Institut Supérieur d'Électronique de Paris
Projet de Fin d'Études

Reponsable: M. Hugueney

Finite State Transducers Just-In-Time Compiling

Do you hear the bytecode ?

Émilien Boulben
Victor Delepine
Corentin Peuvrel

24 juin 2015
Paris

Table des matières

Listings

Liste des tableaux

Table des figures

Introduction

Les FST – Finite State Transducers ou en français Transducteur fini – sont des automates finis particuliers puique possédant une sortie. Ils sont enormément utilisés par les moteurs de recherche puisqu'ils permettent d'associer à un mot une valeur numérique unique. C'est l'indexation.

Nous savons donc déjà transformer le texte en valeur exploitable ensuite (par comparaison avec un disctionnaire par exemple). Mais de part les quantités non négligeable de texte à analyser, et ce le plus rapidement possible, il est essentiel de continuer les recherches afin de réussir à optimiser cette transformation.

Seulement en ingénérie l'optimisation n'est pas tout, il faut aussi pouvoir simplement déployer les solutions sur différents serveurs : une solution trop complexe même performante sera très coûteuse à long termes et donc probablement pas choisie. C'est alors qu'intervient ce projet.

Il s'agit de faire une étude sur la faisabilité et la pertinence d'une nouvelle manière d'indexer le texte en java, L'objectif n'est pas de faire mieux que la référence qu'est Lucène, mais de déterminer s'il est possible de faire mieux en utilisant cette méthode.

Nous allons avec ce projet chercher à déterminer si un interpréteur de FST reposant sur la compilation à la volée est viable en java.

1 Analyse préalable

1.1 Le projet en détail

Pour pallier à des problèmes de performance et de distribution de la solution lors de l'indexation d'un texte en suivant une FST il est important de penser à de nouvelles solutions qui pourront éventuellement challenger la librairie de référence sur ce sujet : Lucène. L'objectif n'est pas d'y parvenir, mais de déterminer si cela est possible avec une solution qui a été imaginée par M. Hugueney et M. Marty.

Dans Lucène lors de la création d'une FST une structure de données est stockée en ram et le texte la parcourt pour connaître le résultat. Afin de gagner en performance en termes de temps d'exécution lors de cette étape, ne serait-il pas mieux de pouvoir avoir un code simple mais conséquent en taille qui soit parfaitement adapté à la FST désirée ? Finalement, plutôt que créer une structure générique, générer un code dédié à la FST en étude et l'optimiser au mieux en bafouant tous les principes de développement afin de gagner en temps d'exécution. La lisibilité est sacrifiée mais ce n'est pas très important compte tenu que ce code n'a pas destination à être lu, seulement compilé puis parcouru.

L'objectif du projet est de réussir à créer un programme Java qui génèrerait le code correspondant à une FST donnée pour pouvoir parser du texte efficacement, d'abord en Java puis en bytecode. Ensuite, faire une étude des performances et si possible les comparer avec les outils déjà existant. Il est très important de documenter les difficultés rencontrées puisque l'objectif reste de pouvoir se prononcer sur la faisabilité ou l'utilité d'un tel produit.

1.2 Premières études

Il nous a été très compliqué de comprendre ce qu'était une FST. Nous avons beaucoup investi de temps à combler ce manque avec des résultats très mitigés. Il était très difficile avec toute la documentation disponible de savoir par quoi commencer, surtout que souvent pour comprendre certains concepts il nous fallut assimiler ce que les explications considéraient acquit.

Cette incompréhension du sujet fût à l'origine de bien des découragements, et il n'a pas toujours été facile de nous remotiver les uns et les autres. Finalement c'est la pratique qui nous a apporté le plus de réponse.

Une FST qui permette l'indexation de texte, qu'est-ce ? Une simple structure de donnée. Un tableau (voir Tableau 5 et Tableau 6) ou un graphe (voir Figure 2 et Figure 3 peut la représenter). Elle est construite par un algorithme à partir d'un dictionnaire qui associe à des mots des valeurs. Enfin cela permet de parser du texte lettre à lettre (voir mots à mots) rapidement et en parallèle.

Une fois cette base essentielle partiellement comprise nous avons pris le courage de nous lancer dans le code.

1.3 Commencer quelque part

Commencer à coder paraît simple et pourtant nous y avons rencontré de trop nombreuses difficultés. Nous nous sommes heurtés à des nombreuses reprises à notre méconnaissances des FST, et n'arrivions pas à dégager un cas simple sur lequel travailler et monter en compétence.

Nous avons aussi perdu du temps à partir sur du code inutile à ce moment du développement : algorithmes de création d'une FST, tests en bytecode... Alors que ce n'était pas la priorité à ce moment.

C'est après une pause dans le projet que nous avons pu le reprendre d'un regard nouveau et l'aborder avec un outil que nous maitrisons mieux pour générer du texte : bash. Par un découragement général nous avons presqu'involontairement trouvé ce dont nous avions besoin pour nous lancer avec efficacité dans le projet : un appui solide mais rapide à construire.

2 Les premiers essais

2.1 Générer du C en bash

L'objectif ici n'était pas de réussir à faire quelque chose de fonctionnel, mais de comprendre ce qu'il nous fallait faire. Pour ce faire nous avons finalement décidé de le faire avec les langages que nous maitrisions le plus et que nous jugions les plus adaptés pour la situation : le bash et le C.

2.1.1 Le jeu de données

Nous avons créé manuellement des FST très basique, reliées à un dictionnaire, afin de disposer d'un jeu de test. Ils se trouvent en Appendice A.

Nous avons légèrement adapter le format défini par AT&T pour décrire des FST dans un fichier texte. Ce jeu de données servira pendant tout le projet en étant réadapté en Java par la suite.

2.1.2 Fonctionnement

Dans ce premier test nous prenons en entrée du script le fichier texte décrivant la FST, puis générons un code C qui permette de parcourir cette FST. Dans ce nouveau code point d'alorithme complexe : simplicité et naïvité sont ici ce que nous cherchons. Nous espérons alors que la pratique nous permettra de mieux comprendre le sujet. De plus nous faisons confiance à gcc pour optimiser le code à la compilation.

La première étape pour construire ce script est bien sûr de prévoir la forme qu'aura le code C généré : nous avons donc concentré nos premiers efforts à la génération d'une fonction contenant de multiples labels correspondant chacun à un état (ou nœud) et un switch qui, suivant la lettre courrante se déplace à la lettre suivante grâce à un goto qui pointe sur le label du state/node suivant. Nous returnons un code d'erreur si le token d'entrée n'est pas compatible avec la FST (le mot n'est pas pris en compte par celle-ci et n'a pas de code associé).

Cette fonction prend comme seul paramètre d'entrée une chaine (tableau de char) – qui sera le token d'entrée dont on veut connaitre la valeur – et retourne le poids cumulé de tous les arcs traversés, ou -1 en cas d'erreur. Il faut remarquer qu'avec cette méthode on ne peut pas supporter de poids négatifs, au risque d'avoir une collision entre le code d'erreur et un poids cumulé effectivement négatif. Pour gérer ce cas, le plus simple serait d'utiliser errno.

Le code a été un peu moins simple que prévu pour pouvoir générer du C valide, en effet, il y a un certain nombre de cas particuliers à prendre en compte afin de gérer correctement les erreurs ou de multiples états de fins.

Au final, un état générera un code semblable à celui présent dans le Listing 1 (pour un état appellé "7", qui possède un arc pour le caractère 'X' avec un poids de 6 et qui va au node "21", et un arc pour le caractère 'Y' avec un poids nulle et allant au node "42").

On remarque la présence d'un compteur "pos" incrémenté de manière inconditionnel, vu que l'on se déplace toujours un caractère par un caractère.

Pour pouvoir facilement lancer le programme généré, nous avons rajouté une fonction main qui appelle juste notre fonction compute_fst sur le premier argument de la command line, rendant ainsi le programme autonome.

```
NODE_7 :
    switch(token[pos++i]) {
        case 'X' :
            total += 6;
            goto NODE_21;
        case 'Y' :
            goto NODE_42;
        default :
            return -1;
    }
```

Listing 1 – Un example de code généré pour un état

Il nous suffit donc, pour générer quelque chose d'utilisable de faire ceci :

$$./gen.sh\ file.fst \mid gcc\ -x\ c\ -o\ fst -$$

Puis :

$$./fst\ LE\_TOKEN$$

Les options sur gcc (on peut aussi rajouter un -O3 pour optimiser au maximum la compilation) servant seulement de prendre l'entrée standard comme "fichier" source, puisque gen.sh affiche le code généré sur la sortie standard.

