Théorie des Langages

Matthew Coyle / Valentin Bayart

January 14, 2018

1 Introduction et concepts de base

- Introduction

Les langages formels ont été étudiés par :
- **les informaticiens** : langages de programmation (Avec une syntaxe (La définir, la vérifier) permettant de traduire ce même langage
- **les linguistes** : langues naturelles

Exemples de Langages :
- Entiers naturels (suite de chiffres parmi 0...9).
- Entiers naturels impairs (même représentation).
- Les mots franais (du dictionnaire).
- Les identificateurs C++.
- Les phrases en franais.
- Les programmes (syntaxiquement corrects) écrits en C++.

Tous ces exemples sont des ensembles ou des sous-ensembles d'un autre ensemble (lettres / alphabet / dictionnaire).

Points communs des langages formels :
- Chaque langages est un ensemble d'éléments appelés **mots** ou "**chaînes**".
- Chaque chaîne est une suite de **symboles** pris parmi un ensemble fini de symboles.
- Chaque chaîne est de longueur finie (même sil ny a pas de limite cette longueur).

On étudie des modèles pour représenter de manière finie des langages :
- automates finis.
- expressions regulières.
- grammaires formelles.
- ...

Application pratiques:
- Recherche de "**motifs**" dans des fichiers.
- Traitement de texte.
- Modélisation de circuits.
- de machines à états.
- Compilation de langages de programmation.
- ...

- Concepts de Base

- Alphabets :
Un **alphabet** est un ensemble fini, non vide, de symboles.
On le note généralement $\sum$.

Exemples d'alphabets :
$\sum_{entiers}$ = {0,1,2,3,4,5,6,7,8,9}
$\sum_{mots}$ = {a,b,c,...,z,',-}
$\sum_{ident}$ = {a,...,z,A,...,Z,0,...,9,_}
$\sum_{prog}$ = {int, float, bool, while, do, for, ..., <, <=, >, >=, =, !=, +, -, /, *, ;, ..., 0, 1, , 25, 26, 27,..., 12.56, ..., a, b, toto, compteur, Tab, ...}

- Chaînes
Un **mot** ou une **chaîne** ω (omega) formé(e) sur un alphabet est une suite finie $s_1s_2...s_n$ de symboles de cet alphabet
La chaîne vide, noté ε (epsilon), est une chaîne ne contenant aucun symbole
La **longueur** dune chaîne ω, notée $|\omega|$, est le nombre de symboles composant la chaîne ω

- Opérations sur les chaînes La concatenation de deux chaînes u et v, notée u.v ou uv est la chaîne obtenue en écrivant les symboles de u suivis de ceux de v.
si $u = a_1a_2...a_n$ et $v = b_1b_2...b_p$
alors $uv = a_1a_2...a_nb_1b_2...b_p$

Propriétés :
- |u.v| = |u| + |v|
- Associativit : (u.v).w = u.(v.w)
- ε est lment neutre : u.ε = ε.u = u

. Puissance d'une chaîne ω ω^k est la chaîne forme par la concatnation de k occurrences de ω
$$\omega^k = \underbrace{\omega\omega\omega\ldots\omega\omega}_{k \text{ fois}}$$
$\omega^0 = \varepsilon$

Un **prefixe** d'une chaîne ω ? est une suite, éventuellement vide, de symboles dbutant ω.
Un **suffixe** de ω est une suite de symboles terminant ω.
$\forall x, y \mid \omega = x.y$, x est un prefixe de ω, y un suffixe.
Une **sous-chaine** d'une chaîne ω est une suite de symboles apparaissant consécutivement dans ω.

Notation : $|\omega|_x$ est le nombre d'occurences de la chaîne x dans la chaîne ω.

- Langage :
Un langage est un ensemble de chaîne.

Exemples de langages:
{toto,titi,tata}
{1,11,101,1001}
$\{1^n \mid n >= 0\} = \{e, 1, 11, 111, 1111, 11111, ...\}$ c'est un langage infini (nombre infini de chanes)dont chaque chaîne est de longueur finie
Nombres binaires impaires: {1,11,101,111,1001,1011,...}
Nombres binaires premiers: {1,10,11,101,111,1011,...}

le Langage vide, noté $\emptyset$, ne contient aucune chaîne (ensemble vide).
Attention : $\emptyset \neq \{\varepsilon\}$

Le **langage plein**, noté $\sum^*$, contient toutes les chaînes que l'on peut former sur l'alphabet $\sum$
Remarque : $\sum^* = \sum^+ \cup \{\varepsilon\}$

- Opérations sur les langages :
l'**Union** de deux langages A et B est le langage, note $A \cup B$, composé de toutes les chaînes qui apparaissent dans l'un au moins de langages A ou B.
$A \cup B = \{\omega \mid \omega \in A \ ou \ \omega \in B\}$

Propriétés :
- Commutativité : $A \cup B = B \cup A$
- Associativité : $(A \cup B) \cup C = A \cup (B \cup C)$
- $\emptyset$ est élément neutre : $A \cup \emptyset = \emptyset \cup A = A$
- Idempotence : $A \cup A = A$

La **concaténation** de deux langages A et B est le langage, note A.B ou AB, composé de toutes les chaînes formées par une chaîne de A concaténée à une chaîne de B.
$A.B = \{u.v \mid u \in A \ ou \ v \in B\}$

Propriétés:
- Associativité: (A.B).C = A.(B.C)
- $\{\varepsilon\}$ est un élément neutre: $A.\{\varepsilon\} = \{\varepsilon\}.A = A$
- $\emptyset$ est element absorbant : $A.\emptyset = \emptyset.A = \emptyset$

Distributivité de la concaténation sur l'union:
- À gauche : $A.(B \cup C) = A.(B.C)$
- À droite : $(B \cup C).A = B.A \cup C.A$

. Puissance d'un langage A
A^k est le langage formé par la concaténation de k occurrences de A.
$A0 = \{\varepsilon\}$
$A^1 = A$
$A^n = \underbrace{AAA \ldots AA}_{\text{n fois}}$
A^k : Mots formés pas la concaténation de k mots de A

. Étoile de Kleene (fermeture ou cloture par .)
- la **fermeture de Kleene** d'un langage A est le langage, noté A^*, défini par : $A^* = A^0 \cup A^1 \cup A^2 \cup A^3$... soit

$$A^* = \bigcup_{i=0}^{\infty} A^i = A^0 \cup A^1 \cup A^2 \cup ...$$

- la **fermeture positive** de A est le langage, noté A^+, défini par : $A^+ = A^1 \cup A^2 \cup A^3 \cup ...$ soit

$$A^+ = \bigcup_{i=1}^{\infty} A^i = A^1 \cup A^2 \cup A^3 \cup ...$$

"mots formés par la concaténation de 1 ou plusieurs mots de A"
Remarque : $A^* = \{\varepsilon\} \cup A^+$ (

Propriétés : $A^+ = A.A^* = A^*.A$ (Posibilité : $A^+ = A^*$)

2 Modeles et Langages

- contextuels
- langages recursivement enumerables