $F \rightarrow (T)$

FIRST(T) =
$$\{id, (\}$$

FIRST(T') = $\{*, \epsilon\}$
FIRST(F) = $\{id, (\}$

	*	()	id	\$
T		$T \rightarrow FT'$		$T \rightarrow F T'$	
T'	T'→* F T'		$T' \rightarrow \epsilon$		Τ'→ε
F		$\mathbf{F} \rightarrow (\mathbf{T})$		$F \rightarrow id$	

Revisit conditions for LL(1)

- A grammar G is LL(1) iff whenever A $\rightarrow \alpha \mid \beta$
 - 1. First(α) \cap First(β) = \emptyset
 - 2. $\alpha \Rightarrow^* \epsilon$ implies !($\beta \Rightarrow^* \epsilon$)
 - 3. $\alpha \Rightarrow^* \epsilon \text{ implies First}(\beta) \cap \text{Follow}(A) = \emptyset$
- No more than one entry per table field

Error Handling

- Reporting & Recovery
 - Report as soon as possible
 - Suitable error messages
 - Resume after error
 - Avoid cascading errors
- Phrase-level vs. Panic-mode recovery

Panic-Mode Recovery

- Skip tokens until synchronizing set is seen
 - Follow(A)
 - garbage or missing things after
 - Higher-level start symbols
 - First(A)
 - garbage before
 - Epsilon
 - if nullable
 - Pop/Insert terminal
 - "auto-insert"
- Add "synch" actions to table

Summary so far

- LL(1) grammars, necessary conditions
 - No left recursion
 - Left-factored
- Not all languages can be generated by LL(1) grammar
- LL(1) Parsing: O(n) time complexity
 - recursive-descent and table-driven predictive parsing
- LL(1) grammars can be parsed by simple predictive recursive-descent parser
 - Alternative: table-driven top-down parser

Extra Slides

ComputeFirst on Left-recursive Grammars

- ComputeFirst as defined earlier loops on leftrecursive grammars
- Here is an alternative algorithm for ComputeFirst
 - 1. Compute non left-recursive cases of FIRST
 - 2. Create a graph of recursive cases where FIRST of a non-terminal depends on another non-terminal
 - 3. Compute Strongly Connected Components (SCC)
 - 4. Compute FIRST starting from root of SCC to avoid cycles

ComputeFirst on Left-recursive Grammars

- Each Strongly Connected Component can have recursion
- But the connections between SCC means that (by defn) what we have now is a directed acyclic graph – hence without left recursion
- Unlike top-down LL parsing, bottom-up LR parsing allows left-recursive grammars, so this algorithm is useful for LR parsing

ComputeFirst on Left-recursive Grammars

- $S \rightarrow BD \mid D$
- $D \rightarrow d \mid Sd$

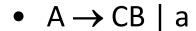
$$FIRST_0[A] := \{a\}$$

$$FIRST_0[C] := \{\}$$

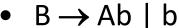
$$FIRST_0[B] := \{b\}$$

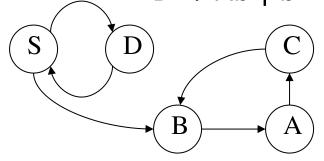
$$FIRST_0[S] := \{b, d\}$$

$$FIRST_0[D] := \{d\}$$



•
$$C \rightarrow Bb \mid \varepsilon$$





Compute
Strongly
Connected
Components

2 SCCs: e.g. consider B-A-C

$$FIRST[B] := FIRST_0[B] + ComputeFirst(A)$$

$$FIRST[A] := FIRST_0[A] + ComputeFirst(C)$$

$$FIRST[A] := FIRST[A] + FIRST_0[B]$$

$$FIRST[C] := FIRST_0[C] + FIRST_0[B]$$

$$FIRST[C] := FIRST[C] + \{\epsilon\}$$

Examples

$$S \rightarrow ABC$$
 $A \rightarrow a \mid \epsilon$
 $B \rightarrow bB \mid \epsilon$
 $C \rightarrow c \mid \epsilon$
Is this LL(1)?

$$S \rightarrow F$$

 $F \rightarrow A (B) | B A$
 $A \rightarrow x | y$
 $B \rightarrow a B | b B | \varepsilon$
Is this LL(1)?

Transition Diagram

$$S \rightarrow cAa$$
 S: $C \rightarrow A \rightarrow a$
 $A \rightarrow cB \mid B$ A: $C \rightarrow B \rightarrow bcB \mid \epsilon$ B: $C \rightarrow B \rightarrow bcB \mid \epsilon$