Les différents codes se situent en sous-section A.3, et plus précisément :
— code bash : Listing 3
— Premier example de code C obtenu : Listing 4
— Second example de code C obtenu : Listing 5

2.1.3 Les avantages

Si le résultat n'avait que peu d'importance ici, ce début à une importance capitale dans ce projet puisque c'est ce petit code qui nous a permis de mieux comprendre ce qui était attendu de nous, comment le faire, et comment utiliser une FST.

2.2 Générer de l'asmX86 en bash

Sachant qu'on devrait sûrement au final généré du bytecode, nous avons décidé que, quitte à avoir du C, autant aller jusqu'à générer directement de l'assembleur. Pas spécialement pour être plus performant que le C (en effet, l'assembleur généré par gcc sera toujours plus efficace que celui que l'on peut faire à la main), mais pour avoir des idées des problèmes que nous rencontrerons en bytecode.

Quand nous parlons d'assembleur, nous entendons "assembleur x86_64" bien sûr, soit l'assembleur qui est généré par gcc sur nos machines.

En donnant l'option "-S" à gcc, on obtient non pas un binaire éxécutable mais un fichier ".s" qui contient le code assembleur généré (avant l'assemblage effectif en binaire). En le générant pour nos sources C, nous avons pu faire du rétro-engineering dessus et comprendre la marche à suivre pour notre deuxième script.

La première partie, pour adapter toutes les parties générées de manière statique, a été relativement aisée. Par exemple, la déclaration de la fonction main, bien que plus longue et moins lisible qu'en C pouvait être plus ou moins copié/collé par rapport à ce que générait gcc, et même si quelques lignes restaient un peu mystérieuses au moment de la mise en place de "l'environnement" de la fonction, ce n'était absolument pas bloquant.

À l'opposé, lorsqu'il a fallu adapter les parties générées dynamiquement, ce fut beaucoup moins simple. Nous avons du, l'espace d'un instant, changer notre façon de programmer. En effet, l'assembleur est tellement bas niveau que l'on ne dispose pas de toutes les "syntactic sugar" dont on a tant l'habitude, nottamment pour le contrôle de flux. Par exemple, ce n'est pas si simple que ce à quoi nous pourrions nous attendre de faire un "if (...) single_instruction ;". Il faut gérer deux sauts, les labels associés, potentiellement préparer un ou deux registres, etc...

Ce fût très amusant à faire, et moins complexe qu'imaginé grâce à l'exhaustivité de la documentation. Néanmoins le temps passé dessus ne s'est révélé être aussi utile qu'escompté car nous le découvrirons plus tard les problèmes rencontrés pour le bytecode sont d'un ordre totalement différents.

Les différents codes se situent en annexe,annexe sous-section A.4, et plus précisément :
— code bash : Listing 6
— Premier example de code C obtenu : Listing 7
— Second example de code C obtenu : Listing 8

2.3 Adapter le travail précédent en Java

Ici le travail fait précédemment avec la génération d'un fichier C s'est montré être très utile. En effet cette étape a été traversée en quelques heures sans difficulté particulière.

Il nous suffit de séparer en trois cas l'étude d'un nœud et de créer une méthode particulière pour gérer intelligemment chaque cas. Le code est ajouté progressivement dans un StringBuffer avant d'être retourné (pour être imprimé en console ou dans un fichier).

Néanmoins cette étape ne fût absolument pas inutile puisqu'elle permit avec un cas simple de s'essayer à la génération d'une classe en Java. Seulement elle ne convenait pas tout à fait puisque nous avions recréé une manière de recréer des états et ce n'est pas ce que nous voulions.

Finalement nous avions obtenu le code présent en annexe dans le Listing 9.

2.4 Générer une structure de switch imbriqués en Java

Le code précédent nous permit de prendre confiance à propos de le génération de code en Java. Mais il s'agissait à présent de générer une structure en switch afin de s'approcher du sujet.

Cette fois nous avons rencontré de nombreuses difficultés à cause d'une problématique qui allait nous poursuivre pendant tout le projet : les switch ne sont pas voisins mais imbriqués les uns dans les autres. Ceci complexifia le problème mais nous avions finalement réussi grâce à un peu d'astuce et des souvenirs de nos cours de programmation en A1 à écrire une méthode récursive qui remplissait le StringBuffer au fur et à mesure et fermait en remontant dans la stack les blocs ouverts.

Il nous fallu aussi prendre en compte le cas supplémentaire suivant : un état final n'étant pas au bout de l'arbre de la FST (exemple : pour 'arbre' et 'arbres' le parcours du dit arbre est le même excepté que pour 'arbre' l'état final est sur le e). Nous n'y avions pas pensé auparavant et ce fut sources de bugs difficiles à identifier jusqu'à ce que nous comprenions notre erreur.

Le code final utilisé pour la génération se trouve dans le Listing 10.

3 Générer une FST

3.1 Le besoin

Avant de compiler nos FST, il nous faut bien sûr les construire à partir de données d'entrée et de données de sortie. Seulement pour faire des tests d'envergure nous ne pouvions plus les construire à la main. Il nous fallait les construire automatiquement.

Pour ce faire, nous avons décidé d'utiliser un algorithme de construction directe de "Minimal acyclic subsequential transducers". Il s'agit de l'algorithme utilisé dans le projet Apache Lucene, pour la construction de leurs FST. Il permet de construire des transducteurs minimums pour des entrées et sorties données. Nous pouvons donc avoir des données de la forme :

In	Out
John	1
Moth	2
Mob	3

Tableau 1 – Exemple de dictionaire

Note : Les mots en entrées doivent être triés dans l'ordre lexicographique pour que l'algorithme fonctionne.

Cet algorithme nous permet d'être certain que le FST construit associera de manière unique la sortie 1 au mot John, la sortie 2 au mot Moth et la sortie 3 au mot Mob. Et ce même pour des millions d'entrées de cette forme.

De plus, étant utilisé par Lucene, cela nous permettra d'avoir une équivalence entre les FST que nous construisons et ceux construits par Lucene, afin qu'il soit possible de comparer les performances de notre méthode et celle de Lucene.

3.2 Fonctionnement de l'algorithme

L'algorithme est issu d'un papier de recherche qui date de 2001, écrit par Denis Maurel et Stoyan Mihov. Il fonctionne pour tout type de données d'entrée/sortie, mais nous l'avons adapté en fixant pour sorties sont des nombres entiers par souci de simplicité (et nous le supposons de performance mais n'en avons pas de preuve). L'algorithme parcourt les mots en entrée un par un, et cherche un préfixe commun entre le mot actuel et le mot du tour de boucle précédent. En déterminant ce préfixe commun, il est ensuite possible de réutiliser des états qui auraient déjà été créés lors du traitement d'un mot précédent (un peu comme dans un jeu de Scrabble). Petit à petit, on arrive donc à construire un Trie correct et à s'assurer que la valeur de sortie de chaque chaîne est unique.

Direct Construction of Minimal Acyclic Subsequential Transducers (2001)

3.3 Difficultés

Il nous a été très difficile de comprendre le pseudo code et de l'implémenter correctement. Nous avons du lors des premiers essais traverser de très nombreuses erreurs qui ne nous permettaient pas d'avancer suffisemment. Finalement nous n'avons pas eu d'autre choix que de l'implémenter en tout trois fois, en essayant à chaque fois d'avoir plus de recul. Nous avons vraiment travaillé tous les trois

ensemble sur le compréhension (et non l'implémentation) par nécessité afin de pouvoir profiter des interprétations de tous pour pouvoir avancer.

Aujourd'hui malgré tous nos efforts l'algorithme n'est pas entièrement fonctionnel :
— Les états ne sont pas restitués dans le bon ordre (bug mineur)
— Les poids ne sont pas correctement distribués sur tous les arcs, particulièrement lorsque les états finaux ne sont pas en bout de branche (bug majeur).
Nous n'avons pas réussi à déterminer l'origine de ces bugs.

