Theory of Computer Science
B7. Context-free Languages: ε-Rules & Chomsky Normal Form

Gabriele Röger

University of Basel

March 27, 2023

B7.1 Context-free Grammars and ε-Rules

B7.2 Chomsky Normal Form

B7.3 Summary

B7.1 Context-free Grammars and ε-Rules

Repetition: Context-free Grammars

> **Definition (Context-free Grammar)**
> A context-free grammar is a 4-tuple $\langle V, \Sigma, P, S \rangle$ with
> 1. V finite set of variables,
> 2. Σ finite alphabet of terminal symbols (with $V \cap \Sigma = \emptyset$),
> 3. $P \subseteq (V \times (V \cup \Sigma)^+) \cup \{\langle S, \varepsilon \rangle\}$ finite set of rules,
> 4. If $S \to \varepsilon \in P$, then all other rules in $V \times ((V \setminus \{S\}) \cup \Sigma)^+$.
> 5. $S \in V$ start variable.

Rule $X \to \varepsilon$ is only allowed if $X = S$
and S never occurs on a right-hand side.

With regular grammars, this restriction could be lifted.
How about context-free grammars?

Short-hand Notation for Rule Sets

We abbreviate several rules with the same left-hand side variable in a single line, using "|" for separating the right-hand sides.

For example, we write

$$X \rightarrow 0Y1 \mid XY$$

for:

$$X \rightarrow 0Y1 \text{ and}$$
$$X \rightarrow XY$$

Context-free Grammars: Exercise

We have used the pumping lemma for regular languages to show that $L = \{a^n b^n \mid n \in \mathbb{N}_0\}$ is not regular.

Show that it is context-free by specifying a suitable grammar G with $\mathcal{L}(G) = L$.

Repetition: Context-free Grammars

> **Definition (Context-free Grammar)**
>
> A context-free grammar is a 4-tuple $\langle V, \Sigma, P, S \rangle$ with
>
> 1. V finite set of variables,
> 2. Σ finite alphabet of terminal symbols (with $V \cap \Sigma = \emptyset$),
> 3. $P \subseteq (V \times (V \cup \Sigma)^+) \cup \{\langle S, \varepsilon \rangle\}$ finite set of rules,
> 4. If $S \rightarrow \varepsilon \in P$, then all other rules in $V \times ((V \setminus \{S\}) \cup \Sigma)^+$.
> 5. $S \in V$ start variable.

Rule $X \rightarrow \varepsilon$ is only allowed if $X = S$
and S never occurs on a right-hand side.

With regular grammars, this restriction could be lifted.
How about context-free grammars?

Repetition: Context-free Grammars

> **Definition (Context-free Grammar)**
>
> A context-free grammar is a 4-tuple $\langle V, \Sigma, P, S \rangle$ with
>
> 1. V finite set of variables,
> 2. Σ finite alphabet of terminal symbols (with $V \cap \Sigma = \emptyset$),
> 3. $P \subseteq (V \times (V \cup \Sigma)^+) \cup \{\langle S, \varepsilon \rangle\}$ finite set of rules,
> 4. If $S \rightarrow \varepsilon \in P$, then all other rules in $V \times ((V \setminus \{S\}) \cup \Sigma)^+$.
> 5. $S \in V$ start variable.

Rule $X \rightarrow \varepsilon$ is only allowed if $X = S$
and S never occurs on a right-hand side.

With regular grammars, this restriction could be lifted.
How about context-free grammars?

Reminder: Start Variable in Right-Hand Side of Rules

For every type-0 language L there is a grammar where the start variable does not occur on the right-hand side of any rule.

Theorem
For every grammar $G = \langle V, \Sigma, P, S \rangle$ there is a grammar $G' = \langle V', \Sigma, P', S \rangle$ with rules $P' \subseteq (V' \cup \Sigma)^+ \times (V' \setminus \{S\} \cup \Sigma)^*$ such that $\mathcal{L}(G) = \mathcal{L}(G')$.

In the proof we constructed a suitable grammar, where the rules in P' were not fundamentally different from the rules in P:

- for rules from $V \times (V \cup \Sigma)^+$, we only introduced additional rules from $V' \times (V' \cup \Sigma)^+$, and
- for rules from $V \times \varepsilon$, we only introduced rules from $V' \times \varepsilon$,

where $V' = V \cup \{S'\}$ for some new variable $S' \notin V$.

ε-Rules

Theorem
For every grammar G with rules $P \subseteq V \times (V \cup \Sigma)^*$ there is a context-free grammar G' with $\mathcal{L}(G) = \mathcal{L}(G')$.

Proof.
Let $G = \langle V, \Sigma, P, S \rangle$ be a grammar with $P \subseteq V \times (V \cup \Sigma)^*$.

Let $G' = \langle V', \Sigma, P', S \rangle$ be a grammar with $\mathcal{L}(G) = \mathcal{L}(G')$ with $P' \subseteq V' \times ((V' \setminus S) \cup \Sigma)^*$.

Let $V_\varepsilon = \{A \in V' \mid A \Rightarrow^*_{G'} \varepsilon\}$. We can find this set V_ε by first collecting all variables A with rule $A \to \varepsilon \in P'$ and then successively adding additional variables B if there is a rule $B \to A_1 A_2 \dots A_k \in P'$ and the variables A_i are already in the set for all $1 \le i \le k$.

