Theory of Computation (III)

Yijia Chen
Fudan University

Review

Regular languages are simple as witnessed by their definition by automata and regular expressions.

But we have also seen that the simple language

$$\left\{0^n 1^n \;\middle|\; n \in \mathbb{N}\right\}$$

is not regular. As a matter of fact, it is a context-free language.

The pumping lemma for regular languages

Lemma

If A is a regular language, then there is a number p (i.e., the pumping length) where if s is any string in A of length at least p, then s may be divided into three pieces, $s = xyz$, satisfying the following conditions:

1. *for each $i \geq 0$, we have $xy^iz \in A$,*
2. *$|y| > 0$, and*
3. *$|xy| \leq p$.*

Any string xyz in A can be pumped along y.

Proof

Let $M = \left(Q, \Sigma, \delta, q_1, F\right)$ be a DFA recognizing A and $p := |Q|$.

Let $s = s_1 s_2 \cdots s_n$ be a string in A with $n \geq p$. Let $r_1, \ldots, r_{n+1}$ be the sequence of states that M enters while processing s, i.e.,

$$r_{i+1} = \delta(r_i, s_i)$$

for $i \in [n]$.

Among the first $p+1$ states in the sequence, two must be the same, say r_j and r_ℓ with $j < \ell \leq p+1$. We define

$$x = s_1 \cdots s_{j-1}, y = s_j \cdots s_{\ell-1}, \text{ and } z = s_\ell \cdots s_n.$$

Example

The language $\left\{0^n 1^n \;\middle|\; n \geq 0\right\}$ is not regular.

Proof.

Choose p be the pumping length and consider $s = 0^p 1^p$. By the Pumping Lemma, $s = xyz$ with $xy^i z \in \left\{0^n 1^n \;\middle|\; n \geq 0\right\}$ for all $i \geq 0$.

1. $y \in 00^*$, then $xyyz$ has more 0s than 1s, a contradiction.
2. $y \in 11^*$, then $xyyz$ has more 1s than 0s, again a contradiction.
3. y consists of both 0s and 1s, then $xyyz$ have 0 and 1 interleaved.

□

Context-Free Languages

An example

The grammar:

$$A \to 0A1$$
$$A \to B$$
$$B \to \#$$

A derivation:

$$A \Rightarrow 0A1 \Rightarrow 00A11 \Rightarrow 000A111 \Rightarrow 000\#111.$$

Context-free grammars

Definition

A context-free grammar (CFL) is a 4-tuple (V, Σ, R, S), where

1. V is a finite set called the variables,
2. Σ is a finite set, disjoint from V, called the terminals.
3. R is a finite set of rules, with each rule being a variable and a string of variables and terminals, and
4. $S \in V$ is the start variable.

Derivations

Let u, v, w be strings of variables and terminals, and

$$A \to w$$

is a rule of the grammar. We say that uAv yields uwv, written $uAv \Rightarrow uwv$.

Say that u derives v, written $u \overset{*}{\Rightarrow} v$, if $u = v$ or if a sequence $u_1, u_2, \ldots, u_k$ exists for $k \geq 0$ and

$$u \Rightarrow u_1 \Rightarrow u_2 \Rightarrow \ldots \Rightarrow u_k \Rightarrow v.$$

The language of the grammar is

$$\left\{ w \in \Sigma^* \mid S \overset{*}{\Rightarrow} w \right\},$$

which is a context-free language (CFL).

Examples

1. The language $\left\{0^n 1^n \;\middle|\; n \geq 0\right\}$ has a grammar:

$$S_1 \to 0S_1 1|\varepsilon.$$

2. The language $\left\{1^n 0^n \;\middle|\; n \geq 0\right\}$ has a grammar:

$$S_2 \to 1S_2 0|\varepsilon.$$

3. The language

$$\left\{0^n 1^n \;\middle|\; n \geq 0\right\} \cup \left\{1^n 0^n \;\middle|\; n \geq 0\right\}$$

has a grammar:

$$\begin{aligned} S &\to S_1 \mid S_2 \\ S_1 &\to 0S_1 1|\varepsilon \\ S_2 &\to 1S_2 0|\varepsilon. \end{aligned}$$

$$\langle \text{EXPR} \rangle \rightarrow \langle \text{EXPR} \rangle + \langle \text{EXPR} \rangle | \langle \text{EXPR} \rangle \times \langle \text{EXPR} \rangle | \big(\langle \text{EXPR} \rangle \big) | \texttt{a}.$$

The string a+a×a have two different derivations:

