Văn phạm phi ngữ cảnh

Trần Vĩnh Đức

Trường Đại Học Bách Khoa Hà Nội

Ngày 12 tháng 3 năm 2019

Thuật ngữ

- Context-Free Language (CFL) : Ngôn ngữ phi ngữ cảnh
- Context-Free Grammar (CFG): Văn phạm phi ngữ cảnh

Nội dung

Nội dung

Giới thiệu

- Ta đã xem xét hai phương pháp khác nhau, nhưng tương đương, để mô tả ngôn ngữ: Các otomat hữu hạn và các biểu thức chính quy.
- Ta đã biết rằng có những ngôn ngữ đơn giản nhưng không thể mô tả theo cách trên. Ví dụ ngôn ngữ

$$B = \{0^n 1^n \mid n \geq 0\}.$$

- Trong chương này ta xem xét một phương pháp mạnh hơn để mô tả ngôn ngữ: Văn phạm phi ngữ cảnh (viết tắt là CFG).

Văn phạm phi ngữ cảnh

- Được sử dụng đầu tiên trong nghiên cứu ngôn ngữ tự nhiên. Nó giúp hiểu quan hệ giữa các khái niệm *câu*, *động từ*, và *mệnh đề*.
- Có thể mô tả một số cấu trúc đệ quy. Điều này rất có ích trong nhiều ứng dụng, đặc biệt trong đặc tả và biên dịch các ngôn ngữ lập trình.
- Lớp ngôn ngữ gắn với CFG gọi là ngôn ngữ phi ngữ cảnh. Chúng chứa thực sự lớp ngôn ngữ chính quy.

Ví dụ

Dưới đây là một CFG

$$A \to 0A1$$
$$A \to B$$
$$B \to \#$$

- bao gồm tập *quy tắc thay thế*, hay còn gọi là *sản xuất*.
- A, B gọi là *biến*
- $0, 1, \#$ gọi là ký hiệu *kết thúc*
- Có một biến *bắt đầu*, ví dụ A.

Mô tả ngôn ngữ bằng CFG

CFG mô tả ngôn ngữ bằng cách sinh ra mỗi xâu theo cách sau đây:

1. Viết ra biến bắt đầu. Thường là biến bên trái của quy tắc đầu tiên.
2. Tìm một biến trong xâu vừa viết và chọn một quy tắc có vế trái là biến đó. Thay thế biến đó bởi vế phải của quy tắc đã chọn.
3. Lặp lại bước **2.** cho đến khi không còn biến nào trong xâu đã viết.

Dãy phép thay thế để đạt được xâu gọi là *dẫn xuất*.

Ví dụ

Xét CFG G_1 dưới đây

$$A \to 0A1$$
$$A \to B$$
$$B \to \#$$

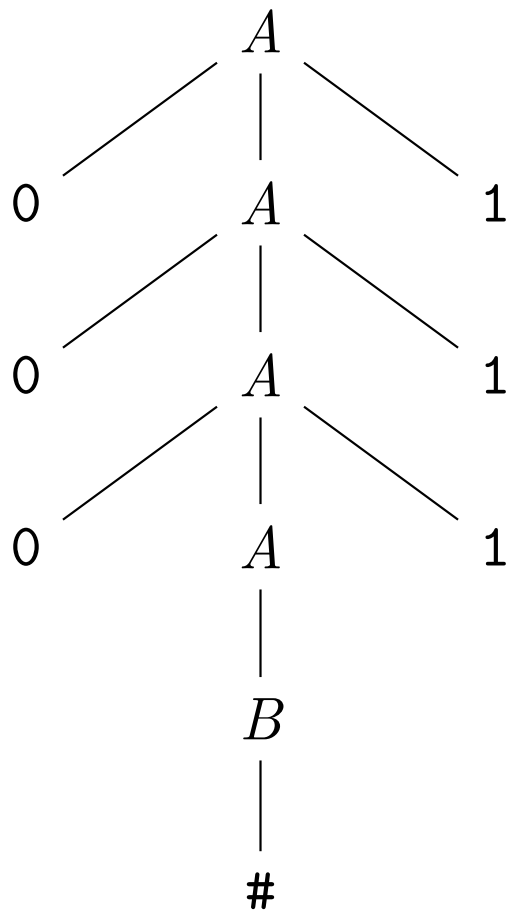
- Một dẫn xuất của xâu $000\#111$ trong G_1 là

