1 Définitions

Les grammaires formelles

Une grammaire formelle est un quadruplet $G = (T, N_t, S, D)$ où :

- T est l'alphabet des terminaux
- N_t est l'alphabet des non terminaux
- $S \in N_t$ est l'axiome
  Notons $\Sigma := N_t \cup T$
- $D \subset \mathcal{P}((\Sigma^\star)^2)$, est l'ensemble des règles de dérivation.

Pour $(a, b) \in D$, soit $\overset{(a,b)}{\to}$ la relation binaire définie sur $\Sigma^\star$ par

$$\forall x, x' \in \Sigma^\star, x \overset{(a,b)}{\to} x' \iff \exists u, v \in \Sigma^\star / x = uav \text{ et } x' = ubv$$

On note $\to := \bigcup_{d \in D} \overset{d}{\to}$ et on note $\overset{*}{\to}$ la cloture transitive et réflexive de $\to$.
Pour $x \in \Sigma^\star$, notons
$\delta(x) := \{y \in \Sigma^\star / x \overset{*}{\to} y\}$
$|x|_l := |\{i \in \mathbb{N} \mid x_i = l\}|$ est le nombre d'occurences de la lettre l dans x.

Alors le langage de la grammaire formelle G est le suivant :

$$\mathcal{L}(G) := \delta(S) \cap T^\star$$

Nous allons supposer que N_t est dénombrable et T est fini.

NB : code typesetutiles debut types

2 Vérification de preuve - grammaire formelle

Une nouvelle structure de données pour les preuves

Nous pouvons nous restreindre à la définition d'un mot qui dérive d'un autre. On va alors fournir, comme preuve, une liste de mots $m_1, \ldots, m_n$ tels que $\forall i \in [1, n-1], m_i \to m_{i+1}$. Nous allons aussi renseigner de quelle manière le mot

m_{i+1} dérive du mot m_i en fournissant l'indice de la règle de dérivation (a, b) et celui du début de a dans m_i qui sera remplacé par b. Nous aboutissons alors à :

```
type 'e preuveformelle = (int*int*('e caractere list)) list
```

Nous prendrons comme exemple la grammaire formelle $G = (T, N_t, S, D)$ où :
$T = \{\underline{a}, \underline{b}, \underline{c}\}$
$N_t = \{\underline{S}, \underline{B}\}$
$D = \{(\underline{S}, \underline{aBSc})_1, (\underline{S}, \underline{abc})_2, (\underline{Ba}, \underline{aB})_3, (\underline{Bb}, \underline{bb})_4\}$,
alors $\underline{\text{ aabbcc }}$ est dans $\mathcal{L}(G)$ car $\underline{S} \rightarrow_1 \underline{aBSc} \rightarrow_2 \underline{aBabcc} \rightarrow_3 \underline{aaBbcc} \rightarrow_4 \underline{aabbcc}$.
En Ocaml :

```
let unepreuve = [(0,1,mot "aBSc");(2,3,mot "abc");(1,3,mot "aB");(2,4,mot "bb")]
```

NB : code verifderiv

NB : code verifpreuve

3 Preuve automatique - grammaire formelle

Nous allons générer successivement les $G_n := \{m \in \Sigma^\star \mid S(\bigcup_{0 \leq k \leq n} \rightarrow^k)m\}$ à l'aide de $G_{n+1} = \bigcup_{x \in G_n} \mathcal{S}(x)$ où $\mathcal{S}(x) := \{m \in \Sigma^\star \mid x \rightarrow m\} = \bigcup_{d \in D} \{m \in \Sigma^\star \mid x \xrightarrow{d} m\}$ Pour ce faire, on trouve les successeurs d'un mot x par $d = (a, b)$ en recherchant le motif a dans x. (L'algorithme de Knuth-Morris-Pratt est adéquat). Nous avons donc un algorithme qui permet théoriquement de prouver que n'importe quel mot de la grammaire appartient bien à cette dernière. Cet algorithme termine si et seulement si le mot à prouver est prouvable (si il est dans la grammaire). Le problème étant la complexité de ce dernier, qui fonctionne comme une machine non déterministe où on ne coupe pas les instances arrivant à un état puit. La suite va donc consister à trouver, de manière heuristique ou non, des méthodes affirmant qu'à partir d'un mot m, on aura des difficultés à dériver en notre mot m' à prouver. Alors, nous allons arrêter les recherches de dérivations de S en m' qui passent par le mot m.

NB : code preprocessgf et succ qui calcule $\mathcal{S}(x)$

Un parcours en largeur particulier

Nous allons utiliser un parcours qui nous permettra de trouver dans un graphe un chemin plus court entre le sommet initial et un sommet 'valide'. Ce parcours devra aussi éviter les chemins passant par un sommet 'interdit'.

Alors on fournira à ce parcours une fonction `interdit` et une fonction `valide` toutes deux de type `sommet -> bool` .

