CIS 341: COMPILERS

Lecture 12
Announcements

• **Homework 4:** Building A Frontend
• **Goal:**
  – Work with lexer & parser
  – Compile a C-like source language to LLVM.
• **Available:** soon (tomorrow?)
• **Due:** Wednesday, March 28th

• **MIDTERM EXAM**
  – Thursday, March 1st in class
  – Coverage: interpreters / program transformers / x86 / calling conventions / IRs / LLVM / Lexing / Parsing
  – See examples on the web pages
LR GRAMMARS
Shift/Reduce Parsing

• Parser state:
  – Stack of terminals and nonterminals.
  – Unconsumed input is a string of terminals
  – Current derivation step is $\text{stack} + \text{input}$

• Parsing is a sequence of $\text{shift}$ and $\text{reduce}$ operations:
  • $\text{Shift}$: move look-ahead token to the stack
  • $\text{Reduce}$: Replace symbols $\gamma$ at top of stack with nonterminal $X$ such that $X \rightarrow \gamma$ is a production. (pop $\gamma$, push $X$)

<table>
<thead>
<tr>
<th>Stack</th>
<th>Input</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E \rightarrow \text{number} \mid ( S )$</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
<td>shift (</td>
</tr>
<tr>
<td>(</td>
<td>$1 + 2 + (3 + 4)) + 5$</td>
<td>shift 1</td>
</tr>
<tr>
<td>(1</td>
<td>$2 + (3 + 4)) + 5$</td>
<td>reduce: $E \rightarrow \text{number}$</td>
</tr>
<tr>
<td>(E</td>
<td>$2 + (3 + 4)) + 5$</td>
<td>reduce: $S \rightarrow E$</td>
</tr>
<tr>
<td>(S</td>
<td>$2 + (3 + 4)) + 5$</td>
<td>shift +</td>
</tr>
<tr>
<td>(S +</td>
<td>$2 + (3 + 4)) + 5$</td>
<td>shift 2</td>
</tr>
<tr>
<td>(S + 2</td>
<td>$+ (3 + 4)) + 5$</td>
<td>reduce: $E \rightarrow \text{number}$</td>
</tr>
</tbody>
</table>
Simple LR parsing with no look ahead.

**LR(0) GRAMMARS**
LR Parser States

- Goal: know what set of reductions are legal at any given point.
- Idea: Summarize all possible stack prefixes $\alpha$ as a finite parser state.
  - Parser state is computed by a DFA that reads the stack $\sigma$.
  - Accept states of the DFA correspond to unique reductions that apply.

- Example: LR(0) parsing
  - Left-to-right scanning, Right-most derivation, zero look-ahead tokens
  - Too weak to handle many language grammars (e.g. the “sum” grammar)
  - But, helpful for understanding how the shift-reduce parser works.
Example LR(0) Grammar: Tuples

• Example grammar for non-empty tuples and identifiers:

$$S \rightarrow (L) \mid \text{id}$$
$$L \rightarrow S \mid L, S$$

• Example strings:
  - x
  - (x, y)
  - (((x)))
  - (x, (y, z), w)
  - (x, (y, (z, w)))

Parse tree for: (x, (y, z), w)
**Shift/Reduce Parsing**

- **Parser state:**
  - Stack of terminals and nonterminals.
  - Unconsumed input is a string of terminals.
  - Current derivation step is stack + input.

- Parsing is a sequence of *shift* and *reduce* operations:

- **Shift**: move look-ahead token to the stack: e.g.

<table>
<thead>
<tr>
<th>Stack</th>
<th>Input</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(x, (y, z), w)</td>
<td>shift (</td>
</tr>
<tr>
<td>(</td>
<td>x, (y, z), w)</td>
<td>shift x</td>
</tr>
</tbody>
</table>

- **Reduce**: Replace symbols $\gamma$ at top of stack with nonterminal $X$ such that $X \rightarrow \gamma$ is a production. (pop $\gamma$, push $X$): e.g.

