بسم الله الرحمن الرحیم

نظریه زبان‌ها و ماشین‌ها

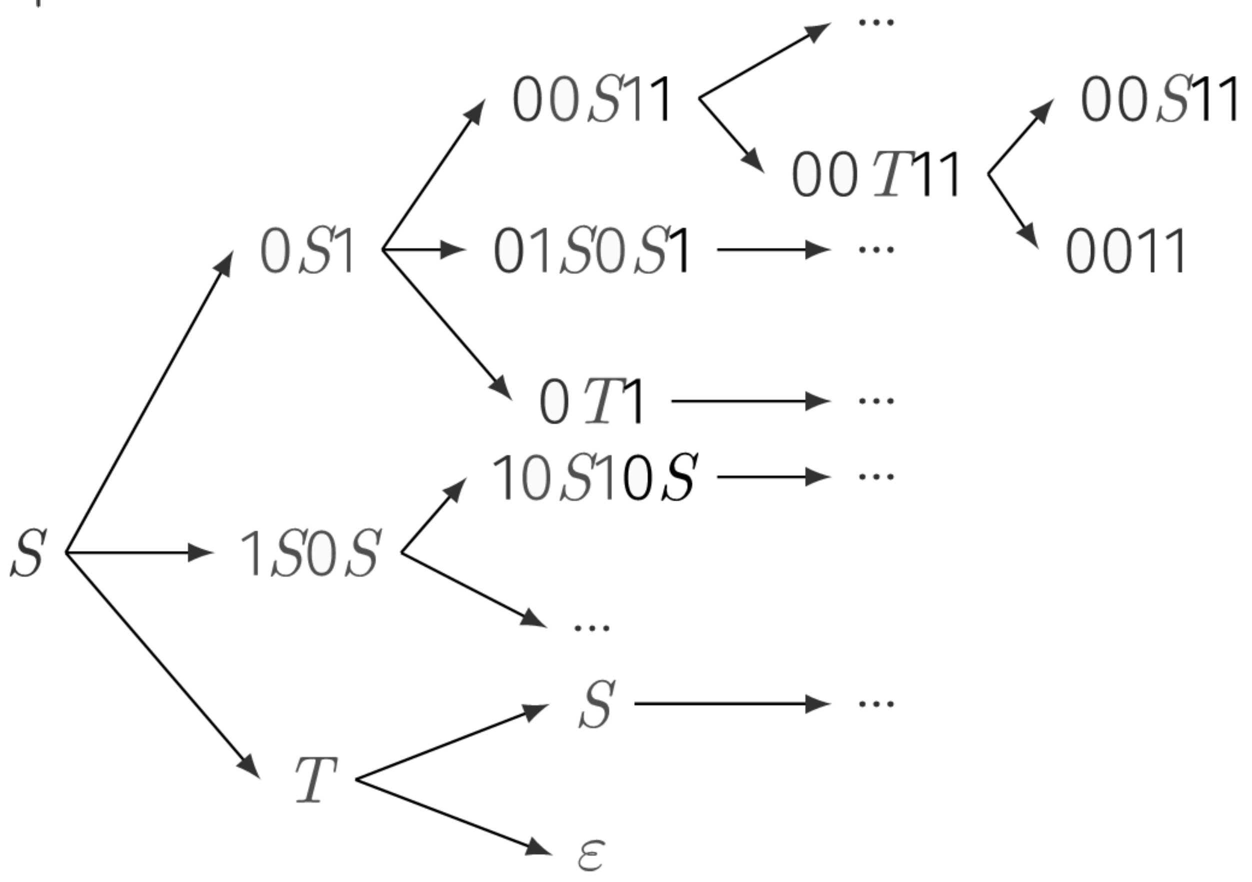
جلسه ۱۶

مجتبی خلیلی
دانشکده برق و کامپیوتر
دانشگاه صنعتی اصفهان

مرور

$$S \to 0S1 \mid 1S0S \mid T$$

$$T \to S \mid \varepsilon$$



مرور

remove all undesirable productions using the following sequence of steps:

1. Remove λ-productions.

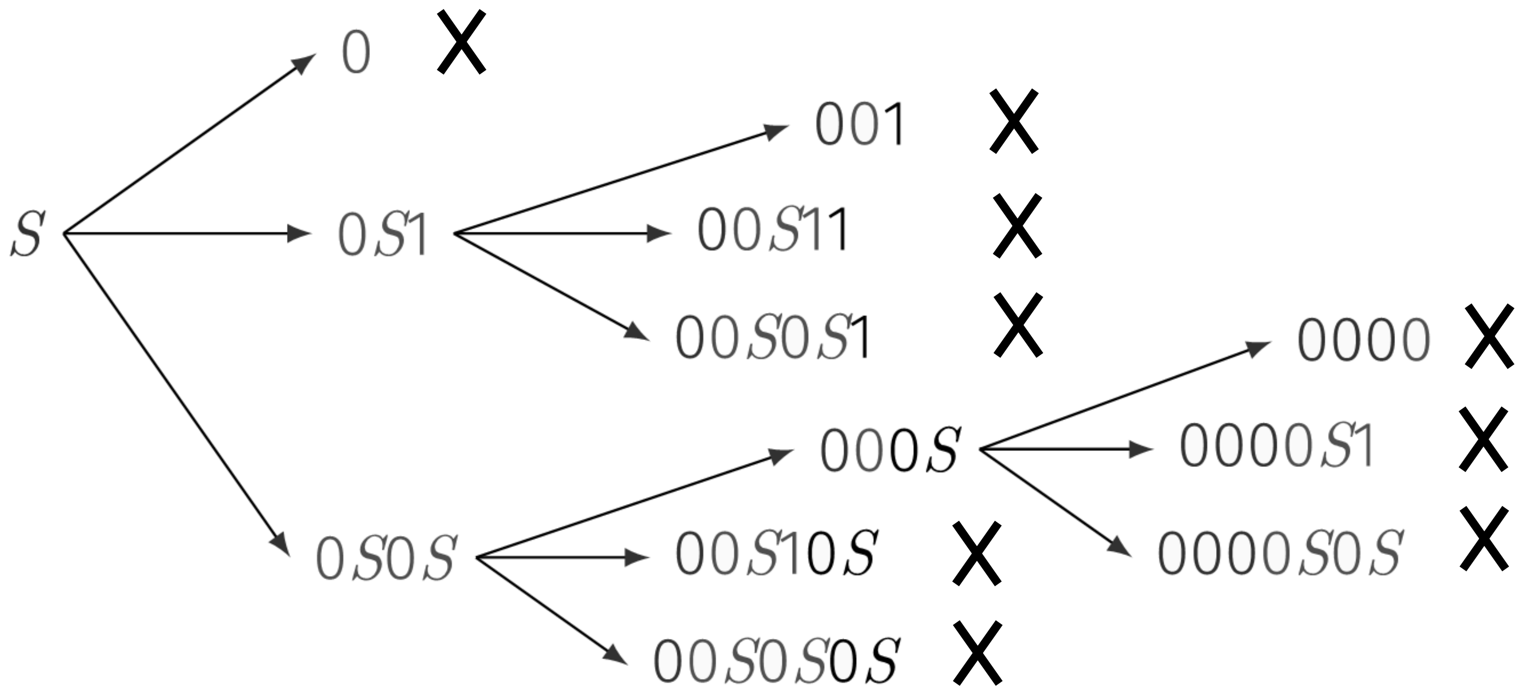
2. Remove unit-productions.

3. Remove useless productions.

تجزیه

$$S \to 0S1 \mid 0S0S \mid 0$$

$0011 \in L(G)?$



تجزیه

For reference below, we will call this **exhaustive search parsing** or **brute force parsing**. It is a form of **top-down parsing**, which we can view as the construction of a derivation tree from the root down.

تجزیه

THEOREM 5.2

Suppose that $G = (V, T, S, P)$ is a context-free grammar that does not have any rules of the form

$$A \rightarrow \lambda,$$

or

$$A \rightarrow B,$$

where $A, B \in V$. Then the exhaustive search parsing method can be made into an algorithm that, for any $w \in \Sigma^*$, either produces a parsing of w or tells us that no parsing is possible.

تجزیه

Proof: For each sentential form, consider both its length and the number of terminal symbols. Each step in the derivation increases at least one of these. Since neither the length of a sentential form nor the number of terminal symbols can exceed $|w|$, a derivation cannot involve more than $2|w|$ rounds, at which time we either have a successful parsing or w cannot be generated by the grammar. ■

total number of sentential forms cannot exceed

$$
\begin{aligned}
M &= |P| + |P|^2 + \cdots + |P|^{2|w|} \\
&= O(P^{2|w|+1}).
\end{aligned}
$$

تجزیه

- به دنبال روش کارآمدتری برای تجزیه هستیم.
- الگوریتم CYK (cocke-younger-kasami)
- الگوریتم‌های دیگری نیز هستند که در درسهای بعدی خواهید دید (گرامر).
- نیاز به معرفی فرم نرمال چامسکی

فرم نرمال چامسکی (CNF)

DEFINITION 2.8

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 not be the start variable. In addition, we permit the rule $S \to \varepsilon$, where S is the start variable.