3.4 Résultat

Pour le dictionnaire suivant :

In	Out
A	1
Aani	2
Aaron	3
Aaronic	4
Aaronical	5
Aaronite	6
Aaronitic	7
Aaru	8
Ab	9
Ababdeh	10

Tableau 2 – Dictionnaire utilisé pour tester l'algorithme

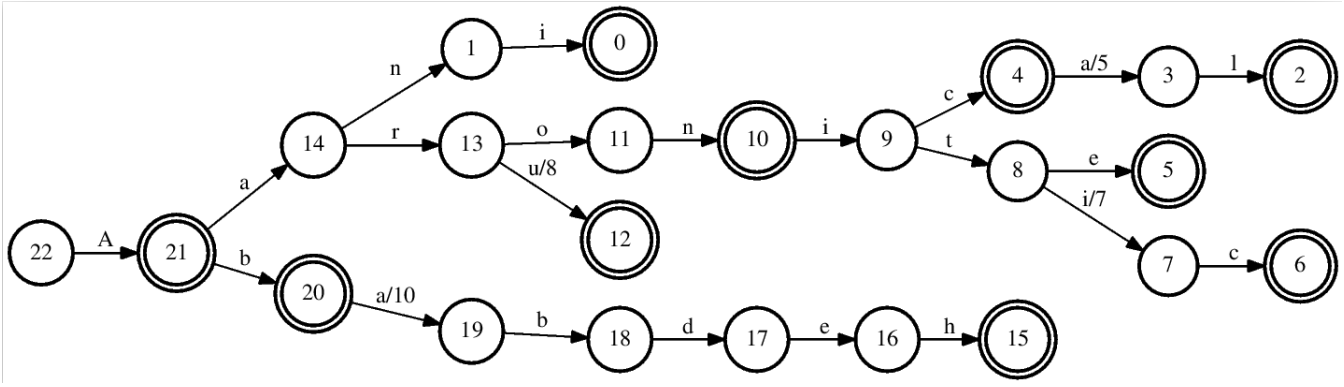

Figure 1 – La FST associée avec le Tableau 2 générée par l'algorithme

4 Idées d'amélioration et difficultés insurmontables

Nous disposions à présent d'un algorithme perfectible mais utilisable qui permettait de créer de grands FST. Nous avions aussi un moyen de générer du code Java pour parcourir la FST donnée en argument. Le gros du travail commençait : utiliser ces outils pour tester dans des conditions plus proches de celles réelles.

4.1 Idées originelles

4.1.1 L'encodage

Nous pensions au début du projet qu'il serait malin de changer l'encodage des caractères afin de contourner les mauvaises performances liées à l'UTF16. Seulement pour des besoins d'implémentation algorithmique nous avons très tôt pris la décision de travailler avec des entiers plutôt que les charactères unicodes. Cela nécessite une transformation de l'entrée avant de l'utiliser, mais cela aurait été nécessaire de toute manière.

Nous supposons en ayant fait ce choix avoir gagné en optimisation. Seulement il est regrettable que nous n'ayons pas réussi à implémenter l'alogorithme avec des charactères ou des string, car cela nous aurait permit de faire un véritable comparatif.

4.1.2 Le traitement d'un texte en multithreading

Nous avons très vite pensé à la manière par laquelle nous pourrions utiliser notre classe générée avec une grande quantité de texte en entrée. Nous avions comme idée de parcourir l'entrée afin d'y dinstinguer les mots (séparés par des espaces) et de lancer le parcourt de la FST sur différentes mots en même temps, dans différents thread. En effet, en divisant l'entrée et en prenant en compte les performances du Java en multi-threading il est imaginable que les performances soient améliorées. Hélas nous n'avons pas arrivé à cette étape à cause des difficultés à venir.

4.2 Les difficultés les plus grandes

4.2.1 Compilation à la volée

Afin d'automatiser le process de génération puis utilisation de classe, il est nécessaire de la compiler. Seulement cela n'est pas simple en Java. Nous avons essayé d'utiliser une des librairies fournies par Google pour le faire mais nous n'avons pas réussi à la faire fonctionner correctement.

De plus, même compilée il nous fallait ensuite utiliser la reflexivité pour utiliser la classe. Si bien que nous avions décidé de nous diriger vers le bytecode le plus vite possible afin de contourner ce problème de compilation.

4.2.2 Méthodes Java et la limitation 64Ko

Lorsque nous commencions à tester le code généré avec des entrées beaucoup plus grandes (un vrai dictionnaire avec quelques dizaines de millions d'entrée) notre première surprise fût la taille de la classe générée : 129Mo. Plutôt gros pour du fichier texte. Avant même d'aller plus loin nous étions très déçus de cette découverte sachant que dans le projet Lucène 8 millions de mots occupaient 90Mo.

Mais notre plus mauvaise surprise eu lieu lors de la compilation. Nous nous attendions à ce qu'il y ait une limite de la taile d'une classe ou méthode en Java et appréhendions le résultat de cette compilation. Mais lorsque le compilateur nous dit explicitement qu'une méthode ne pouvait pas dépasser 64Ko (ce que sera confirmé plus tard par de la lecture de documentions diverses) l'information était difficile à croire pour nous. Une grande désillusion. Et pourtant nous n'étions pas au bout de nos peines.

En effet, de par la très grande variété de caractères présents dans les textes aujourd'hui nous pouvions nous attendre à nous approcher de cette limite dès le premier switch, voir la dépasser avant même d'avoir écrit la première imbrication !

Nous avons décidé de continuer malgré cette problématique qui semblait sans solution viable dans tous les cas de figure afin de pouvoir étudier d'autres aspects du projet. Nous avons donc changé la génération afin qu'une méthode soit créée à chaque 'case' d'un switch. L'objectif est évident : réduire la taille des méthodes. Cette méthode utilisée est inspirée de celle du "trampolining" avec une différence très importance : nous n'avons pas un seul switch qui comporte beaucoup de cas différents. Nous avons une multitude de switch imbriqués les uns dans les autres comportant de très nombreux cas différents.

Le code est présent dans le Listing 11. Il fonctionne très bien pour des cas allant jusqu'à 100K mots : 1s pour 10K, 20s pour 30K, 1mn pour 50K ; puis à partir de 100K mots, le temps peut varier de 2 minutes à trop long, considéré infini. Nous n'avons pas trouvé l'origine de ce comportement et pire : nous n'avons pas la moindre idée de ce qui pourrait en être à l'origine sinon un problème de taille mémoire...

Mais nous étions malgré tout cela satisfait d'avoir su contourner ce problème. Jusqu'à ce qu'on étudie le bytecode de très près et comprenne le prix en Java de la création d'une méthode.

4.2.3 Le bytecode et la librairie asm

Ecriture manuelle du bytecode
L'objectif ici est de garder notre algorithme existant pour créer du java, mais en l'adaptant pour créer du bytecode. Par ailleurs nous ne nous intéresseons qu'au contenu de la méthode compute(), le reste étant juste l'initialisation de la classe.

Notes
La première lettre d'une mnémonique d'instruction bytecode est (dans la grande majorité des cas), le type sur lequelle l'instruction courante se base (i pour integer, f pour float, a pour referance, ...).

Dans la suite, on aura toujours 3 variables locales :
— 0 : int[] token, le token donné en argument de la fonction
— 1 : int pos, la position courante dans le token
— 2 : float result, le résultat final incrémenté lorsqu'on traverse un arc avec un poids non nul

Initialisation des variables
On avait, au début de la méthode :
— int pos=0 ;
— float result=0f ;

Il faut donc faire :
— iconst_0 // pousse 0 (integer) sur la stack
— istore_1 // pop la stack et enregistre la valeur dans la variable local 1 (pos)
— fconst_0 // pousse 0 (float) sur la stack
— fstore_2 // pop la stack et enregistre la valeur dans la variable local 2 (result)

Test d'overflow
Pour chaque state, on avait :

```
1 if( pos >= token.length ) {
2     return -1;
3 }
```

Listing 2 – Test sur la taille du token

On remplace cela par :
— iload_1 // push la valeur de "pos"
— aload_0 // push la reference de "token"
— arraylength // Pop la reference de "token" et en donne sa longeur
— if_cmpge (ERR) // Si (pos>=token.length) alors goto (ERR) (cf ci-dessous)

Erreurs
Lorsque l'on rencontre un cas d'erreur, on veut faire un : return -1 ;

On veut donc avoir quelque part :
— ldc #2 // push la valeur numéro 2 de la "constant pool"
— freturn // retourne la valeur au sommet de la stack
On a donc besoin que la seconde valeur de la constant pool soit "-1" (float). Si ce n'est pas la seconde valeur, il faut changer le #2

Notons qu'utiliser toujours la même addresse pour les cas d'erreurs est plus optimisé en terme de taille de la classe par rapport à ce qui est produit par javac

Switch On a de nombreux switch dans le java.

