COS210 - Theoretical Computer Science
Context-Free Languages (Part 3)

Converting a DFA to a Context-Free Grammar

Recall from the previous lecture that we can construct the grammar $G = (V, \Sigma, R, S)$ from given DFA $M = (Q, \Sigma, \delta, q, F)$ as follows:

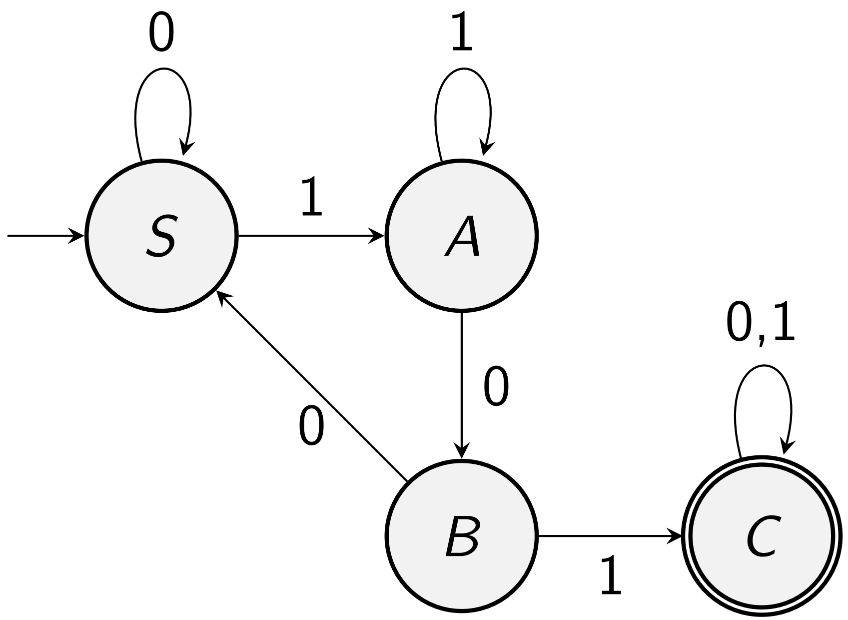
- $V = Q$, the **variables** of grammar G are the **states** of M
- Σ, the set of **terminals** of G is the **alphabet** of M
- $S = q$ the **start variable** of G is the **initial state** of M
- There are two types of rules in G:
  - **non-terminal rules** derived from transitions
    $A \rightarrow aB$, for all $A, B \in Q$, $a \in \Sigma$, and $\delta(A, a) = B$
  - **terminal rules** derived from accepting states
    $A \rightarrow \epsilon$, for all $A \in F$

The language of the constructed G is the same as the language of M

Converting a DFA to a Context-Free Grammar

Consider the following DFA over $\Sigma = \{0, 1\}$ that accepts the language

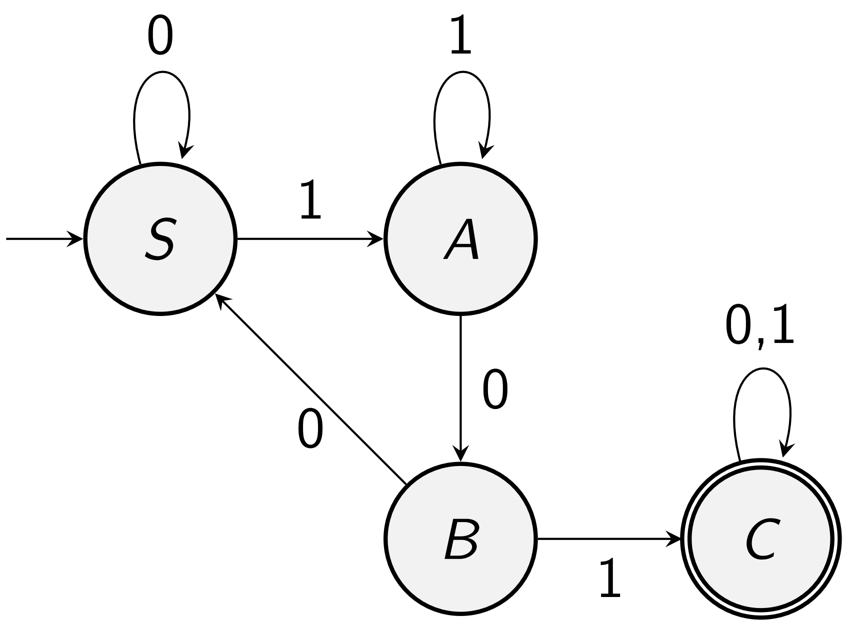
$$L = \{w : 101 \text{ is a substring of } w\}$$