فرم نرمال چامسکی (CNF)

THEOREM 2.9

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

PROOF IDEA We can convert any grammar G into Chomsky normal form. The conversion has several stages wherein rules that violate the conditions are replaced with equivalent ones that are satisfactory. First, we add a new start variable. Then, we eliminate all ***ε-rules*** of the form $A \to \varepsilon$. We also eliminate all ***unit rules*** of the form $A \to B$. In both cases we patch up the grammar to be sure that it still generates the same language. Finally, we convert the remaining rules into the proper form.

فرم نرمال چامسکی

- تبدیل در ۵ گام:

- گام ۱: استارت

First, we add a new start variable S_0 and the rule $S_0 \rightarrow S$, where S was the original start variable. This change guarantees that the start variable doesn't occur on the right-hand side of a rule.

فرم نرمال چامسکی

$$S \rightarrow aSa \mid bSb \mid \epsilon$$

$$S' \rightarrow S$$
$$S \rightarrow aSa \mid bSb \mid \epsilon$$

فرم نرمال چامسکی

- تبدیل در ۵ گام:
  - گام ۲: حذف سمت راستی‌های با متغیر و ترمینال

فرم نرمال چامسکی

$$S' \to S$$
$$S \to \boldsymbol{a}S\boldsymbol{a} \mid \boldsymbol{b}S\boldsymbol{b} \mid \epsilon$$

$$S' \to S$$
$$S \to \boldsymbol{A}S\boldsymbol{A} \mid \boldsymbol{B}S\boldsymbol{B} \mid \epsilon$$
$$\boldsymbol{A} \to \boldsymbol{a}$$
$$\boldsymbol{B} \to \boldsymbol{b}$$

فرم نرمال چامسکی

- تبدیل در ۵ گام:
  - گام ۳: حذف سمت راستی‌های با بیش از دو متغیر

فرم نرمال چامسکی

$$S' \rightarrow S$$
$$S \rightarrow \boldsymbol{ASA} \mid \boldsymbol{BSB} \mid \epsilon$$
$$A \rightarrow a; B \rightarrow b$$

$$S' \rightarrow S$$
$$S \rightarrow \boldsymbol{AX} \mid \boldsymbol{BY} \mid \epsilon$$
$$\boldsymbol{X \rightarrow SA; Y \rightarrow SB}$$
$$A \rightarrow a; B \rightarrow b$$

فرم نرمال چامسکی

- تبدیل در ۵ گام:
  - گام ۴: حذف ϵ-rules

Second, we take care of all ε-rules. We remove an ε-rule $A \to \varepsilon$, where A is not the start variable. Then for each occurrence of an A on the right-hand side of a rule, we add a new rule with that occurrence deleted. In other words, if $R \to uAv$ is a rule in which u and v are strings of variables and terminals, we add rule $R \to uv$. We do so for each *occurrence* of an A, so the rule $R \to uAvAw$ causes us to add $R \to uvAw$, $R \to uAvw$, and $R \to uvw$. If we have the rule $R \to A$, we add $R \to \varepsilon$ unless we had previously removed the rule $R \to \varepsilon$. We repeat these steps until we eliminate all ε-rules not involving the start variable.

فرم نرمال چامسکی

$$S' \rightarrow \boldsymbol{S}$$
$$S \rightarrow AX \mid BY \mid \boldsymbol{\epsilon}$$
$$X \rightarrow \boldsymbol{S}A; Y \rightarrow \boldsymbol{S}B$$
$$A \rightarrow a; B \rightarrow b$$

$$S' \rightarrow S \mid \boldsymbol{\epsilon}$$
$$S \rightarrow AX \mid BY$$
$$X \rightarrow SA \mid \boldsymbol{A}; Y \rightarrow SB \mid \boldsymbol{B}$$
$$A \rightarrow a; B \rightarrow b$$

فرم نرمال چامسکی

- تبدیل در ۵ گام:

- گام ۵: حذف unit rules

Third, we handle all unit rules. We remove a unit rule $A \rightarrow B$. Then, whenever a rule $B \rightarrow u$ appears, we add the rule $A \rightarrow u$ unless this was a unit rule previously removed. As before, u is a string of variables and terminals. We repeat these steps until we eliminate all unit rules.

