Linguagens Formais e Autômatos

Aula 5: Gramática

Prof. Dr. Rodrigo Xavier de Almeida Leão

Cientista de Dados e Big Data

Gramáticas

Também conhecidas como **dispositivos generativos**, **dispositivos de síntese**, ou ainda dispositivos de geração de cadeias, as **gramáticas** constituem sistemas formais baseados em regras de substituição, através dos quais é possível sintetizar, de forma exaustiva, o conjunto das cadeias que compõem uma determinada linguagem.

Para ilustrar esse conceito, nada melhor do que a própria noção intuitiva, adquirida à época do ensino fundamental, do significado do termo "gramática": o livro através do qual são aprendidas as regras que indicam como falar e escrever corretamente um idioma.

Como se sabe, as regras assim definidas especificam combinações válidas dos símbolos que compõem o alfabeto — os diversos verbos, substantivos, adjetivos, advérbios, pronomes, artigos etc. —, e isso é feito com o auxílio de entidades abstratas denominadas classes sintáticas: sujeito, predicado etc. Assim, por exemplo, a frase "O menino atravessou a rua distraidamente" é considerada correta, do ponto de vista gramatical, pois ela obedece a uma das inúmeras regras de formação de frases, baseada no padrão sujeito + predicado + complemento.

Gramáticas

De acordo com tais regras, um sujeito pode ser composto por um *artigo* (**"O"**) seguido de um *substantivo* (**"menino"**), o *predicado* pode conter um *verbo* (**"atravessou"**) e um correspondente *objeto direto* (**"a rua"**), e o *complemento* pode modificar a ação (**"distraidamente"**). O *objeto direto*, por sua vez, pode seguir o mesmo padrão estrutural do *sujeito*: *artigo* (**"a"**) seguido de *substantivo* (**"rua"**).

Naturalmente, o conjunto das regras que formam uma "gramática" deve ser suficiente para permitir a elaboração de qualquer frase ou discurso corretamente construído em um determinado idioma, e não deve permitir a construção de qualquer cadeia que não pertença à linguagem. Convém notar, no exemplo do parágrafo acima, que os termos em *itálico* correspondem às denominadas classes sintáticas do português, e os termos em **negrito**, aos símbolos que efetivamente fazem parte do seu alfabeto.

Gramáticas

Assim como ocorre no caso das linguagens naturais, as linguagens formais também podem ser especificadas através de "gramáticas" a elas associadas. No caso das gramáticas das linguagens formais, que constituem o objeto deste estudo, a analogia com as "gramáticas" das linguagens naturais é muito grande. Tratam-se, as primeiras, de conjuntos de regras que, quando aplicadas de forma recorrente, possibilitam a geração de todas as cadeias pertencentes a uma determinada linguagem.

Diferentemente das gramáticas das linguagens naturais, que são descritas por intermédio de linguagens também naturais (muitas vezes a mesma que está sendo descrita pela gramática), as gramáticas das linguagens formais são descritas utilizando notações matemáticas rigorosas que visam, entre outros objetivos, evitar dúvidas na sua interpretação. Tais notações recebem a denominação de **metalinguagens** — linguagens que são empregadas para definir outras linguagens.

Formalmente, uma gramática G pode ser definida como sendo uma quádrupla[2]:

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

onde:

- V é o **vocabulário** da gramática; corresponde a um conjunto (finito e não-vazio) de símbolos;
- Σ é o conjunto (finito e não-vazio) dos símbolos **terminais** da gramática; também denominado **alfabeto**;
- P é o conjunto (finito e não-vazio) de **produções** ou **regras de substituição** da gramática;
- S é a **raiz** da gramática, $S \in V$.

Gramática

Uma gramática G é uma tupla (V, T, P, S) onde:

- V é um conjunto finito não vazio de variáveis.
- T é um conjunto finito não vazio de símbolos terminais.
- P é um conjunto finito de regras de produção da forma: $\alpha \rightarrow \beta$ onde $\alpha, \beta \in (V \cup T)^*$, onde α tem ao menos uma variável.
- $S \in V$ é o símbolo inicial.

