CS310 : Automata Theory 2019

Lecture 19: Chomsky normal form

Instructor: Ashutosh Gupta

IITB, India

Compile date: 2019-02-21

Normal forms

A class of objects may have multiple objects that have same thing.

Example 19.1

In the set of linear expressions, $x - x + y + 2$ *and* $y + 2$ *have same meaning.*

We prefer to handle simplified versions or normalized versions the objects.

We will look into a normal form of CFGs.

Chomsky normal form

Definition 19.1

A grammar $G = (N, T, P, S)$ is in Chomsky normal form if the production rules in P are in either of the following forms.

- $A \to BC$ *or* (two nonterminals at RHS)
- $A \to a$, (No ϵ derivations)

where $A, B, C \in N$ and $a \in T$.

Converting to Chomsky normal form

We can translate any grammar to a grammar in Chomsky normal form, while preserving (almost) the recognized language.

We may drop ϵ word

We will see a sequence of simplifications first.

- Eliminate useless symbols
- Eliminate ϵ productions
- Eliminate unit productions

Topic 19.1

Eliminate useless symbols

Eliminate useless symbols

Consider grammar $G = (N, T, P, S)$.

Definition 19.2

$X \in N \cup T$ is useful, if there is a derivation $S \overset{*}{\Rightarrow} \alpha X \beta \overset{*}{\Rightarrow} w$ for $w \in T^*$.

Definition 19.3

$X \in N \cup T$ is useless, it X is not useful.

It is clear that eliminating production rules involving useless symbols have no impact on the language of the grammar.

Let us define two classes of symbols to help us remove the useless symbols.

Generating and reachable symbols

Consider grammar $G = (N, T, P, S)$.

Definition 19.4

$X \in N \cup T$ *is generating, if* $X \stackrel{*}{\Rightarrow} w$ *for some* $w \in T^*$.

Definition 19.5

$X \in N \cup T$ *is reachable, if there is some derivation* $S \stackrel{*}{\Rightarrow} \alpha X \beta$.

Example 19.2

Consider grammar

$$S \to AB \mid a$$
$$A \to b$$

A is a generating symbol.

Exercise 19.1

Are the following symbols generating?

- a
- B
- S
- b

Example: removing non generating symbols

Example 19.3

If we drop production rules containing non-generating B from the following grammar

$$S \to AB \mid a$$
$$A \to b,$$

we obtain

$$S \to a$$
$$A \to b.$$

Exercise 19.2

Are the following symbols reachable in the above grammar?

- S
- A
- a
- b

Example: removing unreachable symbols

Example 19.4

If we drop production rules containing unreachable A and b from the following grammar

$$S \to a$$
$$A \to b,$$

we obtain

$$S \to a.$$

Note that removing non-generating symbol first is important otherwise no simplification would have triggered in

$$S \to AB \mid a$$
$$A \to b.$$

Removing useless symbols

Theorem 19.1

Let $G = (N, T, P, S)$ be a CFG such that $L(G) \neq \emptyset$. Let $G_1 = (V_1, T_1, P_1, S)$ be the grammar we obtain by the following steps:

1. *eliminate productions containing nongenerating symbols in G. Let G_2 be this new grammar.*
2. *eliminate productions containing unreachable symbols in G_2*

S is not eliminated.(why?)

Then G_1 has no useless symbols, and $L(G_1) = L(G)$.

Proof.

Suppose $X \in T_1 \cup N_1$.

Therefore $X \overset{*G}{\Longrightarrow} w$ for some $w \in T^*$.

Since all the symbols occurring in the derivation are generating, $X \overset{*G_2}{\Longrightarrow} w$.

Since X survives the second purge, there must be some $S \overset{*G_2}{\Longrightarrow} \alpha X \beta$.

Since all the symbols occurring in the derivation are reachable, $S \overset{*G_1}{\Longrightarrow} \alpha X \beta$.

...

Removing useless symbols

Proof(contd.).

Furthermore, every symbol in α and β is generating in G_2 otherwise they would have been purged in the first round. Therefore, for some x and y

$$\alpha X \beta \overset{*G_2}{\Longrightarrow} xwy$$

Since all in $\alpha X \beta$ reachable, all in $\alpha X \beta \overset{*G_2}{\Longrightarrow} xwy$ is reachable. Therefore,

$$\alpha X \beta \overset{*G_1}{\Longrightarrow} xwy$$

Therefore

$$S \overset{*G_1}{\Longrightarrow} \alpha X \beta \overset{*G_1}{\Longrightarrow} xwy$$

Therefore, X is useful. ...

Removing useless symbols

Proof(contd.).

Clearly $L(G_1) \subseteq L(G)$.(why?)

claim: $L(G) \subseteq L(G_1)$
Let $w \in L(G)$.

There must be a derivation

$$S \stackrel{*G}{\Longrightarrow} w.$$

Since all symbols in the above derivation is reachable and generating, the derivation is also of G_1. □

Exercise 19.3

Why the proof does not work when the eliminations are applied in the reverse order?

