دانشکده‌ی علوم ریاضی

نظریه‌ی زبان‌ها و اتوماتا ۹ آبان ۱۳۹۱

جلسه‌ی ۱۴: زبان‌های مستقل از متن

مدرّس: دکتر شهرام خزائی *نگارنده: حمید پوررربیع رودسری*

۱ مقدمه

گرامر مستقل از متن[۱] یک سیستم نمادگذاری برای کلاسی از زبان‌هاست که با نام زبان‌های مستقل از متن[۲] شناخته می‌شوند. این‌زبان‌ها کلاس بزرگتری را نسبت به کلاس زبان‌های منظم شامل می‌شوند و می‌توانند زبان‌هایی را توصیف کنند که منظم نیستند. این گرامرها به خاطر ساختار ذاتی خود برای توصیف رشته‌هایی که دارای ساختار تودرتو هستند مناسب هستند.

۲ چند مثال

بهتر است برای آشنایی بیشتر با زبان‌های مستقل از متن با چند مثال شروع کنیم:

مثال ۱ *همانطور که در بحث زبان‌های منظم گفته شد زبان L به صورت*

$$L = \{w \in \{0, 1\}^* : w = 0^n 1^n, n \geq 1\}$$

که شامل همه‌ی رشته‌هایی به صورت $0^n 1^n$ است را نمی‌توان با یک ماشین حالت محدود و یا یک عبارت منظم نمایش داد زیرا منظم نیست. ما اینجا نشان می دهیم که این زبان، یک زبان مستقل از متن است و می‌توان آن را با گرامر مستقل از متن تولید کرد. برای تعریف این زبان با استفاده از یک گرامر می‌توان از دو ویژگی زیر استفاده کرد:
۱. رشته‌ی 01 عضو این زبان است؛
۲. در صورتی که رشته‌ی w عضو این زبان باشد، $0w1$ نیز عضو این زبان است.
حال این دو شرط را در گرامر مستقل از متن می‌توان به صورت دو «قانون» زیر نوشت:

$$\begin{aligned} S &\rightarrow 01 \\ S &\rightarrow 0S1 \end{aligned}$$

با بیان بالا می‌توان تمام رشته‌های موجود در زبان را با استفاده از استقرا به دست آورد.

[۱] Contex-Free Grammers
[۲] Contex-Free Languages

مثال ۲ *می خواهیم گرامری برای زبانی که شامل تمام رشته‌هایی که قلب مستوی[۳] است را به دست آوریم. برای راحتی فرض می کنیم تمام رشته‌ها از* $\{۰, ۱\}$ *تشکیل شده‌اند. در این صورت رشته‌هایی مثل* ۱۱۰۱۰۱۱ *و* ۱۰۱۰۱۰۱ *در زبان موجود هستند اما رشته‌هایی مثل* ۱۱۰۰۱ *یا* ۱۰۱۰۱۰ *در این زبان نیستند. این زبان را می‌توان به صورت* $L = \{w \in \{۰, ۱\}^* : w = w^R\}$ *نشان داد.*
با کمی تفکر در رشته‌های این زبان به دو نکته‌ی اساسی پی می‌بریم:
۱. *رشته‌های* ۰ *و* ۱ *و* ϵ *در این زبان موجودند.*
۲. *در صورتی که رشته‌ی* w *در این زبان باشد رشته‌های* $۱w۱$ *و* $۰w۰$ *نیز عضو این زبان هستند.*
با استفاده از نکات بالا می توان این زبان را با استفاده از قوانین زیر تولید کرد:

$$\begin{aligned} S &\rightarrow \epsilon \\ S &\rightarrow ۰ \\ S &\rightarrow ۱ \\ S &\rightarrow ۱S۱ \\ S &\rightarrow ۰S۰ \end{aligned}$$

۳ تعریف رسمی

تعریف ۱ *یک گرامر مستقل از متن یک چهارتایی به صورت* $G = (V, T, P, S)$ *می باشد که در آن:*
۱. V *یک مجموعه‌ی متناهی از نمادها به نام متغیر[۴] است.*
هرکدام از این متغیرها بیانگر یک زبان (مجموعه ای از چند رشته) هستند در مثال‌های بالا V *فقط شامل متغیر* S *است.*
۲. T *مجموعه ای متناهی از نمادها به نام پایانه (ترمینال)[۵] است.*
در مثال‌های ذکر شده T *برابر است با* $\{۰, ۱\}$.
۳. P *یک مجموعه‌ی متناهی از قوانین تولید[۶] به صورت* $A \rightarrow \alpha$ *است که* $A \in V$ *و* $\alpha \in (V \cup T)^*$.
۴. S *متغیر شروع[۷] است که* $S \in V$.