Gramática

Na Seção 1.1 estudamos o conceito de relação. O funcionamento da gramática se dá através da relação $\Rightarrow_G$ ou simplesmente $\Rightarrow$. O domínio e o contradomínio desta relação é o conjunto $(V \cup T)^*$. Se $G = (V, T, P, S)$, dizemos que $\alpha \Rightarrow_G \beta$ quando $\alpha = \delta_1 \alpha_1 \delta_2$ e $\beta = \delta_1 \beta_1 \delta_2$ e a regra $\alpha_1 \rightarrow \beta_1$ está em P. Ou seja, se uma regra da gramática descreve como podemos trocar α_1, um pedaço de α igual ao lado esquerdo de uma regra, pelo lado direito da mesma regra, obtendo β. Neste caso dizemos que α deriva β. Observe que α_1 pode ocorrer em mais de uma posição em α e que a aplicação da regra $\alpha_1 \rightarrow \beta_1$ pode ser aplicada em qualquer destas posições. Devemos aplicar as regras até que não haja mais não variáveis.

Gramática

Vamos considerar uma gramática análoga à gramática vista para números decimais. Entretanto, esta gramática só possui os dígitos 0 e 1, gerando os números de ponto flutuante na base binária. $G = (V, T, P, N)$, onde:

- $T = \{0, 1\}$
- $V = \{N, L, D\}$;
- $P = \{N \to L, \quad N \to L.L, L \to D, L \to LD, D \to 0, D \to 1\}$.

Para esta gramática temos os seguintes exemplos de $\Rightarrow_G$:

Gramática

- $L.L \Rightarrow_G L.LD$
- $L.L \Rightarrow_G LD.L$

Quando a gramática G está subentendida no contexto, nós a suprimimos da notação, escrevendo simplesmente:

- $L.LD \Rightarrow L.L0$

É muito útil escrever mais de um passo da relação $\Rightarrow_G$ com um único símbolo $\Rightarrow_G^*$, por exemplo, se $N \Rightarrow_G L.L \Rightarrow_G D.L \Rightarrow_G 1.L$, podemos escrever $N \Rightarrow_G^* 1.L$. O * indica que $\Rightarrow_G^*$ também é reflexiva, ou seja, $\alpha \Rightarrow^* \alpha$.

Esta notação é lida como "deriva em zero ou mais passos". Por exemplo, $N \Rightarrow_G^* 1.L$ é pronunciado "N deriva em zero ou mais passos $1.L$".

Exemplo 2.21 Seja $G_1 = (V_1, \Sigma_1, P_1, S)$, com:

$$
\begin{aligned}
V_1 &= \{0, 1, 2, 3, S, A\} \\
\Sigma_1 &= \{0, 1, 2, 3\} \\
N_1 &= \{S, A\} \\
P_1 &= \{S \rightarrow 0S33, S \rightarrow A, A \rightarrow 12, A \rightarrow \epsilon\}
\end{aligned}
$$

É fácil verificar que G_1 está formulada de acordo com as regras gerais acima enunciadas para a especificação de gramáticas. □

Denomina-se **forma sentencial** qualquer cadeia obtida pela aplicação recorrente das seguintes regras de substituição:

1. S (a raiz da gramática) é por definição uma forma sentencial;

2. Seja $\alpha\rho\beta$ uma forma sentencial, com α e β cadeias quaisquer de terminais e/ou não-terminais da gramática, e seja $\rho \rightarrow \gamma$ uma produção da gramática. Nessas condições, a aplicação dessa produção à forma sentencial, substituindo a ocorrência de ρ por γ, produz uma nova forma sentencial $\alpha\gamma\beta$.

Denota-se a substituição acima definida, também conhecida como **derivação direta**, por:

$$\alpha\rho\beta \Rightarrow_G \alpha\gamma\beta$$

O índice "G" designa o fato de que a produção aplicada, no caso $\rho \rightarrow \gamma$, pertence ao conjunto de produções que define a gramática G. Nos casos em que a gramática em questão puder ser facilmente identificada, admite-se a eliminação de referências explícitas a ela. Note-se a distinção gráfica e de significado que se faz entre o símbolo empregado na denotação das produções da gramática ($\rightarrow$) e o símbolo utilizado na denotação das derivações ($\Rightarrow$).

