ECE 664
Lecture 22
Avi Kak

Properties of Context-Free Languages

- First we will present some closure properties of CF languages and then show some solvable and unsolvable problems related to them.

Closure Properties:

Theorem: $\boxed{\text{If } L_1 \text{ and } L_2 \text{ are context-free languages, then so is } L_1 \cup L_2.}$

Proof: We need to show there exists a CF grammar that will generate $L_1 \cup L_2$. Let Γ_1 generate L_1 and Γ_2 generate L_2. Without loss of generality, assume the variables of Γ_1 and L_2 are disjoint. The variables of Γ: $\mathcal{V}_1 \cup \mathcal{V}_2 \cup \{S\}$ where S is a new variable for Γ — it will be the start symbol for Γ. The productions of Γ are the productions of Γ_1 and Γ_2 plus the productions $S \rightarrow S_1$ and $S \rightarrow S_2$.

Theorem: $\boxed{\text{There exist context-free languages } L_1 \text{ and } L_2 \text{ such that } L_1 \cap L_2 \text{ is NOT context-free.}}$

This is surprising because as you saw in Lecture 18, regular languages are closed under intersection.

Proof: Because of the word 'exist' in the statement of the theorem, we need to show two CF languages whose intersection is not CF.

The two languages are: $L_1 = \{a^{[n]} b^{[n]} c^{[m]} \mid n, m > 0\}$, Γ_1: $S \rightarrow Sc,\ S \rightarrow Xc,\ X \rightarrow aXb,\ X \rightarrow ab$

$L_2 = \{a^{[m]} b^{[n]} c^{[n]} \mid n, m > 0\}$, Γ_2: $S \rightarrow aS,\ S \rightarrow aX,\ X \rightarrow bXc,\ X \rightarrow bc$

$L_1 \cap L_2 = \{a^{[n]} b^{[n]} c^{[n]} \mid n > 0\}$ that we know is not CF (by Bar-Hillel's pumping lemma)

Theorem: There exists a CF language $L \subseteq A^*$ such that $A^* - L$ is NOT context-free.

Proof: Follows from $L_1 \cap L_2 = A^* - ((A^* - L_1) \cup (A^* - L_2))$. If the set subtraction operation yielded a CF language, then $A^* - L_1$ and $A^* - L_2$ would each be CF. Since we already know that the union operation leads to a CF language, that would imply that $(A^* - L_1) \cup (A^* - L_2)$ is CF. But then that would imply that $L_1 \cap L_2$ is CF. Contradiction.

Theorem: $\boxed{\text{If } R \text{ is a regular language and } L \text{ is a context-free language, then } R \cap L \text{ is a context-free language.}}$

This theorem sounds even more interesting if you remember that every regular language is CF. So, whereas, in general, the intersection of two CF languages is NOT CF, we get a CF intersection if one of the languages is regular.

Proof:
- Let Γ be a positive CF grammar that generates L.
- Let M be the dfa that accepts R.
- We will construct a positive CF grammar $\tilde{\Gamma}$ that generates $L \cap R$.

M: A Q q_1 F δ

Γ: A $\mathcal{V}$ S productions

$\tilde{\Gamma}$: A $\{\sigma^{pq}\}$ $\tilde{S}$ productions

$\sigma \in A \cup \mathcal{V}$ and for every ordered pair $p, q \in Q$

① $\tilde{S} \rightarrow S^{q_1 q}$ for all $q \in F$

② $X^{pq} \rightarrow \sigma_1^{pr} \sigma_2^{r_1 r_2} \ldots \sigma_n^{r_{n-1} q}$ for all productions $X \rightarrow \sigma_1 \sigma_2 \ldots \sigma_n$ of Γ and for all $p, r, r_1, \ldots, r_{n-1}, q \in Q$

③ $a^{pq} \rightarrow a$ for all $a \in A$ and for all $p, q \in Q$ such that $\delta(p, a) = q$.

About the construction of $\tilde{\Gamma}$ on the previous side, note that it generates a terminal symbol a only when the ~~def~~ dfa $\mathcal{M}$ contains a state transition for that symbol. This is controlled by the third type of productions specified for $\tilde{\Gamma}$.

