Théorie des langages et automates

Grammaires formelles – 1

Clément Moreau[1,2]

[1]BRED – Banque Populaire — clement.moreau@bred.fr

[2]Université de Tours

Sommaire

Table of Contents

Grammaire formelle

Dans l'étude des langages formels ou naturels, une *grammaire* est un ensemble de règles qui dicte la nature des phrases ou mots construits.

Grammaire formelle

Dans l'étude des langages formels ou naturels, une *grammaire* est un ensemble de règles qui dicte la nature des phrases ou mots construits.

Par exemple, considérons l'ensemble de règles suivants :

$$
\begin{array}{lcl}
\langle \mathit{phrase} \rangle & \rightarrow & \langle \mathit{sujet} \rangle\ \langle \mathit{predicat} \rangle \\
\langle \mathit{sujet} \rangle & \rightarrow & \text{le chien} \mid \text{le chat} \mid \text{le bébé} \\
\langle \mathit{predicat} \rangle & \rightarrow & \langle \mathit{verbe} \rangle\ \langle \mathit{adverbe} \rangle \mid \langle \mathit{verbe} \rangle\ \langle \mathit{sujet} \rangle \\
\langle \mathit{verbe} \rangle & \rightarrow & \text{mange} \\
\langle \mathit{adverbe} \rangle & \rightarrow & \text{sagement} \mid \text{proprement} \mid \varepsilon
\end{array}
$$

Grammaire formelle

Dans l'étude des langages formels ou naturels, une *grammaire* est un ensemble de règles qui dicte la nature des phrases ou mots construits.

Par exemple, considérons l'ensemble de règles suivants :

$$
\begin{array}{lcl}
\langle \textit{phrase} \rangle & \rightarrow & \langle \textit{sujet} \rangle \ \langle \textit{predicat} \rangle \\
\langle \textit{sujet} \rangle & \rightarrow & \text{le chien} \mid \text{le chat} \mid \text{le bébé} \\
\langle \textit{predicat} \rangle & \rightarrow & \langle \textit{verbe} \rangle \ \langle \textit{adverbe} \rangle \mid \langle \textit{verbe} \rangle \ \langle \textit{sujet} \rangle \\
\langle \textit{verbe} \rangle & \rightarrow & \text{mange} \\
\langle \textit{adverbe} \rangle & \rightarrow & \text{sagement} \mid \text{proprement} \mid \varepsilon
\end{array}
$$

On appelle ici *variable*, les termes entre $\langle\rangle$ et *symboles finaux* les mots en français et police droite.

Grammaire formelle

On peut générer une phrase à l'aide e notre grammaire de la façon suivante :

1. On part de la variable $\langle \textit{phrase} \rangle$.
2. On remplace successivement toutes les variables à droite de la règle jusqu'à obtenir des symboles finaux.

Exemples

$$
\begin{array}{rcl}
\langle \textit{phrase} \rangle & \Rightarrow & \langle \textit{sujet} \rangle\ \langle \textit{predicat} \rangle \\
 & \Rightarrow & \text{le chien } \langle \textit{predicat} \rangle \\
 & \Rightarrow & \text{le chien } \langle \textit{verbe} \rangle\ \langle \textit{adverbe} \rangle \\
 & \Rightarrow & \text{le chien mange } \langle \textit{adverbe} \rangle \\
 & \Rightarrow & \text{le chien mange sagement}
\end{array}
$$

Grammaire formelle

Une grammaire permet de décrire un langage de manière inductive.

Définition

Une *grammaire formelle* G est un quadruplet :

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

avec :

- V — un ensemble fini de variables (symboles non terminaux), chaque variable représente un ensemble de mots.
- Σ— un ensemble fini de symboles (un alphabet) tel que $V \cap \Sigma = \emptyset$ dont les éléments sont appelés les symboles terminaux.
- P — un ensemble fini de règles de production
- $S \in V$ — l'axiome, qui représente tous les mots décrits par la grammaire.

