G52MAL
Machines and Their Langauges
Lecture 17
Recursive-Descent Parsing: Predictive Parsing

Henrik Nilsson

University of Nottingham, UK

This lecture:

- The problem of choice revisited.
- Predictive Parsing and LL(1) grammars.
- Computation of First and Follow Sets.
- Left factoring

Recap: Recursive-Descent Parsing (1)

Recursive-descent parsing is an example of the top-down parsing method:

- One ***parsing function*** associated with each nonterminal; e.g., for nonterminal X, `parseX`:

```
parseX :: [Token] -> Maybe [Token]
```

- A parsing function ***attempts*** to derive a prefix of the current input according to the grammar starting from the nonterminal.
- Other parsing functions invoked (recursively) as needed according to the RHS of the production(s) for the nonterminal.

Recap: Recursive-Descent Parsing (2)

- If successful, a parsing function returns the remainder of the input.
  E.g. if input is $\alpha\beta$, $X \stackrel{*}{\Rightarrow} \alpha$, and `parseX` could carry out this derivation, then:
  $\texttt{parseX}\ \alpha\beta = \texttt{Just}\ \beta$
- If unsuccessful, a parsing function returns `Nothing`.

Recap: Handling Choice (1)

Of course, we want a parsing function to be successful ***exactly*** when a prefix of the input ***can*** be derived from the corresponding nonterminal.

This can be achieved by:

- Adopting a suitable parsing strategy, specifically regarding how to handle ***choice*** between two or more productions for one nonterminal.
- Impose ***restrictions*** on the grammar to ensure success of the chosen parsing strategy.

In particular, ***left recursion*** usually ***not allowed***.

Recap: Handling Choice (2)

Two strategies for handling ***choice***, as in

$$S \to AB \mid CD$$

- Looking at the ***next input symbol*** is sometimes enough; e.g.:

$$S \to aB \mid cD$$

- If not, ***all alternatives*** could be explored through ***backtracking***:

```
parseX :: [Token] -> [[Token]]
```


Predictive Parsing (1)

Today, we are going to look into exactly when the next input symbol, a one symbol ***lookahead***, can be used to make ***all*** parsing decisions.

We note that this ***can*** be the case even if the RHSs start with nonterminals:

$$\begin{aligned} S &\to AB \mid CD \\ A &\to a \mid b \\ C &\to c \mid d \end{aligned}$$

Predictive Parsing (2)

- ***Predictive parsing*** is an example of recursive descent parsing where ***no*** backtracking is needed.
- The grammar must be such that the next input symbol ***uniquely*** determines the next production to use (a grammar restriction).

Productions: $X \to \alpha \mid \beta$

```
parseX (t : ts) =
    | t ??        -> parse α
    | t ??        -> parse β
    | otherwise -> Nothing
```


Predictive Parsing (3)

How to make the choices? Idea:

- Compute the ***set*** of terminal symbols that can ***start*** strings derived from each alternative, the ***first set***.
- If there is a choice between two or more alternatives, insist that the first sets for those are ***disjoint*** (a grammar restriction).
- The right choice can now be made simply by determining to which alternative's first set the next input symbol belongs.

Predictive Parsing (4)

Productions: $X \rightarrow \alpha \mid \beta$

```
parseX (t : ts) =
    | t ∈ first(α) -> parse α
    | t ∈ first(β) -> parse β
    | otherwise    -> Nothing
```


Predictive Parsing (5)

Again, consider: $X \rightarrow \alpha \mid \beta$
What if e.g. $\beta \stackrel{*}{\Rightarrow} \epsilon$?

Clearly, the next input symbol could be a terminal that can ***follow*** a string derivable form X!

```
parseX (t : ts) =
    | t ∈ first(α)             -> parse α
    | t ∈ first(β) ∪ follow(X) -> parse β
    | otherwise    -> Nothing
```

The branches must be mutually exclusive!

First and Follow Sets (1)

Following (roughly) "the Dragon Book" [ASU86]

For a CFG $G = (N,\ T,\ P,\ S)$:

$$\begin{aligned}\text{first}(\alpha) &= \{a \in T \mid \alpha \stackrel{*}{\underset{G}{\Rightarrow}} a\beta\}\\ \text{follow}(A) &= \{a \in T \mid S \stackrel{*}{\underset{G}{\Rightarrow}} \alpha A a \beta\}\\ &\quad \cup \{\$ \mid S \stackrel{*}{\underset{G}{\Rightarrow}} \alpha A\}\end{aligned}$$

where we assume $\alpha,\ \beta \in (N \cup T)^*$, $A \in N$, and where $\$$ is a special "end of input" marker.

