Theoretical Computer Science
Context-Free Languages

Jonas Hübotter

Outline

Overview

Representations of context-free languages

- Context-Free Grammar (CFG)
- Pushdown Automaton (PDA)

Variables

Definition 1

Given a grammar $G = (V, \Sigma, P, S)$, a variable $X \in V$ is

- generative if $\exists X \to_G^* w \in \Sigma^*$;
- reachable if $\exists S \to_G^* X$; and
- helpful if it is generative and reachable.

Inductive Definition

Given a context-free grammar $G = (V, \Sigma, P, S)$ with
$V = \{A_1, \dots, A_k\}$,
productions $A_i \to w_0 A_{i_1} w_1 \dots w_{n-1} A_{i_n} w_n$
correspond to

$$u_1 \in L_G(A_{i_1}) \wedge \dots \wedge u_n \in L_G(A_{i_n}) \\ \implies w_0 u_1 w_1 \dots w_{n-1} u_n w_n \in L_G(A_i).$$

Hence, $L(G) = L_G(S)$.

Productions produce words top-down,
inductive definition *produces* words bottom-up.

Decomposition Lemma

Lemma 2 (Decomposition Lemma)

Any derivation of length n of β from $\alpha_1\alpha_2$ may split β into two separately derivable parts β_1 and β_2 at any position. Formally:

$$\alpha_1\alpha_2 \to_G^n \beta \iff \exists \beta_1, \beta_2, n_1, n_2.\ \beta = \beta_1\beta_2 \land n = n_1 + n_2 \land \alpha_1 \to_G^{n_1} \beta_1 \land \alpha_2 \to_G^{n_2} \beta_2.$$

Syntax Tree

Definition 3

A syntax tree of a derivation $\to_G$ given $G = (V, \Sigma, P, S)$ is a tree where

1. every leaf is labeled with a symbol in $\Sigma \cup \{\epsilon\}$;
2. every inner node is labeled with $A \in V$,
   assuming its children are $X_1, \ldots, X_n \in V \cup \Sigma \cup \{\epsilon\}$,
   $A \to X_1 \ldots X_n \in P$; and
3. a leaf labeled ϵ is an only child of its parent.

The border of a syntax tree is the labels of its leafs concatenated from left to right.

$$\begin{aligned} A \to_G^* w &\iff w \in L_G(A) \\ &\iff \exists \text{ syntax tree with root } A \text{ and border } w. \end{aligned}$$

Syntax Tree

Definition 4

- A CFG G is ambiguous if $\exists w \in L(G)$ that has two distinct syntax trees.
- A CFL L is inherently ambiguous if every CFG G with $L(G) = L$ is ambiguous.

Chomsky Normal Form

Definition 5 (Chomsky Normal Form)

All productions are of the form $A \rightarrow a$ or $A \rightarrow BC$ for $a \in Sigma$ and $A, B, C \in V$.

Algorithm to convert a CFG to Chomsky Normal Form ($\mathcal{O}(|P|^2)$)

1. replace every $a \in \Sigma$ occurring in a production with length > 1 by a non-terminal
2. replace $A \rightarrow B_1 \ldots B_k$ (where $k > 2$) with $A \rightarrow B_1 C_2, C_2 \rightarrow B_2, \ldots, C_k \rightarrow B_k$
3. remove ϵ-productions (i.e. $A \rightarrow \epsilon$)
4. remove chain productions (i.e. $A \rightarrow B$)

Other Normal Forms

Definition 6 (Greibach Normal Form)

All productions are of the form $A \to aA_1 \dots A_n$ for $a \in Sigma$ and $A_1, \dots, A_n \in V$.

Definition 7 (Backus-Naur Normal Form)

Allows the use of regular expressions in productions (in addition to symbols).

Cocke-Younger-Kasami Algorithm (CYK)

Solves the word problem for CFGs.

Algorithm ($\mathcal{O}(|w|^3)$)

Given $G = (V, \Sigma, P, S)$ in Chomsky normal form and $w = a_1 \dots a_n \in \Sigma^*$.
Define $V_{ij} = \{A \in V \mid A \to_G^* a_i \dots a_j\}$ for $i \leq j$ as the set of all initial symbols that may be used to derive $a_i \dots a_j$.
Then $w \in L_G(A) \iff A \in V_{1n}$.

Recursive definition of V_{ij}:

- base: $V_{ii} = \{A \in V \mid (A \to a_i) \in P\}$
- step:

$$V_{ij} = \{A \in V \mid {}^{\exists i \leq k < j, B \in V_{ik}, C \in V_{(k+1)j}.}_{(A \to BC) \in P} \}$$

PDA

Definition 8

A pushdown automaton (PDA) $M = (Q, \Sigma, \Gamma, q_0, Z_0, \delta, F)$ consists of

- a finite set of states Q;
- a (finite) input alphabet Σ;
- a (finite) stack alphabet Γ;
- an initial state $q_0 \in Q$;
- an initial stack element $Z_0 \in \Gamma$;
- a (partial) transition function $\delta : Q \times (\Sigma \cup \{\epsilon\}) \times \Gamma \to 2^{Q \times \Gamma^*}$; and
- a set of terminal (accepting) states $F \subseteq Q$.

Graphically, transitions are denoted as $a, Z/\alpha$ where $a \in \Sigma$ is the input, $Z \in \Gamma$ is the top stack element, and $\alpha \in \Gamma^*$ replaces Z in the new stack.

