Langages réguliers

Montaigne 2023-2024

– mpi23@arrtes.net –

Introduction

Le chapitre précédent a défini la notion de **langage formel**. Si on se contente des trois opérations d'**union**, de **concaténation** et d'**étoile de Kleene**, on construit une classe de langages appelée la classe des **langages réguliers**.

Langages réguliers

Définition 1 (langage régulier)

L'ensemble des **langages réguliers** sur un alphabet Σ est défini inductivement comme suit.

- $\varnothing$ est un langage régulier.
- Pour tout $a \in \Sigma$, le langage singleton $\{a\}$ est régulier.
- Si L et L' sont deux langages réguliers, leur union $L \cup L'$ et leur concaténation LL' sont des langages réguliers.
- Si L est un langage régulier, L^* est un langage régulier.

On note $\mathrm{Reg}(\Sigma)$ l'ensemble des langages réguliers sur Σ.

Deux propriétés

Certaines définitions incluent explicitement le langage $\{\varepsilon\}$. Ce n'est pas nécessaire car ce serait une redondance. En effet, il suffit de se rappeler que $\{\varepsilon\}$, qui est un langage régulier, peut être obtenu par l'application de l'étoile de Kleene au langage $\varnothing$, qui est aussi un langage régulier.

$$\varnothing^* = \{\varepsilon\}$$

Une autre propriété immédiate de cette définition est que **tout langage fini est régulier**.

Vocabulaire

Que ce soit en français ou en anglais, le terme **langage rationnel** (*rational language* en anglais) est souvent utilisé comme synonyme pour langage régulier.

Conformément au programme, nous utilisons systématiquement l'appellation **langage régulier**.

Expressions régulières

En pratique, un langage régulier peut être décrit par ce qu'on appelle une **expression régulière**.

Définition 2 (expression régulière)

Une **expression régulière** sur un alphabet Σ est définie inductivement comme suit.

- $\varnothing$ et ε sont des expressions régulières.
- Pour tout $a \in \Sigma$, le symbole a est une expression régulière.
- Si r_1 et r_2 sont deux expressions régulières, $r_1 | r_2$ et $r_1 r_2$ sont des expressions régulières.
- Si r est une expression régulière, r^* est une expression régulière.

Vocabulaire

Comme pour les langages, on parle d'expressions **rationnelles**.

Les contractions **regex** ou **regexp** (contractions de l'anglais *regular expression*) sont couramment utilisés.

L'opération $|$ est parfois notée $+$: $r_1|r_2 = r_1 + r_2$.

Priorités

L'**opérateur étoile** $*$ est le plus prioritaire. Vient ensuite la **concaténation** puis l'**alternative** $|$.

Les parenthèses sont utilisées pour régler les problèmes de priorités.

Ainsi l'expression :

$$bonjour|aurevoir^*$$

doit être comprise comme :

$$(bonjour)|(aurevoi(r^*))$$

Langage d'une expression régulière

Les expressions régulières permettent de définir des langages réguliers de façon déclarative. À chaque expression régulière peut être associé son langage. On parle d'un langage **dénoté** par une expression régulière.

Définition 3

Soit r une expression régulière sur un alphabet Σ. Le langage $\mathcal{L}(r)$ dénoté par l'expression régulière r est défini inductivement sur la structure de l'expression.

- $\mathcal{L}(\varnothing) = \varnothing$
- $\mathcal{L}(\varepsilon) = \{\varepsilon\}$
- $\forall a \in \Sigma, \mathcal{L}(a) = \{a\}$
- $\mathcal{L}(r_1|r_2) = \mathcal{L}(r_1) \cup \mathcal{L}(r_2)$
- $\mathcal{L}(r_1 r_2) = \mathcal{L}(r_1)\mathcal{L}(r_2)$
- $\mathcal{L}(r^*) = (\mathcal{L}(r))^*$

Exemple

Sur $\Sigma = \{0, 1\}$, le **langage régulier** des mots contenant exactement trois 1 peut être défini en extension par :

$$L = \{111, 0111, 1011, 1101, 1110, 00111, 01011, 01101, 01110, \dots\}$$

On peut encore le décrire sous la forme :

$$L = \{0^m 10^n 10^p 10^q, (m, n, p, q) \in \mathbb{N}^4\}$$

L'**expression régulière** suivante dénote également ce langage.

$$0^* 10^* 10^* 10^*$$

Ainsi :

$$L = \mathcal{L}(0^* 10^* 10^* 10^*)$$

Exemple

Sur $\Sigma = \{0, 1\}$, le **langage régulier** des mots des nombres en base deux sans zéro non significatif est défini par :

$$L = \{0, 1, 10, 100, 101, 110, 111, 1000, \dots\}$$

soit encore :

$$L = \{0\} \cup \{1, 10, 100, 101, 110, 111, 1000, \dots\}$$

On constate que les mots de L sont :

- soit le mot d'une lettre 0 ;
- soit les mots commençant par un 1, suivis de 0 et de 1.

On peut les décrire à l'aide de l'**expression régulière** suivante.

$$0|1(0|1)^*$$

Cette expression rationnelle dénote L.

$$L = \mathcal{L}(0|1(0|1)^*)$$

Expression régulières équivalentes

Définition 4

Deux expressions régulières r_1 et r_2 sur un alphabet Σ sont dites **équivalentes** si elles décrivent le même langage : $\mathcal{L}(r_1) = \mathcal{L}(r_2)$.

Par abus de notation, on note parfois $r_1 \equiv r_2$.

Par exemple, les trois expressions régulières suivantes sont équivalentes.

$$(a|b)^* \qquad (a^* b)^* a^* \qquad (b^* a)^* b^*$$