CSE211-Formal Languages and Automata Theory

U2L11 – Chomsky Normal Form

Dr. P. Saravanan
School of Computing
SASTRA Deemed University

Agenda

- Recap of previous class
  - DPDA, Simplification of rule, Normal forms
- Eliminating unit productions
- Chomsky Normal Forms
- Converting to CNF

Eliminating Useless Symbols

- Eliminate useless symbols in a grammar with the following productions:
  - $S \rightarrow AB \mid a$
  - $A \rightarrow b$.
- *B* is *not generating*, and is so eliminated at first, resulting in $S \rightarrow a$, $A \rightarrow b$, in which *A* is *not reachable*
- and so eliminated too, with $S \rightarrow a$ as the only production left.
- The order of eliminations is *essential*: *eliminate non-generating symbols at first*.

Eliminating ε -Productions

- Given a grammar with productions as follows:

  $S \rightarrow AB$

  $A \rightarrow aAA \mid \varepsilon$

  $B \rightarrow bBB \mid \varepsilon$

- then, we can see the following facts:
  - *A* and *B* are *nullable* because they derive empty strings;
  - *S* is also *nullable* because *A* and *B* are nullable.

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

Eliminating unit productions

- **Definition** --- a unit production is of the form $A \rightarrow B$.
- Unit productions sometimes are useful.
- For example, use of unit productions $E \rightarrow T$ and $T \rightarrow F$ removes ambiguity in the 'expression grammar,' resulting in the following unambiguous grammar

$E \rightarrow T \mid E + T$
$T \rightarrow F \mid T * F$
$F \rightarrow I \mid (E)$
$I \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$

Finding Unit productions

- But unit productions complicate certain proofs.
- A two-step technique to eliminate unit productions without changing the generated language:
  - find all "unit pairs"
  - expand productions using unit pairs until all unit productions disappear.
- **Definition** of *unit pair* ---
  - *Basis*: (A, A) is a unit pair for any nonterminal.
  - *Induction*: If (A, B) is a unit pair and $B \rightarrow C$ is a production, then (A, C) is a unit pair.

Finding unit productions Example

- The unit pairs for the *unambiguous* arithmetic expression grammar may be derived as follows:

$E \rightarrow T \mid E + T$
$T \rightarrow F \mid T * F$
$F \rightarrow I \mid (E)$
$I \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$

- Basis: (E, E), (T, T), (F, F), (I, I) are unit pairs
  - unit pair (E, E) & $E \rightarrow T \Rightarrow$ unit pair (E, T)
  - unit pair (E, T) & $T \rightarrow F \Rightarrow$ unit pair (E, F)
  - unit pair (E, F) & $F \rightarrow I \Rightarrow$ unit pair (E, I)
  - unit pair (T, T) & $T \rightarrow F \Rightarrow$ unit pair (T, F)
  - unit pair (T, F) & $F \rightarrow I \Rightarrow$ unit pair (T, I)
  - unit pair (F, F) & $F \rightarrow I \Rightarrow$ unit pair (F, I)
- Totally, there are 10 unit pairs

Eliminating unit productions

$E \rightarrow T \mid E + T$
$T \rightarrow F \mid T * F$
$F \rightarrow I \mid (E)$
$I \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$

Unit pair	Productions
(E, E)	$E \rightarrow E + T$ (from $E \rightarrow E + T$)
(E, T)	$E \rightarrow T * F$ (from $T \rightarrow T * F$)
(E, F)	$E \rightarrow (E)$
(E, I)	$E \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
(T, T)	$T \rightarrow T * F$
(T, F)	$T \rightarrow (E)$
(T, I)	$T \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
(F, F)	$F \rightarrow (E)$
(F, I)	$F \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
(I, I)	$I \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$

Simplification of Grammar

- **Theorem 7.14** ---
  - If G is a CFG generating a language that contains at least one string other than e, then there is another CFG G_1 such that $L(G_1) = L(G) - \{e\}$, and G_1 has no e-productions, unit productions, or useless symbols.
- Perform eliminations of the following *order* to a grammar G:
  - Elimination of e-productions;
  - Elimination of unit productions;
  - Elimination of useless symbols,

Chomsky Normal Form (CNF)

- **Definition:**

  A grammar *G* is said to be in *Chomsky Normal form* (*CNF*), if the following two conditions hold:

  - all its productions are in one of the following two simple forms:
    - $A \rightarrow BC$
    - $A \rightarrow a$

    where *A*, *B* and *C* are nonterminals and *a* is a terminal; and
  - *G* has no useless symbol.

Ex 1: Converting to CNF

$E \rightarrow T \mid E + T$
$T \rightarrow F \mid T * F$
$F \rightarrow I \mid (E)$
$I \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$

- Convert the expression grammar into CNF.
- Simplify the grammar.