PDA

Definition 9

The configuration of a PDA M is a triple (q, w, α) where $q \in Q$ is its state, $w \in \Sigma^*$ is its remaining input, and $\alpha \in \Gamma^*$ is its stack.

The initial configuration of M on input $w \in \Sigma^*$ is (q_0, w, Z_0).

Definition 10

The transition relation of a PDA M is

$$
\begin{aligned}
(q, aw, Z\alpha) \rightarrow_M (q', w, \beta\alpha) &\quad \text{if } (q', \beta) \in \delta(q, a, Z) \\
(q, w, Z\alpha) \rightarrow_M (q', w, \beta\alpha) &\quad \text{if } (q', \beta) \in \delta(q, \epsilon, Z).
\end{aligned}
$$

PDA

Definition 11

PDA M accepts $w \in \Sigma^*$ with final state if

$$(q_0, w, Z_0) \rightarrow_M^* (f, \epsilon, \gamma) \quad \text{for } f \in F, \gamma in \Gamma^*.$$

So, $L_F(M) = \{w \in \Sigma^* \mid \exists f \in F, \gamma \in \Gamma^*. (q_0, w, Z_0) \rightarrow_M^* (f, \epsilon, \gamma)\}$.

Definition 12

PDA M accepts $w \in \Sigma^*$ with empty stack if

$$(q_0, w, Z_0) \rightarrow_M^* (q, \epsilon, \epsilon) \quad \text{for } q \in Q.$$

So, $L_\epsilon(M) = \{w \in \Sigma^* \mid \exists q \in Q. (q_0, w, Z_0) \rightarrow_M^* (q, \epsilon, \epsilon)\}$.

Both accepting conditions are equally powerful.

Lemmas

Lemma 13 (Extension Lemma)

Every derivation may occur as a sub-derivation of a larger derivation:

$$(q, u, \alpha) \rightarrow^n_M (q', u', \alpha') \implies (q, uv, \alpha\beta) \rightarrow^n_M (q', u'v, \alpha'\beta).$$

Lemma 14 (Decomposition Lemma)

Every derivation that empties the stack can be divided into sub-derivations that each remove a single symbol from the stack:
Given $(q, w, Z_1 \dots Z_k) \rightarrow^n_M (q', \epsilon, \epsilon)$,
then $\forall i \in [1, k].\ \exists u_i, p_i, n_i$ *such that*

$$(p_{i-1}, u_i, Z_i) \rightarrow^{n_i}_M (p_i, \epsilon, \epsilon)$$

with $w = u_1 \dots u_k$, $q = p_0$, $q_k = p_k$, *and* $n = \sum_{i=1}^{k} n_i$.

CFG $\to$ PDA

Given CFG $G = (V, \Sigma, P, S)$,

1. bring all productions into the form

$$A \to bB_1 \dots B_k \quad \text{for } b \in \Sigma \cup \{\epsilon\}$$

2. define the PDA $M = (\{q\}, \Sigma, V, q, S, \delta)$ with

$$\delta(q, b, A) = \{(q, \beta) \mid (A \to b\beta) \in P\}.$$

Then, $L(G) = L_\epsilon(M)$.

PDA $\to$ CFG

Given PDA $G = (Q, \Sigma, \Gamma, q_0, Z_0, \delta, F)$, define CFG $G = (V, \Sigma, P, S)$.

We define $V = Q \times \Gamma \times Q \cup \{S\}$ where each $[q, Z, p] \in V$ describes all possibilities of going from state $q \in Q$ to state $p \in Q$ while $Z \in \Gamma$ is the top stack element.

We define the productions P as

- $\forall q \in Q.\ S \to [q_0, Z_0, q]$ and
- $\forall (r_0, Z_1 \dots Z_k) \in \delta(q, b, Z).\ \forall r_1, \dots, r_k \in Q.$

$$[q, Z, r_k] \to b[r_0, Z_1, r_1][r_1, Z_2, r_2] \dots [r_{k-1}, Z_k, r_k].$$

We observe that

$$[q, Z, r_k] \to_G^* w \iff (q, w, Z) \to_M^* (r_k, \epsilon, \epsilon).$$

So, $L(G) = L_\epsilon(M)$.

Closure Properties

Theorem 15

Given the context-free languages L, L_1, L_2, then the following are also centext-free languages:

- $L_1 L_2$;
- $L_1 \cup L_2$; *and*
- L^*.

Theorem 16

Given the deterministic context-free language L, then $\bar{L}$ is deterministic context-free.

Pumping Lemma

Lemma 17 (Pumping Lemma for context-free languages)

Let $L \subseteq \Sigma^*$ *be context-free. Then there exists some* $n > 0$ *such that every* $z \in L$ *with* $|z| \geq n$ *can be decomposed into* $z = uvwxy$ *such that*

- $vx \neq \epsilon$;
- $|vwx| \leq n$; *and*
- $\forall i \geq 0.\ uv^i wx^i y \in L$.

A necessary condition for context-free languages.

Pumping Lemma

Example 18 (proof structure)

Assume L is context-free.
Let $n > 0$ be a Pumping Lemma number.
Choose $z \in L$ with $|z| \geq n$.
Define $z = uvwxy$ with $vx \neq \epsilon$ and $|vwx| \leq n$.
Then, $\forall i \geq 0.\ uv^i wx^i y \in L$.
Now, use the last statement to find a contradiction separating all possible cases for v and x.