TEHNICI DE COMPILARE – CURSUL 9

Algoritm pentru depistarea circularității in gramaticile atributate

Fie gramatica atributată $GA = (G, Atr, R, B)$, unde $G = (N, \Sigma, S, P)$ este gramatica independentă de context de bază. Fie producția $p: X_0 \to X_1 \dots X_n$. Notăm cu $D_p = (V, E)$ graful de dependență asociat lui p, definit prin $V = \{a \in Atr(X_i) | 0 \leq i \leq n\}$, $(a, b) \in E$ dacă și numai dacă există o regulă semantică $b \leftarrow f(\dots, a, \dots)$ asociată lui p. Pentru un arbore de derivare T din G vom nota cu D_T graful orientat obținut prin combinarea grafurilor D_q, unde q este o producție aplicată în T.

Spunem că GA este bine definită dacă pentru orice arbore de derivare T asociat unei derivări din G, D_T, este aciclic.

```
Dependency − test(GA)
{
  for (X ∈ N ∪ Σ)
        F(X) ← {graful ca are ca varfuri atributele lui X si nicio muchie};
  do{
      change ← false;
     for (fiecare p: A → X1 … Xn ∈ P){
       for(G1 ∈ F(X1), … , Gn ∈ F(Xn)){
         D ← Dp;
         for (fiecare b → c din Gj, 1 ≤ j ≤ n)
                adauga muchia b → c in graful D;
          if(D are un ciclu) error( );
          else {
                  G ← graf nou cu noduri atributele lui A si nicio muchie;
                  for(fiecare b, c ∈ Atr(A))
                      if (exista drum in D de la b la c)
                          adauga muchia b → c la G;
                      if (G ∉ F(A))
                          { adauga G la F(A);
                            change ← true;}
              } // end_else
       } // end_for
     } // end_for
  }while(change == true) // end_do
} // end_Dependency_test
```

Gramatici $S-$ atributate

Definiție. O gramatică atributată care are toate atributele sintetizate se numește gramatică $S-$atributată.

Teoremă. Orice gramatică atributată este echivalentă cu o gramatică $S-$atributată.

Algoritm general de evaluare a atributelor într-o gramatică $S-$atributată

$postorder\ (n{:}\ node)$
{
  $for\ (fiecare\ fiu\ m\ al\ lui\ n, de\ la\ stanga\ la\ dreapta)$
    $postorder(m);$
  $evalueaza\ toate\ atributele\ asociate\ lui\ n;$
}

Evaluarea atributelor sintetizate intr-un parser de tip LR

- Fiecărui simbol X din stiva de lucru i se asociază valorile tuturor atributelor sintetizate (folosind, de exemplu, o structură de tip $union$).
- Atunci când se efectuează o reducere pentru producția $p{:}\ X_0 \to X_1 \dots X_n$, valorile sintetizate asociate neterminalului din membrul stâng, X_0, se vor calcula folosind aserțiunile (regulile de atributare) din $R(p)$, care sunt funcții ce depind de atributele asociate simbolurilor $X_1, \dots, X_n$ ale căror valori sunt pe stivă.
- După ce se face reducerea, se introduc pe stivă valorile asociate lui X_0.
- În $bison$, valorile atributelor sintetizate asociate lui $X_0, X_1, \dots, X_n$ sunt referite cu $$, respectiv $1, ..., $n.

Gramatici $L-$ atributate

Definiție. Spunem că gramatica atributată $GA = (G, A, R, B)$ (unde $G = (N, \Sigma, S, P)$ este gramatică independentă de context de bază) este $L-$atributată dacă pentru orice producție $X_0 \to X_1 \dots X_n \in P$ atributele moștenite asociate lui $X_i, 1 \leq\ i \leq n$, depind de:

- atributele moștenite și/sau sintetizate asociate lui $X_1, \dots, X_{i-1}$
- atributele moștenite asociate lui X_0.