Grammaire formelle

Soient une grammaire $G = (V, \Sigma, P, S)$, des mots $\alpha, \beta \in (V \cup \Sigma)^*$ et une variable $A \in V$. S'il existe une règles de production $A \rightarrow \gamma$, alors

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

. Le symbole $\Rightarrow$ indique une règle de *dérivation* de la grammaire.

- On note $\Rightarrow^*$ la fermeture réflexive et transitive de $\Rightarrow$.
- On note aussi par $\alpha \Rightarrow^n \beta$ s'il existe $\alpha_0, ..., \alpha_n$ tels que $\alpha_0 = \alpha$ et $\alpha_n = \beta$ et $\alpha_i \Rightarrow \alpha_{i+1}$, pour $0 \leq i < n$.
- À chaque étape, il est possible de dériver la variable la plus à gauche ou de dériver la variable la plus à droite.

Langage d'une grammaire

Le langage d'une grammaire est défini tel que :

$$L(G) = \{u | u \in \Sigma^*, S \Rightarrow^* u\}$$

Grammaire formelle

Exemples

1. Soit la grammaire $G = (\{S\}, \{0,1\}, \{S \rightarrow \varepsilon | 0S1\}, S)$.
   Quelle est langage décrit par G ?
2. Donner la grammaire permettant de représenter le langage des palindromes sur l'alphabet $\{0,1\}$.

Grammaire formelle

Grammaire linéaire

Une grammaire G est *linéaire* si toutes les règles de production sont de la forme $A \rightarrow a$ $A \rightarrow aB$ (resp. $A \rightarrow Ba$) ou $S \rightarrow \varepsilon$ où $A, B \in V$ et $a \in \Sigma$.

- $\forall p \in P, p : A \rightarrow aB$, G est linéaire à droite $\Rightarrow$ G est régulière.
- $\forall p \in P, p : A \rightarrow Ba$, G est linéaire à gauche $\Rightarrow$ G est régulière.
- $\forall p \in P, p : A \rightarrow aBb$, G est linéaire $\Rightarrow$ G n'est pas forcément régulière.

Grammaire formelle

Grammaire linéaire

Une grammaire G est *linéaire* si toutes les règles de production sont de la forme $A \to a$ $A \to aB$ (resp. $A \to Ba$) ou $S \to \varepsilon$ où $A, B \in V$ et $a \in \Sigma$.

- $\forall p \in P, p : A \to aB$, G est linéaire à droite $\Rightarrow$ G est régulière.
- $\forall p \in P, p : A \to Ba$, G est linéaire à gauche $\Rightarrow$ G est régulière.
- $\forall p \in P, p : A \to aBb$, G est linéaire $\Rightarrow$ G n'est pas forcément régulière.

Exemples

Proposer une grammaire linéaire capable de générer les langages réguliers suivants du TD1 :

1. L'ensembles des mots qui finissent en 00.
2. L'ensemble des mots qui contiennent un nombre pair de sous-mots 01.
3. L'ensemble des mots sur {a} tel que le mot possède une longueur qui est un multiple de 3 ou de 5 (ex. aaa,aaaaa,aaaaaa)

Grammaire non-contextuelle

Définition

Une grammaire $G = (V, \Sigma, P, S)$ est dite *non-contextuelle* (ou hors-contexte) si toutes les règles de production sont de la forme $A \to \alpha$ où $A \in V$ et $\alpha \in \{V \cup \Sigma\}^*$

Théorème

Pour tout langage $L \in \mathbf{Reg}$, il existe une grammaire non-contextuelle G telle que $L = L(G)$.

Grammaire non-contextuelle

Démonstration.