Considere a gramática $G = (V, T, P, S)$ onde:

- $V = \{S, B\}$
- $T = \{a\}$
- $P = \{S \rightarrow aB, B \rightarrow \epsilon\}$

Esta gramática gera uma única sentença, a saber, a sentença "a". A sentença pode ser obtida pela derivação: $S \Rightarrow aB \Rightarrow a$. Observe que no segundo passo foi aplicada a regra $B \rightarrow \epsilon$. Esta regra indica que B pode ser substituído pela cadeia vazia. Essa gramática é bastante simples e gostaríamos de descrevê-la de forma mais simples do que a apresentada acima. Podemos fazer isso adotando algumas

do que a apresentada acima. Podemos fazer isso adotando algumas convenções. Convencionamos que as variáveis serão sempre letras maiúsculas e que os demais símbolos serão terminais. Além disso, o símbolo inicial é o lado esquerdo da primeira regra apresentada. De posse desta convenção, podemos apresentar a mesma gramática de uma forma muito mais concisa. Poderíamos definir a gramática anterior em duas linhas, simplesmente como:

- $S \rightarrow aB,$
- $B \rightarrow \epsilon$

Podemos agora definir a linguagem gerada pela gramática $G = (V, T, P, S)$ como o conjunto de todas as sentenças $\alpha \in T^*$ tais que $S \Rightarrow_G^* \alpha$. Escrevemos esta linguagem como $L(G)$.

Dada uma gramática G é uma tupla (V, T, P, S), dizemos que a linguagem gerada por G é:

- $L(G) = \{\omega \in T^* \mid S \Rightarrow_G^* \omega\}$.

Análise Sintática

Uma vez que definimos o que é gramática e o que é a linguagem gerada por esta gramática, podemos nos perguntar: dada uma gramática G e a uma sentença ω, como determinar se $\omega \in L(G)$? Este problema é o problema da análise sintática. Podemos dizer que todo o conteúdo

Regras Livres de Contexto

Você deve ter percebido que as regras de uma gramática podem ter diferentes formatos. Por exemplo, considere o formato onde do lado esquerdo só há uma variável, ou seja, da forma $N \rightarrow \gamma$. Tais regras indicam que qualquer ocorrência de N na palavra que está sendo gerada pode ser substituída por γ. Ou seja, o N pode ser substituído por γ em qualquer posição que ele ocorra. Regras com este formato são chamadas de regras livres de contexto, isto é, a regra indica que se pode substituir N por γ independentemente do contexto em que N aparece.

Regras Livres de Contexto

A regra $A \rightarrow aA$ pode ser aplicada na forma sentencial (cadeia de símbolos terminais e não terminais) $ABACAB$ em cada uma das 3 posições do símbolo não terminal A. Para percebermos a diferença das

Regras Livres de Contexto

posições do símbolo não terminal A. Para percebermos a diferença das regras livres de contexto, vamos contrastar com a regra $AB \rightarrow aAB$. Esta regra, em função do B no lado esquerdo, só pode ser aplicada à primeira e à terceira ocorrência de A em $ABACAB$, pois a segunda ocorrência não tem um B justaposto à direita. De fato, a aplicação se dá sobre o AB e só há duas ocorrências de AB na palavra $ABACAB$. Portanto, ao aplicarmos esta regra, que não é livre de contexto, à primeira ocorrência de AB na palavra $ABACAB$ temos $aABACAB$ e caso apliquemos à segunda ocorrência de AB teremos $ABACaAB$. Note

Regras Sensíveis ao Contexto

indefinidamente. A este tipo de regra, que possui mais de um símbolo à esquerda da seta, denominamos regras sensíveis ao contexto, ou dependentes de contexto, sempre que o lado direito tiver mais ou tantos símbolos que o lado esquerdo. Caso contrário, ou seja, se o lado direito possuir menos símbolos que o esquerdo, temos o tipo mais geral de regra. Ou seja, uma regra da forma $\alpha \rightarrow \beta$ é sensível ao contexto se, e somente se, $|\alpha| \leq |\beta|$. Por hora, basta entendermos que isto contempla regras da forma $\alpha_1 A \alpha_2 \rightarrow \alpha_1 \gamma \alpha_2$, com γ não sendo a palavra vazia. Veja, esta última regra basicamente diz que A é γ no contexto de α_1 à esquerda e α_2 à direita.

