Travail de redaction
Greibach normal form

Nguyen Tien Viet
SI 06

02/01/2007

Abstract

Greibach normal form plays a very important role in the theory of context-free language with many applications. In this small work, we're intersted in some extention of the theorem of Greibach normal form, which is a theorem about quadratic double Greibach normal form. The proof of this theorem is, in my opinion, very interesting since it is a very nice composant of several important results and technics in language theory.

1 Some definitions

By convention, all our context-free grammar is over an unique alphabet A, $A^{(n)} = \{w \in A^* \mid |w| \leq n\}$ and $A^n = \{w \in A^* \mid |w| = n\}$.

Definition 1 *A context-free grammar $G = (V, P, S)$ is proper if there isn't any ε-prodution or production of the form $U \to T$ where $U \in V$ and $T \in V$.*

Definition 2 *A context-free grammar $G = (V, P, S)$ is in Greibach normal form* (resp.right Greibach normal form) *if every production in P have its right-hand side in $A(A + V)^*$* (resp.$(A + V)^*A$).

Theoreme 1 *(Greibach normal form) Given a proper context-free grammar G, we can construct an equivalent grammar G' (which means $L_G = L_{G'}$) which is in Greibach normal form.*

We have made a quite nice proof of this very basic theorem in course,so there's no need for doing this again here.Insted, we will interest in some extensions of Greibach normal form, which we will define in a few moments.

Definition 3 *(Double Greibach normal form) A context-free grammar $G = (V, P, S)$ is in double Greibach normal form if every production in P have its right-hand side in $A(A + V)^*A$.*

Definition 4 *(Cubic* (resp.quadratique) *double Greibach normal form) A context-free grammar* $G = (V, P, S)$ *is in cubic* (resp.quadratique) *double Greibach normal form if every production inP have its right-hand side in* $A(A+V)^3A$ (resp.$A(A+V)^2A$).

Theoreme 2 *Given a proper context-free grammar* G*, we can construct an equivalent grammar* G *which is in cubic* resp.quadratique *Greibach normal form.*

2 Some prepairatory lemmas

Lemma 1 *(Theorem of quadratique(Chomsky) normal form) Given a proper context-free grammar* G*, we can construct an equivalent grammar* G' *which is in quadratique normal form.*

This lemma is a very fundamental theorem, since we have proved is in course, the proof of this theorem will not be given here.Thanks to this theorem, from now on we can assume that our given grammar is in quadratique normal form.

Lemma 2 *A proper* (resp.strict) *system of language equations S(that is a system of equations representing a proper* (resp.strict) *grammar G) has a unique solution, which is the language* L_G*.*

This lemma is used to prove the equivalence of 2 grammars by comparing the solutions of the 2 systems of equations representing 2 grammars. The proof of this lemma can be found in

Corrolair : Given a proper(or strict) context-free grammar $G = (V, P, S)$. Let T be a nonterminal in G.Let G' be the grammar obtained by replacing all occurence of T in the right-hand side of each production by all possible productions of T(productions with T in the left-hand side).Then G and G' are equivalent.

Example :For example let $G = (V, P)$ over alphabet $A = \{a, b\}$

$$V = \{S, A\}$$

$$S \longrightarrow SASA \qquad \text{(production in P)}$$

$$S \longrightarrow b \qquad \text{(production in P)}$$

$$A \longrightarrow S \qquad \text{(production in P)}$$

$$A \longrightarrow a \qquad \text{(production in P)}$$

Here let A play the role of T in the corrolair so we will replace each occurence of A in the right-hand side by S and a consecutively to get the new productions:

$$S \longrightarrow SSSS$$

$$S \longrightarrow SaSS$$

$$S \longrightarrow SSSa$$

$$S \longrightarrow SaSa$$
$$S \longrightarrow b$$

and get $G' = (V, P')$

$$V = \{S\}$$
$$S \longrightarrow SSSS$$
$$S \longrightarrow SaSS$$
$$S \longrightarrow SSSa$$
$$S \longrightarrow SaSa$$
$$S \longrightarrow b$$

which is equivalent to G.