Soit $L \in \mathbf{Reg}$, dès lors on sait qu'il existe un AFD $M = (Q, \Sigma, q_0, \delta, F)$ tel que $L(M) = L$.
Posons une grammaire $G = (V, \Sigma, P, S)$ telle que $V = Q$ et $S = q_0$ et

$$P = \{q \to ap | \delta(q, a) = p\} \cup \{f \to \varepsilon | f \in F\}$$

Par induction, on peut prouver que $\forall p, q \in Q$ et $\forall x \in \Sigma^*, \delta(q, x) = p$ si et seulement il existe une dérivation $q \Rightarrow^* xp \in G$.
En particulier, $x \in L$ si et seulement si il existe une dérivation de la forme

$$q_0 \Rightarrow^* xf \Rightarrow x$$

pour tout état $f \in F$. Ce qui montre que $L \subseteq L(G)$. De plus, comme G est linéaire à droite, les dérivations de la forme ci-dessus sont les seules de G. Dès lors, $L = L(G)$. □

Grammaire contextuelle

Définition

Une grammaire $G = (V, \Sigma, P, S)$ est dite *contextuelle* si toutes les règles de production sont de la forme $\phi A \psi \to \phi \alpha \psi$ où $A \in V$ et $\alpha, \phi, \psi \in (V \cup \Sigma)^*$ et $\alpha \neq \varepsilon$.

Grammaire contextuelle

Définition

Une grammaire $G = (V, \Sigma, P, S)$ est dite *contextuelle* si toutes les règles de production sont de la forme $\phi A \psi \to \phi \alpha \psi$ où $A \in V$ et $\alpha, \phi, \psi \in (V \cup \Sigma)^*$ et $\alpha \neq \varepsilon$.

Exemples

Soit le langage $L = \{a^n b^n c^n | n \geq 0\}$, engendré par la grammaire contextuelle :

$$\begin{array}{lcl} S & \to & \varepsilon | abc | aAbc \\ Ab & \to & bA \\ Ac & \to & Bbcc \\ bB & \to & Bb \\ aB & \to & aa | aaA \end{array}$$

Vérifier que le mot *aaabbbccc* est bien produit par la grammaire. On adoptera une dérivation à gauche.

Table of Contents

Problème des parenthèses

Le *problème des parenthèse* revient à déterminer une formalisation à partir d'une grammaire non-contextuelle du langage `PAREN` reconnaissant les expressions correctement parenthésées.

Problème des parenthèses

Le *problème des parenthèse* revient à déterminer une formalisation à partir d'une grammaire non-contextuelle du langage PAREN reconnaissant les expressions correctement parenthésées.

Langage PAREN

Soit $x \in \{[,]\}^*$. On donne les notations suivantes :

- $G(x) = |x|_[=$ Le nombre de parenthèses ouvrantes de x.
- $D(x) = |x|_] =$ Le nombre de parenthèses fermantes de x.

Ainsi, PAREN est le langage des mots x qui respectent les deux axiomes suivants :

1. $G(x) = D(x)$, et
2. Pour tout préfixe $y = x_1 ... x_{k \leq |x|}$ de x, $G(y) \geq D(x)$.

Problème des parenthèses

Le *problème des parenthèse* revient à déterminer une formalisation à partir d'une grammaire non-contextuelle du langage PAREN reconnaissant les expressions correctement parenthésées.

Langage PAREN

Soit $x \in \{[,]\}^*$. On donne les notations suivantes :

- $G(x) = |x|_[=$ Le nombre de parenthèses ouvrantes de x.
- $D(x) = |x|_] =$ Le nombre de parenthèses fermantes de x.

Ainsi, PAREN est le langage des mots x qui respectent les deux axiomes suivants :

1. $G(x) = D(x)$, et
2. Pour tout préfixe $y = x_1 ... x_{k \leq |x|}$ de x, $G(y) \geq D(x)$.

Question : Montrer que PAREN $\notin$ **Reg**.

Problème des parenthèses

La courbe ci-dessous permet d'appréhender intuitivement les propriété de PAREN.