مثال ۳ *گرامر قلب مستوی به صورت زیر تعریف می شود:*

$$G = \{\{S\}, \{۰, ۱\}, P, \{S\}\}$$

که در آن P *بیانگر قوانینی است که در مثال* **۲** *بیان شده است.*

مثال ۴ *می خواهیم زبان عبارات جبری ساده که از* a *و* b *و* ۱ *و* ۰ *تشکیل شده اند و فقط جمع و ضرب در بین*

---

[۳] رشته‌های قلب مستوی ($plaindromes$)، رشته‌هایی هستند که از هر دو طرف به یک صورت خوانده می شوند مانند گرگ در فارسی یا ("$Madam, I'm\ adam$") $madamimadam$ در انگلیسی.

[۴] Variable
[۵] Terminal
[۶] Productions
[۷] Start symbol

آن‌ها مجاز است را بنویسم (متغیرها باید با a یا b شروع شوند و بعد از آن‌ها می تواند عدد یا حرف باشد):

$$\begin{aligned} E &\to I \\ E &\to E+E \\ E &\to E*E \\ E &\to (E) \\ I &\to a \\ I &\to b \\ I &\to Ia \\ I &\to Ib \\ I &\to I\text{۱} \\ I &\to I\text{۰} \end{aligned}$$

قانون اول بیانگر عبارات ساده‌ی جبری است که فقط شامل یک متغیر هستند. قانون دوم عباراتی که جمع در آن‌ها وجود دارد را توصیف می کند، قانون سوم ضرب و قانون چهارم پرانتز دور عبارات را. متغیر کمکی I هم زبانی را تولید می کند که شامل همه‌ی متغیرهای قابل قبول است که در واقع یک زبان منظم به صورت

$$(\mathbf{a}+\mathbf{b})(\mathbf{a}+\mathbf{b}+\mathbf{0}+\mathbf{1})^*$$

است.

نکته: مرسوم است قوانین تولیدی که ابتدای آن‌ها متغیر A است به عنوان قوانین تولید A شناخت. می‌توان قوانین تولید را به صورت خلاصه به این صورت نوشت که همه‌ی متغیرها را یک بار در سمت چپ آورد و تمام قواعد مربوط به هر متغیر را در سمت راست نوشت و آن‌ها را با یک خط عمودی از هم جدا کرد. برای مثال می توان قوانین $A \to \alpha_۱, A \to \alpha_۲, A \to \alpha_۳, \ldots$ را به صورت خلاصه به شکل $A \to \alpha_۱ \mid \alpha_۲ \mid \alpha_۳ \mid \ldots$ نیز نوشت.

شایان ذکر است که می‌توان قوانین نوشته شده در مثال ۴ را به صورت زیر خلاصه کرد:

$$\begin{aligned} E &\to I \mid E+E \mid E*E \mid (E) \\ I &\to a \mid b \mid Ia \mid Ib \mid I\text{۱} \mid I\text{۰} \end{aligned}$$

۴ اشتقاق

تعریف ۲ فرض کنید $G=(V,T,P,S)$ یک گرامر مستقل از متن باشد. به ازای هر $A \in V$ و هر $\alpha, \beta \in (V \cup T)^*$ و هر $(A \to \gamma) \in P$ می گوییم $\alpha\gamma\beta$ یک اشتقاق از $\alpha A\beta$ است و می نویسیم:

$$\alpha A\beta \Rightarrow \alpha\gamma\beta$$

به عنوان مثال:

$$(E*E)*I\text{۰} \Rightarrow (E*E)*a\text{۰} \Rightarrow (I*E)*a\text{۰}$$

تعریف ۳ (اشتقاق مکرر، اشتقاق چند مرحله‌ای) فرض کنید $G=(V,T,P,S)$ یک گرامر مستقل از متن باشد. می‌گوییم β اشتقاق مکرری از α است (که $\alpha, \beta \in (V \cup T)^*$) و با $\alpha \overset{*}{\Rightarrow} \beta$ نشان می‌دهیم، اگر α و β در تعریف استقرایی زیر بگنجند:

پایه:

$$\alpha \overset{*}{\Rightarrow} \alpha$$

استقرا:

$$[\alpha \overset{*}{\Rightarrow} \gamma \wedge \gamma \Rightarrow \beta] \Longrightarrow [\alpha \overset{*}{\Rightarrow} \beta]$$

۵ عرف نمادگذاری

معمولا برای نمایش متغیرهای تولید از حرف‌های ابتدایی و بزرگ الفبا $(A, B, C, ...)$ و همچنین حرف S،
برای ترمینال‌ها از حروف ابتدایی و کوچک الفبا $(a, b, c, ...)$،
برای رشته‌های روی ترمینال‌ها از حروف انتهایی و کوچک الفبا $(..., w, x, y, z)$،
برای متغیر یا پایانه از حروف انتهایی و بزرگ الفبا $(..., X, Y, Z)$
و برای رشته‌های روی ترمینال‌ها و متغیرها از $(\alpha, \beta, ...)$ استفاده می کنیم.