$$A \Rightarrow 0A1 \Rightarrow 00A11 \Rightarrow 000A111 \Rightarrow 000B111 \Rightarrow 000\#111$$

- Mọi xâu sinh ra theo cách này gọi là *ngôn ngữ của văn phạm* G_1, ký hiệu $L(G_1)$.
- Ta có thể kiểm tra

$$L(G_1) = \{0^n\#1^n \mid n \geq 0\}$$

Cây dẫn xuất



Hình: Dẫn xuất

$A \Rightarrow 0A1 \Rightarrow 00A11 \Rightarrow 000A111 \Rightarrow 000B111 \Rightarrow 000\#111$

Mô tả ngôn ngữ Tiếng Việt

- Một câu Tiếng Việt viết hoặc nói ra chỉ gồm các ký tự thuộc bộ chữ cái tiếng Việt, nhưng quá trình viết hoặc nói được hướng dẫn bởi các quy tắc ngữ pháp
- Hướng dẫn về cấu trúc đơn giản của một câu thường có dạng

$$\langle \text{câu} \rangle \rightarrow \langle \text{chủ ngữ} \rangle \langle \text{vị ngữ} \rangle$$

  nghĩa là một câu đúng có thể bắt đầu bằng một chủ ngữ, tiếp theo là vị ngữ.
- về chủ ngữ

$$\langle \text{chủ ngữ} \rangle \rightarrow \langle \text{đại từ} \rangle \mid \langle \text{danh từ} \rangle \mid \langle \text{câu} \rangle$$

  nghĩa là chủ ngữ có thể là đại từ hoặc danh từ hoặc câu.

