Finite automata theory and formal languages (DIT321, TMV027)

Nils Anders Danielsson

2019-02-28

Today

- Grammar transformations.
- Chomsky normal form.
- The pumping lemma for context-free languages.

Grammar transformations

Grammar transformations

- A number of transformations of grammars.
- I have taken some information and terminology from "To CNF or not to CNF? An Efficient Yet Presentable Version of the CYK Algorithm" by Lange and Leiß.

Bin

- Result: No production $A \to \alpha$ where $|\alpha| \geq 3$.
- Replace each production $A \to X_1 X_2 ... X_n$, where $n \geq 3$, with:

$$\begin{aligned} A &\to X_1 A_2 \\ A_2 &\to X_2 A_3 \\ &\vdots \\ A_{n-1} &\to X_{n-1} X_n \end{aligned}$$

  Here $A_2, ..., A_{n-1}$ are new nonterminals.
- $L(\textsc{Bin}(G)) = L(G)$.

$\textsc{Del}$

- Result: No production of the form $A \to \varepsilon$.
- A nonterminal A is *nullable* if $A \Rightarrow^* \varepsilon$.
- Replace each production $A \to X_1 X_2 \ldots X_n$ with $\{ A \to \alpha \mid \alpha \in f(n) \setminus \{ \varepsilon \} \}$:

$$
\begin{aligned}
f &\in \mathbb{N} \to \wp((N \cup \Sigma)^*) \\
f(0) &= \{ \varepsilon \} \\
f(1+k) &= f(k) \{ X_{1+k} \} \\
&\quad \cup \begin{cases} f(k), & \text{if } X_{1+k} \text{ is a nullable} \\ & \text{nonterminal} \\ \emptyset, & \text{otherwise} \end{cases}
\end{aligned}
$$

- $L(\textsc{Del}(G)) = L(G) \setminus \{ \varepsilon \}$.

Del

If Del is applied to the following grammar, how many productions does the resulting grammar contain?

$$(\{\,S, A\,\}, \{\,0\,\}, (S \to (SA)^{10} \mid \varepsilon, A \to 0), S)$$

Del

- The Del transformation can make the grammar much larger.
- If every production $A \to \alpha$ satisfies $|\alpha| \leq 2$, then the blowup is contained.
- Run Bin before Del.

UNIT

- Result: No production of the form $A \to B$.
- (A, B) is a *unit pair* if $A = B$ or $A \Rightarrow C_1 \Rightarrow \cdots \Rightarrow C_n \Rightarrow B$ (where $n \in \mathbb{N}$).
- Include exactly the following productions:

$$\{A \to \alpha \mid (A, B) \text{ is a unit pair, } B \to \alpha \in P, \alpha \text{ is not a single nonterminal}\}$$

- $L(\textsc{Unit}(G)) = L(G)$.

Unit

The resulting grammar could be much larger than the original one:

$$
\begin{aligned}
A_1 &\to A_2 \mid 1 \\
A_2 &\to A_3 \mid 2 \\
A_3 &\to A_4 \mid 3 \\
&\vdots \\
A_n &\to A_1 \mid n
\end{aligned}
$$

Unit

The resulting grammar could be much larger than the original one:

$$
\begin{aligned}
A_1 &\rightarrow 1 \mid 2 \mid 3 \mid \ldots \mid n \\
A_2 &\rightarrow 1 \mid 2 \mid 3 \mid \ldots \mid n \\
A_3 &\rightarrow 1 \mid 2 \mid 3 \mid \ldots \mid n \\
&\vdots \\
A_n &\rightarrow 1 \mid 2 \mid 3 \mid \ldots \mid n
\end{aligned}
$$

Construct a grammar G for which $\textsc{Del}(\textsc{Unit}(G))$ contains a production of the form $A \to B$.

Construct a grammar G for which $\textsc{Del}(\textsc{Unit}(G))$ contains a production of the form $A \to B$.

Run $\textsc{Del}$ before $\textsc{Unit}$.

Term

- Result: No terminals in productions $A \to \alpha$ where $|\alpha| \geq 2$.
- Find all terminals in such productions.
- For each such terminal b, add a new nonterminal B with a single production $B \to b$, and substitute B for b in every production $A \to \alpha$ where $|\alpha| \geq 2$.
- $L(\textsc{Term}(G)) = L(G)$.