*Proof.*The proof can be done in usual(*and heavy*) way,that is , by induction over the number of steps of derivation that $L_G \subset L_{G'}$ and $L_{G'} \subset L_G$.Here we prefer another way which makes use of lemma 2. We can suppose that $V = \{S, S_1, S_2, \ldots, S_n\}$ where S_1 will play the role of T in the corrolair. We consider the system representing the grammar G

$$S = \sum_{w \in (A+V)^*, S \to w \in P} w$$
$$S_i = \sum_{w \in (A+V)^*, S_i \to w \in P} w \quad \text{for each i from 1 to } n$$

Since G is proper(or strict), this system is proper(or strict). So by the lemma 2 it has an unique solution which is $(L_G(S), L_G(S_1), L_G(S_2), \ldots, L_G(S_n))$.

Hence by subtituing S_1 by $\sum_{w \in (A+V)^*, S_1 \to w \in P} w$ in the other equations we will get by defition a system of equations representing G'. This system is also proper(or strict). Again by the lemma 2 this new system have the unique solution $(L_{G'}(S), L_{G'}(S_1), L_{G'}(S_2), \ldots, L_{G'}(S_n))$.

By the property of subtituition we have $(L_G(S), L_G(S_1), L_G(S_2), \ldots, L_G(S_n))$ is also asolution of the new system. By unicity we have $L_G(S) = L_{G'}(S)$.

□

Lemma 3 *Given L be a rational language over an alphabet B. Let a be an element of B ,then $a^{-1}L, La^{-1}$ are also rational language and the set $\mathcal{L} = \{u_{-1}Lv{-}1 \mid u, v \in A^*\}$ is finite.*

The proof of this lemma is given in and will not be given here since rational languages isn't the subject of this work.

3 Proof of the theorem 2

We will prove theorem 2 by showing how to construct an equivalent cubic *(resp.quadratique)* Greibach normal form from a given context-free grammar G which, thanks to lemma 1, can be assumed to be in quadratique normal form.The proof compose of 4 step, where after the third step, we have a cubic Greibach normal form, after the forth step we will have a quadratique Greibach normal form.

First step:New variable

Suppose that $G = (V, P, S)$. The main idea is, for each $T \in V$, to repalce the set of rules having T as the left-hand side by all the rule of the form $T \to aL$ with $L \in V^*$ and $a \in A$ and $S \xrightarrow[G]{*} aL$ without changing the language generated by the grammar. One problem is that these set of a and L can be infinite. Fortunately, this difficulty can be overcome since we have this lemma:

Lemma 4 *$\forall X \in V, a \in A$ we have*

$$L(X) = \{m \in V^* \mid \exists b \in A, \alpha \in V^* : X \xrightarrow[l]{*} \alpha \to bm\}$$
$$L(a, X) = \{m \in V^* \mid \exists \alpha \in V^* : X \xrightarrow[l]{*} \alpha \to am\}$$
$$R(X) = \{m \in V^* \mid \exists b \in A, \alpha \in V^* : X \xrightarrow[r]{*} \alpha \to mb\}$$
$$R(a, X) = \{m \in V^* \mid \exists \alpha \in V^* : X \xrightarrow[r]{*} \alpha \to mb\}$$

then $L(X), L(a, X), R(X), R(a, X)$ are rational languages.

Proof.

For proving $L(X)$ is rational, it's sufficient to show an non-determinist automate $\mathcal{A}(X)$ that recognizes the inversement of $L(X)$(that means set of all inversements of all words in $L(X)$).

Let's consider $\mathcal{A}(X) = (V, V, E, I(X), F)$ in which the set of states is V and over the alphabet V also.

The set of transitions E contain: (M, N, P) iff $M \to NP$ is a production in G and

The set of initial states is $I(X) = \{X\}$.

The set of final states is $F(X) = \{M \in V \mid \exists b \in A : M \to b \text{ is a production of } G\}$.

This automate will read a string from left to right, each transition will simulate a productions of G, and the state of automate is the leftmost variable in the corresponding leftmost derivation. Since $I = \{X\}$ the derivation begin from X and the automate accepts a word iff the leftmost variable in the correspondind leftmost derivation(which we call T) is in F so there must be an $b \in A$ such that $T \to b$ is a production. So

$$X \xrightarrow[l]{*} Tm \to bm$$

then we have $\mathcal{A}(X)$ recognizes the inversement of $L(X)$. So $L(X)$ is rational.