Algoritm general de evaluare a atributelor unei gramatici L −atributate

$dfvisit\ (n{:}\ node)$
{
$for\ (fiecare\ fiu\ m\ al\ lui\ n, de\ la\ stanga\ la\ dreapta)$
{
$evalueaza\ atributele\ mostenite\ ale\ lui\ m;$
$dfvisit(m);$
}
$evalueaza\ atributele\ sintetizate\ ale\ lui\ n;$
}

Implementarea gramaticilor L-atributate într-un parser recursiv descendent

Fie gramatica atributată $GA = (G, A, R, B)$, unde $G = (N, \Sigma, S, P)$ este gramatica independentă de context de bază, de tip $LL(1)$.

Pentru fiecare neterminal $A \in N$ se construiește o funcție cu numele A care ca parametri transmiși prin valoare atributele moștenite și ca parametri transmiși prin referință atributele sintetizate.

Pe baza token-ului curent și a mulțimilor $First\big(\alpha.Follow(A)\big), A \to \alpha \in P$, se decide ce producție a lui A este aplicată (a se vedea algoritmul recursiv descendent pentru gramatici $LL(1)$ din cursul 4).

Codul asociat producției constă din:

- pentru x terminal (token) care are asociat atributul $\uparrow v$ a cărui valoare este furnizată de către analizorul lexical (*lexer* sau *scanner*), se salvează valoarea atributului într-o variabilă locală, după care se avansează la următorul token (se apelează scanner-ul).
- Pentru un neterminal B se apelează funcția corespendentă.
- Pentru o funcție de evaluare a atributelor, se introduce asignarea corespunzătoare.
- Pentru un predicat, se introduce instrucțiune if cu alternativa $error()$.
- Atributele sintetizate care nu apar în lista de parametri ai lui A sunt implementate ca variabile locale.

Implementarea gramaticilor L-atributate într-un parser de tip LR

În acest caz sunt necesare unele condiții suplimentare referitoare la producțiile gramaticii astfel încât atunci când se efectuează reducerile, atributele moștenite să poată fi calculate corect.

Gramatica se modifică în felul următor: pentru o producție în care există atribute moștenite, se introduc neterminali noi de forma $M \to \lambda$ (care nu introduc nimic nou din punct de vedere sintactic) exact înaintea unui neterminal A care are asociat un atribut moștenit, care sunt asociați în mod unic cu producția în care apar. Rolul acestor neterminali este de a furniza modul corect de calcul al atributului moștenit.

În $bison$, acești neterminali noi se numesc marcatori, nu apar în mod explicit.

Exemple

1) Se consideră gramatica atributată

$S \to a\ A \uparrow s\ C \downarrow s \uparrow t$
$S \to b\ A \uparrow s\ BC \downarrow s \uparrow t$
$C \downarrow f \uparrow t \to \quad [t \leftarrow g(f)]$

Neterminalul C moștenește atributul sintetizat de A. Atunci când facem reducerea $C \to c$, valoarea $C \downarrow s$ este fie în $val[top - 1]$, fie in $val[top - 2]$, unde val este un vector în care păstrăm valorile atributelor sintetizate asociate simbolurilor de pe stiva parserului $LR(1)$, iar top este indicele simbolului din vârful stivei.

Pentru a rezolva această ambiguitate, este adăugat înaintea lui C un neterminal nou M, asociat în mod unic cu a doua producție a gramaticii, care va moșteni atributul sintetizat de A.

$S \to a\ A \uparrow s\ C \downarrow s \uparrow t$
$S \to b\ A \uparrow s\ BM \downarrow s \uparrow v\ C \downarrow v \uparrow t$
$C \downarrow s \uparrow t \quad [t \leftarrow g(s)]$
$M \downarrow s \uparrow t \to \quad [t \leftarrow s]$

Atunci când se va face reducerea $M \to \lambda$, deoarece acest neterminal apare doar în producția $S \to bABMC$, atributul moștenit al lui M își va extrage valoarea de la atributul sintetizat aflat pe stivă cu 2 poziții înaintea lui. Astfel, atunci când se va face reducerea $C \to \lambda$, atributul moștenit asociat lui C va avea ca valoare atributul sintetizat al simbolului ce apare înaintea sa pe stivă, indiferent dacă C apare în producția $C \to aAC$ sau în $C \to bABMC$.