Algorithm for generating and reachable symbols

Exercise 19.4

Give an algorithm for finding generating symbols?

Exercise 19.5

Give an algorithm for finding reachable symbols?

Topic 19.2

Eliminate ϵ productions

Eliminating ϵ productions

$A \to \epsilon$ are called ϵ productions.

Like ϵ-NFA, ϵ productions do not add in expressive power, but provide a convenient device to write CFGs.

Let us see the method to remove ϵ productions in CFGs.

Nullable symbols

Consider grammar $G = (N, T, P, S)$.

Definition 19.6

$X \in N$ is nullable, if there is a derivation if $X \stackrel{*}{\Rightarrow} \epsilon$.

Example 19.5

Consider the following grammar

$$S \to AB$$
$$A \to aAA \mid \epsilon$$
$$B \to bBB \mid \epsilon$$

A is nullable.

Exercise 19.6

Are the following symbols nullable in the above grammar?

- B
- S
- a
- b

Exercise 19.7

Give an algorithm to find nullable symbols in a grammar.

Removing epsilon productions

We apply following two transformations.

1. Let $A \to Y_1 .... Y_k$ be a production rule and $Y_{i_1} ... Y_{i_m}$ are nullable.

   For each $SubNull \subseteq \{Y_{i_1}, ..., Y_{i_m}\}$, we add a rule $A \to Z_1 .. Z_k$, where

$$Z_i = \begin{cases} \epsilon & Y_i \in SubNull \\ Y_i & \text{otherwise.} \end{cases}$$

2. We remove all epsilon transitions.

Exercise 19.8

Can we reverse the above two operations?

Example: epsilon removal

Example 19.6

Consider the following grammar

$$S \to AB$$
$$A \to aAA \mid \epsilon$$
$$B \to bBB \mid \epsilon$$

Let us apply the first transformation on the production rule $B \to bBB$.
Both Bs are nullable symbols.
We introduce rules for all subset of nullable symbols. Therefore, we add

- $B \to b$
- $B \to bB$
- $B \to bBB$

Exercise 19.9

What rules will be added due the production rules for A and S?

Example: epsilon removal

Example 19.7

After the first transformation

$$S \to AB \mid A \mid B \mid \epsilon$$
$$A \to aAA \mid aA \mid a \mid \epsilon$$
$$B \to bBB \mid bB \mid b \mid \epsilon$$

Now we drop al ϵ productions.

$$S \to AB \mid A \mid B$$
$$A \to aAA \mid aA \mid a$$
$$B \to bBB \mid bB \mid b$$

Correctness of ϵ removal

Theorem 19.2

If G_1 be the grammar obtained from G after applying the transformations, then $L(G) - \{\epsilon\} = L(G_1)$.

Proof.

Skipped in the class, but it is part of the course. Read the book! □

Topic 19.3

Eliminate unit productions

Eliminating unit productions

Definition 19.7

$A \to B$ is called unit production, if $A, B \in N$.

The unit productions do not contribute in progress in derivations.

However, very useful in writing down CFGs.

Example 19.8

Consider the following grammar

$$I \to a \mid b \mid Ia \mid Ib$$
$$F \to I \mid (S)$$
$$M \to F \mid M \times F$$
$$S \to M \mid S + M$$

Let us see the method to remove unit productions in CFGs.

Eliminating unit productions by substitutions

Let us suppose we have production rules $A \to B$ and $B \to \alpha$.

We can drop $A \to B$ and add a new rule $A \to \alpha$.

Example 19.9

Consider the following grammar

$$I \to a \mid b \mid Ia \mid Ib$$
$$F \to I \mid (S)$$
$$M \to F \mid M \times F$$
$$S \to M \mid S + M$$

We can remove $F \to I$ and add all the productions of I and obtain

$$I \to a \mid b \mid Ia \mid Ib$$
$$F \to a \mid b \mid Ia \mid Ib \mid (S)$$
$$M \to F \mid M \times F$$
$$S \to M \mid S + M$$

Exercise 19.10

Apply the substitutions aggressively and obtain unit-production-free grammar?

Circular unit production rules

Example 19.10

Consider the following grammar

$$A \to B \mid a$$
$$B \to C$$
$$C \to A$$

Exhaustive substitutions will run into circular substitutions and will never finish

We need to be aware of the potential circularity and needs a systematic way of avoiding them

Unit pairs

Definition 19.8

(A, B) is a unit pairs if $A \stackrel{*}{\Rightarrow} B$ via only unit production rules.

Example 19.11

Consider the following

$$I \to a \mid b \mid Ia \mid Ib$$
$$F \to I \mid (S)$$
$$M \to F \mid M \times F$$
$$S \to M \mid S + M$$

(I, I), (F, F), (M, M), (S, S) are unit pairs.

(F, I) is unit pair.