Ce parcours est donc un parcours en largeur depuis un sommet x_0 qui garde en mémoire le chemin parcouru et qui s'arrête dès qu'un sommet s, tel que `valide s` , est parcouru. Il renvoit alors un chemin de x_0 à s. Lorsque le parcours passe par un sommet s tel que `interdit s` , il ne parcourt pas ensuite ses voisins.

NB : code parcoursmagique

Enfin, un prouveur automatique d'appartenance d'un mot à une grammaire formelle

Il ne nous reste plus qu'à appliquer notre parcours sur un graphe où les sommets sont les mots de $\Sigma^\star$, et les arêtes sont $\{(a,b) \in \Sigma^\star \mid a \to b\} = \{(a,b) \in \Sigma^\star \mid b \in \mathcal{S}(a)\}$.

NB : code chercherderivationnaif

Par la suite, nous allons apporter des améliorations, en fournissant des fonctions `interdit` .

4 Amélioration pour les grammaires croissantes

Une grammaire croissante est une grammaire telle que :

$$\forall (a,b) \in D, |a| \leq |b| \tag{1}$$

On a alors une première propriété très simple, $\forall m, m' \in \delta(S), m \overset{*}{\to} m' \implies |m| \leq |m'|$.

Alors, dans la recherche de dérivations de S vers m, on peut supprimer les "branches de recherche" qui partent d'un mot de longueur $> |m|$. Il ne reste qu'à faire le même parcours que précédement avec la fonction suivante comme fonction `interdit` .

```
let interditcroiss x = Array.length x > Array.length m
```

5 Amélioration dans le cas général : déduction sur le nombre d'occurence de chaque lettre

Rappelons, le problème : étant donné un mot $m \in \delta(S)$, comment le dériver en un mot $m' \in \delta(m)$ donné ? Pour tout $l \in \Sigma$, on va chercher un ensemble $q(l)$ qui majore $|\delta(m)|_l$ $(= \{|x|_l \mid x \in \delta(m)\})$, l'ensemble des nombres d'occurence de l dans les successeurs de m.

Ainsi, $m \xrightarrow{*} m' \implies |m'|_l \in q(l)$, la contraposée nous sera utile :

$$|m'|_l \notin q(l) \implies \neg(m \xrightarrow{*} m')$$

Ce qui pourra nous permettre de réduire notre champ de recherche. En effet,

$\boxed{\text{si pour un mot } x \in \delta(m),\ \exists l \in \Sigma, |\delta(x)|_l \cap q(l) = \emptyset, \text{ alors } m' \notin \delta(x)}$.

Pour calculer $|\delta(x)|_l \cap q(l)$, nous allons utiliser un graphe semblable à un automate que nous noterons A_0.

Les sommets sont des états de $Q \subset (\mathcal{P}(\mathbb{N}))^\Sigma$ tels que :

$$\forall q, q' \in Q, \exists l \in \Sigma / q(l) \cap q'(l) = \emptyset$$
$$\text{et } \forall m \in \delta(S), \exists q \in Q / \forall l \in \Sigma, |m|_l \in q(l)$$

Ainsi à tout mot $x \in \delta(S)$, on peut associer un unique état q tel que $\forall l \in \Sigma, |x|_l \in q(l)$, notons cet état $cat(x)$

Les arêtes du graphe sont les dérivation possibles d'un famille majorante q à une autre q'.

$$\begin{aligned}(q, q') \in A_0 &\iff \exists x, x' \in \Sigma^\star / x \to x' \text{ et } \forall l \in \Sigma, |x|_l \in q(l) \text{ et } |x'|_l \in q(l)\\ &\iff \exists x, x' \in \Sigma^\star / x \to x' \text{ et } cat(x) = q \text{ et } cat(x') = q'\end{aligned}$$

Ainsi, si x correspond à un état q, et x' à un état q', alors $x \xrightarrow{*} x' \implies q'$ est accessible depuis q dans tout graphe A majorant A_0.

Si m' n'est pas accessible depuis x dans un graphe A majorant A_0, alors on peut interdire le parcours passant par x, `interdit x` renverra vrai.

Les états q sous la forme $q \in Q = \{\{0\}, \mathbb{N}^*\}^\Sigma$

Exemple avec la grammaire suivante :

$D := \{(S, \underline{abc})_0, (\underline{abc}, \underline{ab})_1, (\underline{b}, \underline{k})_2, (\underline{c}, \underline{ak})_3, (\underline{kak}, \underline{aa})_4, (\underline{a}, \underline{aaa})_5\}$

Voici un graphe A majorant A_0 (les arêtes sont étiquetées par les indices de règles de dérivation donnant l'arête):

Et voici ce à quoi correspondent les différents états (sommets) du graphe :

Le graphe réduit suivant sera utile :

Ainsi, nous pouvons tout de suite affirmer qu'il est impossible de dériver $\underline{\text{ aaabakab }}$ en $\underline{\text{ akkcckaaakck }}$ (en effet, ψ n'est pas accessible depuis δ)

Intéressons nous au calcul des arrêtes A du graphe.