2) Fie producția

$S \to aA \uparrow sC \downarrow m \quad [m \leftarrow h(s)]$.

În acest caz introducem înaintea lui C marcatorul N, asociat unic cu producția $S \to aANC$, și producția $N \to \lambda$, astfel ca producțiile atributate să fie:

$S \to aA \uparrow sN \downarrow s_1 \uparrow v\ C \downarrow m \quad [s_1 \leftarrow h(s), m \leftarrow v]$

$N \downarrow s_1 \uparrow v \to [v \leftarrow s_1]$

Analiza sintactică și evaluarea atributelor moștenite într-un parser $LR(1)$

Input: gramatica atributată $GA = (G, A, R, B)$, unde $G = (N, \Sigma, S, P)$ este gramatica de bază de tip $\underline{LL(1)}$.

Output. Parser $LR(1)$ care calculează valorile tuturor atributelor în stiva asociată.

- Presupunem că fiecare neterminal X are asociat cel mult un atribut moștenit, m_X, și cel mult unul sintetizat, s_X.
- În cele ce urmează vom considera producția $A \to X_1 \dots X_n$ care va fi înlocuită cu și producțiile $A \to M_1 X_1 \dots M_n X_n, M_j \to \lambda$, unde $M_1, \dots, M_n$ sunt neterminali noi asociați în mod unic cu producția $A \to X_1 \dots X_n$.
- Asociem pe stivă fiecărui simbol (terminal sau neterminal) valoarea atributului sintetizat (dacă există).
- Atributul sintetizat s_{X_j}, dacă există, apare în stiva parserului LR asociat lui X_j, iar cel moștenit, m_{X_j}, dacă există, va apărea pe stivă asociat cu M_j. Producția atributată va fi $M_j \downarrow m_{M_j} \uparrow s_{M_j} \to [s_{M_j} \leftarrow f(m_{M_j})]$. Dacă atributul m_A există, $A \neq S$, atunci valoarea sa se găsește pe stivă, fiind asociată cu simbolul dinaintea lui M_1. Dacă $A = S$ are asociat un atribut moștenit, atunci valoarea acestuia trebuie plasată pe stivă inițial, înaintea începerii procesului de parsare, asociat eventual cu simbolul inițial al stivei (să ne reamintim că inițial pe stivă este plasat simbolul 0, corespunzător stării inițiale a AFD care recunoaște mulțimea prefixelor viabile, urmat de perechi $simbol, stare$, astfel că întotdeauna în vârful stivei apare o stare).
- Pentru reducerea numărului de marcatori: dacă X_j nu are asociat niciun atribut moștenit, atunci nu se introduce M_j. De asemeni, dacă m_{X_1} există și este egal cu m_A, atunci putem omite M_1.
- Atunci cînd se efectuează reducerea $M_j \to \lambda$, știm cu ce producție, unică, este asociat M_j, deci știm unde găsim valorile care intervin în calculul (funcția) valorii atributului asociat lui M_j, s_{M_j}. Deoarece gramatica este L-atributată, valoarea lui s_{M_j} este calculată în funcție de atributele sintetizate și/sau

moștenite care apar înaintea lui M_j în producția asociată, deci le regăsim pe stivă.

- Pentru reducerea $A \to M_1X_1 \dots M_nX_n$, vom calcula doar atributul sintetizat s_A, în funcție de atributele asociate lui $M_1, X_1, \dots, M_n, X_n$ care sunt pe stivă.
- Atributul m_A este deja plasat pe stivă, după reducere chiar înaintea lui A.

Propoziție. Dacă gramatica de bază $G = (N, \Sigma, S, P)$ a gramaticii atributate $GA = (G, A, R, B)$ este de tip $LL(1)$, atunci prin introducerea marcatorilor ca în algoritmul explicat mai sus se obține o gramatică $LR(1)$.