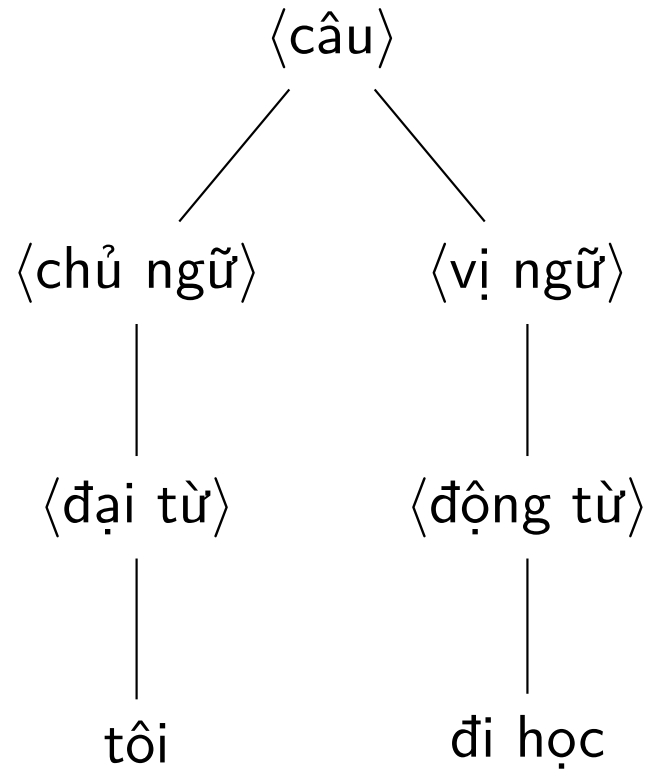
Văn phạm Tiếng Việt giản lược

Ví dụ

Văn phạm sau đây hướng dẫn cách viết đúng nhiều câu Tiếng Việt đơn giản

$$\begin{aligned}
\langle \text{câu} \rangle &\to \langle \text{chủ ngữ} \rangle \langle \text{vị ngữ} \rangle \\
\langle \text{chủ ngữ} \rangle &\to \langle \text{đại từ} \rangle \mid \langle \text{danh từ} \rangle \mid \langle \text{câu} \rangle \\
\langle \text{vị ngữ} \rangle &\to \langle \text{động từ} \rangle \mid \langle \text{tính từ} \rangle \\
\langle \text{đại từ} \rangle &\to \text{tôi} \mid \text{anh} \mid \text{nó} \mid \ldots \\
\langle \text{động từ} \rangle &\to \text{đi học} \mid \text{đi chơi} \mid \ldots
\end{aligned}$$

Hãy đưa ra một cây dẫn xuất cho câu "tôi đi học".



Hình: Cây dẫn xuất cho câu “tôi đi học”.

Nội dung

Định nghĩa

Một Văn phạm phi ngữ cảnh là bộ bốn (V, Σ, R, S) trong đó

1. *V là một tập hữu hạn, mỗi phần tử thuộc nó được gọi là biến,*
2. *Σ là một tập hữu hạn, phân biệt với V, mỗi phần tử thuộc nó được gọi là ký hiệu kết thúc,*
3. *R là một tập hữu hạn, mỗi phần tử của R gọi là một quy tắc có dạng $A \to u$ trong đó $A \in V$ và u là một xâu trên $V \cup \Sigma$, và*
4. *$S \in V$ là biến bắt đầu.*

- Ta đã xét CFG $G_1 = (\{A, B\}, \{0, 1, \#\}, R, A)$ với tập quy tắc

$$R = \{A \to 0A1,\ A \to B,\ B \to \#\}.$$

Ngôn ngữ của văn phạm

- Xét u, v, và w là xâu trên $V \cup \Sigma$, và $A \to w$ là một quy tắc của văn phạm. Khi đó ta nói uAv *dẫn trực tiếp* uwv và viết $uAv \Rightarrow uwv$.
- Ta nói rằng u *dẫn* ra v và viết

$$u \Rightarrow^* v$$

nếu $u = v$ hoặc nếu tồn tại một dãy $u_1, u_2, \ldots, u_k$, với $k \geq 0$ và

$$u \Rightarrow u_1 \Rightarrow u_2 \Rightarrow \cdots \Rightarrow u_k = v.$$

- *Ngôn ngữ của văn phạm* G là

$$L(G) = \{w \in \Sigma^* \mid S \Rightarrow^* w\}.$$

Nội dung

Ví dụ

Xét văn phạm $G_3 = (\{S\}, \{\mathtt{a}, \mathtt{b}\}, R, S\})$. Tập quy tắc R là

$$S \to \mathtt{a}S\mathtt{b} \mid SS \mid \varepsilon.$$

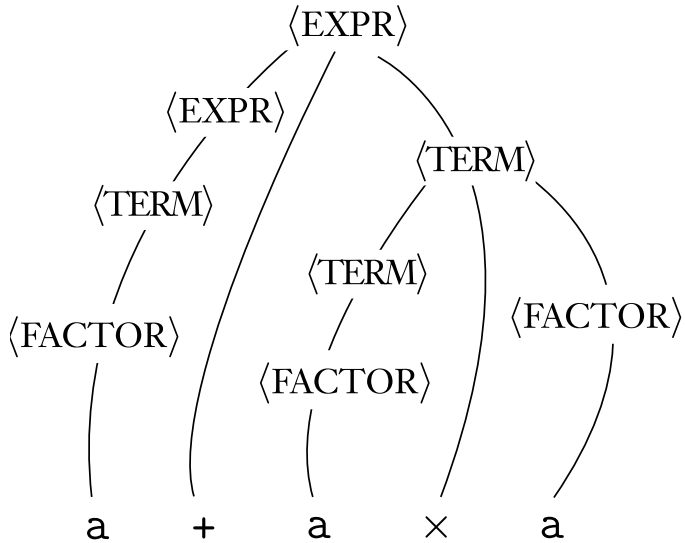
- Văn phạm này sinh ra xâu như $\mathtt{abab}, \mathtt{aaabbb}$, và $\mathtt{aababb}$.
- Xem $\mathtt{a}$ như dấu ngoặc trái "(" và $\mathtt{b}$ như dấu ngoặc phải ")".
- $L(G_3)$ là ngôn ngữ gồm mọi xâu "cân bằng ngoặc".