Considérons d'abord les arêtes partant d'un état q donné, puis celles correspondant à une dérivation d donnée.

$A_q(\Sigma) := \{(a, b) \in A \mid a = q\}$

$A_{q,d}(\Sigma) := \{(a, b) \in A_q \mid \exists x, x' \in \Sigma^\star / x \xrightarrow{d} x' \text{ et } cat(x) = q \text{ et } cat(x') = q'\}$

Soit $l \in \Sigma$,

$\Sigma' := \Sigma \backslash \{l\}$

On a alors $A_q(\Sigma) = \bigcup_{d \in D} A_{q,d}(\Sigma)$ et $A = \bigcup_{q \in Q} A_q(\Sigma)$.

Et les $A_{q,d}(\Sigma)$ sont majorables de la manière suivante :

Si $cat(b)(l) = \mathbb{N}^\star$, alors

$$A_{q,d}(\Sigma) \subset \{(q,b) \in Q^2 \mid (q_{\Sigma'}, b_{\Sigma'}) \in A_{q,d}(\Sigma') \text{ et } b(l) = \mathbb{N}^\star\}$$

En effet, si d produit la lettre l, elle est forcément dans le mot produit.

Si $cat(b)(l) = \{0\}$ et $q(l) = \{0\}$, alors

$$A_{q,d}(\Sigma) \subset \{(q,b) \in Q^2 \mid (q_{\Sigma'}, b_{\Sigma'}) \in A_{q,d}(\Sigma') \text{ et } b(l) = \{0\}\}$$

En effet, si x ne contient pas la lettre l, et que d ne produit pas la lettre l, le mot produit x' n'aura pas de l.

Si $cat(b)(l) = \{0\}$ et $q(l) = \Sigma^\star$, alors

$$\begin{aligned} A_{q,d}(\Sigma) \subset & \{(q,b) \in Q^2 \mid (q_{\Sigma'}, b_{\Sigma'}) \in A_{q,d}(\Sigma') \text{ et } b(l) = \{0\}\} \\ & \cup \{(q,b) \in Q^2 \mid (q_{\Sigma'}, b_{\Sigma'}) \in A_{q,d}(\Sigma') \text{ et } b(l) = \Sigma^\star\} \end{aligned}$$

On ne peut rien dire, alors les deux prolongements possibles (sur la valeur de $b(l)$) seront explorés.

On peut alors majorer l'ensemble des arêtes du graphe.

Maintenant, une fois le graphe majorant associé à notre grammaire calculé, on peut faire notre fonction `interdit` , `interdit x` renvoit vrai si et seulement si x n'est pas accessible depuis S dans le graphe majorant. (On peut précalculer le graphe réduit où les sommets sont les composantes fortement connexes et sa matrice d'accessibilité avec l'algorithme Floyd Warshall par exemple, pour réduire les temps de calcul).

6 ANNEXE 1 : Notations et rappel des définitions

- On considère la grammaire formelle $G = (T, N_t, S, D)$ où T est l'ensemble des terminaux, N_t l'ensemble de non-terminaux, S, l'axiome et D l'ensemble des règles de dérivation.
- $\Sigma := N_t \cup T$
- $x \overset{(a,b)}{\to} x' \iff \exists u, v \in \Sigma^\star / x = uav$ et $x' = ubv$ si x' dérive directement depuis x par (a, b).
- $\to := \bigcup_{d \in D} \overset{d}{\to}$, $x \to x'$ si x' dérive directement depuis x. ($x' \in \mathcal{S}(x)$)
- Pour $x \in \Sigma^\star$, $\mathcal{S}(x) := \{y \in \Sigma^\star \mid x \to y\}$ ensemble des mots directements dérivables depuis x.
- $\overset{*}{\to} := \bigcup_{n \in \mathbb{Z}}$ est la cloture transitive réflexive et symétrique de $\to$, $x \overset{*}{\to} x'$ si x' dérive depuis x. ($x' \in \delta(x)$)
- Pour $x \in \Sigma^\star$, $\delta(x) := \{y \in \Sigma^\star \mid x \overset{*}{\to} y\}$ ensemble des mots dérivables depuis x.
- $\mathcal{L}(G) := \delta(S) \cap T^\star$ le langage de la grammaire.
- Pour $x \in \Sigma^\star$, $|x|_l := |\{1 \le i \le |x| \mid x_i = l\}|$ le nombre d'occurences de la lettre l dans x.
- $Q \subset (\mathcal{P}(\mathbb{N}))^\Sigma$ ensemble d'états / familles majorantes / sommets.
- $cat(x)$ unique état du graphe A_0 contenant le nombre d'occurence de chaque lettre de x.
- A_0 graphe des dérivations possibles entre familles majorantes.
- A graphe majorant A_0. ($(q, q') \in A_0 \implies (q, q') \in A$).

oui