CSE322
Normal forms:CNF & GNF

Lecture #28

Normal Forms

In this section, we shall introduce context-free grammars to two special forms, one is **Chomsky normal form**(CNF), the other is **Greiback normal form**(GNF).

Definition 5: A CFG G = (V, T, P, S) is in Chomsky normal form if for each product of G is of the forms

$A \rightarrow BC$, or $A \rightarrow a$.

Here, A, B and C are in V, and a is in T.

Definition 6: A CFG G = (V, T, P, S) is in Greiback normal form if for each product of G is of the form

$A \rightarrow a\alpha$, where $A \in V$, $a \in T$ and $\alpha \in V^*$.

ε-productions

An ε-production is of the form $A \to \varepsilon$.

But, if the language L generated by a grammar G does not include ε, we can modify the grammar G to an equivalent grammar G' so that there are no ε-productions in G'.

The algorithm to modify a grammar G to G' is as follows.

while (there is a an ε-production $A \to \varepsilon$ in G) do

remove the ε-production $A \to \varepsilon$ from G

If there is a production $B \to xAy$ from G,

then add a production $B \to xy$ to G

Example : Consider the grammar G

$$S \rightarrow 0T0 \mid 1T1$$

$$T \rightarrow 0T0 \mid 1T1 \mid \varepsilon$$

Then $L(G) = L = \{ x x^R \mid x \in T^*, \text{ and } x \neq \varepsilon \}$.

We can modify the above grammar to a grammar without ε-productions as the following and generating the same language L.

$$S \rightarrow 0T0 \mid 1T1 \mid 00 \mid 11$$

$$T \rightarrow 0T0 \mid 1T1 \mid 00 \mid 11$$

The above grammar is equivalent to the following grammar.

$$S \rightarrow 0S0 \mid 1S1 \mid 00 \mid 11$$

Useless variables

A is a useful variable if $A \Rightarrow^* w$ for some $w \in T^*$.

For a CFG G, we can modify the grammar G to an equivalent grammar G' so that there are no useless variables in G'.

The algorithm to modify a grammar G to G' is as follows.

First, let the set $S \leftarrow \emptyset$, flag $\leftarrow$ TRUE

while (flag) do // get useful variables

flag $\leftarrow$ FALSE

If $B \rightarrow \alpha \in T^*$ in G, or $B \rightarrow x_1 A_1 x_2 A_2 \ldots x_k A_k y$ in G, where $y, x_i \in T^*$ and $A_i \in S$, $i = 1,2, \ldots, k$,

then $S \leftarrow S \cup \{B\}$, flag $\leftarrow$ TRUE

The set **V \ S** is a set of useless variables.

Next, if there is a useless production

$B \rightarrow x_1A_1x_2A_2\ldots x_kA_ky$ in G, where y, $x_i \in T^*$ and one of B and A_i, i = 1,2, …, k, is in V \ S,

then remove the production $B \rightarrow x_1A_1x_2A_2\ldots x_kA_ky$ from G.

The algorithm to remove useless variables and useless productions from the grammar G to get an equivalent grammar G' without useless variables and useless productions.

Unit productions

A unit production is of the form $A \rightarrow B$, where A and B are variables.

A unit production basically is a redundant production. Therefore, we can eliminate the unit productions by the following method:

For each unit production $A \rightarrow B$, remove the production from the grammar, and add the following productions:

For each non-unit production $B \rightarrow w$ in G, add the production $A \rightarrow w$ to G.

By the above algorithms, we are able to obtain the following theorem.

Theorem 5 : For each CFL L without ε, there is a CFG G with no useless variables, ε-productions or unit productions such that L(G) = L.

Theorem 6 : For each CFL L without ε, there is a CFG G in Chomsky normal form such that L(G) = L, i.e., each production in G is of the form $A \to BC$, or $A \to a$,where A, B and C are in V, and a is in T.

Proof :

For each CFL L without ε, there is a CFG G with no useless variables, ε-productions or unit productions such that L(G) = L.

First step, for each production $A \to \alpha$ in G, if $|\alpha| = 1$, then $\alpha \in T$.

Otherwise, $|\alpha| > 1$, say $\alpha = X_1 X_2 \ldots X_k$, $k>1$.

If X_i, say i =1and $X_1 = a$, is a terminal and there is a production $B \to a$, then replace X_1 by B to get a new production $A \to BX_2 \ldots X_k$, and remove the production $A \to \alpha$.

If X_i, say i =1and X_1 = a, is a terminal and there is no production $B \rightarrow a$, then add a new variable C, and replace X_1 by C to get a new production $A \rightarrow CX_2 \ldots X_k$, remove the production $A \rightarrow \alpha$ and add a new production $C \rightarrow a$.

After the first step, all productions are of the form either

$A \rightarrow a$, or

$A \rightarrow \alpha$, where $|\alpha| > 1$ and $\alpha \in V^+$

Second step, for each production of the form $A \rightarrow \alpha$, where $|\alpha| > 1$, we know that $\alpha \in V^+$. If $|\alpha| = 2$, then we do not need to modify the production.