On a donc d'une part le switch, et d'autre part le pos++ :
— aload_0 // push la reference de "token"
— iload_1 // push la valeur de "pos"
— iinc 1 1 // incrémente de "1" la variable locale "1" (pos)
— iaload // pop les 2 dernière valeur de la stack pour push la i-ème (n-1) case du tableau (n-2), en l'occurence token[pos]

lookupswitch {
    (val_X) : (pos_label_X) // correspond aux "case"
    (val_Y) : (pos_label_Y) // exemple : "65 : 42" si le goto pour la lettre 'A' (65) est l'instruction numéro 42
    ...
    default : (ERR) // default : return -1
}

Incrément du total
Lorsque le poids d'un arc est != 0, on fait en java : result+=21.0f;

On va remplacer ça par :
fload_2 // push "result"
(
fconst_1 // si le poids est 1
OR
ldc #N // sinon, et il faut ajouter cela dans la constant pool si ce n'est pas déjà présent
)
fadd // additionne les deux dernières valeurs de la stack et push le result
fstore_2 // save le résultat

State finaux
Lorsque l'on est sur un state final, on fait en java : return (pos != token.length) ? -1 : result;

On va utiliser notre (ERR) pour factoriser le code :
— iload_1 // push la valeur de "pop"
— aload_0 // push la référence de "token"
— arraylength // pop et push la taille de "token"
— if_cmpne (ERR) // si (pos != token.length) goto (ERR)
— fload_2 // push la valeur de "result"
— freturn // retourne la valeur au sommet de la stack

Résultats
Malgré tous nos efforts nous n'avons pas réussi à implémenter la génération en bytecode. Nous n'avons pas réussi à maîtriser la bibliothèque ASM, par manque de documentation, d'expertise mais aussi de contrôle ! Nous avons été en effet extrêmement surpris d'avoir si peu de contrôle sur les variables ou sur les structures dans le bytecode, ce qui nous a beaucoup frustré.

Nous espérions que l'expérience accumulée en générant de l'assembleur nous serait utile mais elle ne l'a pas été car ce sont contre les outils même de manipulation de la JVM que nous nous sommes échoués. Il est étrange qu'il soit plus simple de générer de l'assembleur x86 que du code pour une Machine Virtuelle.

5 Analyse

A Annexe : tests avec un script shell

A.1 Dictionnaires

Value	Word
0	mop
1	moth
2	pop
3	star
4	stop
5	top

Tableau 3 – Dictionnaire à utiliser avec la FST dans le Tableau 5

Value	Word
0	mop
1	moth
2	pop
3	slop
4	sloth
5	stop
6	top

Tableau 4 – Dictionnaire à utiliser avec la FST dans le Tableau 6

A.2 FST

Nœu	0	0	0	0	1	2	3	2	4	6	7	5	7	8	
Nœu suivant	1	4	4	6	2	3	9	9	5	7	5	9	8	9	
Nœu final															9
Caractère	M	P	T	S	O	T	H	P	O	T	O	P	A	R	
Poids		2	5	3			1				1				

TABLEAU 5 – FST utilisée avec le dictionnaire Tableau 3, voir Figure 2 page 16

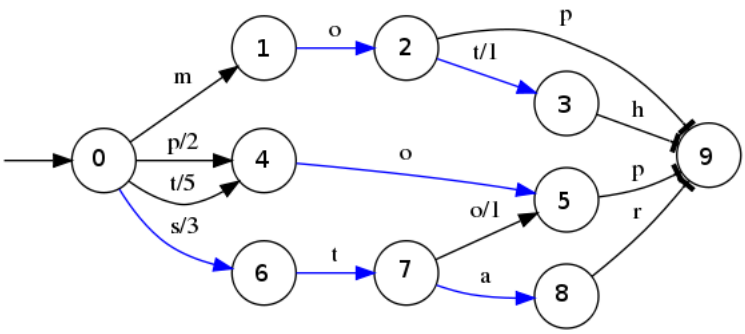

FIGURE 2 – La FST associée avec le Tableau 5 page 16

Nœu	0	0	0	0	3	3	1	2	4	5	6	5	
Nœu suivant	1	1	3	4	1	4	2	7	5	6	7	7	
Nœu final													7
Caractère	P	T	S	M	T	L	O	P	O	T	H	P	
Poids	2	6	3		2					1			

TABLEAU 6 – FST utilisée avec le dictionnaire Tableau 4, voir Figure 3 page 16

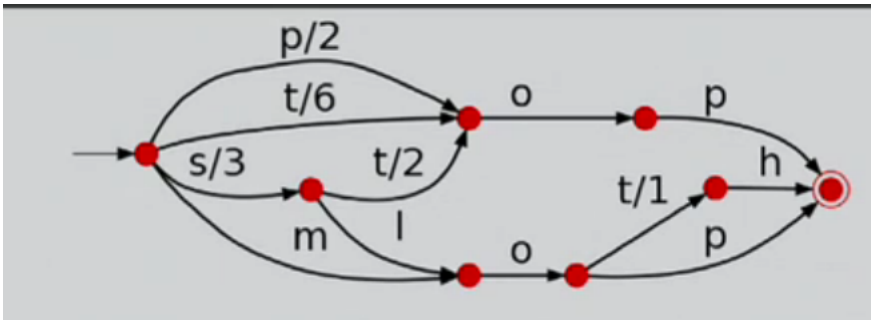

FIGURE 3 – La FST associée avec le Tableau 6 page 16

Les

A.3 Générer à la volée du code C

A.3.1 Script shell

```
#!/bin/bash

FST="$1"

sort "$FST" > "$FST.sort"

FIRST_CALL=1

cat <<EOF
#include <stdio.h>

int compute_fst(const char* token)
{
    int pos=0;
    int total=0;

EOF

while read DEP ARR CHAR WEIGHT ; do
    # If it's a final node
    if [[ ! "$CHAR" ]]; then
        WEIGHT=${ARR:-0}
        cat <<EOF
    default:
        return -1;
    }

NODE_$DEP :
EOF
    (( WEIGHT != 0 )) &&
        echo "    total += $WEIGHT;"
    echo "    if (token[pos] != '\\0') return -1;"
    echo "    goto END;"

        continue
    fi

    : ${WEIGHT:=0}

    if [[ $DEP != $PREV_DEP ]]; then
        if [[ ! "$FIRST_CALL" ]]; then
            cat <<EOF
    default:
        return -1;
    }

EOF
        fi

        cat <<EOF
NODE_$DEP :
    switch (token[pos++]) {
EOF
    fi

    echo "    case '$CHAR':"
```

```
    (( WEIGHT != 0 )) &&
        echo "        total += $WEIGHT;"
    echo "        goto NODE_$ARR;"

    PREV_DEP=$DEP
    FIRST_CALL=
done < "$FST.sort"

cat <<EOF

END :
    return total;
}

int main(int argc, const char *argv[])
{
    if (argc < 2)
        return 1;

    printf("%d\n", compute_fst(argv[1]));
    return 0;
}
EOF

rm "$FST.sort"
```

Listing 3 – Script pour générer un code C à la volée d'une FST

A.3.2 Code C généré pour la FST définie dans le Tableau 5

```c
#include <stdio.h>

int compute_fst(const char* token)
{
    int pos=0;
    int total=0;

NODE_0 :
    switch (token[pos++]) {
    case 'M':
        goto NODE_1;
    case 'P':
        total += 2;
        goto NODE_4;
    case 'T':
        total += 5;
        goto NODE_4;
    case 'S':
        total += 3;
        goto NODE_6;
    default:
        return -1;
    }

NODE_1 :
    switch (token[pos++]) {
    case 'O':
        goto NODE_2;
    default:
        return -1;
    }

NODE_2 :
    switch (token[pos++]) {
    case 'T':
        total += 1;
        goto NODE_3;
    case 'P':
        goto NODE_9;
    default:
        return -1;
    }

NODE_3 :
    switch (token[pos++]) {
    case 'H':
        goto NODE_9;
    default:
        return -1;
    }

NODE_4 :
    switch (token[pos++]) {
    case 'O':
        goto NODE_5;
    default:
        return -1;
    }
```

```
59
60  NODE_5 :
61      switch (token[pos++]) {
62      case 'P':
63          goto NODE_9;
64      default:
65          return -1;
66      }
67
68  NODE_6 :
69      switch (token[pos++]) {
70      case 'T':
71          goto NODE_7;
72      default:
73          return -1;
74      }
75
76  NODE_7 :
77      switch (token[pos++]) {
78      case 'O':
79          total += 1;
80          goto NODE_5;
81      case 'A':
82          goto NODE_8;
83      default:
84          return -1;
85      }
86
87  NODE_8 :
88      switch (token[pos++]) {
89      case 'R':
90          goto NODE_9;
91      default:
92          return -1;
93      }
94
95  NODE_9 :
96      goto END;
97
98  END :
99      return total;
100 }
101
102 int main(int argc, const char *argv[])
103 {
104     if (argc < 2)
105         return 1;
106
107     printf("%d\n", compute_fst(argv[1]));
108     return 0;
109 }
```