Converting a DFA to a Context-Free Grammar

Consider the following DFA over $\Sigma = \{0, 1\}$ that accepts the language

$$L = \{w : 101 \text{ is a substring of } w\}$$



We can construct the corresponding grammar as follows:

$G = (V, \Sigma, R, S)$

- $V = \{S, A, B, C\}$
- $\Sigma = \{0, 1\}$
- $S = S$
- Rules of R:
  - $S \to 0S|1A$
  - $A \to 0B|1A$
  - $B \to 0S|1C$
  - $C \to 0C|1C|\epsilon$

Chomsky Normal Form

From our definition of grammars $G = (V, \Sigma, R, S)$, the rules are of the form $A \rightarrow w$ where A is a variable and w is an **arbitrary** string over $V \cup \Sigma$

We now show that every grammar G can be converted to a G', such that $L(G) = L(G')$ and the rules of G' are **restricted** by the following definition

Definition (Chomsky Normal Form (CNF))

A context-free grammar $G = (V, \Sigma, R, S)$ is in **Chomsky normal form**, if every rule in R has one of the following forms:

1. $A \rightarrow BC$, where $A \in V$, and $B, C \in V \backslash \{S\}$
2. $A \rightarrow a$, where $A \in V$ and $a \in \Sigma$
3. $S \rightarrow \epsilon$, where S is the start variable

Chomsky Normal Form: Advantages

- For context-free grammars of arbitrary form there is no general limit on the number of steps it may take to derive a string w
- For CNF grammars there exists an algorithm that decides worst case runtime $O(|w|^3 \cdot |G|)$ whether w is derivable or not
- CNF makes proofs easier: We will show that we can construct a pushdown automaton with the same language as a given CNF grammar

Chomsky Normal Form: Theorem

Theorem

Let Σ be an alphabet and let $L \subseteq \Sigma^*$ be a context-free language. There exists a context-free grammar in **Chomsky normal form** *whose language is L*

Proof:

- We prove this by showing that we can transform an arbitrary context-free grammar G into a grammar G_5 in Chomsky normal form such that $L(G) = L(G_5)$
- The transformation consists of five steps, which must be performed in order
- We will build the intermediate grammars G_1 to G_5 with

$$L(G) = L(G_1) = L(G_2) = L(G_3) = L(G_4) = L(G_5)$$

Chomsky Normal Form: Step 1

Eliminate the start variable from the right-hand side of all rules

- Given $G = (V, \Sigma, R, A)$
- Define $G_1 = (V_1, \Sigma, R_1, S)$ where
  - S is the new start variable
  - $V_1 = V \cup \{S\}$ adding S to set of variables
  - $R_1 = R \cup \{S \rightarrow A\}$ adding rule that links new to old start variable
- G_1 has the following properties:
  - S does not occur on the right-hand side of any rule
  - $L(G_1) = L(G)$

Chomsky Normal Form: Step 1 – Example

Consider the $G = (V, \Sigma, R, A)$, where $V = \{A, B\}$, $\Sigma = \{0, 1\}$, A is the start variable and R consists of the rules:

$$A \rightarrow BAB|B|\epsilon$$
$$B \rightarrow 00|\epsilon$$

Eliminate the start variable from the right-hand side of all rules

- Introduce a new start variable S and add the rule $S \rightarrow A$:

$$S \rightarrow A$$
$$A \rightarrow BAB|B|\epsilon$$
$$B \rightarrow 00|\epsilon$$

Chomsky Normal Form: Step 2

Given $G_1 = (V_1, \Sigma, R_1, S)$**, eliminate all rules** $A \to \epsilon$ **where** $A \neq S$

For each such $A \to \epsilon$ in R_1:

- **Remove** $A \to \epsilon$
- For each rule in R_1 of the form
  (a) $B \to A$:
  Add $B \to \epsilon$ (unless this rule has been already removed)
  replaces $B \implies A \implies \epsilon$ by $B \implies \epsilon$
  (b) $B \to uAv$ (where u, v non-empty strings):
  Add $B \to uv$
  replaces $B \implies uAv \implies uv$ by $B \implies uv$
  (c) $B \to uAvAw$ (where u, v arbitrary strings):
  Add $B \to uvw$ (unless $u = v = w = \epsilon$ and $B \to \epsilon$ already removed)
  $B \to uAvw$
  $B \to uvAw$
  (d) treat rules in which A occurs more than twice on the right-hand side similar as in (c)