Marcatorii sunt neterminali din care se derivă doar λ și care apar o singură dată în corpul unei producții unice. Dacă gramatica de bază este $LL(1)$, putem să determinăm dacă șirul w din intrare este derivat din S într-o derivare care începe cu $S \to \alpha$ prin analiza primului simbol al lui w (sau a următorului, dacă $w = \lambda$). Astfel, dacă parsăm bottom-up pe w, atunci atunci faptul că un prefix al lui w poate fi redus la α și apoi la S este cunoscut îndată ce începutul lui w apare pe stivă. În particular, dacă inserăm marcatori oriunde în α, stările LR vor încorpora faptul că acel marcator trebuie să fie acolo, și se va face reducerea lui λ la marcatorul respectiv în momentul potrivit.

Exemplu de gramatică $LR(1)$ pentru care introducerea marcatorilor conduce la o gramatică care nu este $LR(1)$

Fie gramatica atributată cu producțiile:

$S \to L \downarrow c \quad [c \leftarrow 0]$

$L \downarrow c \to L \downarrow c_1 a \quad [c_1 \leftarrow c + 1]$

$L \downarrow c \to \lambda \quad [print\ c]$

Introducem marcatorul M, astfel că gramatica de mai sus devine:

$S \to L \downarrow c \quad [c \leftarrow 0]$

$L \downarrow c \to M \downarrow m \uparrow s\ L \downarrow s\ a \quad [m \leftarrow c + 1]$

$L \downarrow c \to \lambda \quad [print\ c]$

$M \downarrow m \uparrow s \to \lambda \quad [s \leftarrow m]$

Gramatica $S \to L,\ L \to La|\lambda$ este $LR(1)$, dar gramatica $S \to L, L \to MLa|\lambda, M \to \lambda$ nu este $LR(1)$.

Un exemplu mai complex de atributare a gramaticii sintactice pentru un mini-limbaj like-Pascal

$Z \rightarrow$ "PROGRAM" ID ";" "VAR" VH "BEGIN" B "END" "."
$V \rightarrow ID$ ":" T ";" V // declararea variabilelor
$\rightarrow \lambda$
$T \rightarrow$ "INTEGER"
$\rightarrow$ "BOOLEAN"
$H \rightarrow$ "FUNCTION" ID ":" T "VAR" VH "BEGIN" B "END" ";" H
$\rightarrow \lambda$
$B \rightarrow ID$ " := " E ";" B // corpul programului/functiei
$\rightarrow$ "RETURN" E // apare doar in corpul unei functii, la final
$\rightarrow \lambda$ // folosita doar in programul principal
$E \rightarrow SC$ // expresie
$C \rightarrow$ " = " S // comparatie (expresie booleana)
$\rightarrow \lambda$ // expresie simpla
$S \rightarrow FR$ // Factor urmat de Rest
$R \rightarrow$ " * " FR // expresie de tip INTEGER
$\rightarrow$ "AND" FR // expresie de tip BOOLEAN
$\rightarrow \lambda$
$F \rightarrow$ "(" E ")"
$\rightarrow ID\ A$ // referinta la variabila/functie
$\rightarrow NUM$ // constanta numerica
$\rightarrow$ "TRUE"
$\rightarrow$ "FALSE"
$A \rightarrow$ "(" ")" // referinta la functie
$\rightarrow \lambda$ // referinta la identificator

În acest mini-limbaj există funcții fără parametri, două tipuri de instrucțiuni (asignare și return), două tipuri pentru variabile și funcții (INTEGER, BOOLEAN). Referința la funcție se face prin adăugarea "()" după numele funcției.

Vom atributa această gramatică pentru a verifica dacă:

- variabilele sunt declarate înainte de utilizare
- operanzii care apar în expresii au tipuri compatibile între ele și cu operatorii
- tipul returnat de expresia ce apare în instrucțiunea $RETURN\ E$ coincide cu tipul declarat al funcției în corpul căreia se află.