Listing 4 – Code C généré pour la FST définie dans Tableau 5

A.3.3 Code C généré pour la FST définie dans le Tableau 6

```c
#include <stdio.h>

int compute_fst(const char* token)
{
    int pos=0;
    int total=0;

NODE_0 :
    pos++;
    switch (token[pos-1]) {
    case 'P':
        total += 2;
        goto NODE_1;
    case 'T':
        total += 6;
        goto NODE_1;
    case 'S':
        total += 3;
        goto NODE_3;
    case 'M':
        goto NODE_4;
    default:
        return -1;
    }

NODE_1 :
    pos++;
    switch (token[pos-1]) {
    case 'O':
        goto NODE_2;
    default:
        return -1;
    }

NODE_2 :
    pos++;
    switch (token[pos-1]) {
    case 'P':
        goto NODE_7;
    default:
        return -1;
    }

NODE_3 :
    pos++;
    switch (token[pos-1]) {
    case 'T':
        total += 2;
        goto NODE_1;
    case 'L':
        goto NODE_4;
    default:
        return -1;
    }

NODE_4 :
    pos++;
    switch (token[pos-1]) {
```

```
59      case 'O':
60          goto NODE_5;
61      default:
62          return -1;
63      }
64
65  NODE_5 :
66      pos++;
67      switch (token[pos-1]) {
68      case 'T':
69          total += 1;
70          goto NODE_6;
71      case 'P':
72          goto NODE_7;
73      default:
74          return -1;
75      }
76
77  NODE_6 :
78      pos++;
79      switch (token[pos-1]) {
80      case 'H':
81          goto NODE_7;
82      default:
83          return -1;
84      }
85
86  NODE_7 :
87      goto END;
88
89  END :
90      return total;
91  }
92
93  int main(int argc, const char *argv[])
94  {
95      if (argc < 2)
96          return 1;
97
98      printf("%d\n", compute_fst(argv[1]));
99      return 0;
100 }
```

Listing 5 – Code C généré pour la FST définie dans Tableau 6

A.4 Générer à la volée du code assembleur x86

A.4.1 Script shell

```
#!/bin/bash

FST="$1"
oIFS=$IFS
: ${ASM_SWITCH:=1}
(( ASM_SWITCH == 0 )) && unset ASM_SWITCH

DO_SWITCH=
: ${SWITCH_LIMIT:=2}

sort "$FST" > "$FST.sort"

FIRST_CALL=1

compute_switch() {
    ARR_ASCII=( $(awk 'BEGIN{for(n=0;n<256;n++)ord[sprintf("%c",n)]=n}
    /^'$DEP'/{print ord[$3]}' "$FST.sort" | sort) )
}

next() {
    PREV_DEP=$DEP
    FIRST_CALL=

    continue
}

cat <<EOF
        .file   "${FST%.*}.c"
        .text
        .globl  compute_fst
        .type   compute_fst, @function

compute_fst:
        pushq   %rbp            # remember old base pointer
        movq    %rsp, %rbp      # set new base pointer
        movq    %rdi, -24(%rbp) # put content of rdi (token) in -24(%rbp)
        movl    \$0, -8(%rbp)    # pos   (-8(%rbp)) = 0
        movl    \$0, -4(%rbp)    # total (-4(%rbp)) = 0

EOF

while read DEP ARR CHAR WEIGHT ; do
    # If it's a final node
    if [[ ! "$CHAR" ]]; then
        if [[ ${#tmp[@]} != 0 ]] ; then
            IFS=$'\n'
            echo "${tmp[*]}"
            IFS=$oIFS
            tmp=()
        fi

        WEIGHT=${ARR:-0}

        [[ ! "$DO_SWITCH" ]] &&
            cat <<EOF
        movl \$-1, %eax          # default : return -1
```

```
57          jmp .RET
58  EOF
59          echo -e ".NODE_$DEP:"
60
61      (( WEIGHT != 0 )) &&
62          echo "  addl  \$$WEIGHT, -4(%rbp)          # total += Weight"
63      echo "      jmp .END                # goto END"
64
65          continue
66      fi
67
68      : ${WEIGHT:=0}
69
70      # Change of node
71      if [[ $DEP != $PREV_DEP ]]; then
72          if [[ ! "$FIRST_CALL" && ! "$DO_SWITCH" ]]; then
73              cat <<EOF
74          movl \$-1, %eax           # default : return -1
75          jmp .RET
76
77  EOF
78          fi
79
80          if [[ ${#tmp[@]} != 0 ]] ; then
81              IFS=$'\n'
82              echo "${tmp[*]}"
83              IFS=$oIFS
84              tmp=()
85          fi
86
87          cat <<EOF
88  .NODE_$DEP:
89          addl    \$1, -8(%rbp)     # pos++
90          movl    -8(%rbp), %eax    # eax = pos
91          leaq    -1(%rax), %rdx    # rdx = pos - 1
92          movq    -24(%rbp), %rax   # load token (address) in rax
93          addq    %rdx, %rax        # rax = &(token[pos-1])
94          movzbl  (%rax), %eax      # eax = token[pos-1]
95
96  EOF
97          if [[ "$ASM_SWITCH" ]] ; then
98              DO_SWITCH=
99              compute_switch
100             (( ${#ARR_ASCII[@]} > SWITCH_LIMIT )) &&
101                 DO_SWITCH=1
102         fi
103     fi
104
105     printf -v CHAR_INT '%d' "\"$CHAR"
106
107
108     [[ ! "$DO_SWITCH" ]] &&
109         echo "  cmpl \$$CHAR_INT, %eax           # case '$CHAR'"
110
111     if (( WEIGHT != 0 || DO_SWITCH == 1 )) ; then
112         tmp+=(
113             ".NODE_${DEP}_$CHAR:"
114             "   addl  \$$WEIGHT, -4(%rbp)       # total += $WEIGHT"
115             "   jmp .NODE_$ARR"
116             ""
```

```
117             )
118
119         if [[ "$DO_SWITCH" ]]; then
120             [[ $DEP == $PREV_DEP ]] && next
121
122             echo "        subl    \$${ARR_ASCII[0]}, %eax        # eax -= '$(
                    printf '%c' "$CHAR")'"
123             # MAX - MIN
124             CHAR_RANGE=$(( ARR_ASCII[${#ARR_ASCII[@]}-1] - ARR_ASCII[0] ))
125
126             cat <<EOF
127         cmpl \$$CHAR_RANGE, %eax            # eax = '$(printf '%c' "$(printf "\
                x$(printf "%x" ${ARR_ASCII[${#ARR_ASCII[@]}-1]})")")' - '$(printf
                "\x$(printf "%x" ${ARR_ASCII[0]})")' (max - min)
128         ja .END
129         movq .NODE_${DEP}_SW(,%rax,8), %rax
130         jmp *%rax
131
132     .section .rodata
133 .NODE_${DEP}_SW:
134 EOF
135             j=0
136             for (( i = 0; i <= (ARR_ASCII[${#ARR_ASCII[@]}-1] - ARR_ASCII[0])
                    ; i++ )); do
137                 if (( ARR_ASCII[0] + i == ARR_ASCII[j] )) ; then
138                     echo "        .quad .NODE_${DEP}_$(printf "\x$(printf "%x" $
                        {ARR_ASCII[j]})")          # '$(printf "\x$(printf "%x"
                        ${ARR_ASCII[j]})")'"
139                     (( j++ ))
140                 else
141                     echo "        .quad .ERR                # '$(printf "\x$(
                        printf "%x" $(( ARR_ASCII[0] + i )))")'"
142                 fi
143             done
144
145             echo -e "    .text\n"
146         else
147             echo "        je .NODE_${DEP}_$CHAR"
148         fi
149     else
150         echo "  je .NODE_$ARR"
151     fi
152
153     next
154 done < "$FST.sort"
155
156 cat <<EOF
157
158 .ERR:
159         movl \$-1, %eax           # return -1
160         jmp .RET
161 .END:
162         movl    -4(%rbp), %eax   # Put return value in eax
163 .RET:
164         popq    %rbp
165         ret
166         .size   compute_fst, .-compute_fst
167         .section        .rodata
168
169 .PRINTF_FMT:
```

```
170         .string "%d\n"
171         .text
172         .globl  main
173         .type   main, @function
174 main:
175         pushq   %rbp
176         movq    %rsp, %rbp
177         subq    \$16, %rsp
178         movl    %edi, -4(%rbp)
179         movq    %rsi, -16(%rbp)
180         cmpl    \$1, -4(%rbp)         # if (argc < 2)
181         jg      .DO_MAIN
182         movl    \$1, %eax             # return 1;
183         jmp     .END_MAIN
184 .DO_MAIN:
185         movq    -16(%rbp), %rax      # rax = &(argv[0])
186         addq    \$8, %rax             # rax = &(argv[1])
187         movq    (%rax), %rax         # rax = argv[1]
188         movq    %rax, %rdi           # rdi = argv[1]
189         call    compute_fst
190         movl    %eax, %esi           # esi = compute_fst()
191         movl    \$.PRINTF_FMT, %edi
192         movl    \$0, %eax
193         call    printf
194         movl    \$0, %eax
195 .END_MAIN:
196         leave
197         ret
198         .size   main, .-main
199         .ident  "GCC: (GNU) 4.9.2 20150212 (Red Hat 4.9.2-6)"
200         .section        .note.GNU-stack,"",@progbits
201 EOF
202
203 rm "$FST.sort"
```