فرم نرمال چامسکی

$$\boldsymbol{S'} \rightarrow \boldsymbol{S} \mid \epsilon$$
$$S \rightarrow AX \mid BY$$
$$\boldsymbol{X} \rightarrow SA \mid \boldsymbol{A};\boldsymbol{Y} \rightarrow SB \mid \boldsymbol{B}$$
$$A \rightarrow a; B \rightarrow b$$

$$\boldsymbol{S'} \rightarrow \boldsymbol{AX} \mid \boldsymbol{BY} \mid \epsilon$$
$$S \rightarrow AX \mid BY$$
$$X \rightarrow SA \mid \boldsymbol{a};Y \rightarrow SB \mid \boldsymbol{b}$$
$$A \rightarrow a; B \rightarrow b$$

مثال (تبدیل به CNF)

$$A \to BAB \mid B \mid \varepsilon$$
$$B \to 00 \mid \varepsilon$$

$$\Rightarrow$$

$$S \to A$$
$$A \to BAB \mid B \mid \varepsilon$$
$$B \to 00 \mid \varepsilon$$

$$S \to A$$
$$A \to BA_1 \mid B \mid \varepsilon$$
$$B \to 00 \mid \varepsilon$$
$$A_1 \to AB$$

$$\Rightarrow$$

$$S \to A \mid \varepsilon$$
$$A \to BA_1 \mid B \mid A_1$$
$$B \to 00$$
$$A_1 \to AB \mid B \mid A$$

مثال (تبدیل به CNF)

IUT-ECE

$$S \to A \mid \varepsilon$$
$$A \to BA_1 \mid B \mid A_1$$
$$B \to 00$$
$$A_1 \to AB \mid B \mid A$$

$$S \to BA_1 \mid \varepsilon \mid 00 \mid AB$$
$$A \to BA_1 \mid 00 \mid AB$$
$$B \to 00$$
$$A_1 \to AB \mid 00 \mid BA_1$$

$$S \to BA_1 \mid \varepsilon \mid ZZ \mid AB$$
$$A \to BA_1 \mid ZZ \mid AB$$
$$B \to ZZ$$
$$A_1 \to AB \mid ZZ \mid BA_1$$
$$Z \to 0$$

فرم نرمال چامسکی (CNF): مثال

$L = \{w \mid w \in \{a, b\}^* \text{ and } w = w^R\}$

$S \rightarrow a|b|aSa|bSb|\epsilon$

فرم نرمال چامسکی (CNF): مثال

IUT-ECE

$L = \{w \mid w \in \{a,b\}^* \text{ and } w = w^R\}$

baaab

$$
\begin{aligned}
S &\to AU \mid BV \mid \mathtt{a} \mid \mathtt{b} \mid \varepsilon \\
T &\to AU \mid BV \mid \mathtt{a} \mid \mathtt{b} \\
U &\to TA \\
V &\to TB \\
A &\to \mathtt{a} \\
B &\to \mathtt{b}
\end{aligned}
$$

$$
\begin{aligned}
S &\Rightarrow BV \\
&\Rightarrow \mathtt{b}V \\
&\Rightarrow \mathtt{b}TB \\
&\Rightarrow \mathtt{b}AUB \\
&\Rightarrow \mathtt{ba}UB \\
&\Rightarrow \mathtt{ba}TAB \\
&\Rightarrow \mathtt{baa}AB \\
&\Rightarrow \mathtt{baaa}B \\
&\Rightarrow \mathtt{baaab}
\end{aligned}
$$

2|w|-1

مثال

1. The original CFG G_6 is shown on the left. The result of applying the first step to make a new start variable appears on the right.

$$\begin{aligned} S &\to ASA \mid \mathtt{a}B \\ A &\to B \mid S \\ B &\to \mathtt{b} \mid \varepsilon \end{aligned}$$

$$\begin{aligned} \boldsymbol{S_0} &\to \boldsymbol{S} \\ S &\to ASA \mid \mathtt{a}B \\ A &\to B \mid S \\ B &\to \mathtt{b} \mid \varepsilon \end{aligned}$$

مثال

IUT-ECE

2. Remove ε-rules $B \to \varepsilon$, shown on the left, and $A \to \varepsilon$, shown on the right.

$$\begin{aligned} S_0 &\to S \\ S &\to ASA \mid \mathtt{a}B \mid \mathbf{a} \\ A &\to B \mid S \mid \boldsymbol{\varepsilon} \\ B &\to \mathtt{b} \mid \varepsilon \end{aligned}$$

$$\begin{aligned} S_0 &\to S \\ S &\to ASA \mid \mathtt{a}B \mid \mathtt{a} \mid \boldsymbol{SA} \mid \boldsymbol{AS} \mid \boldsymbol{S} \\ A &\to B \mid S \mid \varepsilon \\ B &\to \mathtt{b} \end{aligned}$$

مثال

3a. Remove unit rules $S \to S$, shown on the left, and $S_0 \to S$, shown on the right.