If $|\alpha| > 2$, say $\alpha = X_1X_2\ldots X_k$, k>2, then we need to introduce new variables $Y_1, Y_2, \ldots, Y_{k-2}$, and add the new productions as follows.

$$A \rightarrow X_1 Y_1$$
$$Y_1 \rightarrow X_2 Y_2$$

$$Y_i \rightarrow X_{i+1} Y_{i+1},\ i = 1, 2, \ldots, k-3,$$

$$Y_{k-3} \rightarrow X_{k-2} Y_{k-2}$$

$$Y_{k-2} \rightarrow X_{k-1} X_k$$

Therefore, we can modify the grammar to a new equivalent grammar in Chomsky normal form.

Theorem 7 : For each CFL L without ε, there is a CFG G in Greiback normal form such that L(G) = L.

Proof :

For each CFL L without ε, there is a CFG G'' with no useless variables, ε-productions or unit productions such that L(G'') = L.

By theorem 6, there is a CFG G'=(V', T, P', S') in Chomsky normal form such that L(G') = L(G'') = L.

Rewrite the k variables in V' with indices 1, 2, …, k. Let it be V = $\{A_1, A_2, \ldots, A_k\}$ and the start variable is A_1.

First step, modify the productions into the forms

$A_i \rightarrow a\alpha$, where $a \in T$ and $\alpha \in V^*$, or

$A_i \rightarrow A_j\,\alpha$, where $j > i$

To achieve the result, we start from A_1.

If there is a production $A_1 \rightarrow a\alpha$, where $a \in T$ and $\alpha \in V^*$, then we keep the production.

If there is a production $A_1 \rightarrow A_1\alpha$, then apply theorem 4 in section 4.2 to revise the left recursive production to a right recursive production until there is no A_1 recursive production.

Next we revise the A_2 productions until A_k.

A possible algorithm for this step is as follows.

for i=1 to k do

for j=1 to i –1 do

for each production $A_i \rightarrow A_j\alpha$ do

for each production $A_j \rightarrow \beta$ do

add production $A_i \rightarrow \beta\alpha$

remove production $A_i \rightarrow A_j\alpha$

for each production $A_i \rightarrow A_i\alpha$ do

add productions $B_i \rightarrow \alpha$ and $B_i \rightarrow \alpha B_i$

remove productions $A_i \rightarrow A_i\alpha$

for each production $A_i \rightarrow \beta$, where β does not begin with A_i do

add production $A_i \rightarrow \beta B_i$

After the first step, the productions are of the forms:

$A_i \rightarrow a\alpha$, where $a \in T$ and $\alpha \in V^*$,

$A_i \rightarrow A_j\, \alpha$, where $k \geq j > i$, or

$B_i \rightarrow \alpha$, where $\alpha \in (V \cup \{B_1, B_2, \ldots, B_{i-1}\})^*$.

The A_k production must be of the form $A_k \rightarrow a\alpha$, where $a \in T$ and $\alpha \in V^*$,

The B_i production must be of the form

$B_i \rightarrow A_j\, \alpha$, where $\alpha \in (V \cup \{B_1, B_2, \ldots, B_{i-1}\})^*$, or

$B_i \rightarrow a\alpha$, where $a \in T$ and $\alpha \in (V \cup \{B_1, B_2, \ldots, B_{i-1}\})^*$, Second step, modify each A_i production into the forms

$$A_i \to a\alpha, \text{ where } a \in T \text{ and } \alpha \in V^*.$$

To achieve the result, we start from A_{k-1} to modify the A_{k-1} productions so that the right side of the production starting with a terminal. Then proceed the same process until A_1.

A possible algorithm for this step is as follows.

for i=k –1 to 1 do

for j=k to i +1 do

for each production $A_i \to A_j \alpha$ do

for each production $A_j \to \beta$ do

add production $A_i \to \beta\alpha$

remove production $A_i \to A_j \alpha$

Third step, modify each B_i production into the forms

$B_i \to a\alpha$, where $a \in T$ and $\alpha \in V^*$.

A possible algorithm for this step is as follows.

for i=1 to k do

for each production $B_i \to A_j \alpha$ do

for each production $A_j \to \beta$ do

add production $A_j \to \beta\alpha$

remove production $B_i \to A_j \alpha$

Example : Consider the grammar G

$$S \rightarrow 0S0 \mid 1S1 \mid 00 \mid 11$$

An equivalent CFG in GNF is

$$S \rightarrow 0SZ \mid 1SY \mid 0Z \mid 1Y$$

$$Y \rightarrow 1 \qquad\qquad Z \rightarrow 0$$

An equivalent CFG in CNF is

$$S \rightarrow ZX \mid YW \mid ZZ \mid YY$$

$$X \rightarrow SZ \qquad\qquad W \rightarrow SY$$

$$Y \rightarrow 1 \qquad\qquad Z \rightarrow 0$$

Example : Convert the grammar G in CNF to a grammar in GNF :

$$A_1 \to A_1 A_2 | A_3 A_1$$
$$A_2 \to A_1 A_3 | 0$$
$$A_3 \to A_1 A_2 | 1$$

Solution:

First stage:

Step 1: Introduce variable B_1 to modify the productions $A_1 \to A_1 A_2 | A_3 A_1$ to the following productions:

$$A_1 \to A_3 A_1 | A_3 A_1 B_1$$

$$B_1 \to A_2 | A_2 B_1$$

Step 2: Modify the productions $A_2 \rightarrow A_1 A_3 \mid 0$ to the following productions:

$$A_2 \rightarrow A_3 A_1 A_3 \mid A_3 A_1 B_1 A_3 \mid 0$$

Step 3: Modify the productions $A_3 \rightarrow A_1 A_2 \mid 1$ to the following productions:

$$A_3 \rightarrow A_3 A_1 A_2 \mid A_3 A_1 B_1 A_2 \mid 1$$

Step 4: Introduce variable B_3 to modify the productions $A_3 \rightarrow A_3 A_1 A_2 \mid A_3 A_1 B_1 A_2 \mid 1$ to the following production:

$$A_3 \rightarrow 1 \mid 1 B_3$$

$$B_3 \rightarrow A_1 A_2 B_3 \mid A_1 B_1 A_2 B_3 \mid A_1 A_2 \mid A_1 B_1 A_2$$

After the first stage we have the following productions:

$$A_1 \rightarrow A_3 A_1 \mid A_3 A_1 B_1$$

$$B_1 \rightarrow A_2 \mid A_2 B_1$$

$$A_2 \rightarrow A_3 A_1 A_3 \mid A_3 A_1 B_1 A_3 \mid 0$$

$$A_3 \rightarrow 1 \mid 1 B_3$$

$$B_3 \rightarrow A_1 A_2 B_3 \mid A_1 B_1 A_2 B_3 \mid A_1 A_2 \mid A_1 B_1 A_2$$

Second stage:

Step 5: By A_3 productions $A_3 \rightarrow 1 \mid 1B_3$, modify the A_2 productions to the following productions:

$$A_2 \rightarrow 1A_1A_3 \mid 1B_3A_1A_3 \mid 1A_1B_1A_3 \mid 1B_3A_1B_1A_3 \mid 0$$

Step 6: Modify the A_1 productions to the following productions:

$$A_1 \rightarrow 1A_1 \mid 1B_3A_1 \mid 1A_1B_1 \mid 1B_3A_1B_1$$

After the second stage, we have the following productions:

$$A_1 \rightarrow 1A_1 \mid 1B_3A_1 \mid 1A_1B_1 \mid 1B_3A_1B_1$$

$$A_2 \rightarrow 1A_1A_3 \mid 1B_3A_1A_3 \mid 1A_1B_1A_3 \mid 1B_3A_1B_1A_3 \mid 0$$

$$A_3 \rightarrow 1 \mid 1B_3 \qquad\qquad B_1 \rightarrow A_2 \mid A_2B_1$$

$$B_3 \rightarrow A_1A_2B_3 \mid A_1B_1A_2B_3 \mid A_1A_2 \mid A_1B_1A_2$$

Third stage:

Step 7: Modify the B productions to the following productions:

$$A_1 \rightarrow 1A_1 \mid 1B_3A_1 \mid 1A_1B_1 \mid 1B_3A_1B_1$$

$$A_2 \rightarrow 1A_1A_3 \mid 1B_3A_1A_3 \mid 1A_1B_1A_3 \mid 1B_3A_1B_1A_3 \mid 0$$

$$A_3 \rightarrow 1 \mid 1B_3$$

$$B_1 \rightarrow 1A_1A_3 \mid 1B_3A_1A_3 \mid 1A_1B_1A_3 \mid 1B_3A_1B_1A_3 \mid 0$$

$$B_1 \rightarrow 1A_1A_3B_1 \mid 1B_3A_1A_3B_1 \mid 1A_1B_1A_3B_1 \mid 1B_3A_1B_1A_3B_1 \mid 0B_1$$

$$B_3 \rightarrow 1A_1A_2B_3 \mid 1B_3A_1A_2B_3 \mid 1A_1B_1A_2B_3 \mid 1B_3A_1B_1A_2B_3$$

$$B_3 \rightarrow 1A_1B_1A_2B_3 \mid 1B_3A_1B_1A_2B_3 \mid 1A_1B_1B_1A_2B_3 \mid 1B_3A_1B_1B_1A_2B_3$$

$$B_3 \rightarrow 1A_1A_2 \mid 1B_3A_1A_2 \mid 1A_1B_1A_2 \mid 1B_3A_1B_1A_2$$

$$B_3 \rightarrow 1A_1B_1A_2 \mid 1B_3A_1B_1A_2 \mid 1A_1B_1B_1A_2 \mid 1B_3A_1B_1B_1A_2$$