Regras Sensíveis ao Contexto

Existem casos particulares de regras livres de contexto que são importantes por serem mais simples e ainda permitem a geração de linguagens infinitas. Regras nas formas $A \rightarrow aB$, $A \rightarrow \epsilon$ ou $A \rightarrow b$ são chamadas de regulares. Entende-se que A e B representam duas variáveis quaisquer (podendo ser a mesma), enquanto a e b representam dois terminais quaisquer.

Notação

- $A \rightarrow aA$
- $A \rightarrow b$

Podemos ver que $A \Rightarrow aA$, portanto $A \Rightarrow^{*} aaA$, $A \Rightarrow^{*} aaaA$ etc. Usando a notação a^k para indicar uma sequência de k a´s, teremos que $A \Rightarrow^{*} a^k A$, com $k > 0$ e, portanto, a linguagem gerada por A é formada pelas palavras $a^k b$, com $0 < k$, pois uma vez aplicada a regra $A \rightarrow b$ não teremos mais variáveis na palavra. Existe, entretanto, um fato bastante relevante nas derivações de $a^k b$: A cadeia sendo gerada só possui uma ocorrência de variável, e esta está sempre no final da cadeia. Dada uma

Regras

- Regra regular: $A \rightarrow b$, $A \rightarrow \epsilon$ ou $A \rightarrow aB$, com $A, B \in V$ e $a, b \in T$;
- Regra livre de contexto: $A \rightarrow \gamma$, com $A \in V$ e $\gamma \in (V \cup T)^*$;
- Regra sensível ao contexto: $\alpha \rightarrow \beta$, com $\alpha, \beta \in (V \cup T)^*$ e $|\alpha| \leq |\beta|$; ou $S \rightarrow \epsilon$, se S não aparece do lado direito de nenhuma regra;
- Regra geral: Não possui restrição além daquela na definição de gramática, que obriga a existir ao menos uma variável no lado esquerdo da regra.

Regras

Das denominações acima podemos constatar que toda regra regular é livre de contexto, toda livre de contexto cujo lado direito é não vazio é sensível ao contexto e toda sensível ao contexto é geral (GARCIA, 2017). Por outro lado, existem regras gerais que não são sensíveis ao contexto, regras sensíveis ao contexto que não são livres de contexto e finalmente regras livres de contexto que não são regulares. Temos uma hierarquia de regras. Vejamos como esta hierarquia de regras dá origem a uma hierarquia de linguagens.

Linguagem

Uma gramática é dita ser de certo tipo (regular, livre de contexto, sensível ao contexto, geral), se, e somente se, todas as suas regras são daquele tipo. Uma linguagem é de certo tipo Ti, se, e somente se, existe uma gramática do tipo Ti que gera a linguagem. Dizer então que uma linguagem não é de certo tipo é equivalente a dizer que não existe gramática daquele tipo capaz de gerar a linguagem.

Linguagem Regular

Seja a gramática G_1:

- $A \rightarrow aA$
- $A \rightarrow b$

G_1 só possui regras regulares, portanto é uma gramática regular. A linguagem $L = \{a^k b, k > 0\}$ é gerada pela gramática regular G_1, portanto L é regular. Seja a gramática G_2:

Linguagem Regular

- $A \rightarrow aaA$
- $A \rightarrow aA$
- $A \rightarrow b$

G_2 possui uma regra que não é regular, portanto a gramática G_2 não é regular. Entretanto, a linguagem $L(G_2)$ é regular, pois $L(G_2) = L(G_1)$, e basta que exista uma gramática regular que a gere para que a linguagem seja regular.