First and Follow Sets (2)

Consider:

$$\begin{aligned} S &\rightarrow ABC & B &\rightarrow b \mid \epsilon\\ A &\rightarrow aA \mid \epsilon & C &\rightarrow c \mid d\end{aligned}$$

$$\begin{aligned}\text{first}(C) &= \{c,\ d\}\\ \text{first}(B) &= \{b\}\\ \text{first}(A) &= \{a\}\\ \text{first}(S) &= \text{first}(ABC)\\ &= [\text{because } A \stackrel{*}{\Rightarrow} \epsilon \text{ and } B \stackrel{*}{\Rightarrow} \epsilon]\\ &\quad \text{first}(A) \cup \text{first}(B) \cup \text{first}(C)\\ &= \{a,\ b,\ c,\ d\}\end{aligned}$$

First and Follow Sets (3)

Same grammar:

$$\begin{aligned} S &\rightarrow ABC & B &\rightarrow b \mid \epsilon\\ A &\rightarrow aA \mid \epsilon & C &\rightarrow c \mid d\end{aligned}$$

Follow sets:

$$\begin{aligned}\text{follow}(C) &= \{\$\}\\ \text{follow}(B) &= \text{first}(C) = \{c,\ d\}\\ \text{follow}(A) &= [\text{because } B \stackrel{*}{\Rightarrow} \epsilon]\\ &\quad \text{first}(B) \cup \text{first}(C)\\ &= \{b,\ c,\ d\}\end{aligned}$$

LL(1) Grammars (1)

Consider all productions for a nonterminal A in some grammar:

$$A \rightarrow \alpha_1 \mid \alpha_2 \mid \ldots \mid \alpha_n$$

In the parsing function for A, on input symbol t, we parse according to α_i if $t \in \text{first}(\alpha_i)$.

If $\alpha_i \stackrel{*}{\Rightarrow} \epsilon$, we should parse according to α_i also if $t \in \text{follow}(A)$!

LL(1) Grammars (2)

Thus, if:

- $\text{first}(\alpha_i) \cap \text{first}(\alpha_j) = \emptyset$ for $1 \leq i < j \leq n$, and
- if $\alpha_i \stackrel{*}{\Rightarrow} \epsilon$ for some i, then, for all $1 \leq j \leq n$, $j \neq i$,
  - $\alpha_j \stackrel{*}{\not\Rightarrow} \epsilon$, and
  - $\text{follow}(A) \cap \text{first}(\alpha_j) = \emptyset$

then it is always clear what do do!

A grammar satisfying these conditions is said to be an ***LL(1)*** grammar.

Nullable Nonterminals (1)

In order to compute the first and follow sets for a grammar $G = (N,\ T,\ P,\ S)$, we first need to know all nonterminals $A \in N$ such that $A \stackrel{*}{\Rightarrow} \epsilon$; i.e. the set $N_\epsilon \subseteq N$ of ***nullable*** nonterminals.

Let $\text{syms}(\alpha)$ denote the ***set*** of symbols in a string α:

$$\begin{aligned}\text{syms} &\in (N \cup T)^* \rightarrow \mathcal{P}(N \cup T)\\ \text{syms}(\epsilon) &= \emptyset\\ \text{syms}(X\alpha) &= \{X\} \cup \text{syms}(\alpha)\end{aligned}$$

Nullable Nonterminals (2)

The set N_ϵ is the ***smallest*** solution to the equation

$$N_\epsilon \ = \ \{A \mid A \rightarrow \alpha \in P \ \wedge \ \forall X \in \text{syms}(\alpha)\ .\ X \in N_\epsilon\}$$

(Note that $A \in N_\epsilon$ if $A \rightarrow \epsilon \in P$ because $\text{syms}(\epsilon) = \emptyset$ and $\forall X \in \emptyset\ .\ \ldots$ is trivially true.)

We can now define a predicate nullable on ***strings*** of grammar symbols:

$$\begin{aligned}\text{nullable} &\in (N \cup T)^* \rightarrow \text{Bool}\\ \text{nullable}(\epsilon) &= \text{true}\\ \text{nullable}(X\alpha) &= X \in N_\epsilon \wedge \text{nullable}(\alpha)\end{aligned}$$

Nullable Nonterminals (3)

The equation for N_ϵ can be solved iteratively as follows:

1. Initialize N_ϵ to $\{A \mid A \to \epsilon \in P\}$.
2. If there is a production $A \to \alpha$ such that $\forall X \in \text{syms}(\alpha) \,.\, X \in N_\epsilon$, then add A to N_ϵ.
3. Repeat step 2 until no further nullable nonterminals can be found.