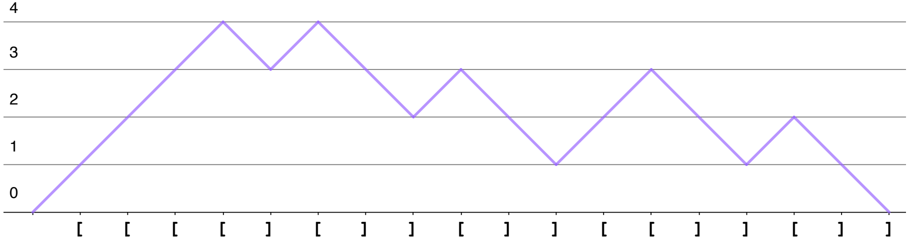


Figure – Fonction $G(y) - D(y)$ sur l'intervalle des préfixes de x

Un mot $x \in \mathtt{PAREN}$ si et seulement si $G(x) - D(x) = 0$ et si, la courbe ne descend jamais en dessous de 0.

Exemples

Déterminer une grammaire non-contextuelle qui génère PAREN.

Problème des parenthèses

Soit grammaire $G_{\texttt{PAREN}}$ suivante :

$$S \quad \rightarrow \quad [S]|SS|\varepsilon$$

On a $L(G_{\texttt{PAREN}}) = \texttt{PAREN}$.

Exemples

Vérifier que le mot représenté sur la figure précédente peut être généré par la grammaire G.

Grammaire des expressions arithmétiques

Soit grammaire $G_{\texttt{ARI}}$ suivante :

$$E \quad \rightarrow \quad E + E | E - E | E \times E | (E) | a | b$$

Exemples

Donner la dérivation à gauche, puis à droite du mot $a + a \times a$.

Grammaire des expressions arithmétiques

Soit grammaire $G_{\texttt{ARI}}$ suivante :

$$E \quad \rightarrow \quad E + E | E - E | E \times E | (E) | a | b$$

Exemples

Donner la dérivation à gauche, puis à droite du mot $a + a \times a$.

À gauche :

$$\begin{aligned} E &\Rightarrow E + E \\ &\Rightarrow a + E \\ &\Rightarrow a + E \times E \\ &\Rightarrow a + a \times E \\ &\Rightarrow a + a \times a \end{aligned}$$

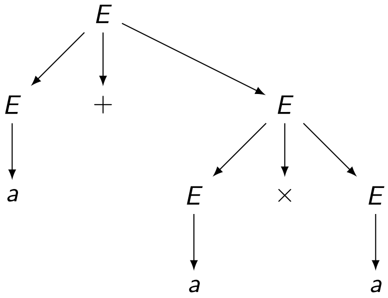
À droite :

$$\begin{aligned} E &\Rightarrow E \times E \\ &\Rightarrow E \times a \\ &\Rightarrow E + E \times a \\ &\Rightarrow E + a \times a \\ &\Rightarrow a + a \times a \end{aligned}$$

Arbre de dérivation

Il est possible de représenter la dérivation sous forme arborescente.

$$
\begin{aligned}
E &\Rightarrow E + E \\
&\Rightarrow a + E \\
&\Rightarrow a + E \times E \\
&\Rightarrow a + a \times E \\
&\Rightarrow a + a \times a
\end{aligned}
$$



Arbre de dérivation

Définition

Soit $G = (V, \Sigma, P, S)$ une grammaire hors-contexte. Un arbre Δ est un *arbre de dérivation* pour G si et seulement si :

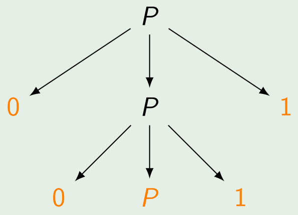
1. la racine de Δ est étiquetée par l'axiome S ;
2. chaque nœud interne est étiqueté par une variable dans V ;
3. chaque feuille est étiquetée par un symbole dans $V \cup \Sigma \cup \{\varepsilon\}$;
4. toute feuille étiquetée par ε est l'unique nœud de son parent ;
5. si un nœud interne est étiqueté par A, et tous ses nœuds fils (de gauche à droite) sont étiquetés par $X_1, X_2, ..., X_k$, alors $\rightarrow X_1 X_2 X_k$ est une règle de production dans P.