Hierarquia Chomsky

A hierarquia de Chomsky é justamente a afirmação que as linguagens formam uma hierarquia a partir dos tipos das gramáticas que são capazes de gerá-las. Cada tipo gramatical também nos indicará o mecanismo/ autômato capaz de resolver o problema da análise sintática para aquele tipo de gramática. Na sua forma menos detalhada, a hierarquia de Chomsky é mostrada de acordo com a Tabela 1.1.

Tipos de Gramáticas	Regras	Exemplos de Linguagens geradas por estas gramáticas
Regulares	$A \rightarrow b$, $A \rightarrow \epsilon$ ou $A \rightarrow aB$, $A, B \in V^*$, $a, b \in T^*$	$a^n b^m, n, m > 0$

Hierarquia Chomsky

Livres de Contexto	$A \rightarrow \gamma$ $\gamma \in (V \cup T)^*$	$a^n b^n, n > 0$
Sensíveis ao Contexto	$\alpha \rightarrow \beta$, com $\alpha, \beta \in (V \cup T)^*$ e $\lvert \alpha \rvert \leq \lvert \beta \rvert$; ou $S \rightarrow \epsilon$, se S não aparece do lado direito de nenhuma regra;	$a^n b^n c^n, n > 0$

Hierarquia Chomsky

Irrestrita ou geral	$\alpha \rightarrow \beta$, com $\alpha, \beta \in (V \cup T)^*$ α tem pelo menos símbolo de V.	----------

Binário

Você aprendeu que gramáticas são formalismos/ferramentas capazes de gerar, ou descrever, linguagens formais. Observamos também que gramáticas são objetos formais finitos. De um ponto de vista bem básico, uma gramática é uma palavra em um alfabeto que inclui o símbolo $\rightarrow$ e o alfabeto da própria linguagem que ela gera. Para simplificar nossa discussão, vamos considerar somente gramáticas que geram palavras sobre o alfabeto $\{0,1\}$.

Bin

Podemos estabelecer uma notação para as variáveis das gramáticas e com isso fixamos o alfabeto para escrever gramáticas em um conjunto fixo de símbolos. Por exemplos se temos n variáveis na gramática, usamos as variáveis com nomes $<V1>$, $<V11>$, $<V111>$, ..., $<V1^n>$. Vamos usar a ausência de símbolos para o ϵ. Observe o uso dos $<>$ para delimitar a notação de variáveis e não misturar o símbolo 1 do alfabeto da linguagem descrita/gerada pela gramática com a notação para variáveis. Só para exemplificar, uma gramática como:

$$S \rightarrow 1S0, S \rightarrow \epsilon$$

é representada pela cadeia $<V1>;<V1> \rightarrow 1<V1>0;<V1> \rightarrow$. O primeiro $<V1>$ indica que ele é o símbolo inicial da gramática. Em seguida vêm as regras,

Binário

Você se convenceu que qualquer gramática que gera uma linguagem sobre $\{0,1\}$ pode ser escrita nesta forma? Demonstre que sim descrevendo a palavra associada a uma gramática tal como $S; S \to ASBC;\ S \to; AB \to BA;\ BC \to CB;\ AC \to CA;$ $A \to 10;\ B \to 00;\ C \to 11$, com variáveis A, B e C, no formato recém-discutido.

Binário

Veja que a gramática que pedimos para codificar:

$$S; S \to ASBC; \quad S \to; AB \to BA; \quad BC \to CB; \quad AC \to CA;$$
$$A \to 10;\ B \to 00;\ C \to 11.$$

Pode ser codificada em $\{<, > V, 1, 0, \to, ;\}$ como:

$$< V1 > ; < V1 > \to < V11 >< V1 >< V111 >< V1111 >;$$
$$< V1 > \to; < V11 >< V111 > \to < V111 >< V11 >;$$
$$< V111 >< V1111 > \to < V1111 >< V111 >;$$
$$< V11 >< V1111 > \to < V1111 >< V11 >;$$
$$< V11 > \to 10; < V111 > \to 00; < V1111 > \to 11$$