<table>
<thead>
<tr>
<th>Stack</th>
<th>Input</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x</td>
<td>, (y, z), w)</td>
<td>reduce $S \rightarrow$ id</td>
</tr>
<tr>
<td>(S</td>
<td>, (y, z), w)</td>
<td>reduce $L \rightarrow$ S</td>
</tr>
</tbody>
</table>
## Example Run

<table>
<thead>
<tr>
<th>Stack</th>
<th>Input</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x, (y, z), w)</td>
<td>(x, (y, z), w)</td>
<td>shift (</td>
</tr>
<tr>
<td>(x, (y, z), w)</td>
<td>x, (y, z), w)</td>
<td>shift x</td>
</tr>
<tr>
<td>(S, (y, z), w)</td>
<td>(S, (y, z), w)</td>
<td>reduce S → id</td>
</tr>
<tr>
<td>(L, (y, z), w)</td>
<td>(L, (y, z), w)</td>
<td>reduce L → S</td>
</tr>
<tr>
<td>(L, (L, z), w)</td>
<td>(L, (L, z), w)</td>
<td>shift ,</td>
</tr>
<tr>
<td>(L, (L, z), w)</td>
<td>(L, (L, z), w)</td>
<td>shift (</td>
</tr>
<tr>
<td>(L, (L, z), w)</td>
<td>(L, (L, z), w)</td>
<td>shift y</td>
</tr>
<tr>
<td>(L, (L, z), w)</td>
<td>(L, (L, z), w)</td>
<td>reduce S → id</td>
</tr>
<tr>
<td>(L, (L, S), w)</td>
<td>(L, (L, S), w)</td>
<td>reduce L → S</td>
</tr>
<tr>
<td>(L, (L), w)</td>
<td>(L, (L), w)</td>
<td>shift ,</td>
</tr>
<tr>
<td>(L, (L), w)</td>
<td>(L, (L), w)</td>
<td>shift z</td>
</tr>
<tr>
<td>(L, (L), w)</td>
<td>(L, (L), w)</td>
<td>reduce S → id</td>
</tr>
<tr>
<td>(L, (L), w)</td>
<td>(L, (L), w)</td>
<td>reduce L → L, S</td>
</tr>
<tr>
<td>(L, (L), w)</td>
<td>(L, (L), w)</td>
<td>shift )</td>
</tr>
<tr>
<td>(L, (L), w)</td>
<td>(L, (L), w)</td>
<td>reduce S → ( L )</td>
</tr>
<tr>
<td>(L, (L), w)</td>
<td>(L, (L), w)</td>
<td>reduce L → L, S</td>
</tr>
<tr>
<td>(L, (L), w)</td>
<td>(L, (L), w)</td>
<td>shift ,</td>
</tr>
<tr>
<td>(L, (L), w)</td>
<td>(L, (L), w)</td>
<td>shift ,</td>
</tr>
<tr>
<td>(L, (L), w)</td>
<td>(L, (L), w)</td>
<td>shift ,</td>
</tr>
</tbody>
</table>

**Production Rules:**

- **S** → ( L ) | id
- **L** → S | L, S

CIS 347: Compilers
Action Selection Problem

• Given a stack $\sigma$ and a look-ahead symbol $b$, should the parser:
  – Shift $b$ onto the stack (new stack is $\sigma b$)
  – Reduce a production $X \rightarrow \gamma$, assuming that $\sigma = \alpha\gamma$ (new stack is $\alpha X$)?

• Sometimes the parser can reduce but shouldn’t
  – For example, $X \rightarrow \varepsilon$ can always be reduced

• Sometimes the stack can be reduced in different ways

• Main idea: decide what to do based on a prefix $\alpha$ of the stack plus the look-ahead symbol.
  – The prefix $\alpha$ is different for different possible reductions since in productions $X \rightarrow \gamma$ and $Y \rightarrow \beta$, $\gamma$ and $\beta$ might have different lengths.