Pentru aceasta vom implementa ca atribut moștenit tabela de simboluri, în care vom introduce pentru fiecare identificator numărul unic asociat de scanner și o valoare împachetată, $value = tip + ref$, unde:

$$tip = \begin{cases} 1, daca\ tipul\ este\ INTEGER \\ 2, daca\ tipul\ este\ BOOLEAN \end{cases}$$

$$ref = \begin{cases} 30, daca\ este\ referita\ o\ variabila \\ 40, daca\ este\ referita\ o\ functie \end{cases}$$

Astfel, dacă identificatorul $ab1$ are asignată valoarea 42, aceasta înseamnă că $ab1$ este o funcție care returnează o valoare de tip boolean.

Pentru a introduce identificatori în tabelă vom folosi funcția $into$:

$[into \downarrow symbtab \downarrow idn \downarrow value \uparrow newtab]$

Astfel, în tabela curentă, $symbtab$, se introduce identificatorul cu numărul unic asociat idn și valoarea $value$. Se obține o nouă tabelă, $newtab$.

Pentru a extrage identificatori din tabelă vom folosi funcția $from$:

$[from \downarrow symbtab \downarrow idn \uparrow value]$

Din tabela curentă, $symbtab$, se extrage pentru identificatorul cu numărul unic asociat, idn, valoarea asociată, $value$. Tabela rămâne neschimbată.

Vom considera o variabilă globală, lex, care nu apare explicit în gramatica atributată și care reprezintă nivelul lexical al variabilelor din tabelă. Astfel:

- variabilele și funcțiile declarate în programul principal au nivel lexical 0;
- variabilele și funcțiile declarate în funcțiile din programul principal au nivel lexical 1;
- ș.a.m.d.

Pentru limbaje precum C, C++ există doar două nivele lexicale, 0 pentru main și 1 pentru variabilele declarate în funcții.

Atunci când se trece la un nou nivel lexical se va utiliza funcția

$[open \downarrow symbtab \uparrow newtab]$

în care $newtab$ diferă de $symbtab$ doar prin nivelul lexical asociat, care este cu 1 mai mare decât nivelul lexical al lui $symbtab$. Aceste nivele lexicale diferite

permit ca identificatori cu același nume să fie utilizați la nivele lexicale diferite, deși este vorba de entități diferite (ca tipuri, adrese în memorie etc.).

Neterminalii gramaticii vor avea asociate atributele:

- $V \downarrow tabIn \uparrow tabOut$; V primește ca atribut moștenit tabela curentă, $tabIn$, și după ce se fac toate declarațiile de variabile se sintetizează $tabOut$.
- $T \uparrow type; type \in \{1,2\}; T$ este neterminalul pentru tip.
- $B \downarrow symbtab \downarrow typeR$; B este neterminalul pentru corpul ($body$) al programului principal sau al unei funcții. B primește ca atribut moștenit tabela de simboluri curentă, $symbtab$, și tipul întors de funcția respectivă (1 sau 2) dacă este utilizat pentru corpul unei funcții sau valoarea 0 când este utilizat pentru programul principal.
- $E \downarrow symbtab \uparrow type$; neterminalul pentru expresii, E, primește ca atribut moștenit tabela de simboluri curentă, $symbtab$, și sintetizează tipul expresiei respective (1 sau 2).
- $S \downarrow symbtab \uparrow type$; neterminalul S este folosit pentru introducerea unei expresii numerice și booleene.
- $C \downarrow symbtab \downarrow typeIn \uparrow typeOut$; neterminalul C este folosit pentru a introduce o expresie simplă ($C \rightarrow \lambda$) sau o comparație ($C \rightarrow$ " = "S); $typeIn$ este tipul moștenit de la primul termen al comparației, care este o expresie simplă.
- $R \downarrow symbtab \downarrow type$; neterminalul pentru restul expresiei, R, primește ca atribut moștenit tipul de la factorul anterior.
- $F \downarrow symbtab \uparrow type$; neterminalul pentru factor.
- $A \downarrow value \uparrow type; A$ primește ca atribut moștenit valoarea extrasă din tabelă pentru identificatorul ce apare înaintea lui A; sintetizează $type$.
- $H \downarrow tabIn \uparrow tabOut$;
- $ID \uparrow idn$; idn este numărul unic asociat de scanner identificatorului.