By changing a litle on $\mathcal{A}(X)$ we can obtain an automate which recognizes the inversment of $L(a, X)$. Let $\mathcal{A}(a, X) = (V, V, E, I(X), F(a))$ where

$$F(a) = \{M \in V \mid M \to a \text{ is a production of } G\}$$

By an almost similar reasoning with the last case with a small change that $\mathcal{A}(a, X)$ accepts m iff the leftmost variable T in the corresponding leftmost derivation is in $F(a)$, so by definition of $F(a)$ we must have $T \to a$ is aproduction. So

$$X \xrightarrow[l]{*} Tm \to am$$

which show that $\mathcal{A}(a, X)$ recognizes the inversement of $L(a, X)$. So $L(a, X)$ is rational.

Similarly we can prove that $R(X)$, $R(a, X)$ are rational. But beware, this time the derivation is the rightmost derivation, so the states of automate will be the rightmost variable in the corresponding rightmost derivation and this time the automates will directly recognize $R(X)$ and $R(a, X)$.

Now we can complete the proof of this lemma.

□

Thanks to this lemma, instead of consider a possibly infinite number of productions, we need only consider a finite classes of productions. To do that we consider

$$\begin{aligned}
\mathcal{L} &= \{L(a, X) \mid a \in A, X \in V\} \\
\mathcal{R} &= \{R(a, X) \mid a \in A, X \in V\} \\
\mathcal{K} &= \mathcal{L} \cup \mathcal{R} \\
\mathcal{H} &= \{U \mid \exists u, v \in V^*, \exists T \in \mathcal{K} : U = u^{-1}Tv^{-1}\} \\
&= \bigcup_{K \in \mathcal{K}} \{U \mid \exists u, v \in V^* : U = u^{-1}Tv^{-1}\}
\end{aligned}$$

We have $\mathcal{K}$ is finite since $\mathcal{L}, \mathcal{R}$ are finite. So by the lemma 3 we have $\mathcal{K}$ is a finite sets of rational languages and $\forall L \in \mathcal{K}, \forall X \in V$ we have $X^{-1}L, LX^{-1} \in \mathcal{K}$.

Now for each language $H \in \mathcal{H}$ we introduce a new variable called V_H and let $V_{\mathcal{H}}$ be the set of all new variables.

Second step:Construct new grammar

Consider the grammar $G_1 = (V \cup V_{\mathcal{H}}, P_1, S)$ The set of productions contains:

$X \longrightarrow a$ if $X \in V$, $a \in A$, $X \to a \in P$ (P is the set of producions of G)
$X \longrightarrow aV_L$ if $L = L(a, X)$, $a \in A$
$V_L \longrightarrow XV_{L'}$ if $L' = X_{-1}L$, $L \in \mathcal{H}$
$V_L \longrightarrow \varepsilon$ if $\varepsilon \in L$, $L \in \mathcal{H}$

We eliminate the $\varepsilon - rule$ by adding the new rules in which the variable having $\varepsilon - rule$ is obmited, that means replace all the rules of the last type by the new rules:

$$V_L \longrightarrow X \text{ if } \varepsilon \in X^{-1}L,\ L \in \mathcal{H}$$

with the remark that each of the rules

$$X \longrightarrow a \text{ if } \varepsilon \in L(a, X),\ a \in A$$

is equivalent to a rule of the first type, so we don't need to add again to our grammar. Now we need to prove that $L_G = L_{G_1}$. Let's consider the system of language equations that represent G_1

$$\begin{aligned} X &= \sum_{X \to a \in P} a + \sum_{a \in A} a V_{L(a,X)} && (X \in V) \\ V_L &= \sum_{X \in V, \varepsilon \in X^{-1}L} X + \sum_{X \in V} X V_{X^{-1}L} && (L \in \mathcal{H}) \end{aligned}$$