Listing 6 – Script pour générer un code assembleur x86 à la colée d'une FST

A.4.2 Code assembleur x86 généré pour la FST définié dans le Tableau 5

```
        .file   "3.c"
        .text
        .globl  compute_fst
        .type   compute_fst, @function

compute_fst:
        pushq   %rbp            # remember old base pointer
        movq    %rsp, %rbp      # set new base pointer
        movq    %rdi, -24(%rbp) # put content of rdi (token) in -24(%rbp)
        movl    $0, -8(%rbp)    # pos   (-8(%rbp)) = 0
        movl    $0, -4(%rbp)    # total (-4(%rbp)) = 0

.NODE_0:
        addl    $1, -8(%rbp)    # pos++
        movl    -8(%rbp), %eax  # eax = pos
        leaq    -1(%rax), %rdx  # rdx = pos - 1
        movq    -24(%rbp), %rax # load token (address) in rax
        addq    %rdx, %rax      # rax = &(token[pos-1])
        movzbl  (%rax), %eax    # eax = token[pos-1]

        cmpl $77, %eax          # case 'M'
        je .NODE_1
        cmpl $80, %eax          # case 'P'
        je .NODE_0_P
        cmpl $84, %eax          # case 'T'
        je .NODE_0_T
        cmpl $83, %eax          # case 'S'
        je .NODE_0_S
        movl $-1, %eax          # default : return -1
        jmp .RET

.NODE_0_P:
        addl  $2, -4(%rbp)      # total += 2
        jmp .NODE_4

.NODE_0_T:
        addl  $5, -4(%rbp)      # total += 5
        jmp .NODE_4

.NODE_0_S:
        addl  $3, -4(%rbp)      # total += 3
        jmp .NODE_6

.NODE_1:
        addl    $1, -8(%rbp)    # pos++
        movl    -8(%rbp), %eax  # eax = pos
        leaq    -1(%rax), %rdx  # rdx = pos - 1
        movq    -24(%rbp), %rax # load token (address) in rax
        addq    %rdx, %rax      # rax = &(token[pos-1])
        movzbl  (%rax), %eax    # eax = token[pos-1]

        cmpl $79, %eax          # case 'O'
        je .NODE_2
        movl $-1, %eax          # default : return -1
        jmp .RET

.NODE_2:
        addl    $1, -8(%rbp)    # pos++
```

```
59          movl    -8(%rbp), %eax   # eax = pos
60          leaq    -1(%rax), %rdx   # rdx = pos - 1
61          movq    -24(%rbp), %rax  # load token (address) in rax
62          addq    %rdx, %rax       # rax = &(token[pos-1])
63          movzbl  (%rax), %eax     # eax = token[pos-1]
64
65          cmpl $84, %eax           # case 'T'
66          je .NODE_2_T
67          cmpl $80, %eax           # case 'P'
68          je .NODE_9
69          movl $-1, %eax           # default : return -1
70          jmp .RET
71
72  .NODE_2_T:
73          addl  $1, -4(%rbp)       # total += 1
74          jmp .NODE_3
75
76  .NODE_3:
77          addl    $1, -8(%rbp)     # pos++
78          movl    -8(%rbp), %eax   # eax = pos
79          leaq    -1(%rax), %rdx   # rdx = pos - 1
80          movq    -24(%rbp), %rax  # load token (address) in rax
81          addq    %rdx, %rax       # rax = &(token[pos-1])
82          movzbl  (%rax), %eax     # eax = token[pos-1]
83
84          cmpl $72, %eax           # case 'H'
85          je .NODE_9
86          movl $-1, %eax           # default : return -1
87          jmp .RET
88
89  .NODE_4:
90          addl    $1, -8(%rbp)     # pos++
91          movl    -8(%rbp), %eax   # eax = pos
92          leaq    -1(%rax), %rdx   # rdx = pos - 1
93          movq    -24(%rbp), %rax  # load token (address) in rax
94          addq    %rdx, %rax       # rax = &(token[pos-1])
95          movzbl  (%rax), %eax     # eax = token[pos-1]
96
97          cmpl $79, %eax           # case 'O'
98          je .NODE_5
99          movl $-1, %eax           # default : return -1
100         jmp .RET
101
102 .NODE_5:
103         addl    $1, -8(%rbp)     # pos++
104         movl    -8(%rbp), %eax   # eax = pos
105         leaq    -1(%rax), %rdx   # rdx = pos - 1
106         movq    -24(%rbp), %rax  # load token (address) in rax
107         addq    %rdx, %rax       # rax = &(token[pos-1])
108         movzbl  (%rax), %eax     # eax = token[pos-1]
109
110         cmpl $80, %eax           # case 'P'
111         je .NODE_9
112         movl $-1, %eax           # default : return -1
113         jmp .RET
114
115 .NODE_6:
116         addl    $1, -8(%rbp)     # pos++
117         movl    -8(%rbp), %eax   # eax = pos
118         leaq    -1(%rax), %rdx   # rdx = pos - 1
```

```
119         movq    -24(%rbp), %rax # load token (address) in rax
120         addq    %rdx, %rax      # rax = &(token[pos-1])
121         movzbl  (%rax), %eax    # eax = token[pos-1]
122
123         cmpl $84, %eax          # case 'T'
124         je .NODE_7
125         movl $-1, %eax          # default : return -1
126         jmp .RET
127
128 .NODE_7:
129         addl    $1, -8(%rbp)    # pos++
130         movl    -8(%rbp), %eax  # eax = pos
131         leaq    -1(%rax), %rdx  # rdx = pos - 1
132         movq    -24(%rbp), %rax # load token (address) in rax
133         addq    %rdx, %rax      # rax = &(token[pos-1])
134         movzbl  (%rax), %eax    # eax = token[pos-1]
135
136         cmpl $79, %eax          # case 'O'
137         je .NODE_7_O
138         cmpl $65, %eax          # case 'A'
139         je .NODE_8
140         movl $-1, %eax          # default : return -1
141         jmp .RET
142
143 .NODE_7_O:
144         addl  $1, -4(%rbp)      # total += 1
145         jmp .NODE_5
146
147 .NODE_8:
148         addl    $1, -8(%rbp)    # pos++
149         movl    -8(%rbp), %eax  # eax = pos
150         leaq    -1(%rax), %rdx  # rdx = pos - 1
151         movq    -24(%rbp), %rax # load token (address) in rax
152         addq    %rdx, %rax      # rax = &(token[pos-1])
153         movzbl  (%rax), %eax    # eax = token[pos-1]
154
155         cmpl $82, %eax          # case 'R'
156         je .NODE_9
157         movl $-1, %eax          # default : return -1
158         jmp .RET
159
160 .NODE_9:
161         jmp .END                # goto END
162
163 .END:
164         movl    -4(%rbp), %eax  # Put return value in eax
165 .RET:
166         popq    %rbp
167         ret
168         .size   compute_fst, .-compute_fst
169         .section        .rodata
170
171 .PRINTF_FMT:
172         .string "%d\n"
173         .text
174         .globl  main
175         .type   main, @function
176 main:
177         pushq   %rbp
178         movq    %rsp, %rbp
```

```
179         subq    $16, %rsp
180         movl    %edi, -4(%rbp)
181         movq    %rsi, -16(%rbp)
182         cmpl    $1, -4(%rbp)        # if (argc < 2)
183         jg      .DO_MAIN
184         movl    $1, %eax            # return 1;
185         jmp     .END_MAIN
186 .DO_MAIN:
187         movq    -16(%rbp), %rax     # rax = &(argv[0])
188         addq    $8, %rax            # rax = &(argv[1])
189         movq    (%rax), %rax        # rax = argv[1]
190         movq    %rax, %rdi          # rdi = argv[1]
191         call    compute_fst
192         movl    %eax, %esi          # esi = compute_fst()
193         movl    $.PRINTF_FMT, %edi
194         movl    $0, %eax
195         call    printf
196         movl    $0, %eax
197 .END_MAIN:
198         leave
199         ret
200         .size   main, .-main
201         .ident  "GCC: (GNU) 4.9.2 20150212 (Red Hat 4.9.2-6)"
202         .section        .note.GNU-stack,"",@progbits
```

