Langages et Automates : LA3
Partie 1 : Langages, Expressions Rationnelles

Théorie des Langages ?

Origine : Linguistique - Etude des langages naturels (ceux que nous parlons) en vue de faire de la traduction automatique par exemple.

Aujourd'hui : Analyse de texte au sens large, par exemple

- Recherche de motifs (ex : séquençage du génome)
- Analyse lexicale, grammaticale
- Compilation de Programmes

Compilateur

Programme qui prend en entrée un texte dans un langage A dans le but de le traduire dans un langage B. Pour cela il doit connaître et reconnaitre pour ces deux langages

- leur vocabulaire (les mots autorisés)
- la syntaxe (la structure des phrases autorisées)
- leur sémantique (le sens des phrases autorisées)

Alphabet

Définition

Un alphabet est un ensemble **fini** *de symboles.*

Définition

Un mot sur l'alphabet Σ *est une séquence* **finie et ordonnée**, *éventuellement vide, d'éléments de* Σ.
Le mot vide est dénoté ε.

Exemple

aabac est un mot sur l'alphabet $\{a, b, c\}$.
00101 *et* 00011 *sont deux mots sur l'alphabet* $\{0, 1\}$.
Un brin d'adn est un mot sur l'alphabet $\{A, C, G, T\}$.

Définition (Longueur, Occurence)

Un mot w est donc une séquence de lettres $w_1 w_2 \dots w_n$. L'entier n est appelé longueur du mot w et est noté $|w|$.

La longueur du mot vide ε est 0.

Si w_i est la lettre x, il s'agit d'une occurence de la lettre x en position i.

Le nombre d'occurences de la lettre x dans le mot w sera notée $|w|_x$.

Exemple

Le mot abaac est de longueur 5 et contient trois occurrences de la lettre a, en position 1, 3 et 5.

Définition (Concaténation)

Si u et v sont deux mots, on définit un nouveau mot, noté $u.v$, ou uv, appelé produit de concaténation de u et v en mettant bout à bout les lettres de u et v (sa longueur vérifie donc $|uv| = |u| + |v|$).

Exemple

Si $u = abb$ et $v = ba$, alors $uv = abbba$.

Bien sur, à moins que l'alphabet n'ait taille 0 ou 1, cette opération n'est pas commutative

Exemple

Si $u = abb$ et $v = ba$, alors $uv = abbba$ et $vu = baabb$.

Monoïde Libre

Définition

Si Σ est un alphabet, on dénote par Σ^*, l'ensemble de tous les mots sur Σ. Muni du produit de concaténation, c'est un monoïde, puisque cette opération

- *est associative :* $\forall u, v, w \in \Sigma^* \ (u.v).w = u.(v.w)$
- *possède un élément neutre, le mot vide* ε *:* $\forall u \in \Sigma^* \ u.\varepsilon = \varepsilon.u = u$

L'ensemble Σ^* est appelé monoïde libre engendré par X.

Facteurs, Sous-Mots

Définition (Facteur, Sous-Mot)

- *Un mot u est un facteur d'un mot v si il existe des mots w_1 et w_2 tels que $v = w_1.u.w_2$.*
  Si $w_1 = \varepsilon$, on dit que u est un prefixe de v.
  Si $w_2 = \varepsilon$, on dit que u est un suffixe de v.
- *Un mot u est un sous-mot de v si il peut être obtenu a partir de v en supprimant une ou plusieurs lettres.*

Remarque

Tout facteur est un un sous-mot (mais la réciproque n'est pas vraie).

Exemple

- *aba est un facteur de aababbaba*
- *bbbb est un sous mot de aababbaba*

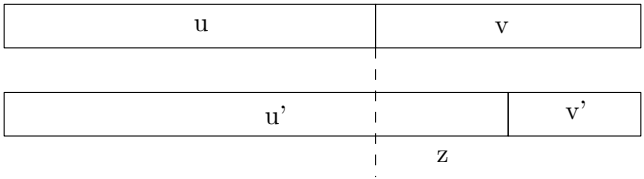
Un Lemme de Factorisation

Lemme (Lemme de Levi)

Si u, v, u', v' sont tels que $uv = u'v'$, alors il existe un mot z tel que

- *soit $u' = uz$ et $v = zv'$*
- *soit $u = u'z$ et $v' = zv$*

Autrement dit soit u est un prefixe de u' (et v' suffixe de v), soit l'inverse.



Une Application

On dit que deux mots u et v commutent si $uv = vu$.

Théoreme

u et v commutent si et seulement si il existe un mot w et deux entiers k et l tels que $u = w^k$ et $v = w^l$.

Preuve :
On prouve le résultat par récurrence sur $n = |u| + |v|$.
Si $n = 0$ le résultat est trivial.
Soit $n \geq 1$, supposons le résultat vrai pour tout $n' < n$.
D'apres le lemme de Levi, il existe z tel que $u = vz$ ou $v = uz$.
Supposons $u = vz$ (lautre cas se résout de la même manière).
On a alors $vzv = vvz$ et donc $zv = vz$. Si $v = \varepsilon$ le résultat est trivial ($v = u^0$), donc on peut supposer $|v| > 0$ et donc $|z| < |u|$ et on peut appliquer l'hypothèse de récurrence à z et v.
Il existe donc w, k' et l' tels que $v = w^{k'}$ et $z = w^{l'}$ et donc $u = w^{k'+l'}$.

Définition (Langage)

On appelle langage sur un alphabet Σ tout sous-ensemble L de Σ^*.