$$\begin{aligned} S_0 &\to S \\ S &\to ASA \mid \mathtt{a}B \mid \mathtt{a} \mid SA \mid AS \mid S \\ A &\to B \mid S \\ B &\to \mathtt{b} \end{aligned}$$

$$\begin{aligned} S_0 &\to S \mid \boldsymbol{ASA} \mid \mathbf{a}\boldsymbol{B} \mid \mathbf{a} \mid \boldsymbol{SA} \mid \boldsymbol{AS} \\ S &\to ASA \mid \mathtt{a}B \mid \mathtt{a} \mid SA \mid AS \\ A &\to B \mid S \\ B &\to \mathtt{b} \end{aligned}$$

مثال

3b. Remove unit rules $A \to B$ and $A \to S$.

$$
\begin{aligned}
S_0 &\to ASA \mid \mathtt{a}B \mid \mathtt{a} \mid SA \mid AS \\
S &\to ASA \mid \mathtt{a}B \mid \mathtt{a} \mid SA \mid AS \\
A &\to B \mid S \mid \mathtt{b} \\
B &\to \mathtt{b}
\end{aligned}
\qquad
\begin{aligned}
S_0 &\to ASA \mid \mathtt{a}B \mid \mathtt{a} \mid SA \mid AS \\
S &\to ASA \mid \mathtt{a}B \mid \mathtt{a} \mid SA \mid AS \\
A &\to S \mid \mathtt{b} \mid \boldsymbol{ASA} \mid \mathbf{a}\boldsymbol{B} \mid \mathbf{a} \mid \boldsymbol{SA} \mid \boldsymbol{AS} \\
B &\to \mathtt{b}
\end{aligned}
$$

مثال

4. Convert the remaining rules into the proper form by adding additional variables and rules. The final grammar in Chomsky normal form is equivalent to G_6. (Actually the procedure given in Theorem 2.9 produces several variables U_i and several rules $U_i \to \texttt{a}$. We simplified the resulting grammar by using a single variable U and rule $U \to \texttt{a}$.)

$$\begin{aligned}
S_0 &\to AA_1 \mid UB \mid \texttt{a} \mid SA \mid AS \\
S &\to AA_1 \mid UB \mid \texttt{a} \mid SA \mid AS \\
A &\to \texttt{b} \mid AA_1 \mid UB \mid \texttt{a} \mid SA \mid AS \\
A_1 &\to SA \\
U &\to \texttt{a} \\
B &\to \texttt{b}
\end{aligned}$$

تجزیه

- به دنبال روش کارآمدتری برای تجزیه هستیم.
- الگوریتم CYK (cocke-younger-kasami)
- الگوریتم‌های دیگری نیز هستند که در درسهای بعدی خواهید دید.
- نیاز به معرفی فرم نرمال چامسکی

الگوریتم CYK

Assume that we have a grammar $G = (V, T, S, P)$ in Chomsky normal form and a string

$$w = a_1 a_2 \cdots a_n.$$

We define substrings

$$w_{ij} = a_i \cdots a_j,$$

and subsets of V

$$V_{ij} = \left\{ A \in V : A \overset{*}{\Rightarrow} w_{ij} \right\}.$$

Clearly, $w \in L(G)$ if and only if $S \in V_{1n}$.

الگوریتم CYK

IUT-ECE

$$V_{ij} = \left\{ A \in V : A \stackrel{*}{\Rightarrow} w_{ij} \right\}.$$

1. To compute V_{ij}, observe that $A \in V_{ii}$ if and only if G contains a production $A \to a_i$.

2. for $j > i$, A derives w_{ij} if and only if there is a production $A \to BC$, with $B \stackrel{*}{\Rightarrow} w_{ik}$ and $C \stackrel{*}{\Rightarrow} w_{k+1j}$ for some k with $i \leq k, k < j$. In other words,

$$V_{ij} = \bigcup_{k \in \{i, i+1, \ldots, j-1\}} \{A : A \to BC, \text{ with } B \in V_{ik}, C \in V_{k+1,j}\}. \qquad (6.8)$$

الگوریتم CYK

$$V_{ij} = \left\{ A \in V : A \stackrel{*}{\Rightarrow} w_{ij} \right\}.$$

An inspection of the indices in (6.8) shows that it can be used to compute all the V_{ij} if we proceed in the sequence

1. Compute $V_{11}, V_{22}, ..., V_{nn}$,
2. Compute $V_{12}, V_{23}, ..., V_{n-1,n}$,
3. Compute $V_{13}, V_{24}, ..., V_{n-2,n}$,

and so on.

Dynamic programming