(1) create new nonterminals for the terminals to produce the following productions:

- $A \rightarrow a$ $B \rightarrow b$
  $Z \rightarrow 0$ $O \rightarrow 1$
  $P \rightarrow +$ $M \rightarrow *$
  $L \rightarrow ($ $R \rightarrow)$

Unit pair	Productions
(E, E)	$E \rightarrow E + T$ (from $E \rightarrow E + T$)
(E, T)	$E \rightarrow T * F$ (from $T \rightarrow T * F$)
(E, F)	$E \rightarrow (E)$
(E, I)	$E \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
(T, T)	$T \rightarrow T * F$
(T, F)	$T \rightarrow (E)$
(T, I)	$T \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
(F, F)	$F \rightarrow (E)$
(F, I)	$F \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
(I, I)	$I \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$

Ex 2: Converting to CNF

(2) transformation of $E \rightarrow E + T \mid T * F \mid (E) \mid a \mid b \mid Ia \mid Ib \mid I0 \mid I1$

- $\Rightarrow E \rightarrow EPT \mid TMF \mid LER \mid a \mid b \mid IA \mid IB \mid IZ \mid IO$
- $T \rightarrow TMF \mid LER \mid a \mid b \mid IA \mid IB \mid IZ \mid IO$
- $F \rightarrow LER \mid a \mid b \mid IA \mid IB \mid IZ \mid IO$
- $I \rightarrow a \mid b \mid IA \mid IB \mid IZ \mid IO$

Unit pair	Productions
(E, E)	$E \rightarrow E + T$ (from $E \rightarrow E + T$)
(E, T)	$E \rightarrow T * F$ (from $T \rightarrow T * F$)
(E, F)	$E \rightarrow (E)$
(E, I)	$E \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
(T, T)	$T \rightarrow T * F$
(T, F)	$T \rightarrow (E)$
(T, I)	$T \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
(F, F)	$F \rightarrow (E)$
(F, I)	$F \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
(I, I)	$I \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$

Ex 1: Converting to CNF

(2) transformation of $E \rightarrow E + T \mid T * F \mid (E) \mid a \mid b \mid Ia \mid Ib \mid I0 \mid I1$

- $\Rightarrow E \rightarrow EPT \mid TMF \mid LER \mid a \mid b \mid IA \mid IB \mid IZ \mid IO$
- $T \rightarrow TMF \mid LER \mid a \mid b \mid IA \mid IB \mid IZ \mid IO$
- $F \rightarrow LER \mid a \mid b \mid IA \mid IB \mid IZ \mid IO$
- $I \rightarrow a \mid b \mid IA \mid IB \mid IZ \mid IO$

- $\Rightarrow E \rightarrow EC_1, C_1 \rightarrow PT,$
  $E \rightarrow TC_2, C_2 \rightarrow MF,$
  $E \rightarrow LC_3, C_3 \rightarrow \text{ER},$
- $\Rightarrow T \rightarrow TC_2, C_2 \rightarrow MF,$
  $T \rightarrow LC_3, C_3 \rightarrow \text{ER},$
- $\Rightarrow \text{F} \rightarrow LC_3, C_3 \rightarrow \text{ER},$

The grammar in CNF

- $\Rightarrow E \rightarrow EC_1 \mid TC_2 \mid LC_3 \mid a \mid b \mid IA \mid IB \mid IZ \mid IO$
- $T \rightarrow TC_2 \mid LC_3 \mid a \mid b \mid IA \mid IB \mid IZ \mid IO$
- $F \rightarrow LC_3 \mid a \mid b \mid IA \mid IB \mid IZ \mid IO$
- $I \rightarrow a \mid b \mid IA \mid IB \mid IZ \mid IO$
- $C_1 \rightarrow PT,$
- $C_2 \rightarrow MF,$
- $C_3 \rightarrow \text{ER},$
- $A \rightarrow a \qquad B \rightarrow b$
  $Z \rightarrow 0 \qquad O \rightarrow 1$
  $P \rightarrow + \qquad M \rightarrow *$
  $L \rightarrow (\qquad R \rightarrow)$

Ex 2: Converting to CNF

- Find the CNF for the following grammar

$S \rightarrow bA / aB$
$A \rightarrow bAA / aS / a$
$B \rightarrow aBB / bS / b$

Ex 3: Converting to CNF

- Find the CNF for the following grammar

$S \rightarrow AB$
$A \rightarrow aAA \mid \varepsilon$
$B \rightarrow bBB \mid \varepsilon$

Summary

- Simplification of CFG
- Chomsky Normal Form (CNF)
  - All the productions of the forms
    - $A \rightarrow BC$
    - $A \rightarrow a$
  - No useless symbol
- Converting the given grammar to CNF
  - Eliminate ε-productions, unit productions and useless symbols
  - Introduce new set of nonterminals

References

- John E. Hopcroft, Rajeev Motwani and Jeffrey D. Ullman, *Introduction to Automata Theory*, Languages, and Computation, Pearson, 3rd Edition, 2011.
- Peter Linz, An Introduction to Formal Languages and Automata, Jones and Bartle Learning International, United Kingdom, 6th Edition, 2016.

Next Class

Greibach Normal Form

Thank you.