Context-Free Grammars

Sipser 2.1 (pages 99 – 109)

What are we missing?

- So far:
- We know how to *recognize* languages
  - With finite state automata
  - As people…
- We know how to *generate* languages
  - With regular expressions
  - As people…
- Finite state automata and regular expressions are limited, though!

Bring back memories?

- In English, a grammar tells us whether a particular sentence is well formed or not
- For instance, "a sentence can consist of a noun phrase followed by a verb phrase"
- More concisely, we could write

$$\langle \textit{sentence} \rangle \rightarrow_G \langle \textit{noun\_phrase} \rangle \langle \textit{verb\_phrase} \rangle$$

Great, but what's a noun phrase?

- A sentence is
  - *<sentence>* $\rightarrow_G$ *<noun_phrase><verb_phrase>*
- We need to provide definitions for the newly introduced constructs *<noun_phrase>* and *<verb_phrase>*
  - *<noun_phrase>* $\rightarrow_G$ *<article><noun>*
  - *<verb_phrase>* $\rightarrow_G$ *<verb>*

Generating well-formed sentences

- Grammar rules so far:
  - *<sentence> $\rightarrow_G$ <noun_phrase><verb_phrase>*
  - *<noun_phrase> $\rightarrow_G$ <article><noun>*
  - *<verb_phrase> $\rightarrow_G$ <verb>*
- *To complete our simple grammar, we associate actual words with the terms <article>, <noun>, and <verb>*
  - *<article> $\rightarrow_G$ a*
  - *<article> $\rightarrow_G$ the*
  - *<noun> $\rightarrow_G$ student*
  - *<verb> $\rightarrow_G$ relaxes*
  - *<verb> $\rightarrow_G$ studies*

Context-free grammars

- A context-free grammar G is a quadruple (V, Σ, R, S), where
  - V is a finite set called the **variables**
  - **Σ is a finite set, disjoint from V, called the terminals**
  - **R is a finite subset of $V \times (V \cup \Sigma)^*$ called the rules**
  - **$S \in V$ is called the start symbol**
- **For any $A \in V$ and $u \in (V \cup \Sigma)^*$, we write $A \to_G u$ whenever $(A, u) \in R$**

The language of a grammar

- If
  - $u, v, w \in (V \cup \Sigma)^*$
  - $A \rightarrow_G w$ is a rule

  then
  - We say uAv **yields** uwv
  - **Write** $uAv \Rightarrow_G uwv$
- **If**
  - $u \Rightarrow_G u1 \Rightarrow_G u2 \Rightarrow_G \ldots \Rightarrow_G uk \Rightarrow_G v$

  then
  - **We write** $u \Rightarrow^*_G v$
- **The language of the grammar** G **is**

  $L(G) = \{w \in \Sigma^* \mid S \Rightarrow^*_G w\}$

For example…

- Consider $G = (V, \Sigma, R, S)$, where
  - $V = \{S\}$
  - $\Sigma = \{a, b\}$
  - $R = \{ S \rightarrow_G aSa \mid bSb \mid aSb \mid bSa \mid \varepsilon \}$
- Is there a grammar whose language is $PAL = \{w \in \Sigma^* \mid w = reverse(w)\}$?

Arithmetic expressions and parse trees

- Consider $G = (V, \Sigma, R, S)$, *where*
  - $V = \{<EXPR>, <TERM>, <FACTOR>\}$
  - $\Sigma = \{a, +, \times, (, \ )\}$
  - $R = \{ <EXPR> \rightarrow_G \ <EXPR>+<TERM> \mid <TERM>,$

    $<TERM> \rightarrow_G \ <TERM>\times<FACTOR> \mid <FACTOR>,$

    $<FACTOR> \rightarrow_G \ <EXPR> \mid a \ \}$
  - $S = <EXPR>$
- What about $a \times a + a$?

Leftmost derivation

- A derivation of a string in a grammar is a leftmost derivation if:

  at every step the *leftmost* remaining variable is the one replaced

Needlessly complicated?

- How about just

  $$\langle EXPR\rangle \rightarrow_G \langle EXPR\rangle + \langle EXPR\rangle \mid \langle EXPR\rangle \times \langle EXPR\rangle \mid a$$

- A grammar G is **ambiguous if some string w has two or more different leftmost derivations**

Chomsky normal form

- A context-free grammar G is in **Chomsky normal form**
  - **If every rule is of the form**
    - $A \longrightarrow BC$
    - $A \longrightarrow a$
    - **where** $A,B,C \in V$, $B \neq S \neq C$**, and** $a \in \Sigma$
  - **We permit S → ε**

Chomsky normal form

- **Theorem 2.9: Any context-free language is generated by a context-free grammar in Chomsky normal form**
- **Proof:**
  1. **Make sure S appears only on the left**
  2. **Remove empty rules:** $A \longrightarrow \varepsilon$
  3. **Handle unit rules:** $A \longrightarrow B$
  4. **Fix all the rest…**
- For example:
  - $S \longrightarrow_G ASA \mid aA$
  - $A \longrightarrow_G b \mid \varepsilon$