Chomsky Normal Form: Step 2 Continued

- Given $G_1 = (V_1, \Sigma, R_1, S)$
- Define $G_2 = (V_2, \Sigma, R_2, S)$ where
  - $V_2 = V_1$
  - R_2 corresponds to R_1 after eliminating all rules $A \rightarrow \epsilon$ according to the procedure on the previous slide
- G_2 has the following properties:
  - S does not occur on the right-hand side of any rule
  - R_2 does not contain any rule $A \rightarrow \epsilon$ where $A \neq S$
  - $L(G_2) = L(G_1) = L(G)$

Chomsky Normal Form: Step 2 – Example

Consider $G_1 = (V_1, \Sigma, R_1, S)$ with set of rules R_1:

$$S \rightarrow A$$
$$A \rightarrow BAB|B|\epsilon$$
$$B \rightarrow 00|\epsilon$$

Eliminate all rules of the form $A \rightarrow \epsilon$ where $A \neq S$

Remove $A \rightarrow \epsilon$, then consider all rules with A on the right-hand side:

- $S \rightarrow A$ is a rule, **add** $S \rightarrow \epsilon$ (a)
- $A \rightarrow BAB$ is a rule, **add** $A \rightarrow BB$ (b)

$$S \rightarrow A|\epsilon$$
$$A \rightarrow BAB|B|BB$$
$$B \rightarrow 00|\epsilon$$

Chomsky Normal Form: Step 2 – Example Continued

$$S \to A|\epsilon$$
$$A \to BAB|B|BB$$
$$B \to 00|\epsilon$$

Remove $B \to \epsilon$, then consider all rules with B on the right-hand side:

- $A \to BAB$ is a rule, **add** $A \to A$, $A \to AB$, $A \to BA$ (c)
- $A \to B$ is a rule, do not add $A \to \epsilon$ since rule already removed (a)
- $A \to BB$ is a rule, **add** $A \to B$ but not $A \to \epsilon$ (c)

$$S \to A|\epsilon$$
$$A \to A|AB|BA|BAB|B|BB$$
$$B \to 00$$

Chomsky Normal Form: Step 3

Given $G_2 = (V_2, \Sigma, R_2, S)$**, eliminate all unit rules** $A \to B$ **where** $A, B \in V_2$

For each such $A \to B$ in R_2:

- **Remove** $A \to B$
- For each rule in R_2 of the form $B \to u$ where $u \in (V_2 \cup \Sigma)^*$:
  - **Add** $A \to u$ (unless this rule has been already removed) replaces $A \implies B \implies u$ by $A \implies u$

Chomsky Normal Form: Step 3 Continued

- Given $G_2 = (V_2, \Sigma, R_2, S)$
- Define $G_3 = (V_3, \Sigma, R_3, S)$ where
  - $V_3 = V_2$
  - R_3 corresponds to R_2 after eliminating all unit rules $A \to B$ according to the procedure on the previous slide
- G_3 has the following properties:
  - S does not occur on the right-hand side of any rule
  - R_3 does not contain any rule $A \to \epsilon$ where $A \neq S$
  - R_3 does not contain any unit rule
  - $L(G_3) = L(G_2) = L(G_1) = L(G)$

Chomsky Normal Form: Step 3 – Example

Consider $G_2 = (V_2, \Sigma, R_2, S)$ with set of rules R_2:

$$S \rightarrow A|\epsilon$$
$$A \rightarrow A|AB|BA|BAB|B|BB$$
$$B \rightarrow 00$$

Eliminate all unit rules $A \rightarrow B$ **where** $A, B \in V_2$

Remove $A \rightarrow A$, then consider all rules with A on the left-hand side:

- $A \rightarrow AB|BA|BAB|B|BB$ are such rules,

  add $A \rightarrow AB|BA|BAB|B|BB$ (already contained)

$$S \rightarrow A|\epsilon$$
$$A \rightarrow AB|BA|BAB|B|BB$$
$$B \rightarrow 00$$

Chomsky Normal Form: Step 3 – Example Continued

$$S \to A|\epsilon$$
$$A \to AB|BA|BAB|B|BB$$
$$B \to 00$$

Remove $S \to A$, then consider all rules with A on the left-hand side:

- $A \to AB|BA|BAB|B|BB$ are such rules, **add** $S \to AB|BA|BAB|B|BB$

$$S \to \epsilon|AB|BA|BAB|B|BB$$
$$A \to AB|BA|BAB|B|BB$$
$$B \to 00$$

Chomsky Normal Form: Step 3 – Example Continued

$$S \to \epsilon | AB | BA | BAB | B | BB$$
$$A \to AB | BA | BAB | B | BB$$
$$B \to 00$$

Remove $S \to B$, then consider all rules with B on the left-hand side:

- $B \to 00$ is such a rule,
  add $S \to 00$

$$S \to \epsilon | AB | BA | BAB | BB | 00$$
$$A \to AB | BA | BAB | B | BB$$
$$B \to 00$$

Chomsky Normal Form: Step 3 – Example Continued

$$S \to \epsilon|AB|BA|BAB|BB|00$$
$$A \to AB|BA|BAB|B|BB$$
$$B \to 00$$

Remove $A \to B$, then consider all rules with B on the left-hand side:

- $B \to 00$ is such a rule,
  add $A \to 00$

$$S \to \epsilon|AB|BA|BAB|BB|00$$
$$A \to AB|BA|BAB|BB|00$$
$$B \to 00$$

All unit rules have been eliminated

Chomsky Normal Form: Step 4

Given $G_3 = (V_3, \Sigma, R_3, S)$, eliminate all rules that have more than two symbols on the right-hand side

For each $A \to u_1 \ldots u_k$ where $k \geq 3$ and each $u_i \in V_3 \cup \Sigma$:

- **Remove** $A \to u_1 \ldots u_k$
- **Add** the rules

$$\begin{aligned} A &\to u_1 A_1 \\ A_1 &\to u_2 A_2 \\ A_2 &\to u_3 A_3 \\ &\vdots \\ A_{k-3} &\to u_{k-2} A_{k-2} \\ A_{k-2} &\to u_{k-1} u_k \end{aligned}$$

where $A_1, \ldots, A_{k-2}$ are new variables that are added to V_3

(replaces a 1-step derivation by a $(k-1)$-step derivation)

Examples of rules to be eliminated:

$A \to abcd$
$A \to ABA$
$A \to AcdA$

$A \to abcd$
gets replaced by
$A \to aA_1$
$A_1 \to bA_2$
$A_2 \to cd$

Chomsky Normal Form: Step 4 Continued

- Given $G_3 = (V_3, \Sigma, R_3, S)$
- Define $G_4 = (V_4, \Sigma, R_4, S)$ where
  - V_4 corresponds to V_3 after adding new variables according to the procedure on the previous slide
  - R_4 corresponds to R_3 after eliminating all rules that have more than two symbols on the right-hand side, according to the procedure on the previous slide
- G_4 has the following properties:
  - S does not occur on the right-hand side of any rule
  - R_4 does not contain any rule $A \to \epsilon$ where $A \neq S$
  - R_4 does not contain any unit rule
  - R_4 does not contain any rule with more than two symbols on the right
  - $L(G_4) = L(G_3) = L(G_2) = L(G_1) = L(G)$

Chomsky Normal Form: Step 4 – Example

Consider $G_3 = (V_3, \Sigma, R_3, S)$ with set of rules R_3:

$$S \to \epsilon|AB|BA|BAB|BB|00$$
$$A \to AB|BA|BAB|BB|00$$
$$B \to 00$$

Eliminate all rules with more than two symbols on the right

- **Remove** $S \to BAB$ and **add** $S \to BA_1$ and $A_1 \to AB$
- **Remove** $A \to BAB$ and **add** $A \to BA_2$ and $A_2 \to AB$

$$S \to \epsilon|AB|BA|BB|00|BA_1$$
$$A \to AB|BA|BB|00|BA_2$$
$$B \to 00$$
$$A_1 \to AB$$
$$A_2 \to AB$$

Chomsky Normal Form: Step 5

Given $G_4 = (V_4, \Sigma, R_4, S)$**, eliminate all rules of the form** $A \to u_1 u_2$**, where** u_1 **and** u_2 **are not both variables**

For each such $A \to u_1 u_2$:

Remove $A \to u_1 u_2$

1. **If** $u_1 \in \Sigma$ and $u_2 \in V_4$, then
   add $A \to U_1 u_2$ and $U_1 \to u_1$ where U_1 is a new variable
   (replaces $A \implies u_1 u_2$ by $A \implies U_1 u_2 \implies u_1 u_2$)

2. **If** $u_1 \in V_4$ and $u_2 \in \Sigma$, then
   add $A \to u_1 U_2$ and $U_2 \to u_2$ where U_2 is a new variable
   (replaces $A \implies u_1 u_2$ by $A \implies u_1 U_2 \implies u_1 u_2$)

Examples of rules to be eliminated:

1) $A \to bC$
2) $A \to Cb$
3) $A \to bc$
4) $A \to bb$

$A \to bC$
gets replaced by
$A \to BC$
$B \to b$

Chomsky Normal Form: Step 5 Continued

$A \rightarrow bc$
gets replaced by
$A \rightarrow BC$
$B \rightarrow b$
$C \rightarrow c$

Remove $A \rightarrow u_1 u_2$

3. **If** $u_1 \in \Sigma$, $u_2 \in \Sigma$, and $u_1 \neq u_2$, then
   add $A \rightarrow U_1 U_2$, $U_1 \rightarrow u_1$, and $U_2 \rightarrow u_2$
   (replaces $A \implies u_1 u_2$ by $A \implies U_1 U_2 \implies U_1 u_2 \implies u_1 u_2$)

4. **If** $u_1 \in \Sigma$, $u_2 \in \Sigma$, and $u_1 = u_2$, then
   add $A \rightarrow U_1 U_1$ and $U_1 = u_1$
   (replaces $A \implies u_1 u_2 = u_1 u_1$ by $A \implies U_1 U_1 \implies U_1 u_1 \implies u_1 u_1$)

$A \rightarrow bb$
gets replaced by
$A \rightarrow BB$
$B \rightarrow b$

Chomsky Normal Form: Step 5 Continued

- Given $G_4 = (V_4, \Sigma, R_4, S)$, define $G_5 = (V_5, \Sigma, R_5, S)$ where
  - V_5 corresponds to V_4 after adding new variables according to the procedure on the previous slide
  - R_5 corresponds to R_4 after eliminating all rules of the form $A \to u_1 u_2$, where u_1 and u_2 are not both variables, according to the procedure on the previous slides

- G_5 has the following properties:
  - S does not occur on the right-hand side of any rule
  - R_5 does not contain any rule $A \to \epsilon$ where $A \neq S$
  - R_5 does not contain any unit rule
  - R_5 does not contain any rule with more than two symbols on the right
  - R_5 does not contain any rule of the form $A \to u_1 u_2$, where u_1 and u_2 are not both variables
  - $L(G_5) = L(G_4) = L(G_3) = L(G_2) = L(G_1) = L(G)$

Since the grammar G_5 is in Chomsky normal form, the proof is complete

Chomsky Normal Form: Step 5 – Example

Consider $G_4 = (V_4, \Sigma, R_4, S)$ with set of rules R_4:

$$
\begin{aligned}
S &\rightarrow \epsilon | AB | BA | BB | 00 | BA_1 \\
A &\rightarrow AB | BA | BB | 00 | BA_2 \\
B &\rightarrow 00 \\
A_1 &\rightarrow AB \\
A_2 &\rightarrow AB
\end{aligned}
$$

Eliminate all rules of the form $A \rightarrow u_1 u_2$, where u_1 and u_2 are not both variables:

- Replace the rule $S \rightarrow 00$ by the rules $S \rightarrow A_3 A_3$ and $A_3 \rightarrow 0$
- Replace the rule $A \rightarrow 00$ by the rules $A \rightarrow A_4 A_4$ and $A_4 \rightarrow 0$
- Replace the rule $B \rightarrow 00$ by the rules $B \rightarrow A_5 A_5$ and $A_5 \rightarrow 0$

Chomsky Normal Form: Step 5 – Example Continued

This gives us a grammar with the following rules

$$
\begin{aligned}
S &\rightarrow BB|AB|BA|A_3A_3|BA_1|\epsilon \\
A &\rightarrow BB|AB|BA|A_4A_4|BA_2 \\
B &\rightarrow A_5A_5 \\
A_1 &\rightarrow AB \\
A_2 &\rightarrow AB \\
A_3 &\rightarrow 0 \\
A_4 &\rightarrow 0 \\
A_5 &\rightarrow 0
\end{aligned}
$$

which is in Chomsky Normal form