(M, F) and (M, I) are unit pairs.

(S, M), (S, F), and (M, I) are unit pairs.

Exercise 19.11

Give an algorithm to compute all unit pairs.

Unit production elimination via unit pairs

Given a CFG $G = (V, T, P, S)$, construct CFG $G_1 = (V, T, P_1, S)$:

1. Identify all unit pairs in P
2. For each unit pair (A, B) and nonunit $B \to \alpha \in P$, add rule $A \to \alpha$ in P_1.

Exercise 19.12

Does P_1 contains all nonunit rules of P?

Example: substitutions via to unit pairs

Example 19.12

Consider the following

$$I \to a \mid b \mid Ia \mid Ib$$
$$F \to I \mid (S)$$
$$M \to F \mid M \times F$$
$$S \to M \mid S + M$$

Unit pair	*Production rules added*
(I, I)	$I \to a \mid b \mid Ia \mid Ib$
(F, F)	$F \to (S)$
(M, M)	$M \to M \times F$
(S, S)	$S \to S + M$
(F, I)	$F \to a \mid b \mid Ia \mid Ib$
(M, I)	$M \to a \mid b \mid Ia \mid Ib$
(M, F)	$M \to (S)$
(S, I)	$S \to a \mid b \mid Ia \mid Ib$
(S, F)	$S \to (S)$
(S, M)	$S \to M \times F$

Finally the grammar is

$$I \to a \mid b \mid Ia \mid Ib$$
$$F \to a \mid b \mid Ia \mid Ib \mid (S)$$
$$M \to a \mid b \mid Ia \mid Ib \mid (S) \mid M \times F$$
$$S \to a \mid b \mid Ia \mid Ib \mid (S) \mid M \times F \mid S + M$$

Correctness of transformation

Theorem 19.3

If grammar G_1 is obtained from grammar G by the algorithm described eariler for eliminating unit productions, then $L(G_1) = L(G)$.

Proof.

Skipped in the class, but it is part of the course! □

Order of eliminations

We must follow the following order of eliminations.

- Eliminate ϵ productions
- Eliminate unit productions
- Eliminate useless symbols

Exercise 19.13

Why later eliminations do not introduce earlier eliminated forms?

Exercise 19.14

Give examples when any other order of elimination reintroduces already eliminated forms?

Topic 19.4

Chomsky normal form

Chomsky normal form (CNF)

Definition 19.9

A grammar $G = (N, T, P, S)$ is in Chomsky normal form if the production rules in P are in either of the following forms.

- $A \to BC$ *or*
- $A \to a$,

where $A, B, C \in N$ and $a \in T$, and G has no useless symbols.

Getting to the CNF

Let us suppose we have applied our earlier eliminations, we will be left with the following two problems in $A \to \alpha \in P$

1. $|\alpha| > 1$ and α has terminals
2. $|\alpha| > 2$

Both can be avoided with simple transformations.

Exercise 19.15

How to handle the first problem?

Removing long derivations

Let $A \to B_1 \ldots B_k \in P$, where $k > 2$.

We replace the rule by the following chain of rules.

$$A \to B_1 C_1$$
$$C_1 \to B_2 C_2$$
$$\vdots$$
$$C_{k-2} \to B_{k-1} B_k$$

where $C_1, \ldots, C_{k-2}$ are fresh nonterminals.

Example: CNF step 1. removing terminals

Example 19.13

Consider our running example

$$I \to a \mid b \mid Ia \mid Ib$$
$$F \to a \mid b \mid Ia \mid Ib \mid (S)$$
$$M \to a \mid b \mid Ia \mid Ib \mid (S) \mid M \times F$$
$$S \to a \mid b \mid Ia \mid Ib \mid (S) \mid M \times F \mid S + M$$

After removing terminals from non-unit derivations

$$I \to a \mid b \mid IA \mid IB$$
$$F \to a \mid b \mid IA \mid IB \mid LSR$$
$$M \to a \mid b \mid IA \mid IB \mid LSR \mid MTF$$
$$S \to a \mid b \mid IA \mid IB \mid LSR \mid MTF \mid SPM$$
$$A \to a \qquad L \to (\qquad T \to \times$$
$$B \to b \qquad R \to) \qquad P \to +$$

Example: CNF step 1. removing long derivations

Example 19.14

After removing long derivations

$$I \to a \mid b \mid IA \mid IB$$
$$F \to a \mid b \mid IA \mid IB \mid LC_1$$
$$M \to a \mid b \mid IA \mid IB \mid LC_1 \mid MC_2$$
$$S \to a \mid b \mid IA \mid IB \mid LC_1 \mid MC_2 \mid SC_3$$
$$C_1 \to SR$$
$$C_2 \to TF$$
$$C_3 \to PM$$
$$A \to a \qquad L \to (\qquad T \to \times$$
$$B \to b \qquad R \to) \qquad P \to +$$

End of Lecture 19