• Main goal: know what set of reductions are legal at any point.
  – How do we keep track?
LR(0) States

- An LR(0) state is a set of items keeping track of progress on possible upcoming reductions.
- An LR(0) item is a production from the language with an extra separator “.” somewhere in the right-hand-side

```
S ⟷ ( L ) | id
L ⟷ S | L , S
```

- Example items:  S ⟷ .( L )  or  S ⟷ (. L)  or  L ⟷ S.
- Intuition:
  - Stuff before the ‘.’ is already on the stack  
    (beginnings of possible γ’s to be reduced)
  - Stuff after the ‘.’ is what might be seen next
  - The prefixes α are represented by the state itself
Constructing the DFA: Start state & Closure

• First step: Add a new production $S' \rightarrow S$ to the grammar
• Start state of the DFA = empty stack, so it contains the item: $S' \rightarrow .S$

• Closure of a state:
  – Adds items for all productions whose LHS nonterminal occurs in an item in the state just after the ‘.’
  – The added items have the ‘.’ located at the beginning (no symbols for those items have been added to the stack yet)
  – Note that newly added items may cause yet more items to be added to the state… keep iterating until a fixed point is reached.
• Example: $\text{CLOSURE}([S' \rightarrow .S$]) = {S' $\rightarrow .S$, S $\rightarrow .(L)$, S $\rightarrow .id}$

• Resulting “closed state” contains the set of all possible productions that might be reduced next.
Example: Constructing the DFA

- First, we construct a state with the initial item $S' \rightarrow .S$
Example: Constructing the DFA

- Next, we take the closure of that state:
  \[ \text{CLOSURE}\{ S' \rightarrow .S$ \} = \{ S' \rightarrow .S$, \( S \rightarrow .( L ) \), \( S \rightarrow .\text{id} \} \]

- In the set of items, the nonterminal \( S \) appears after the \( .' \).
- So we add items for each \( S \) production in the grammar
Example: Constructing the DFA

- Next we add the transitions:
- First, we see what terminals and nonterminals can appear after the ‘.’ in the source state.
  - Outgoing edges have those label.
- The target state (initially) includes all items from the source state that have the edge-label symbol after the ‘.’, but we advance the ‘.’ (to simulate shifting the item onto the stack)
Example: Constructing the DFA

- Finally, for each new state, we take the closure.
- Note that we have to perform two iterations to compute $\text{CLOSURE}({S \rightarrow (\ . \ L )})$
  - First iteration adds $L \rightarrow .S$ and $L \rightarrow .L, S$
  - Second iteration adds $S \rightarrow .(L)$ and $S \rightarrow .id$
Full DFA for the Example

S' ⟷ .S$
S ⟷ .( L )
S ⟷ .id

S ⟷ id.

L ⟷ L, . S
S ⟷ .( L )
S ⟷ .id

L ⟷ L, S.

S' ⟷ .S$
$↓

Done!

Reduce state: ‘.’ at the end of the production

Current state: run the DFA on the stack.

If a reduce state is reached, reduce.

Otherwise, if the next token matches an outgoing edge, shift.

If no such transition, it is a parse error.
Using the DFA

• Run the parser stack through the DFA.
• The resulting state tells us which productions might be reduced next.
  – If not in a reduce state, then shift the next symbol and transition according to DFA.
  – If in a reduce state, $X \mapsto \gamma$ with stack $\alpha\gamma$, pop $\gamma$ and push $X$.