Listing 7 – Code assembleur x86 généré pour la FST définie dans Tableau 5

A.4.3 Code assembleur x86 généré pour la FST définié dans le Tableau 6

```
        .file   "3.c"
        .text
        .globl  compute_fst
        .type   compute_fst, @function

compute_fst:
        pushq   %rbp            # remember old base pointer
        movq    %rsp, %rbp      # set new base pointer
        movq    %rdi, -24(%rbp) # put content of rdi (token) in -24(%rbp)
        movl    $0, -8(%rbp)    # pos   (-8(%rbp)) = 0
        movl    $0, -4(%rbp)    # total (-4(%rbp)) = 0

.NODE_0:
        addl    $1, -8(%rbp)    # pos++
        movl    -8(%rbp), %eax  # eax = pos
        leaq    -1(%rax), %rdx  # rdx = pos - 1
        movq    -24(%rbp), %rax # load token (address) in rax
        addq    %rdx, %rax      # rax = &(token[pos-1])
        movzbl  (%rax), %eax    # eax = token[pos-1]

        cmpl $80, %eax          # case 'P'
        je .NODE_0_P
        cmpl $84, %eax          # case 'T'
        je .NODE_0_T
        cmpl $83, %eax          # case 'S'
        je .NODE_0_S
        cmpl $77, %eax          # case 'M'
        je .NODE_4
        movl $-1, %eax          # default : return -1
        jmp .RET

.NODE_0_P:
        addl  $2, -4(%rbp)      # total += 2
        jmp .NODE_1

.NODE_0_T:
        addl  $6, -4(%rbp)      # total += 6
        jmp .NODE_1

.NODE_0_S:
        addl  $3, -4(%rbp)      # total += 3
        jmp .NODE_3

.NODE_1:
        addl    $1, -8(%rbp)    # pos++
        movl    -8(%rbp), %eax  # eax = pos
        leaq    -1(%rax), %rdx  # rdx = pos - 1
        movq    -24(%rbp), %rax # load token (address) in rax
        addq    %rdx, %rax      # rax = &(token[pos-1])
        movzbl  (%rax), %eax    # eax = token[pos-1]

        cmpl $79, %eax          # case 'O'
        je .NODE_2
        movl $-1, %eax          # default : return -1
        jmp .RET

.NODE_2:
        addl    $1, -8(%rbp)    # pos++
```

```
59          movl    -8(%rbp), %eax   # eax = pos
60          leaq    -1(%rax), %rdx   # rdx = pos - 1
61          movq    -24(%rbp), %rax  # load token (address) in rax
62          addq    %rdx, %rax       # rax = &(token[pos-1])
63          movzbl  (%rax), %eax     # eax = token[pos-1]
64
65          cmpl $80, %eax           # case 'P'
66          je .NODE_7
67          movl $-1, %eax           # default : return -1
68          jmp .RET
69
70  .NODE_3:
71          addl    $1, -8(%rbp)     # pos++
72          movl    -8(%rbp), %eax   # eax = pos
73          leaq    -1(%rax), %rdx   # rdx = pos - 1
74          movq    -24(%rbp), %rax  # load token (address) in rax
75          addq    %rdx, %rax       # rax = &(token[pos-1])
76          movzbl  (%rax), %eax     # eax = token[pos-1]
77
78          cmpl $84, %eax           # case 'T'
79          je .NODE_3_T
80          cmpl $76, %eax           # case 'L'
81          je .NODE_4
82          movl $-1, %eax           # default : return -1
83          jmp .RET
84
85  .NODE_3_T:
86          addl  $2, -4(%rbp)       # total += 2
87          jmp .NODE_1
88
89  .NODE_4:
90          addl    $1, -8(%rbp)     # pos++
91          movl    -8(%rbp), %eax   # eax = pos
92          leaq    -1(%rax), %rdx   # rdx = pos - 1
93          movq    -24(%rbp), %rax  # load token (address) in rax
94          addq    %rdx, %rax       # rax = &(token[pos-1])
95          movzbl  (%rax), %eax     # eax = token[pos-1]
96
97          cmpl $79, %eax           # case 'O'
98          je .NODE_5
99          movl $-1, %eax           # default : return -1
100         jmp .RET
101
102 .NODE_5:
103         addl    $1, -8(%rbp)     # pos++
104         movl    -8(%rbp), %eax   # eax = pos
105         leaq    -1(%rax), %rdx   # rdx = pos - 1
106         movq    -24(%rbp), %rax  # load token (address) in rax
107         addq    %rdx, %rax       # rax = &(token[pos-1])
108         movzbl  (%rax), %eax     # eax = token[pos-1]
109
110         cmpl $84, %eax           # case 'T'
111         je .NODE_5_T
112         cmpl $80, %eax           # case 'P'
113         je .NODE_7
114         movl $-1, %eax           # default : return -1
115         jmp .RET
116
117 .NODE_5_T:
118         addl  $1, -4(%rbp)       # total += 1
```

```
119         jmp .NODE_6
120
121 .NODE_6:
122         addl    $1, -8(%rbp)     # pos++
123         movl    -8(%rbp), %eax   # eax = pos
124         leaq    -1(%rax), %rdx   # rdx = pos - 1
125         movq    -24(%rbp), %rax  # load token (address) in rax
126         addq    %rdx, %rax       # rax = &(token[pos-1])
127         movzbl  (%rax), %eax     # eax = token[pos-1]
128
129         cmpl $72, %eax           # case 'H'
130         je .NODE_7
131         movl $-1, %eax           # default : return -1
132         jmp .RET
133
134 .NODE_7:
135         jmp .END                 # goto END
136
137 .END:
138         movl    -4(%rbp), %eax   # Put return value in eax
139 .RET:
140         popq    %rbp
141         ret
142         .size   compute_fst, .-compute_fst
143         .section        .rodata
144
145 .PRINTF_FMT:
146         .string "%d\n"
147         .text
148         .globl  main
149         .type   main, @function
150 main:
151         pushq   %rbp
152         movq    %rsp, %rbp
153         subq    $16, %rsp
154         movl    %edi, -4(%rbp)
155         movq    %rsi, -16(%rbp)
156         cmpl    $1, -4(%rbp)         # if (argc < 2)
157         jg      .DO_MAIN
158         movl    $1, %eax             # return 1;
159         jmp     .END_MAIN
160 .DO_MAIN:
161         movq    -16(%rbp), %rax      # rax = &(argv[0])
162         addq    $8, %rax             # rax = &(argv[1])
163         movq    (%rax), %rax         # rax = argv[1]
164         movq    %rax, %rdi           # rdi = argv[1]
165         call    compute_fst
166         movl    %eax, %esi           # esi = compute_fst()
167         movl    $.PRINTF_FMT, %edi
168         movl    $0, %eax
169         call    printf
170         movl    $0, %eax
171 .END_MAIN:
172         leave
173         ret
174         .size   main, .-main
175         .ident  "GCC: (GNU) 4.9.2 20150212 (Red Hat 4.9.2-6)"
176         .section        .note.GNU-stack,"",@progbits
```