1. $\langle \text{EXPR} \rangle \rightarrow \langle \text{EXPR} \rangle \times \langle \text{EXPR} \rangle \Rightarrow \langle \text{EXPR} \rangle + \langle \text{EXPR} \rangle \times \langle \text{EXPR} \rangle \overset{*}{\Rightarrow} \texttt{a+a} \times \texttt{a}.$
2. $\langle \text{EXPR} \rangle \rightarrow \langle \text{EXPR} \rangle + \langle \text{EXPR} \rangle \Rightarrow \langle \text{EXPR} \rangle + \langle \text{EXPR} \rangle \times \langle \text{EXPR} \rangle \overset{*}{\Rightarrow} \texttt{a+a} \times \texttt{a}.$

Leftmost derivations

A derivation of a string w in a grammar G is a leftmost derivation if at every step the leftmost remaining variable is the one replaced.

Ambiguity

Definition

A string w is derived ambiguously is context free grammar G if it has two or more different leftmost derivations.

Grammar G is ambiguous if it generates some string ambiguously.

$\{\mathrm{a}\}$ has two different grammars $S_1 \to S_2|\mathrm{a}; S_2 \to \mathrm{a}$ and $S \to \mathrm{a}$. The first is ambiguous, while the second is not.

$$\left\{\mathrm{a}^i\mathrm{b}^j\mathrm{c}^k \;\middle|\; i = j \text{ or } j = k\right\}$$

is inherently ambiguous, i.e., its every grammar is ambiguous.

Chomsky normal form

Definition

A context-free grammar is in Chomsky normal form if every rule is of the form

$$A \to BC$$
$$A \to a$$

where a is any terminal and A, B, and C are any variables – except that B and C may be not the start variable.

In addition, we permit the rule $S \to \varepsilon$, where S is the start variable.

Theorem

Any context-free language is generated by a context-free grammar in Chomsky normal form.

Proof of the theorem

1. Add a new start variable S_0 with the rule $S_0 \to S$, where S is the original start variable.
2. Remove every $A \to \varepsilon$, where $A \neq S$. Then for each occurrence of A on the right-hand side of a rule, we add a new rule with that occurrence deleted. For $R \to A$, we add $R \to \varepsilon$ unless we had previously removed $R \to \varepsilon$.
3. Remove every $A \to B$. Then whenever a rule $B \to u$ appears, where u is a string of variables and terminals, we add the rule $A \to u$ unless this was previously removed.
4. Replace each rule $A \to u_1 u_2 \cdots u_k$ with $k \geq 3$ and each u_i is a variable or terminal with the rules

   $$A \to u_1 A_1,\ A_1 \to u_2 A_2,\ A_2 \to u_2 A_3,\ \ldots,\ \text{and}\ A_{k-2} \to u_{k-1} u_k.$$

   The A_i's are new variables. We replace any terminal u_i with the new variable U_i and add $U_i \to u_i$.

Pushdown Automata

Pushdown automata

Definition

A pushdown automaton (PDA) is a 6-tuple $(Q, \Sigma, \Gamma, \delta, q_0, F)$, where

1. Q is a finite set of states,
2. Σ is a finite set of input alphabet,
3. Γ is a finite set of stack alphabet,
4. $\delta : Q \times \Sigma_\varepsilon \times \Gamma_\varepsilon \to \mathcal{P}(Q \times \Gamma_\varepsilon)$ is the transition function,
5. $q_0 \in Q$ is the start state, and
6. $F \subseteq Q$ is the set of accept states.

Formal definition of computation

Let $M = (Q, \Sigma, \delta, q_0, F)$ be a pushdown automaton. M accepts input w if w can be written as $w = w_1 \dots w_m$ with $w_1, \dots, w_m \in \Sigma_\varepsilon$, and sequences of states $r_0, r_1, \dots, r_m \in Q$ and strings $s_0, s_1, \dots, s_m \in \Gamma^*$ exist that satisfy the following three conditions.

1. $r_0 = q_0$ and $s_0 = \varepsilon$.
2. For $i = 0, \dots, m-1$, we have $(r_{i+1}, b) \in \delta(r_i, w_{i+1}, a)$, where $s_i = at$ and $s_{i+1} = bt$ for some $a, b \in \Gamma_\varepsilon$ and $t \in \Gamma^*$.
3. $r_m \in F$.

PDA for $\left\{0^n 1^n \mid n \geq 0\right\}$

$Q = \left\{q_1, q_2, q_3, q_4\right\}$,
$\Sigma = \{0, 1\}$,
$\Gamma = \{0, \$\}$,
$F = \left\{q_1, q_4\right\}$, and

δ is given by the following table, wherein blank entries signify $\emptyset$.

