Syntax Analysis

Sudakshina Dutta

IIT Goa

15^{th} February, 2022

Top-down parsing without backtracking

- Essentially we want to predict the production rule unambiguously
- For the moment, assume that the parser has an oracle that picks the correct production at each point in the parsing process
- For transforming the grammar so that it can have oracular choice, we need to apply left factoring and left recursion elimination
- With a single focus symbol and the lookahead symbol parser can say which production to apply
  - The process is called predictive parsing and the grammar is called predictive grammar

FIRST

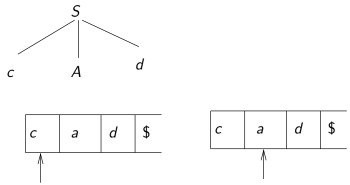
- The construction of top-down and bottom-up parsers is aided by two functions, FIRST and FOLLOW
- They allow us to choose which production to apply for top-down parsing
- $FIRST(\alpha)$ is the set of terminals that begin the strings derived from α

  In other words, It is the set of all possible first letters in the strings derived from ALPHA
- For predictive parsing, the set $FIRST(\alpha)$ and $FIRST(\beta)$ are two disjoint sets for a production $A \rightarrow \alpha|\beta$

  — the next production to be applied can be chosen by looking at the next input symbol *a*, where *a* is in either $FIRST(\alpha)$ or in $FIRST(\beta)$

Why *FIRST* ?

- Consider the grammar
  1. $S \rightarrow cAd$
  2. $A \rightarrow eb|a$
- Consider the input string to be “cad”
- We choose $A \rightarrow a$ instead of $A \rightarrow eb$ to accept the give string
- Hence, if the parser knows, the FIRST set, it can correctly apply (predict) the appropriate production rule

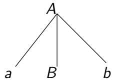


FOLLOW

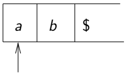
- $FOLLOW(A)$ is the set of terminals a that can appear immediately to the right of A in some sentential form
- It is the set of terminals a such that there exists a derivation of the form $S \xrightarrow{*} \alpha A a \beta$
- If A is the rightmost symbol of a derivation, then \$ is in $FOLLOW(A)$

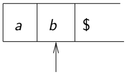
Why *FOLLOW* ?

- Consider the grammar
  1. $A \rightarrow aBb$
  2. $B \rightarrow c|\epsilon$
- Consider the input string to be “ab”
- First, the rule $A \rightarrow aBb$ is applied
- No production which is derived from B has b as the first character
- However, $B \rightarrow \epsilon$ is present
- Hence, if the parser knows, the FOLLOW set, it can correctly apply (predict) the appropriate production rule



Here we knew that follow of B is b. So we take B -> Epsilon.





Computation of the set FIRST

- To compute $FIRST(X)$ for all grammar symbol X, apply following rules until no more terminals or ϵ can be added to any $FIRST$ set
  1. If X is a terminal, then $FIRST(X) = \{X\}$
  2. If $X \to \epsilon$ is a production, then add ϵ to $FIRST(X)$
  3. If X is a non-terminal and $X \to Y_1 Y_2 \cdots Y_k$ is a production for some $k \geq 1$, then place a in $FIRST(X)$ if for some i, a is in $FIRST(Y_i)$, and ϵ is in all of $FIRST(Y_1), \cdots, FIRST(Y_{i-1})$; that is, $Y_1 \cdots Y_{i-1} \Rightarrow^* \epsilon$. If ϵ is in $FIRST(Y_j)$ for all $j = 1, 2, \cdots, k$, then add ϵ to $FIRST(X)$

In Step 3, We need to start computing FIRST(Yi) from i = 1 to k

Computation of the set FOLLOW

- To compute $FOLLOW(A)$ for all non-terminal A, apply following rules until nothing can be added to any $FOLLOW$ set
  1. Place \$ in $FOLLOW(S)$, where S is the start symbol, and \$ is the input right end marker
  2. If there is a production $A \to \alpha B \beta$, then everything in $FIRST(\beta)$ except ϵ is in $FOLLOW(B)$
  3. If there is a production $A \to \alpha B \beta$, or a production $A \to \alpha B \beta$, where $FIRST(\beta)$ contains ϵ, then everything in $FOLLOW(A)$ is in $FOLLOW(B)$

Typo in point3 ?

Example

- Consider the following example
  1. $S \rightarrow Bb|Cd$
  2. $B \rightarrow aB|\epsilon$
  3. $C \rightarrow cC|\epsilon$
- $FIRST(S)$ is $\{a, b, c, d\}$, $FIRST(B)$ is $\{a, \epsilon\}$, $FIRST(C)$ is $\{c, \epsilon\}$
- $FOLLOW(S)$ is $\{\$\}$, $FOLLOW(B)$ is $\{b\}$, $FOLLOW(C)$ is $\{d\}$

Example

Don't forget to treat even Parenthesis as Letters of the word in Production rules.

- Consider the following example
  1. $A \rightarrow aB$
  2. $B \rightarrow (C)|id$
  3. $C \rightarrow *T|\epsilon$
- $FIRST(A)$ is $\{a\}$, $FIRST(B)$ is $\{(, id\}$ and $FIRST(C)$ is $\{*, \epsilon\}$
- $FOLLOW(A)$ is $\{\$\}$, $FOLLOW(B)$ is $\{\$\}$ and $FOLLOW(C)$ is $\{)\}$

A -> aB   -> a(C)
Here in no sentential form, a character appears after B. So, $ is the Follow(B)

But, We have ")" appearing after C. So, ")" is the Follow(C)