• Optimization: No need to re-run the DFA from beginning every step
  – Store the state with each symbol on the stack: e.g. $1(3(3 L_5)_6$
  – On a reduction $X \mapsto \gamma$, pop stack to reveal the state too:
    e.g. From stack $1(3(3 L_5)_6$ reduce $S \mapsto ( L )$ to reach stack $1(3$
  – Next, push the reduction symbol: e.g. to reach stack $1(3S$
  – Then take just one step in the DFA to find next state: $1(3S_7$
Implementing the Parsing Table

Represent the DFA as a table of shape:
state * (terminals + nonterminals)

- Entries for the “action table” specify two kinds of actions:
  - Shift and goto state \( n \)
  - Reduce using reduction \( X \mapsto \gamma \)
    - First pop \( \gamma \) off the stack to reveal the state
    - Look up \( X \) in the “goto table” and goto that state
## Example Parse Table

<table>
<thead>
<tr>
<th></th>
<th>(  )</th>
<th>id</th>
<th>,</th>
<th>$</th>
<th>S</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>s3</td>
<td>s2</td>
<td></td>
<td></td>
<td>g4</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>s→id</td>
<td>s→id</td>
<td>s→id</td>
<td>s→id</td>
<td>s→id</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>s3</td>
<td>s2</td>
<td></td>
<td></td>
<td>g7</td>
<td>g5</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>DONE</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>s6</td>
<td>s8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>s→(L)</td>
<td>s→(L)</td>
<td>s→(L)</td>
<td>s→(L)</td>
<td>s→(L)</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>L→S</td>
<td>L→S</td>
<td>L→S</td>
<td>L→S</td>
<td>L→S</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>s3</td>
<td>s2</td>
<td></td>
<td></td>
<td>g9</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>L→L,S</td>
<td>L→L,S</td>
<td>L→L,S</td>
<td>L→L,S</td>
<td>L→L,S</td>
<td></td>
</tr>
</tbody>
</table>

sx = shift and goto state x

gx = goto state x
### Example

- **Parse the token stream**: $(x, (y, z), w)$

<table>
<thead>
<tr>
<th>Stack</th>
<th>Stream</th>
<th>Action (according to table)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon_1$</td>
<td>$(x, (y, z), w)$</td>
<td>s3</td>
</tr>
<tr>
<td>$\varepsilon_1(3)$</td>
<td>$x, (y, z), w)$</td>
<td>s2</td>
</tr>
<tr>
<td>$\varepsilon_1(3x_2)$</td>
<td>, $(y, z), w)$</td>
<td>Reduce: $S \rightarrow \text{id}$</td>
</tr>
<tr>
<td>$\varepsilon_1(3S)$</td>
<td>, $(y, z), w)$</td>
<td>g7  (from state 3 follow $S$)</td>
</tr>
<tr>
<td>$\varepsilon_1(3S_7)$</td>
<td>, $(y, z), w)$</td>
<td>Reduce: $L \rightarrow S$</td>
</tr>
<tr>
<td>$\varepsilon_1(3L)$</td>
<td>, $(y, z), w)$</td>
<td>g5  (from state 3 follow $L$)</td>
</tr>
<tr>
<td>$\varepsilon_1(3L_5)$</td>
<td>, $(y, z), w)$</td>
<td>s8</td>
</tr>
<tr>
<td>$\varepsilon_1(3L_5,8)$</td>
<td>$(y, z), w)$</td>
<td>s3</td>
</tr>
<tr>
<td>$\varepsilon_1(3L_5,8(3)$</td>
<td>$y, z), w)$</td>
<td>s2</td>
</tr>
</tbody>
</table>
LR(0) Limitations

• An LR(0) machine only works if states with reduce actions have a *single* reduce action.
  – In such states, the machine *always* reduces (ignoring lookahead)

• With more complex grammars, the DFA construction will yield states with shift/reduce and reduce/reduce conflicts:

  OK | shift/reduce | reduce/reduce
  ---------------|-------------|--------------
  S → ( L ).     | S → ( L ).  | S → L , S.
  L → .L , S     | L → .L , S  | S → , S.