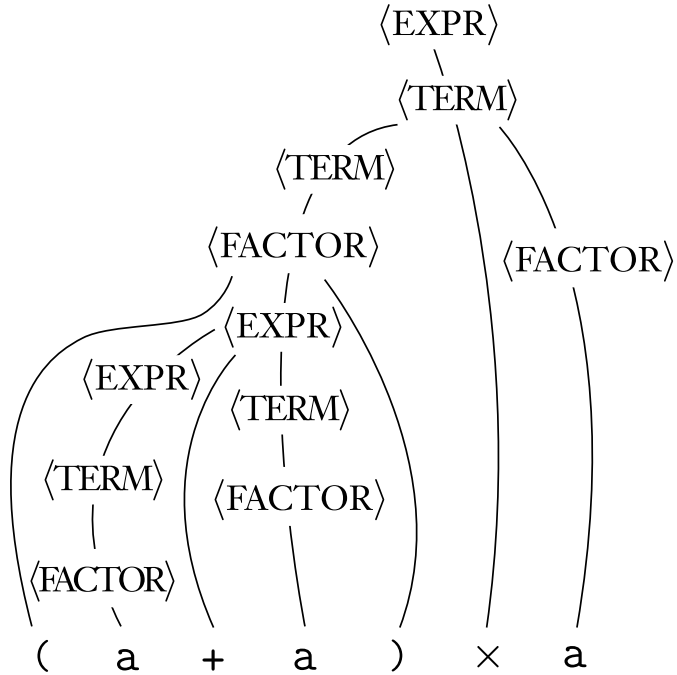
Ví dụ

Xét văn phạm $G_4 = (V, \Sigma, R, \langle \texttt{EXPR} \rangle)$ trong đó

- $V = \{\langle \texttt{EXPR} \rangle, \langle \texttt{TERM} \rangle, \langle \texttt{FACTOR} \rangle\}$
- $\Sigma = \{\texttt{a}, +, \times, (,)\}$
- Các quy tắc là

$$\begin{aligned} \langle \texttt{EXPR} \rangle &\to \langle \texttt{EXPR} \rangle + \langle \texttt{TERM} \rangle \mid \langle \texttt{TERM} \rangle \\ \langle \texttt{TERM} \rangle &\to \langle \texttt{TERM} \rangle \times \langle \texttt{FACTOR} \rangle \mid \langle \texttt{FACTOR} \rangle \\ \langle \texttt{FACTOR} \rangle &\to (\langle \texttt{EXPR} \rangle) \mid \texttt{a} \end{aligned}$$


⟨EXPR⟩
⟨EXPR⟩
⟨TERM⟩
⟨FACTOR⟩
a
+
⟨TERM⟩
⟨TERM⟩
⟨FACTOR⟩
a
×
⟨FACTOR⟩
a


⟨EXPR⟩
⟨TERM⟩
⟨TERM⟩
⟨FACTOR⟩
⟨FACTOR⟩
⟨EXPR⟩
⟨EXPR⟩
⟨TERM⟩
⟨TERM⟩
⟨FACTOR⟩
⟨FACTOR⟩
(
a
+
a
)
×
a

Nội dung

Hợp của nhiều văn phạm

Ví dụ

Xây dựng CFG cho ngôn ngữ $\{0^n1^n \mid n \geq 0\} \cup \{1^n0^n \mid n \geq 0\}$.