- For the proof in the forward direction, we start with a word u that is in $L(\Gamma)$ and also in R. That is, $u \in L(\mathcal{M})$ and $u \in L(\Gamma)$. We now want to prove that $u \in L(\tilde{\Gamma})$. So we are given $S \overset{*}{\underset{\Gamma}{\Longrightarrow}} u = a_1 a_2 \ldots a_n$ with $\delta(q_1, a_1) = q_2$, $\delta(q_2, a_2) = q_3, \ldots, \delta(q_n, a_n) = q_{n+1} \in F$. Now we must prove

(A) $\tilde{S} \overset{*}{\underset{\tilde{\Gamma}}{\Longrightarrow}} a_1 a_2 \ldots a_n$. In terms of the productions specified for $\tilde{\Gamma}$ on the previous page, we obviously have

$$\tilde{S} \Longrightarrow S^{q_1 q_{n+1}} \overset{*}{\underset{\tilde{\Gamma}}{\Longrightarrow}} a_1^{q_1 q_2} a_2^{q_2 q_3} \ldots a_n^{q_n q_{n+1}} \Longrightarrow a_1 a_2 \ldots a_n$$

because by (3) $a_i^{pq} \to a$ when $\delta(p, a_i) = q$ which is indeed true by virtue of (A) above.

This establishes the proof in one direction.

- The proof in the other direction is based on the lemma that when $\sigma^{pq} \overset{*}{\underset{\tilde{\Gamma}}{\Longrightarrow}} u \in A^*$ then $\delta^*(p, u) = q$ and when $\sigma \in V$ then we also have $\sigma \overset{*}{\underset{\Gamma}{\Longrightarrow}} u$. See the text for a proof of the lemma.

Sub-Alphabet Erasure:

Let A be the alphabet of a ~~CF~~ language ~~L(Γ)~~ L. Now consider a subset $P \subseteq A$. We want to delete all of the symbols of P from the words of L. Let $x \in L$. We denote by $Er_P(x)$ what remains of x after erasing from x the symbols of P. We now ~~w~~ write

$$Er_P(L) = \{ Er_P(x) \mid x \in L \}$$

Now consider the more specific case of a language generated by a CF grammar Γ. In this case, we are interested in erasing a subset of <u>terminals</u>. We can prove the following:

> If $L \subseteq A^*$ is a CF language and $P \subseteq A$, then $Er_P(L)$ is also CF.

The Word Problem for CF languages:

(word problem means figuring out whether $u \in L$ for a $u \in A^*$)

- You'll recall that the word problem was <u>unsolvable</u> for a semi-Thue process.
- The word problem was trivially <u>solvable</u> for languages accepted by finite automata (and therefore for languages generated by regular grammars).
- The good news is that the word problem is <u>solvable</u> for context-free languages. This is because of the following theorem:

THEOREM: Let Δ be a Chomsky normal form grammar with terminals T. For a variable V of Δ let $V \overset{*}{\underset{\Delta}{\Longrightarrow}} u \in T^*$ then there is a derivation of u from V in Δ of length $2|u|$.

(see Lecture 20 for definition of the length of a derivation)

Proof: We can prove this by induction on $|u|$. When $|u| = 1$, then u is a terminal. So we must have $V \to a$ for some $a \in T$. The length of this derivation is 2. Now lets consider u with $|u| > 1$. We must have $V \Longrightarrow XY \overset{*}{\Longrightarrow} u$. We should therefore expect $X \overset{*}{\Longrightarrow} v$ and $Y \overset{*}{\Longrightarrow} w$ with $u = vw$. By the <u>induction hypothesis</u>, we must have $X = \alpha_1 \Rightarrow \alpha_2 \Rightarrow \cdots \Rightarrow \alpha_{2|v|} = v$ and $Y = \beta_1 \Rightarrow \beta_2 \Rightarrow \cdots \Rightarrow \beta_{2|w|} = w$. Therefore, $V \Longrightarrow XY \Rightarrow \alpha_1 Y \Rightarrow \alpha_2 Y \Rightarrow \cdots \Rightarrow \alpha_{2|v|} Y \Rightarrow v\beta_1 \Rightarrow v\beta_2 \Rightarrow \cdots \Rightarrow v\beta_{2|w|} = vw = u$. But the length of this derivation is $2|v| + 2|w| = 2|u|$.

- $L(\Gamma) \neq \emptyset$ iff there is a derivation tree $\mathcal{T}$ in Γ of a word $u \in T^*$ such that each path in $\mathcal{T}$ contains fewer than $n+2$ nodes. ~~(From Bar-Hillel's Pumping Lemma)~~ Let u be the shortest possible string in L whose $\mathcal{T}$ * longest possible path in $\mathcal{T}$ has $n+2$ nodes.