• Such conflicts can often be resolved by using a look-ahead symbol: LR(1)
Examples

• Consider the left associative and right associative “sum” grammars:

  left
  \[
  S \rightarrow S + E \mid E \\
  E \rightarrow \text{number} \mid (S)
  \]

  right
  \[
  S \rightarrow E + S \mid E \\
  E \rightarrow \text{number} \mid (S)
  \]

• One is LR(0) the other isn’t… which is which and why?
• What kind of conflict do you get? Shift/reduce or Reduce/reduce?

• Ambiguities in associativity/precedence usually lead to shift/reduce conflicts.
LR(1) Parsing

- Algorithm is similar to LR(0) DFA construction:
  - LR(1) state = set of LR(1) items
  - An LR(1) item is an LR(0) item + a set of look-ahead symbols:
    \[ A \rightarrow \alpha \beta , \mathcal{L} \]

- LR(1) closure is a little more complex:
- Form the set of items just as for LR(0) algorithm.
- Whenever a new item \( C \rightarrow \gamma \) is added because \( A \rightarrow \beta . C \delta , \mathcal{L} \) is already in the set, we need to compute its look-ahead set \( \mathcal{M} \):
  1. The look-ahead set \( \mathcal{M} \) includes FIRST(\( \delta \))
     (the set of terminals that may start strings derived from \( \delta \))
  2. If \( \delta \) is or can derive \( \varepsilon \) (i.e. it is nullable), then the look-ahead \( \mathcal{M} \) also contains \( \mathcal{L} \)
Example Closure

\[
S' \rightarrow S$
\]

\[
S \rightarrow E + S \mid E
\]

\[
E \rightarrow \text{number} \mid ( S )
\]

• Start item: \( S' \rightarrow .S$ \ , \ {} \)
• Since S is to the right of a ‘.’, add:
  \[
  S \rightarrow .E + S \ , \ {}$
  \]
  Note: \{\$\} is FIRST($$
  S \rightarrow .E \ , \ {}$
  \]
• Need to keep closing, since E appears to the right of a ‘.’ in
  ‘.E + S’:
  \[
  E \rightarrow .\text{number} \ , \ {}+$
  \]
  Note: + added for reason 1
  \[
  E \rightarrow .( S ) \ , \ {}+$
  \]
  FIRST(+ S) = {+}
• Because E also appears to the right of ‘.’ in ‘.E’ we get:
  \[
  E \rightarrow .\text{number} \ , \ {}$
  \]
  Note: $ added for reason 2
  \[
  E \rightarrow .( S ) \ , \ {}$
  \]
  \( \delta \) is \( \epsilon \)
• All items are distinct, so we’re done
The behavior is determined if:

- There is no overlap among the look-ahead sets for each reduce item, and
- None of the look-ahead symbols appear to the right of a ‘.’

Fragment of the Action & Goto tables:

<table>
<thead>
<tr>
<th>State</th>
<th>Action</th>
<th>Goto</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+</td>
<td>g2</td>
</tr>
<tr>
<td>2</td>
<td>s3</td>
<td>S → E</td>
</tr>
</tbody>
</table>
LR variants

- LR(1) gives maximal power out of a 1 look-ahead symbol parsing table
  - DFA + stack is a push-down automaton (recall 262)
- In practice, LR(1) tables are big.
  - Modern implementations (e.g. menhir) directly generate code

- LALR(1) = “Look-ahead LR”
  - Merge any two LR(1) states whose items are identical except for the look-ahead sets:

```
S' → .S$   {}  
S   → .E + S  {$}  
S   → .E     {$}  
E   → .num   {+}  
E   → .( S )  {+}  
E   → .num   {$}  
E   → .( S )  {$}  
```

  - Such merging can lead to nondeterminism (e.g. reduce/reduce conflicts), but
  - Results in a much smaller parse table and works well in practice
  - This is the usual technology for automatic parser generators: yacc, ocamlyacc

- GLR = “Generalized LR” parsing
  - Efficiently compute the set of all parses for a given input
  - Later passes should disambiguate based on other context
Classification of Grammars

LR(1)
LALR(1)
LL(1)
SLR
LR(0)
Debugging parser conflicts.
Disambiguating grammars.
Practical Issues

• Dealing with source file location information
  – In the lexer and parser
  – In the abstract syntax

  – See range.ml, ast.ml

• Lexing comments / strings
Menhir output

• You can get verbose ocamlyacc debugging information by doing:
  – menhir --explain ...
  – or, if using ocamlbuild:
    ocamlbuild --use-menhir --yaccflag --explain ...