- Xây dựng CFG

$$S_1 \to 0S_1 1 \mid \varepsilon$$

cho ngôn ngữ $\{0^n1^n \mid n \geq 0\}$ và CFG

$$S_2 \to 1S_2 0 \mid \varepsilon$$

cho ngôn ngữ $\{1^n0^n \mid n \geq 0\}$.

- Kết hợp lại ta được CFG

$$S \to S_1 \mid S_2$$
$$S_1 \to 0S_1 1 \mid \varepsilon$$
$$S_2 \to 1S_2 0 \mid \varepsilon$$

Với ngôn ngữ chính quy

Ta xây dựng DFA đoán nhận ngôn ngữ đó. Sau đó xây dựng CFG tương đương như sau:

- Các biến R_i tương ứng với các trạng thái q_i của DFA
- Thêm quy tắc

$$R_i \to aR_j$$

  nếu $\delta(q_i, a) = q_j$.
- Thêm quy tắc

$$R_i \to \varepsilon$$

  nếu q_i là trạng thái chấp nhận của DFA.
- Đặt R_0 là trạng thái bắt đầu, với q_0 là trạng thái bắt đầu của DFA.

Nội dung

Sự nhập nhằng

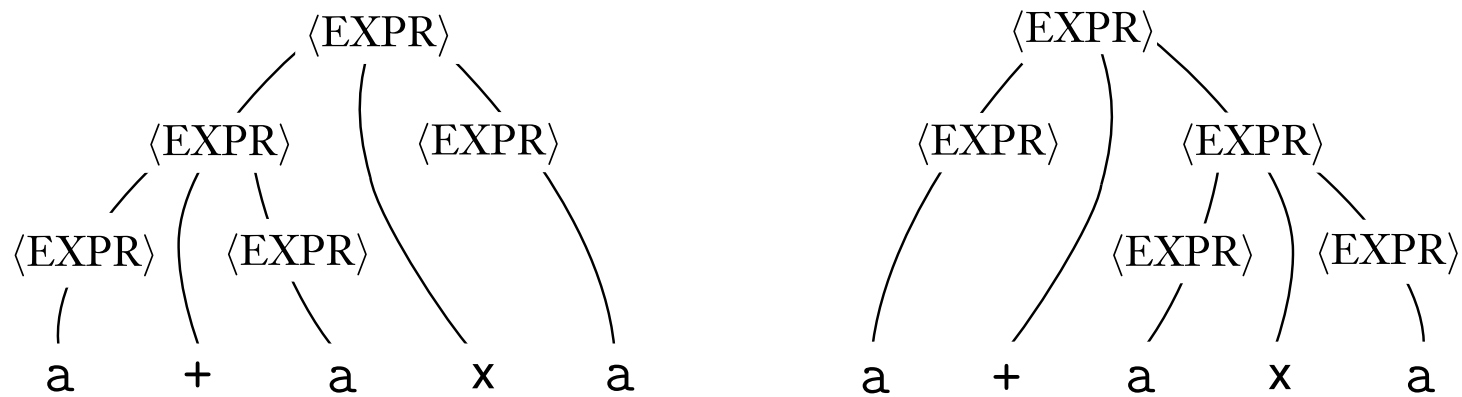
- Một CFG có thể sinh cùng một xâu theo nhiều cách khác nhau.
- Xâu này có nhiều cây dẫn xuất, nên nó có nhiều nghĩa.
- Điều này không mong muốn trong nhiều ứng dụng, như trong các ngôn ngữ lập trình: Các chương trình phải được diễn dịch một cách duy nhất.

Ví dụ

- Xét văn phạm G_5

$$\langle \texttt{EXPR} \rangle \rightarrow \langle \texttt{EXPR} \rangle + \langle \texttt{EXPR} \rangle \mid \langle \texttt{EXPR} \rangle \times \langle \texttt{EXPR} \rangle \mid (\langle \texttt{EXPR} \rangle) \mid \texttt{a}$$

- Văn phạm này sinh ra xâu $\texttt{a} + \texttt{a} \times \texttt{a}$ một cách nhập nhằng.