Listing 8 – Code assembleur x86 généré pour la FST définie dans Tableau 6

B Annexe : générer du code java dynamiquement

B.1 Dans une seule méthode

B.1.1 Avec différents états représentés par des méthodes

```
1  import java.util.List;
2
3  class FstGenerator {
4
5      private StringBuffer strBuff;
6      private List<State> fstStates;
7
8      public FstGenerator(List<State> fstStates) {
9          this.fstStates = fstStates;
10         this.strBuff = new StringBuffer();
11     }
12
13     public StringBuffer compute() {
14         append("class FstCompute {");
15         generateMain();
16
17         for ( State state : fstStates ) {
18             generateStateCase(state);
19         }
20
21         append("}");
22
23         return strBuff;
24     }
25
26     private void generateMain() {
27         appendWithTab("public static float compute(int[] token) {"
               , 1);
28
29         appendWithTab("return node_0(token, 0, 0f);", 2);
30
31         appendWithTab("}", 1);
32     }
33
34     private void generateStateCase(State state) {
35         switch (state.getNumArcs()) {
36             case 0:
37                 generateLastStateCase(state);
38                 break;
39             case 1:
40                 generateStateWithOneArc(state);
41                 break;
42             default:
43                 generateGeneralState(state);
```

```
44                  break;
45              }
46          }
47
48          private void generateLastStateCase(State state) {
49              appendWithTab("private static float node_" + state.getId()
50                      + "(int[] token, int pos, float result) {", 1);
51              appendWithTab("return result;", 2);
52              appendWithTab("}", 1);
53          }
54
55          private void generateStateWithOneArc(State state) {
56              appendWithTab("private static float node_" + state.getId()
57                      + "(int[] token, int pos, float result) {", 1);
58
59              appendWithTab("if(pos>=token.length || token[pos]!=" +
60                      state.getArc(0).getIlabel() + ")",2);
61              appendWithTab("return -1;", 3);
62              appendWithTab("return node_" + state.getArc(0).
                  getNextState().getId()
63                      + "(token, pos+1, result+" +
64                      state.getArc(0).getWeight() + "f);", 2);
65
66              appendWithTab("}", 1);
67          }
68
69          private void generateGeneralState(State state) {
70              appendWithTab("private static float node_" + state.getId()
71                      + "(int[] token, int pos, float result) {", 1);
72              appendWithTab("if(pos>=token.length) {return -1;}", 2);
73              appendWithTab("switch(token[pos]) {", 2);
74              for (int i = 0; i < state.getNumArcs(); i++) {
75                  appendWithTab("case " + state.getArc(i).getIlabel()  +
                        ":", 3);
76              appendWithTab("return node_" + state.getArc(i).
                  getNextState().getId()
77                      + "(token, pos+1, result+" +
78                      state.getArc(i).getWeight() + "f);", 4);
79              }
80              appendWithTab("default:", 3);
81              appendWithTab("return -1;", 4);
82              appendWithTab("}",2);
83              appendWithTab("}", 1);
84          }
85
86          private void append(String strToAppend) {
87              strBuff.append(strToAppend);
88              strBuff.append("\n");
89          }
90
```

```
91      private void appendWithTab(String strToAppend, int numberOfTab
           ) {
92          for (int i = 0; i < numberOfTab; i++) {
93              strBuff.append("\t");
94          }
95          strBuff.append(strToAppend);
96          strBuff.append("\n");
97      }
98
99  }
```

Listing 9 – Code Java générant une structure avec des nœuds

B.1.2 Uniquement avec des switch

```
1  package generator;
2
3  import util.State;
4
5  public class FstGenerator {
6
7      private StringBuffer strBuff;
8
9      public FstGenerator() {
10     }
11
12     public StringBuffer compute(State initState, String className)
           {
13         strBuff = new StringBuffer();
14         System.out.println("[FstGenerator] Transforming Fst to a
               java class");
15         append("package generated;");
16
17         append("public class " + className + " {");
18
19         appendWithTab("public static float compute(int[] token) {"
               , 1);
20
21         appendWithTab("int pos=0;", 2);
22         appendWithTab("float result=0f;", 2);
23
24         generateCases(initState, 2);
25
26         appendWithTab("}", 1);
27
28         append("}");
29         System.out.println("[FstGenerator] Successfully
               transformed fst to " + className + ".java");
30         return strBuff;
31     }
32
33
34     private void generateCases(State currentState, int tab) {
35
36         if( currentState.getNumArcs() > 0) {
37             generateTokenLengthTest(tab);
38             appendWithTab("switch(token[pos++]) {", tab);
39             for (int i = 0; i < currentState.getNumArcs(); i++) {
40                 appendWithTab("case " + currentState.getArc(i).
                       getIlabel()  + ":", tab+1);
41                 if (currentState.getArc(i).getOlabel() != 0) {
42                     appendWithTab("result+=" + currentState.getArc
                           (i).getOlabel() + "f;", tab+2);
```

```
43                  }
44                  if (currentState.getArc(i).getNextState().
                        isFinalState()) {
45                      appendWithTab("if(pos==token.length) {return
                            result;}", tab+2);
46                  }
47                  generateCases(currentState.getArc(i).getNextState
                        (), tab+2);
48              }
49              appendWithTab("default:", tab+1);
50              appendWithTab("return -1;", tab+2);
51              appendWithTab("}", tab);
52          } else {
53              appendWithTab("return (pos!=token.length) ? -1 :
                    result;", tab);
54          }
55      }
56
57      private void generateTokenLengthTest(int tab) {
58          appendWithTab("if(pos>=token.length) {return -1;}", tab);
59      }
60
61      private void append(String strToAppend) {
62          strBuff.append(strToAppend);
63          strBuff.append("\n");
64      }
65
66      private void appendWithTab(String strToAppend, int numberOfTab
            ) {
67          for (int i = 0; i < numberOfTab; i++) {
68              strBuff.append("\t");
69          }
70          strBuff.append(strToAppend);
71          strBuff.append("\n");
72      }
73
74  }
```

Listing 10 – Code Java générant une structure switch

B.2 Dans différentes méthodes

```
1  package generator;
2
3  import util.State;
4  import java.util.List;
5  import java.util.ArrayList;
6
7  public class BigFstGenerator {
8
9      private StringBuffer strBuff;
10
11     public BigFstGenerator() {
12     }
13
14     public StringBuffer compute(State initState, String className)
           {
15         strBuff = new StringBuffer();
16         System.out.println("[FstGenerator] Transforming Fst to a
               java class");
17         append("package generated;");
18
19         append("public class " + className + " {");
20
21         appendWithTab("public static float compute(int[] token) {"
               , 1);
22
23         appendWithTab("int pos=0;", 2);
24         appendWithTab("float result=0f;", 2);
25         appendWithTab("return state_" + initState.getId() + "(
               token, pos, result);", 2);
26
27         appendWithTab("}", 1);
28
29         List<State> doneStates = new ArrayList<>();
30         generateCases(initState, doneStates);
31
32         append("}");
33         System.out.println("[FstGenerator] Successfully
               transformed fst to " + className + ".java");
34         return strBuff;
35     }
36
37
38     private void generateCases(State currentState, List<State>
           doneStates) {
39         List<State> nextStates = new ArrayList<>();
40
41         append("\n\tprivate static float state_" + currentState.
               getId() +
```

```
42                  "(int[] token, int pos, float result) {");
43
44          generateTokenLengthTest(2);
45          appendWithTab("switch(token[pos++]) {", 2);
46          for (int i = 0; i < currentState.getNumArcs(); i++) {
47              appendWithTab("case " + currentState.getArc(i).
                    getIlabel()  + ":", 3);
48
49              if (currentState.getArc(i).getOlabel() != 0) {
50                  appendWithTab("result+=" + currentState.getArc(i).
                        getOlabel() + "f;", 4);
51              }
52
53              if (currentState.getArc(i).getNextState().isFinalState
                    ()) {
54                  appendWithTab("if(pos==token.length) {return
                        result;}", 4);
55              }
56
57              if( currentState.getArc(i).getNextState().getNumArcs()
                    <= 0) {
58                  appendWithTab("return (pos!=token.length) ? -1 :
                        result;", 4);
59                  continue;
60              }
61
62              appendWithTab("return state_" + currentState.getArc(i)
                    .getNextState().getId() +
63                      "(token, pos, result);", 4);
64
65              nextStates.add(currentState.getArc(i).getNextState());
66          }
67          appendWithTab("default:", 3);
68          appendWithTab("return -1;", 4);
69          appendWithTab("}", 2);
70          appendWithTab("}", 1);
71
72          for (State next: nextStates) {
73              if (doneStates.contains(next)) {
74                  continue;
75              }
76              generateCases(next, doneStates);
77              doneStates.add(next);
78          }
79      }
80
81      private void generateTokenLengthTest(int tab) {
82          appendWithTab("if(pos>=token.length) {return -1;}", tab);
83      }
84
```

```
85      private void append(String strToAppend) {
86          strBuff.append(strToAppend);
87          strBuff.append("\n");
88      }
89
90      private void appendWithTab(String strToAppend, int numberOfTab
            ) {
91          for (int i = 0; i < numberOfTab; i++) {
92              strBuff.append("\t");
93          }
94          strBuff.append(strToAppend);
95          strBuff.append("\n");
96      }
97
98  }
```

Listing 11 – Code Java générant une structure switch avec fonctions