Chomsky normal form

Chomsky normal form

- A context-free grammar is in *Chomsky normal form* if every production is of the form $A \to BC$ or $A \to a$.
- For any context-free grammar G the grammar $G' = \textsc{Term}(\textsc{Unit}(\textsc{Del}(\textsc{Bin}(G))))$ is in Chomsky normal form and satisfies $L(G') = L(G) \setminus \{ \varepsilon \}$.

Chomsky normal form

- A context-free grammar is in *Chomsky normal form* if every production is of the form $A \to BC$ or $A \to a$.
- For any context-free grammar G the grammar $G' = \textsc{Term}(\textsc{Unit}(\textsc{Del}(\textsc{Bin}(G))))$ is in Chomsky normal form and satisfies $L(G') = L(G) \setminus \{ \varepsilon \}$.

I dropped the text book's requirement that there should be no useless symbols.

Consider the grammar $G = (\{ S, A, B \}, \{ 0, 1, 2 \}, P, S)$, where P is defined in the following way:

$$S \to 0A \mid B$$
$$A \to 1B1 \mid \varepsilon$$
$$B \to S \mid 2$$

- Is G ambiguous?
- Is $\textsc{Term}(\textsc{Unit}(\textsc{Del}(\textsc{Bin}(G))))$ ambiguous?

The pumping lemma

The pumping lemma for CFLs

For every context-free language L
over the alphabet Σ:

$$
\begin{array}{l}
\exists m \in \mathbb{N}. \\
\quad \forall w \in L.\ |w| \geq m \Rightarrow \\
\quad\quad \exists r, s, t, u, v \in \Sigma^*. \\
\quad\quad\quad w = rstuv \wedge |stu| \leq m \wedge su \neq \varepsilon \wedge \\
\quad\quad\quad \forall n \in \mathbb{N}.\ rs^n tu^n v \in L
\end{array}
$$

The pumping lemma for CFLs

For every context-free language L over the alphabet Σ:

$$
\begin{aligned}
&\exists m \in \mathbb{N}.\\
&\quad \forall w \in L.\ |w| \geq m \Rightarrow\\
&\qquad \exists r, s, t, u, v \in \Sigma^*.\\
&\qquad\quad w = rstuv \land |stu| \leq m \land su \neq \varepsilon \land\\
&\qquad\quad \forall n \in \mathbb{N}.\ rs^n tu^n v \in L
\end{aligned}
$$

Height

The height of a parse tree:

$$\begin{aligned}
&\mathit{height} \in P_N(G, A) \to \mathbb{N} \\
&\mathit{height}(\mathsf{leaf}(A)) &&= 0 \\
&\mathit{height}(\mathsf{node}(A, \mathit{ts})) &&= 1 + \mathit{height}^*(\mathit{ts})
\end{aligned}$$

$$\begin{aligned}
&\mathit{height}^* \in P_N^*(G, \alpha) \to \mathbb{N} \\
&\mathit{height}^*(\mathsf{nil}) &&= 0 \\
&\mathit{height}^*(\mathsf{term}(a, \mathit{ts})) &&= \mathit{height}^*(\mathit{ts}) \\
&\mathit{height}^*(\mathsf{nonterm}(t, \mathit{ts})) &&= \mathit{max}(\mathit{height}(t), \mathit{height}^*(\mathit{ts}))
\end{aligned}$$

For parse trees in $P(G, A)$ the height is equal to the largest number of nonterminals encountered on any path from the root to a leaf.

Height

For context-free grammars in Chomsky normal form:

$$\forall p \in P(G, A).\ |yield(p)| \leq 2^{height(p)-1}$$

What is the smallest value of $W \in \mathbb{N}$ for which

$$\forall p \in P(G, A).\ |yield(p)| \leq W^{height(p)-1}$$

holds for the grammar G below?

$$(\{\,S\,\}, \{\,0\,\}, (S \to SSS \mid 0), S)$$

The pumping lemma for CFLs

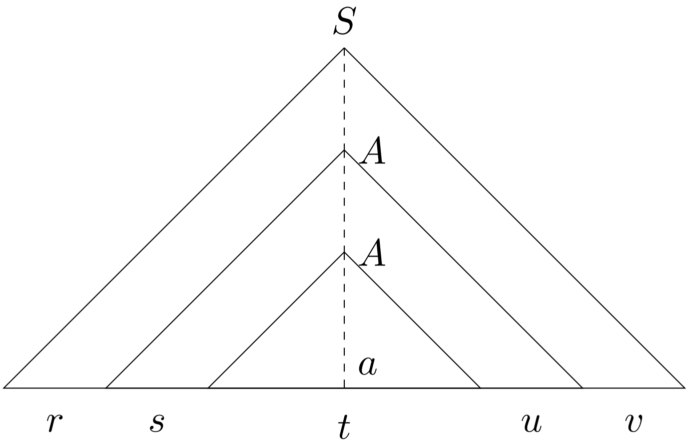
Proof sketch:

- Take any context-free grammar G for L.
- Let $G' = \textsc{Term}(\textsc{Unit}(\textsc{Del}(\textsc{Bin}(G))))$.
- If $G' = (N, \Sigma, P, S)$, let $m = 2^{|N|}$.
- Given a string $w \in L$ with $|w| \geq m$ we know that $w \neq \varepsilon$, so we have $w \in L \setminus \{ \varepsilon \} = L(G')$.

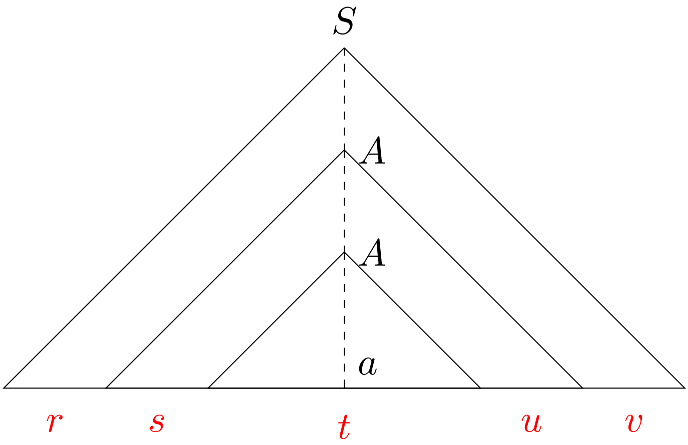
The pumping lemma for CFLs

- Take any parse tree p for w with respect to G'.
- We know that $2^{|N|} \leq |w| \leq 2^{height(p)-1}$, so $height(p) > |N|$.
- Take a path of maximal length from the root of p to a leaf.
- Such a path must contain at least $|N| + 1$ nonterminals.
- By the pigeonhole principle the path must contain two instances of the same nonterminal, at most $|N| + 1$ steps from the leaf.