- Văn phạm không nhập nhằng G_4 có $L(G_4) = L(G_5)$.

Định nghĩa văn phạm nhập nhằng

- Hai dẫn xuất có thể khác nhau chỉ do thứ tự thay thế các biến trong xâu. Ta sẽ chỉ quan tâm đến dẫn xuất trái nhất.
- Một dẫn xuất là *trái nhất* nếu tại mọi bước ta luôn thay thế biến còn lại bên trái nhất.

Định nghĩa

- *Một xâu w được dẫn xuất một cách nhập nhằng trong CFG G nếu nó có ít nhất hai dẫn xuất trái nhất.*
- *Văn phạm G là nhập nhằng nếu nó sinh một cách nhập nhằng một xâu nào đó.*

Nhập nhằng cố hữu

- Đôi khi, khi cho một văn phạm nhập nhằng ta có thể tìm một văn phạm không nhập nhằng sinh ra cùng ngôn ngữ.
- Tuy vậy, có những CFL chỉ sinh bởi những văn phạm nhập nhằng. Những ngôn ngữ như vậy gọi là *nhập nhằng cố hữu*.
- Ví dụ, ngôn ngữ

$$\{\mathsf{a}^i\mathsf{b}^j\mathsf{c}^k \mid i = j \text{ hoặc } j = k\}$$

là nhập nhằng cố hữu.

Nội dung

Dạng chuẩn Chomsky

Định nghĩa

Một CFG là ở dạng chuẩn Chomsky nếu mọi quy tắc của nó có dạng

$$A \to BC$$
$$A \to a$$

trong đó a là một ký hiệu kết thúc và A, B, và C là các biến–nhưng B và C không là biến bắt đầu. Thêm nữa, ta cho phép quy tắc

$$S \to \varepsilon$$

nếu S là biến bắt đầu.

Đưa một CFG về dạng chuẩn Chomsky

Định lý

Mỗi ngôn ngữ phi ngữ cảnh đều sinh bởi một văn phạm phi ngữ cảnh ở dạng chuẩn Chomsky.

Biến đổi CFG theo bốn bước sau:

1. Thêm biến bắt đầu mới để biến bắt đầu không xuất hiện ở vế phải của bất cứ quy tắc nào.
2. Loại bỏ các ε-quy tắc.
3. Loại bỏ các quy tắc đơn có dạng $A \to B$.
4. Chuyển các quy tắc còn lại về dạng đúng.

Bước 1: Xử lý biến bắt đầu

- Để biến bắt đầu S không xuất hiện vế phải của quy tắc nào.
- Ta thêm biến mới S_0 và quy tắc

$$S_0 \to S,$$

với S là biến khởi đầu.

Bước 2: Loại bỏ ε-quy tắc

Với mỗi quy tắc dạng $A \to \varepsilon$, ta thực hiện

- Loại bỏ quy tắc $A \to \varepsilon$
- Nếu có quy tắc dạng $R \to uAv$, thì ta thêm quy tắc $R \to uv$. Ta làm bước này với mọi xuất hiện của A. Có nghĩa rằng nếu có quy tắc dạng $R \to uAvAw$ thì ta phải thêm các quy tắc

$$R \to uAvw \mid uvAw \mid uvw$$

- Nếu có quy tắc $R \to A$, ta thêm quy tắc $R \to \varepsilon$ trừ khi ở các bước trước ta đã loại bỏ quy tắc $R \to \varepsilon$.

Bước 3: Loại bỏ quy tắc đơn

Với mỗi quy tắc đơn dạng $A \to B$, ta thực hiện

- Loại bỏ quy tắc $A \to B$.
- Với mỗi quy tắc dạng $B \to u$, ta thêm quy tắc

$$A \to u$$

trừ khi đây là quy tắc đơn đã bị loại bỏ ở các bước trước đó.