Input:	0			1			ε		
Stack:	0	\$	ε	0	\$	ε	0	\$	ε
q_1									$\{(q_2, \$)\}$
q_2			$\{(q_2, 0)\}$	$\{(q_3, \varepsilon)\}$					
q_3				$\{(q_3, \varepsilon)\}$				$\{(q_4, \varepsilon)\}$	
q_4									

Theorem

A language is context free if only if some pushdown automaton recognizes it.

Every context-free language can be recognized by a PDA

1. Place the marker symbol \$ and the start variable on the stack.

2. Repeat the following steps forever.

   2.1 If the top of stack is a variable symbol A, nondeterministically select one of the rules for A and substitute A by the string on the right-hand side of the rule.

   2.2 If the top of stack is a terminal symbol a, read the next symbol from the input and compare it to a. If they match, repeat. If they do not match, reject on this branch of the nondeterminism.

   2.3 If the top of stack is the symbol \$, enter the accept state. Doing so accepts the input if it has all been read.

Push a long string in "one step"

Let q and r be states of the PDA and let $a \in \Sigma_\varepsilon$ and $s \in \Gamma_\varepsilon$.

We want the PDA to go from q to r when it reads a and pops s. Furthermore, we want it to push the entire string $u = u_1 \cdots u_\ell$ on the stack at the same time.

$$
\begin{aligned}
&(q_1, u_\ell) \in \delta(q, a, s) \\
&\delta(q_1, \varepsilon, \varepsilon) = \{(q_2, u_{\ell-1})\}, \\
&\delta(q_2, \varepsilon, \varepsilon) = \{(q_3, u_{\ell-2})\}, \\
&\qquad \vdots \\
&\delta(q_{\ell-1}, \varepsilon, \varepsilon) = \{(r, u_1)\}.
\end{aligned}
$$

We use the abbreviation

$$(r, u) \in \delta(q, a, s).$$

Proof

We construct a pushdown automaton P as follows.

The states of P are

$$Q = \{q_{\text{start}}, q_{\text{loop}}, q_{\text{accept}}\} \cup E,$$

where E is the set of states we need for the construction in the previous slide.

For the transition function,

- $\delta(q_{\text{start}}, \varepsilon, \varepsilon) = \{(q_{\text{loop}}, S\$)\}$.
- $\delta(q_{\text{loop}}, \varepsilon, A) = \{(q_{\text{loop}}, w) \mid A \to w \text{ is a rule in the given grammar}\}$.
- $\delta(q_{\text{loop}}, a, a) = \{(q_{\text{loop}}, \varepsilon)\}$.
- $\delta(q_{\text{loop}}, \varepsilon, \$) = \{(q_{\text{accept}}, \varepsilon)\}$.

Every language recognized by a PDA is context free

Let P be a PDA. For each pair of states p and q, the grammar has a variable A_{pq} which generates

all strings taking P from p with an empty stack to q with an empty stack.

We modify P such that:

1. It has a single accept state q_{accept}.
2. It empties its stack before accepting.
3. Each transition either pushes a symbol onto the stack or pops one off the stack, but it does not do both at the same time.

Inductive definition of A_{pq}

Two possibilities occur during P's computation on an input string x.

1. The symbol popped at the end is the symbol that was pushed at the beginning. Then, we have a rule $A_{pq} \to aA_{rs}b$.

2. Otherwise, we have a rule $A_{pq} \to A_{pr}A_{rq}$.

Proof (1)

Assume $P = \big(Q, \Sigma, \Gamma, \delta, q_0, \{q_{\text{accept}}\}\big)$.

The variables of the desired context-free grammar G are

$$\big\{A_{pq} \;\big|\; p, q \in Q\big\},$$

in which the start variable is $A_{q_0, q_{\text{accept}}}$.

For the rules:

(R1) For each $p, q, r, s \in Q$, $u \in \Gamma$, and $a, b \in \Sigma_\varepsilon$, if $(r, u) \in \delta(p, a, \varepsilon)$ and $(q, \varepsilon) \in \delta(s, b, u)$, then G has the rule

$$A_{pq} \to aA_{rs}b.$$

(R2) For each $p, q, r \in Q$, G has the rule

$$A_{pq} \to A_{pr}A_{rq}.$$

(R3) For each $p \in Q$, G has the rule

$$A_{pp} \to \varepsilon.$$

Proof (2)

Claim

If A_{pq} generates x, the x can bring P from p with empty stack to q with empty stack.

Basis: The derivation has 1 step.

A derivation with a single step must use a rule whose right-hand side contains no variables. The only rules in G where no variables occur on the right-hand side are $A_{pp} \to \varepsilon$.