Nullable Nonterminals (4)

Consider the following grammar:

$$\begin{array}{lcl@{\qquad}lcl} S & \to & ABC \mid AB & B & \to & b \mid \epsilon \\ A & \to & aA \mid BB & C & \to & c \mid d \end{array}$$

- Because $B \to \epsilon$ is a production, $B \in N_\epsilon$.
- Because $A \to BB$ is a production and $B \in N_\epsilon$, additionally $A \in N_\epsilon$.
- Because $S \to AB$ is a production, and $A, B \in N_\epsilon$, additionally $S \in N_\epsilon$.
- No more production with nullable RHS. The set of nullable symbols $N_\epsilon = \{S, A, B\}$.

Computing First Sets (1)

For a CFG $G = (N, T, P, S)$, the sets $\text{first}(A)$ for $A \in N$ are the smallest sets satisfying:

$$\begin{array}{lcl} \text{first}(A) & \subseteq & T \\ \text{first}(A) & = & \displaystyle\bigcup_{A \to \alpha \,\in\, P} \text{first}(\alpha) \end{array}$$

Computing First Sets (2)

For strings, first is defined as (note the ***overloaded*** notation):

$$\begin{array}{rcl} \text{first} & \in & (N \cup T)^* \to \mathcal{P}(T) \\ \text{first}(\epsilon) & = & \emptyset \\ \text{first}(a\alpha) & = & \{a\} \\ \text{first}(A\alpha) & = & \text{first}(A) \cup \begin{cases} \text{first}(\alpha), & \text{if } A \in N_\epsilon \\ \emptyset, & \text{if } A \notin N_\epsilon \end{cases} \end{array}$$

where $a \in T$, $A \in N$, and $\alpha \in (N \cup T)^*$.

Computing First Sets (3)

The solutions can often be obtained directly by expanding out all definitions.

If necessary, the equations can be solved by iteration in a similar way to how N_ϵ is computed.

Note that the smallest solution to set equations of the type

$$A = A \cup B$$

is simply

$$A = B$$

Computing First Sets (4)

Consider (again):

$$\begin{array}{lcl@{\qquad}lcl} S & \to & ABC & B & \to & b \mid \epsilon \\ A & \to & aA \mid \epsilon & C & \to & c \mid d \end{array}$$

First compute the nullable nonterminals: $N_\epsilon = \{A, B\}$.

Then compute first sets:

$$\begin{array}{rcl} \text{first}(A) & = & \text{first}(aA) \cup \text{first}(\epsilon) \\ & = & \{a\} \cup \emptyset = \{a\} \end{array}$$

Computing First Sets (5)

$$\begin{array}{lcl@{\qquad}lcl} S & \to & ABC & B & \to & b \mid \epsilon \\ A & \to & aA \mid \epsilon & C & \to & c \mid d \end{array}$$

$$\begin{array}{rcl} \text{first}(B) & = & \text{first}(b) \cup \text{first}(\epsilon) \\ & = & \{b\} \cup \emptyset = \{b\} \\ \\ \text{first}(C) & = & \text{first}(c) \cup \text{first}(d) \\ & = & \{c\} \cup \{d\} = \{c, d\} \end{array}$$

Computing First Sets (6)

$$\begin{array}{lcl@{\qquad}lcl} S & \to & ABC & B & \to & b \mid \epsilon \\ A & \to & aA \mid \epsilon & C & \to & c \mid d \end{array}$$

$$\begin{array}{rcl} \text{first}(S) & = & \text{first}(ABC) \\ & = & [A \in N_\epsilon] \\ & & \text{first}(A) \cup \text{first}(BC) \\ & = & [B \in N_\epsilon \wedge C \notin N_\epsilon] \\ & & \text{first}(A) \cup \text{first}(B) \cup \text{first}(C) \cup \emptyset \\ & = & \{a\} \cup \{b\} \cup \{c, d\} = \{a, b, c, d\} \end{array}$$

Computing Follow Sets (1)

For a CFG $G = (N, T, P, S)$, the sets $\text{follow}(A)$ are the smallest sets satisfying:

- $\{\$\} \subseteq \text{follow}(S)$
- If $A \to \alpha B \beta \in P$, then $\text{first}(\beta) \subseteq \text{follow}(B)$
- If $A \to \alpha B \beta \in P$, and $\text{nullable}(\beta)$ then $\text{follow}(A) \subseteq \text{follow}(B)$

$A, B \in N$, and $\alpha, \beta \in (N \cup T)^*$.

(It is assumed that there are no ***useless*** symbols; i.e., all symbols can appear in the derivation of some sentence.)