$\cdots$

ε-Rules

Theorem
For every grammar G with rules $P \subseteq V \times (V \cup \Sigma)^*$ there is a context-free grammar G' with $\mathcal{L}(G) = \mathcal{L}(G')$.

Proof (continued).
Let P'' be the rule set that is constructed from P' by

- adding rules that obviate the need for $A \to \varepsilon$ rules:
  for every existing rule $B \to w$ with $B \in V', w \in (V' \cup \Sigma)^+$, let I_ε be the set of positions where w contains a variable $A \in V_\varepsilon$. For every non-empty set $I' \subseteq I_\varepsilon$, add a new rule $B \to w'$, where w' is constructed from w by removing the variables at all positions in I'.
- removing all rules of the form $A \to \varepsilon$ ($A \ne S$).

Then $G'' = \langle V', \Sigma, P'', S \rangle$ is context-free and $\mathcal{L}(G) = \mathcal{L}(G'')$. $\qquad\square$

Example

Consider $G = \langle \{X, Y, Z, S\}, \{a, b\}, R, S \rangle$ with rules:

$$S \to \varepsilon \mid XY$$
$$X \to aXYbX \mid YZ$$
$$Y \to \varepsilon \mid b$$
$$Z \to \varepsilon \mid a$$

$\rightsquigarrow$ blackboard

B7.2 Chomsky Normal Form

Chomsky Normal Form: Motivation

As in logical formulas (and other kinds of structured objects),
normal forms for grammars are useful:

▶ they show which aspects are critical for defining grammars
  and which ones are just syntactic sugar

▶ they allow proofs and algorithms to be restricted
  to a limited set of grammars (inputs): those in normal form

Hence we now consider a normal form for context-free grammars.

Chomsky Normal Form: Definition

Definition (Chomsky Normal Form)
A context-free grammar G is in Chomsky normal form
(CNF) if all rules have one of the following three forms:

▶ $A \to BC$ with variables A, B, C, or

▶ $A \to a$ with variable A, terminal symbol a, or

▶ $S \to \varepsilon$ with start variable S.

in short:
rule set $P \subseteq (V \times (V'V' \cup \Sigma)) \cup \{\langle S, \varepsilon \rangle\}$ with $V' = V \setminus \{S\}$

Chomsky Normal Form: Theorem

Theorem
For every context-free grammar G there is a context-free grammar
G' in Chomsky normal form with $\mathcal{L}(G) = \mathcal{L}(G')$.

Proof.
The following algorithm converts the rule set of G into CNF:

Step 1: Eliminate rules of the form $A \to B$ with variables A, B.

If there are sets of variables $\{B_1, \ldots, B_k\}$ with rules
$B_1 \to B_2, B_2 \to B_3, \ldots, B_{k-1} \to B_k, B_k \to B_1$,
then replace these variables by a new variable B.

Define a strict total order $<$ on the variables such that $A \to B \in P$
implies that $A < B$. Iterate from the largest to the smallest
variable A and eliminate all rules of the form $A \to B$ while adding
rules $A \to w$ for every rule $B \to w$ with $w \in (V \cup \Sigma)^+$.   . . .

Chomsky Normal Form: Theorem

Theorem
For every context-free grammar G there is a context-free grammar G' in Chomsky normal form with $\mathcal{L}(G) = \mathcal{L}(G')$.

Proof (continued).
Step 2: Eliminate rules with terminal symbols on the
right-hand side that do not have the form $A \rightarrow a$.

For every terminal symbol $a \in \Sigma$ add a new variable A_a
and the rule $A_a \rightarrow a$.

Replace all terminal symbols in all rules that do not have
the form $A \rightarrow a$ with the corresponding newly added variables. $\ldots$

Chomsky Normal Form: Theorem

Theorem
For every context-free grammar G there is a context-free grammar G' in Chomsky normal form with $\mathcal{L}(G) = \mathcal{L}(G')$.

Proof (continued).
Step 3: Eliminate rules of the form $A \rightarrow B_1 B_2 \ldots B_k$ with $k > 2$

For every rule of the form $A \rightarrow B_1 B_2 \ldots B_k$ with $k > 2$, add new
variables $C_2, \ldots, C_{k-1}$ and replace the rule with

$$A \rightarrow B_1 C_2$$
$$C_2 \rightarrow B_2 C_3$$
$$\vdots$$
$$C_{k-1} \rightarrow B_{k-1} B_k$$

Example

Consider $G = \langle \{Y, Z, S\}, \{a, b\}, R, S \rangle$ with rules:

$$S \rightarrow aZbY \mid Y \mid ab$$
$$Y \rightarrow Z \mid b$$
$$Z \rightarrow Y \mid bSa$$

$\rightarrow$ blackboard

Chomsky Normal Form: Length of Derivations

Observation
Let G be a grammar in Chomsky normal form,
and let $w \in \mathcal{L}(G)$ be a non-empty word generated by G.

Then all derivations of w have exactly $2|w| - 1$ derivation steps.

Why?

B7.3 Summary

Summary

► The restriction of ε-occurrences in rules is not necessary to characterize the set of context-free languages.

► Every context-free language has a grammar in Chomsky normal form. All rules have form
  ► $A \to BC$ with variables A, B, C, or
  ► $A \to a$ with variable A, terminal symbol a, or
  ► $S \to \varepsilon$ with start variable S.