Induction step: The derivation has $k+1$ step with $A_{pq} \stackrel{*}{\Rightarrow} x$. Thus, either $A_{pq} \Rightarrow aA_{rs}b$ or $A_{pq} \Rightarrow A_{pr}A_{rq}$.

In case $A_{pq} \Rightarrow aA_{rs}b$ the claim follows from (R1) and the induction hypothesis.

For $A_{pq} \Rightarrow A_{pr}A_{rq}$, there exist y and z with $x = yz$ such that $A_{pr} \stackrel{*}{\Rightarrow} y$ and $A_{qr} \stackrel{*}{\Rightarrow} z$ both in at most k steps. The claim then again follows from the induction hypothesis.

Proof (3)

Claim

If x can bring P from p with empty stack to q with empty stack, then A_{pq} generates x.

Basis: The computation has 0 steps.

$x = \varepsilon$ and we have $A_{pp} \to \varepsilon$.

Induction step:

If the stack is always non-empty in the middle of the computation, then:

- There is a u which is pushed in the first move and popped in the last move.
- In the first move, let a be the input and r be the state after; in the last move let b be the input and s be the state before.
- We deduce $(r, u) \in \delta(p, a, \varepsilon)$ and $(q, \varepsilon) \in \delta(s, b, u)$. Hence, G has the rule $A_{pq} \to aA_{rs}b$.

We can conclude by the induction hypothesis.

If the stack becomes empty in the middle of the computation, the claim then again follows from the induction hypothesis.

Non-Context-Free Languages

The pumping lemma for context-free languages

Lemma

If A is a context-free language, then there is a number p (the pumping length) where, if s is any string in A of length at least p, then s may be divided as $s = uvxyz$ satisfying the conditions

1. *for each $i \geq 0$, $uv^i xy^i z \in A$.*
2. $|vy| > 0$, *and*
3. $|vxy| \leq p$.

Proof (1)

Let G be a CFG for CFL A. Let b be the maximum number of symbols in the right-hand side of a rule. In any parse tree using this grammar, every node can have no more than b children. So, if the height of the parse tree is at most h, the length of the string generated is at most b^h. Conversely, if a generated string is at lest $b^h + 1$ long, each of its parse trees must be at least $h + 1$ high.

We choose the pumping length

$$p = b^{|V|+1}.$$

For any string $s \in A$ with $|s| \geq p$, any of its parse trees must be at least $|V| + 1$ high.

Proof (2)

Let τ be one parse tree of s with smallest number of nodes, whose height is at least $|V| + 1$. So τ has a path from the root to a leaf of length $|V| + 1$ with $|V| + 2$ nodes. One variable R must appear at least twice in the last $|V| + 1$ variables nodes on this path.

We divide s into $uvxyz$:

- u from the leftmost leaf of τ to the leaf left next to the leftmost leaf of the subtree hanging on the first R,
- v from the leftmost leaf of the subtree hanging on the first R to the leaf left next to the leftmost leaf of the subtree hanging on the second R,
- x for all the leaves of the subtree hanging on the second R,
- y from the leaf right next to the rightmost leaf of the subtree hanging on the second R to the rightmost leaf of the subtree hanging on the first R,
- z from the leaf right next to the rightmost leaf of the subtree hanging on the first R to the rightmost leaf of τ.

Proof (3)

If $|vy| = 0$, i.e., $v = y = \varepsilon$, then τ cannot have the smallest number of nodes.

To see $|vxy| \leq p = b^{|V|+1}$, note that vxy is generated by the second R, whose height is at most $|V| + 1$.

Example

$\{a^n b^n c^n \mid n \geq 0\}$ is not context free.

Proof.

Assume otherwise, and let p be the pumping length. Consider $s = a^p b^p c^p$ and divide it to $uvxyz$ according to the Pumping Lemma.

- When both v and y contain only one type of symbols, i.e., one of a, b, c, then uv^2xy^2z cannot contain equal number of a's, b's, and c's.
- If either v or y contains more than one type of symbols, then uy^2wx^2y would have symbols interleaved.

□

Example

$\{ww \mid w \in \{0,1\}^*\}$ is not context free.

Proof.

Assume otherwise, and let p be the pumping length. Consider $s = 0^p1^p0^p1^p$ and divide it to $s = uvxyz$ with $|vxy| \leq p$.

- If vxy occurs only in the first half of s, then the second half of uv^2xy^2z must start with a 1. This is impossible.
- Similarly vxy cannot occur only in the second half of s.
- If vxy straddles the midpoint of s, then pumping s to uyz with the form $0^p1^i0^j1^p$ cannot ensure $i = j = p$.

□