This system is proper, applying the lemma 2 we have the system have a unique solution. We can test that

$$\begin{aligned} L_{G_1}(X) &= L_G(X) && (X \in V) \\ L_{G_1}(V_L) &= \sum_{w \in L} L_G(w) && (L \in \mathcal{H}) \end{aligned}$$

is the solution of the system representing G_1. In fact:

$$\begin{aligned} L_{G_1}(X) &= L_G(X) \\ &= \sum_{a \in A} a \sum_{w \in L(a,X)} L_G(w) \\ &= \underbrace{\sum_{a \in A, \varepsilon \in L(a,X)} a}_{\alpha} + \underbrace{\sum_{a \in A} a \left(\sum_{w \in L(a,X)} L_G(w) \right)}_{\beta} \quad (\text{since } \alpha \subset \beta) \\ &= \sum_{a \in A, X \to a \in P} a + \sum_{a \in A} a \sum_{w \in L(a,X)} L_G(w) \\ &= \sum_{a \in A, X \to a \in P} a + \sum_{a \in A} L_{G_1}(V_{L(a,X)}) \\ L_{G_1}(V_L) &= \sum_{w \in L} L_G(w) \\ &= \sum_{X \in V} \left(\sum_{u \in X^{-1}L} L_G(X) L_G(u) \right) \\ &= \sum_{X \in V} L_G \left(\sum_{u \in X^{-1}L} (X) L_G(u) \right) \\ &= \sum_{X \in V} L_{G_1}(X) L_{G_1}(V_{X^{-1}L}) \end{aligned}$$

So by unnicity $L_G = L_G(S) = L_{G_1}(S) = L_{G_1}$, then G and G_1 are equivalent.

Now we will make a little change.For each variable X of G_1, in each production that have X in the right hand-side we replace X by its productions, that means for each rule of the form $V_L \to XV_{X^{-1}L}$ we reaplace X by an element of the set $\{a \in A \mid X \to a \in P_1\} \cup \{aV_{L'} \mid X \to aV_{L'} \in P_1\}$ to get new rules. By applying the corrolair of the lemma 2 for each $X \in V$ we have the new grammar G_2 obtained after this change is equivalent to G and G_1. Moreover, this grammar is in Greibach normal form. For convinient we will call G_2 by $G^L = (V \cup V_{\mathcal{H}}, P^L, S)$ (L for left).

By the same way we can construct an equivalent grammar $G^R = (V \cup V_{\mathcal{H}}, P^R, S)$ which is in right Greibach normal form of G.Notice that G^L and GR have the same set of variables which is $V \cup V_{\mathcal{K}}$ and the same initial vaiable S(the same as the initial variable S of G)(R for right).

And we have also:

$$\begin{aligned} L_{G^R}(X) = L_{G^L}(X) &= L_G(X) \quad (X \in V) \\ L_{G^R}(V_L) = L_{G^L}(V_L) &= \sum_{w \in L} L_G(w) \quad (L \in \mathcal{H}) \end{aligned}$$

is the common solution of the systems of language equations that represent G^R and G^L.

Third step:Cubic Greibach normal form Now the idea is to replace each rightmost variable $V_L (L \in \mathcal{H})$ in the right-hand side of each production by its productions to get a new grammar $\mathbb{G} = (V \cup V_{\mathcal{H}}, \mathbb{P}, S)$ which is in cubic double Greibach normal form.Concretelyly we can write as follow:

If $X \overset{G^L}{\to} aV_L$ with $a \in A, L \in \mathcal{H}$ then we will replace V_L by each element of the set $\{w \in (V \cup V_{\mathcal{H}})^{(2)}A \mid V_L \overset{G^R}{\to} w\}$ to get new rules which have the right-hand side is an element in $A(V \cup V_{\mathcal{H}})^{(2)}A$.

If $V_L \overset{G^L}{\to} aV_KV_H$ with $a \in A$, $L, H, K \in \mathcal{H}$ then we will replace VH by each element of the set $\{w \in (V \cup V_{\mathcal{H}})^{(2)}A \mid V_H \overset{G^R}{\to} w\}$ to get new rules which have the right-hand side is an element in $A(V \cup V_{\mathcal{H}})^{(3)}A$.

If $V_L \overset{G^L}{\to} aV_H$ with $a \in A$, $L, H, \in \mathcal{H}$ then we will replace VH by each element of the set $\{w \in (V \cup V_{\mathcal{H}})^{(2)}A \mid V_H \overset{G^R}{\to} w\}$ to get new rules which have the right-hand side is an element in $A(V \cup V_{\mathcal{H}})^{(2)}A$.

In fact,in those agument, we can take w just in $V_{\mathcal{H}}$ since elements of V never occur in the right-hand side of the productions of G^R.