Computing Follow Sets (2)

$$S \to ABC \qquad B \to b \mid \epsilon$$
$$A \to aA \mid \epsilon \qquad C \to c \mid d$$

Constraints for $\text{follow}(S)$:

$$\{\$\} \subseteq \text{follow}(S)$$

Constraints for $\text{follow}(A)$ (note: $\neg\text{nullable}(BC)$):

$$\begin{aligned} \text{first}(BC) &\subseteq \text{follow}(A) \\ \text{first}(\epsilon) &\subseteq \text{follow}(A) \\ \text{follow}(A) &\subseteq \text{follow}(A) \end{aligned}$$

Computing Follow Sets (3)

$$S \to ABC \qquad B \to b \mid \epsilon$$
$$A \to aA \mid \epsilon \qquad C \to c \mid d$$

Constraints for $\text{follow}(B)$ (note: $\neg\text{nullable}(C)$):

$$\text{first}(C) \subseteq \text{follow}(B)$$

Constraints for $\text{follow}(C)$ (note: $\text{nullable}(\epsilon)$):

$$\begin{aligned} \text{first}(\epsilon) &\subseteq \text{follow}(C) \\ \text{follow}(S) &\subseteq \text{follow}(C) \end{aligned}$$

Computing Follow Sets (4)

In general:

$$A \subseteq C \land B \subseteq C \iff A \cup B \subseteq C$$

Also, constraints like $A \subseteq A$ are trivially satisfied and can be omitted.
The constraints can thus be written as:

$$\begin{aligned} \{\$\} &\subseteq \text{follow}(S) \\ \text{first}(BC) \cup \text{first}(\epsilon) &\subseteq \text{follow}(A) \\ \text{first}(C) &\subseteq \text{follow}(B) \\ \text{first}(\epsilon) \cup \text{follow}(S) &\subseteq \text{follow}(C) \end{aligned}$$

Computing Follow Sets (5)

Using

$$\begin{aligned} \text{first}(\epsilon) &= \emptyset \\ \text{first}(C) &= \{c, d\} \\ \text{first}(BC) &= \text{first}(B) \cup \text{first}(C) \cup \emptyset \\ &= \{b\} \cup \{c, d\} = \{b, c, d\} \end{aligned}$$

the constraints can be simplified further:

$$\begin{aligned} \{\$\} &\subseteq \text{follow}(S) \\ \{b, c, d\} &\subseteq \text{follow}(A) \\ \{c, d\} &\subseteq \text{follow}(B) \\ \text{follow}(S) &\subseteq \text{follow}(C) \end{aligned}$$

Computing Follow Sets (6)

Looking for the smallest sets satisfying these constraints, we get:

$$\begin{aligned} \text{follow}(S) &= \{\$\} \\ \text{follow}(A) &= \{b, c, d\} \\ \text{follow}(B) &= \{c, d\} \\ \text{follow}(C) &= \text{follow}(S) = \{\$\} \end{aligned}$$

LL(1), Left-Recursion, Ambiguity (1)

No left-recursive or ambiguous grammar can be LL(1)! For example, consider:

$$A \to Aa \mid \beta$$

First assume $\text{first}(\beta) \neq \emptyset$.

Note that

- $\text{first}(\beta) \subseteq \text{first}(A)$
- $\text{first}(A) \subseteq \text{first}(Aa)$
  ($\text{first}(A) = \text{first}(Aa)$ if $A \not\overset{*}{\Rightarrow} \epsilon$)
- Thus $\text{first}(Aa) \cap \text{first}(\beta) \neq \emptyset$. Not LL(1)!

LL(1), Left-Recursion, Ambiguity (2)

Now assume $\text{first}(\beta) = \emptyset$

This can only be the case if $\beta \overset{*}{\Rightarrow} \epsilon$ and nothing else.

Assuming $S \overset{*}{\Rightarrow} \alpha A \gamma$, we note

- $a \in \text{first}(Aa)$ because $A \Rightarrow \beta \overset{*}{\Rightarrow} \epsilon$, and
- $a \in \text{follow}(A)$ because $S \overset{*}{\Rightarrow} \alpha A \gamma \Rightarrow \alpha A a \gamma$
- Because $\beta \overset{*}{\Rightarrow} \epsilon$, the LL(1) conditions require that $\text{first}(Aa)$ and $\text{follow}(A)$ be disjoint. But that is clearly not the case!

Left Factoring (1)

Left factoring means factoring out a common prefix among a group of productions. This can help making a grammar suitable for predictive recursive descent parsing.

Example:

$$S \to aXbY \mid aXbYcZ$$

Not suitable for predictive parsing!

But note common prefix! Let's try to postpone the choice!

Left Factoring (2)

Before left factoring:

$$S \to aXbY \mid aXbYcZ$$

After left factoring:

$$\begin{aligned} S &\to aXbYS' \\ S' &\to \epsilon \mid cZ \end{aligned}$$

Now suitable for predictive parsing!