Bước 4: Chuyển các quy tắc còn lại về dạng đúng

- Thay thế các quy tắc dạng

$$A \to u_1 u_2 \dots u_k, \qquad \text{với} \quad k \geq 3 \text{ và } u_i \in V \cup \Sigma$$

bởi các quy tắc

$$A \to u_1 A_1,\ A_1 \to u_2 A_2,\ \dots, A_{k-2} \to u_{k-1} u_k.$$

- Thay thế mọi ký hiệu kết thúc u_i trong các quy tắc vừa thay thế ở trên bởi biến mới U_i và thêm quy tắc $U_i \to u_i$.

Ví dụ

1. Đảm bảo biến bắt đầu của văn phạm bên trái dưới đây không xuất hiện ở vế phải của quy tắc nào.

$$S \to ASA \mid \mathtt{a}B$$
$$A \to B \mid S$$
$$B \to \mathtt{b} \mid \varepsilon$$

$$S_0 \to S$$
$$S \to ASA \mid \mathtt{a}B$$
$$A \to B \mid S$$
$$B \to \mathtt{b} \mid \varepsilon$$

2. Loại bỏ các ε-quy tắc: Loại bỏ $B \to \varepsilon$ sau đó loại bỏ $A \to \varepsilon$

$$S_0 \to S$$
$$S \to ASA \mid \mathtt{a}B \mid \mathtt{a}$$
$$A \to B \mid S \mid \varepsilon$$
$$B \to \mathtt{b}$$

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

3a. Loại bỏ các quy tắc đơn: Loại bỏ $S \to S$ và sau đó là $S_0 \to S$

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

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

3b. Loại bỏ quy tắc $A \to B$ và sau đó là $A \to S$

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

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

3. Sau khi đã loại bỏ hết quy tắc đơn, ta được

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

4. Đưa các quy tắc còn lại về dạng đúng

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

Nội dung

Bài toán

Cho CFG G và một xâu w. Kiểm tra xem G có sinh ra xâu w.

Ý tưởng thuật toán

- Xét G là văn phạm ở dạng chuẩn Chomsky:

$$A \to a \qquad \text{hoặc} \qquad A \to BC.$$

- Thuật toán dựa trên kỹ thuật quy hoạch động.
- Xét xâu vào

$$w = w_1 w_2 \cdots w_n.$$

- Ta xây dựng bảng $n \times n$: Phần tử (i, j) (với $i \leq j$) của bảng là tập các biến sinh ra xâu

$$w_i w_{i+1} \cdots w_j.$$

Ví dụ

Xâu $w =$ baba có sinh bởi văn phạm sau không?

$$\begin{aligned} S &\to RT \\ R &\to TR \mid \mathtt{a} \\ T &\to TR \mid \mathtt{b} \end{aligned}$$

	1	2	3	4
1	T	R, T	S	S, R, T
2		R	S	S
3			T	R, T
3				R

Với xâu vào $w = w_1 \cdots w_n$:

1. Nếu $w = \varepsilon$ và có luật $S \to \varepsilon$, chấp nhận. ⟦ xử lý ε ⟧
2. For $i = 1$ đến n: ⟦ xử lý xâu con độ dài 1 ⟧
3. For mỗi biến A:
4. Kiểm tra xem có luật $A \to$ b với $w_i =$ b.
5. Nếu có, đặt A vào $table(i, i)$.
6. For $l = 2$ đến n: ⟦ l là độ dài xâu con ⟧
7. For $i = 1$ đến $n - l + 1$: ⟦ i là vị trí bắt đầu ⟧
8. Đặt $j = i + l - 1$. ⟦ j là vị trí kết thúc ⟧
9. For $k = i$ đến $j - 1$: ⟦ k là vị trí cắt xâu ⟧
10. For mỗi luật $A \to BC$:
11. Nếu $B \in table(i, k)$ và $C \in table(k + 1, j)$, thì đặt A vào $table(i, j)$.
12. Nếu $S \in table(1, n)$, chấp nhận. Ngược lại, bác bỏ.