• The result is a <basename>.conflicts file that contains a description of the error
  – The parser items of each state use the ‘.’ just as described above

• The flag --dump generates a full description of the automaton

• Example: see start-parser.mly
Precendence and Associativity Declarations

- Parser generators, like menhir often support precedence and associativity declarations.
  - Hints to the parser about how to resolve conflicts.
  - See: good-parser.mly

- Pros:
  - Avoids having to manually resolve those ambiguities by manually introducing extra nonterminals (as seen in parser.mly)
  - Easier to maintain the grammar

- Cons:
  - Can’t as easily re-use the same terminal (if associativity differs)
  - Introduces another level of debugging

- Limits:
  - Not always easy to disambiguate the grammar based on just precedence and associativity.
Example Ambiguity in Real Languages

- Consider this grammar:
  \[ S \rightarrow \text{if} \ (E) \ S \]
  \[ S \rightarrow \text{if} \ (E) \ S \text{ else } S \]
  \[ S \rightarrow X = E \]
  \[ E \rightarrow \ldots \]

- Is this grammar OK?

- Consider how to parse:
  \[ \text{if} \ (E_1) \ \text{if} \ (E_2) \ S_1 \]
  \[ \text{else} \ S_2 \]

- This is known as the “dangling else” problem.

- What should the “right” answer be?

- How do we change the grammar?
How to Disambiguate if-then-else

• Want to rule out:

\[
\text{if } (E_1) \begin{cases} \text{if } (E_2) S_1 \end{cases} \text{ else } S_2
\]

• Observation: An un-matched ‘if’ should not appear as the ‘then’ clause of a containing ‘if’.

\[
\begin{align*}
S & \mapsto M \mid U & \text{ // } M = \text{“matched”}, \ U = \text{“unmatched”} \\
U & \mapsto \text{if } (E) S & \text{ // Unmatched ‘if’} \\
U & \mapsto \text{if } (E) \text{ M else } U & \text{ // Nested if is matched} \\
M & \mapsto \text{if } (E) \text{ M else } M & \text{ // Matched ‘if’} \\
M & \mapsto X = E & \text{ // Other statements}
\end{align*}
\]

• See: else-resolved-parser.mly
Alternative: Use { }

- Ambiguity arises because the ‘then’ branch is not well bracketed:

\[
\text{if } (E_1) \{ \text{if } (E_2) \{ S_1 \} \} \text{ else } S_2 \quad \text{\texttt{// unambiguous}}
\]

\[
\text{if } (E_1) \{ \text{if } (E_2) \{ S_1 \} \text{ else } S_2 \} \quad \text{\texttt{// unambiguous}}
\]

- So: could just require brackets
  - But requiring them for the else clause too leads to ugly code for chained if-statements:

\[
\text{if } (c_1) \{
\quad \ldots
\} \text{ else } \{
\quad \text{if } (c_2) \{
\quad \} \text{ else } \{
\quad \text{if } (c_3) \{
\quad \} \text{ else } \{
\quad \}
\quad \}
\quad \}
\]

So, compromise? Allow unbracketed else block only if the body is ‘if’:

\[
\text{if } (c_1) \{
\quad \} \text{ else if } (c_2) \{
\quad \} \text{ else if } (c_3) \{
\quad \} \text{ else } \{
\}
\]

Benefits:
- Less ambiguous
- Easy to parse
- Enforces good style