Gramatica atributată este:

$Z \rightarrow$ "PROGRAM" $ID \uparrow idn$ ";" "VAR" $V \downarrow \emptyset \uparrow tabnext$

$H \downarrow tabnext \uparrow tabH$ "BEGIN" $B \downarrow tabH \downarrow 0$ "END" "."

$V \downarrow tabIn \uparrow tabOut \rightarrow ID \uparrow idn$ ":" $T \uparrow type$ ";"
$[into \downarrow tabIn \downarrow idn + 30 \uparrow tabNext]$
$V \downarrow tabNext \uparrow tabOut$
$\rightarrow [tabOut \leftarrow tabIn]$

$T \uparrow type \rightarrow$ "INTEGER" $[type \leftarrow 1]$
$\rightarrow$ "BOOLEAN"$[type \leftarrow 2]$

$H \downarrow tabIn \uparrow tabOut \rightarrow$"FUNCTION"$\uparrow idn$ ": " $T \uparrow type$";"
$[into \downarrow tabIn \downarrow idn \downarrow type + 40 \uparrow tabNext]$
$[open \downarrow tabNext \uparrow newTab]$
"VAR" $V \downarrow newTab \uparrow vtab$ $H \downarrow vtab \uparrow bodyTab$
"BEGIN" $B \downarrow bodyTab \downarrow type$ "END" ";"
$H \downarrow tabNext \uparrow tabOut$ // acelasi nivel lexical ca inainte
// de $open$
$[tabOut \leftarrow tabIn]$

$B \downarrow symbtab \downarrow typeR \rightarrow ID \uparrow idn$ " := " $E \downarrow symbtab \uparrow type$ ";"
$[from \downarrow symbtab \downarrow idn \uparrow value]$
$[type == value - 30]$
$B \downarrow symbtab \downarrow typeR$
$\rightarrow$ "RETURN" $E \downarrow symbtab \uparrow type$
$[type == typeR]$
$\rightarrow [typeR == 0]$ // apare doar in progr principal

$E \downarrow symbtab \uparrow type \rightarrow S \downarrow symbtab \uparrow stype$ $C \downarrow symbtab \downarrow stype \uparrow type$

$C \downarrow symbtab \downarrow typeIn \uparrow typeOut \rightarrow$ " = "$S \downarrow symbtab \uparrow stype$
$[stype == typeIn; typeOut \leftarrow 2]$
$\rightarrow [typeOut \leftarrow typeIn]$

$S \downarrow symbtab \uparrow type \rightarrow F \downarrow symbtab \uparrow type$ $R \downarrow symbtab \downarrow type$

$R \downarrow symbtab \downarrow type \rightarrow$ " $*$ "$F \downarrow symbtab \uparrow ftype$
$[ftype == type; type == 1]$
$R \downarrow symbtab \downarrow ftype$
$\rightarrow$ "AND" $F \downarrow symbtab \uparrow ftype$
$[ftype == type; type == 2]$
$R \downarrow symbtab \downarrow ftype$
$\rightarrow \lambda$

$F \downarrow symbtab \uparrow type \rightarrow$ "(" $E \downarrow symbtab \uparrow type$ ")"
$\rightarrow ID \uparrow idn$
$[from \downarrow symbtab \downarrow idn \uparrow value]$
$A \downarrow value \uparrow type$
$\rightarrow NUM \uparrow val$ $[type \leftarrow 1]$
$\rightarrow$ "TRUE" $[type \leftarrow 2]$
$\rightarrow$ "FALSE" $[type \leftarrow 2]$

$A \downarrow value \uparrow type \rightarrow$ "(" ")" $[value\ div\ 10 == 4;\ type \leftarrow value - 40]$
$\rightarrow [value\ div\ 10 == 3;\ type \leftarrow value - 30]$

Observație. Gramatica mini-limbajului like-Pascal de mai sus este $LL(1)$, deci este și $LR(1)$. Prin adăugarea marcatorilor gramatica obținută este $LR(1)$ și deci este posibilă implementarea unui parser LR în care pot fi păstrate pe stivă și evaluate toate atributele, moștenite sau sintetizate.