The set of rules $\mathbb{P}$ contains only the 2 above kind of rules and the terminal rules of G (rules of the form $X \to a$ with $X \in V, a \in A$). This grammar is in cubic double Greibach normal form.Our work is to prove it is equivalent to G. To do this, we use again our usual methode: introduce the system of equation representing $\mathbb{G}$.This system is a proper system so it have an unique solution (lemma 2).This unique solution is, however, the same as the solution of the system that repesenting G^L and G^R.In fact,since the system representing $\mathbb{G}$ can be obtain from the system representing G^L by subtituing some ocurrence of V_L in the right-hand side of its equations by its expression in the system

representing G^R,and we have:

$$\forall L \in \mathcal{H} L_G(V_L) = \sum_{V_L \to w \in P^R} L_G(w) \qquad \text{(since the systems representing } G^L \text{ and } G^R \text{ have the same solution)}$$

So we must have

$$\begin{aligned} \sum_{X \to w \in \mathbb{P}} L_G(w) &= \sum_{X \to w \in P^L} L_G(w) \\ &= L_G(X) \qquad (\forall X \in V \cup V_{\mathcal{H}}) \end{aligned}$$

This fact means that

$$\begin{aligned} L_{\mathbb{G}}(X) &= L_G(X) \qquad (X \in V) \\ L_{\mathbb{G}}(V_K) &= L_G(K) \qquad (K \in V_H) \end{aligned}$$

is also the solution of the system of equations representing $\mathbb{G}$.

Forth step:Quadratique Greibach normal form

Now we introduce a set of new variables: $V_{\mathcal{H}\mathcal{H}} = \{V_{<L_1,L_2>} \mid L_1, L_2 \in \mathcal{H}\}$ and let:

$$\mathcal{G}^Q = (V \cup V_{\mathcal{H}} \cup V_{\mathcal{H}\mathcal{H}}, \mathcal{P}^Q, S)$$

in which we have the following rules in $\mathcal{G}^Q$:

–Each rules of the form $V_L \to aV_{L_1}V_{L_2}V_{L_3}b$ give raise to the rule of the form $V_L \to aV_{<L_1L_2>}V_{L_3}b$

–Each new variable have their set of rules:

$$V_{<L_1L_2>} \to aV_{<L_{11}L_{12}>}V_{<L_{21}L_{22}>}$$

if we have:

$$\begin{aligned} V_{L_1} &\xrightarrow{G^L} aV_{L_{11}}V_{L_{12}} \\ V_{L_2} &\xrightarrow{G^R} aV_{L_{21}}V_{L_{22}} \end{aligned}$$

–Copy all the remain rules of $\mathbb{G}$ (rules that are already in form quadratique double Greibach).

This new grammar is already in quadratique double Greibach normal form. So its representing system is proper and strict.By lemma **??** it has a unique solution. Now it's easy to check that:

$$\begin{aligned} L_{\mathcal{G}^Q}(X) &= L_G(X) \qquad X \in V \\ L_{\mathcal{G}^Q}(V_L) &= L_G(L) \qquad L \in \mathcal{H} \\ L_{\mathcal{G}^Q}(V_{<L_1L_2>}) &= L_G(L_1L_2) \qquad L_1, L_2 \in \mathcal{H} \end{aligned}$$

is this unique solution.

So $L_{\mathcal{G}^Q}(S) = L_G(S)$, which means that $\mathcal{G}^Q$ is the equivalent quadratique double Greibach normal form of G. Proof terminated.

□

4 Some comment

This proof is, in my opinion, very interesting because it make use of many fundamental methods and results in language theory(rational language, system of language equations, subtituition...), and compose them very nicely.

We first notice that we can obtain an equivalent left or right Greibach normal form of a given grammar in an easier way(for instant, the way we have done in the course).But in order to get an equivalent double Greibach normal form, this is not enough. The difficulty is that we can hardly find a strong relationship between the left and right Greibach normal form in order to combine it to a double Greibach normal form. Herewe have overcome this dificulty by considering the set $V_{\mathcal{H}}$, so that the left and right Greibach normal form have the same set of variables. This relation is quite strong, and in fact, strong enough for us to use.

Second, the method we have used to pass from cubic duble Greibach normal form to quadratique double Greibach normal form is a very common technique to get a quadratique form which can also be used to get an equivalent Chomsky normal form from a given grammar.

The last point, this is the method we have used to prove 2 grammars is equivalent to each other, considering their representing systems of equations. Using this method make the proof more clear and less heavy than using the method induction over the number of derivations.