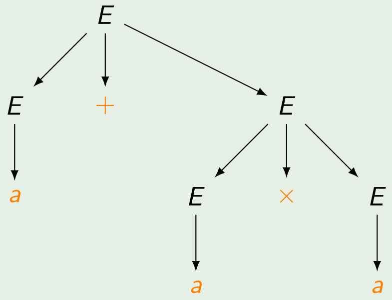
Arbre de dérivation

Définition

La *frontière* $B(\Delta)$ est la chaîne formée de toutes les feuilles de gauche à droite d'un arbre de dérivation Δ (i.e. parcours préfixe des feuilles).

Exemples





Arbre de dérivation

Définition

Un arbre de dérivation Δ est dit *terminal* si $\forall x \in B(\Delta), x \in \Sigma$, i.e. tous les symboles de la frontière sont des symboles terminaux.

Corollaire

Soit une grammaire G et un arbre de dérivation terminal Δ_G issu de G, alors $B(\Delta_G) \in L(G)$.

Table of Contents

Grammaires ambiguës

Définition

Soit $G = (V, \Sigma, P, S)$ une grammaire hors-contexte.
On dit que G est *ambiguë* si : $\exists \aleph \in \Sigma^*$ et Δ_1, Δ_2, deux arbres de dérivation terminaux tels que $\Delta_1 \neq \Delta_2$ et $B(\Delta_1) = B(\Delta_2) = \aleph$.

Exemples

Montrer que la grammaire des expressions arithmétiques vue précédemment est une grammaire ambiguë.

Analyse de l'ambiguïté

- Généralement, l'ambiguïté provient de dérivations parallèles formant une même séquences de symboles non terminaux.
- Par exemple, dans la grammaire $G_{\mathtt{ARI}}$, E produit $E + E$ ou $E\times$, et chacune permet d'arriver à la séquence $E + E \times E$.
- Afin d'éviter ce problème, et dans ce cas précis, il est possible d'utiliser une variable supplémentaire afin de différencier les termes des différents opérateurs.

Analyse de l'ambiguïté

- Généralement, l'ambiguïté provient de dérivations parallèles formant une même séquences de symboles non terminaux.
- Par exemple, dans la grammaire $G_{\texttt{ARI}}$, E produit $E + E$ ou $E\times$, et chacune permet d'arriver à la séquence $E + E \times E$.
- Afin d'éviter ce problème, et dans ce cas précis, il est possible d'utiliser une variable supplémentaire afin de différencier les termes des différents opérateurs.

On donne la nouvelle grammaire $G_{\texttt{ARI}}$ suivante :

$$\begin{array}{lcl} E & \rightarrow & E + T | E - T | T \\ T & \rightarrow & T \times F | F \\ F & \rightarrow & (E) | a | b \end{array}$$

Exemples

Vérifier que cette grammaire ne provoque plus d'ambiguïté pour le mot $a + a \times a$.

Suppression et détection de l'ambiguïté

Bien

- Il est parfois possible de lever manuellement l'ambiguïté d'une grammaire.
- Il existe des conditions suffisantes pour garantir la non-ambiguïté.

Suppression et détection de l'ambiguïté

Bien

- Il est parfois possible de lever manuellement l'ambiguïté d'une grammaire.
- Il existe des conditions suffisantes pour garantir la non-ambiguïté.

Moins bien...

- Il n'existe pas d'algorithme permettant de le faire de façon automatique.
- Le problème de savoir si une grammaire est ambiguë ou non est indécidable.
- Il existe des langages hors-contexte qui ne peuvent être générés que par des grammaires ambigües.