Exemple

Sur l'alphabet $\Sigma = \{a, b\}$

- $\{a, aa, aba, bbbab\}$
- $\emptyset$
- $\{\varepsilon\}$
- $\{\epsilon, ab, a^2b^2, a^3b^3, \ldots\} = \{a^n b^n,\ n \in \mathbb{N}\}$.
- *L'ensemble des mots de Σ^* qui commencent et finissent par un a.*
- *Les palindromes : l'ensemble des mots qui se lisent de la meme façon dans un sens et dans l'autre comme abba ou babab.*

Description d'un langage

Elle peut être par exemple :

- en langue naturelle :
  Ex : "l'ensemble des mots sur l'alphabet qui commencent par un a"
  ou "l'ensemble des palindromes sur un alphabet donné".
- de façon énumérative : on énumère tous les mots.
  Ex : tous les langages finis, mais aussi certains langages comme
  $\{a^n b^n,\ n \in \mathbb{N}\}$.
- par une "machine" qui prend en entrée un mot et répond Oui ou Non.
  Ex : automates
- par mécanismes génératifs : on définit des briques de bases et des règles de production.
  Ex : On définit recursivement le langage L sur $\{a, b\}$ par :
  - a et b sont dans L
  - Pour tout mot u de L aua et bub sont encore dans L.
  (Quel est ce langage ?)

Opérations sur les Langages

Etant donnés deux langages L et L' sur l'alphabet Σ on peut construire :

- l'union $L \cup L'$, l'intersection $L \cap L'$.
- le complément $\overline{\Sigma} = \{u \in \Sigma^* | \, u \notin L\}$.
- Le produit de concaténation $L.L' = \{w \in \Sigma^* | \, \exists u \in L \text{ et } \exists v \in L' \text{ tq } w = u.v\}$
- La puissance, définie par récurrence par
  - $L^0 = \{\varepsilon\}$
  - $L^n = L.L^{n-1} \, \forall n \geq 1$.
- Le passage à l'étoile : $L^* = L^0 \cup L \cup L^2 \cup L^3 \cup L^4 \ldots = \bigcup_{n \geq 0} L^n$

Exercice

L'alphabet est $\Sigma = \{a, b\}$.

- $L_1 = \{a, ab\}$ *et* $L_2 = \{b, \varepsilon\}$. *Que vaut* $L_1.L_2$ *?*
- $L = \{ab\}$. *Decrire* L^*
- $L_1 = \{a\}$ *et* $L_2 = \{a, b\}^*$ *Decrire* $L_1.L_2$ *en langage naturel*
- *Donner une expression pour le langage "L'ensemble des mots contenant le facteur aba"*

Proposition

Pour tous langages L_1, L_2, L_3 on a :

- $L_1.(L_2 \cup L_3) = (L_1.L_2) \cup (L_1.L_3)$
- *Attention faux pour intersection ! ! !*
  Seulement $L_1.(L_2 \cap L_3) \subset (L_1.L_2) \cap (L_1.L_3)$ *est vrai*

Quotient gauche d'un Langage

Définition

Le quotient gauche d'un langage L par le mot u est défini par :

$$u^{-1}L = \{v \in \Sigma^* | \, uv \in L\}$$

Proposition

- $\varepsilon^{-1}L = L$
- *Pour tous mots u et v,* $(uv)^{-1}L = v^{-1}(u^{-1}L)$

Définition (Langage rationnel)

Soit Σ un alphabet. Les langages rationnels sur Σ sont définis inductivement par :

i) $\{\varepsilon\}$ *et* $\emptyset$ *sont des langages rationnels*

ii) $\forall a \in \Sigma, \{a\}$ *est un langage rationnel*

iii) *si* L, L_1 *et* L_2 *sont des langages rationnels, alors* $L_1 \cup L_2$, $L_1.L_2$, *et* L^* *sont également des langages rationnels.*

Est alors rationnel tout langage construit par un nombre fini d'applications de la récurrence iii).

Langages Rationnels - Exemples

- L'ensemble des mots sur l'alphabet $\{a, b\}$ qui commencent par a ou terminent par b : $\{a\}.\{a, b\}^* \cup \{a, b\}^*\{b\}$
- L'ensemble des mots sur l'alphabet $\{a, b\}$ qui contiennent le facteur aba : $\{a, b\}^*.(aba).\{a, b\}^*$

Expressions Rationnelles

Pour représenter un langage rationnel on utilise une expression rationnelle

On utilise le symbole $+$ pour symboliser l'union et on oublie les accolades autour des singletons.

Ainsi $\{a\}.\{a,b\}^* \cup \{a,b\}^*\{b\}$ est représenté par l'expression rationnelle $a(a+b)^* + (a+b)^*b$

Expressions Rationnelles - Exemples

Donner des ER pour les langages suivants

- L'ensemble des mots de longueur multiple de 3
- L'ensemble des mots ne contenant pas le sous-mot bb
- L'ensemble des mots ne contenant pas le facteur bb
- L'ensemble des mots ne contenant pas le sous-mot ab
- L'ensemble des mots ne contenant pas le facteur ab
- L'ensemble des mots contenant un nombre pair de a

Expressions Rationnelles - Exemples

A une expression rationnelle correspond un unique langage rationnel mais un langage rationnel peut être représenté par plusieurs ER différentes.

Exemple

- $(a + b)^*$ *et* $(a^*b^*)^*$
- $a(ba)^*$ *et* $(ab)^*a$

Une question algorithmique peut alors se poser : comment décider si deux expressions rationnelles représentent le même langage ? Nous résoudrons ce problème au cours de ce cours.

Expressions Rationnelles - Arbre syntaxique

Il est commode de représenter une ER par un arbre dont les noeuds sont étiquetés par les opérandes et les feuilles par les lettres

$(a + ba)^*(aa + b)$