The pumping lemma for CFLs

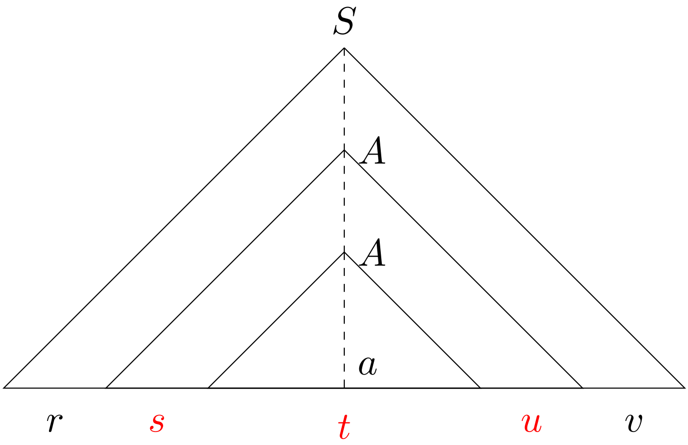


The pumping lemma for CFLs



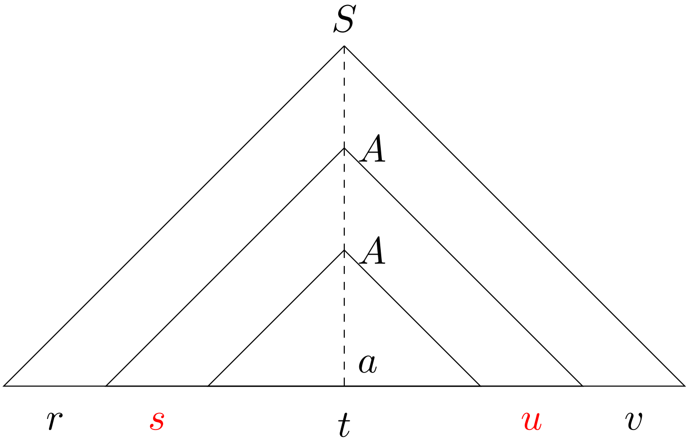
$w = rstuv$

The pumping lemma for CFLs



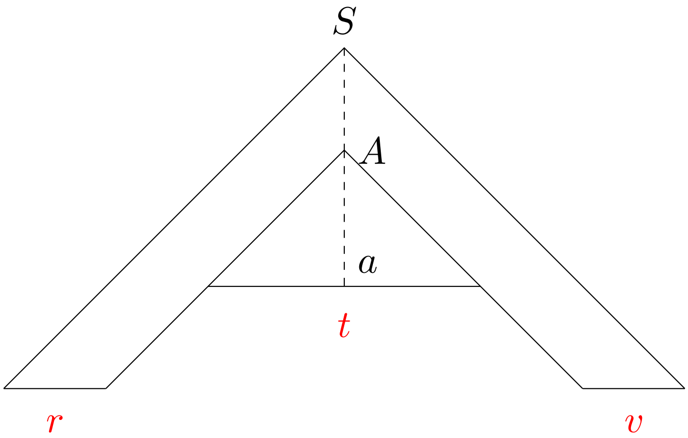
$$|stu| \leq 2^{(|N|+1)-1} = 2^{|N|} = m$$

The pumping lemma for CFLs



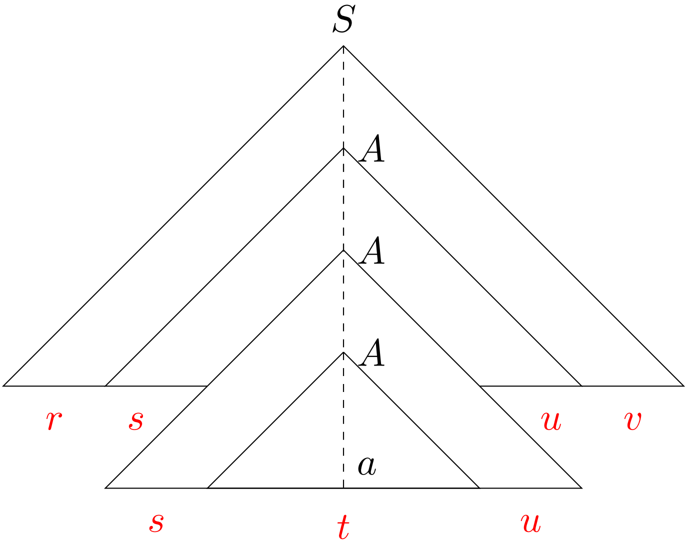
No nonterminal is nullable, $A \to BC \Rightarrow$
$s \neq \varepsilon \vee u \neq \varepsilon \Rightarrow su \neq \varepsilon$

The pumping lemma for CFLs



$rtv \in L(G') \subseteq L$

The pumping lemma for CFLs



$rs^2tu^2v \in L(G') \subseteq L$

The pumping lemma for CFLs

The language $L = \{\, 0^n 1^n 2^n \mid n \in \mathbb{N} \,\}$ over $\Sigma = \{\, 0, 1, 2 \,\}$ is not context-free. Proof sketch:

- Assume that L is context-free.
- Take the constant $m \in \mathbb{N}$ that we get from the pumping lemma.
- Consider the string $w = 0^m 1^m 2^m \in L$.
- Because $|w| \geq m$ we get some information:

$$\begin{aligned} &\exists r, s, t, u, v \in \Sigma^*. \\ &\quad w = rstuv \wedge |stu| \leq m \wedge su \neq \varepsilon \wedge \\ &\quad \forall n \in \mathbb{N}.\ rs^n tu^n v \in L \end{aligned}$$

The pumping lemma for CFLs

- Because $|w| \geq m$ we get some information:

$$\begin{aligned} &\exists r, s, t, u, v \in \Sigma^*. \\ &\quad w = rstuv \wedge |stu| \leq m \wedge su \neq \varepsilon \wedge \\ &\quad \forall n \in \mathbb{N}.\ rs^n tu^n v \in L \end{aligned}$$

- Because $|stu| \leq m$ this substring cannot contain both 0 and 2.
- Because $su \neq \varepsilon$ either s or u must contain at least one symbol from Σ.
- Thus rtv does not contain the same number of each symbol from Σ.
- This is a contradiction, because $rtv \in L$.

What is the smallest possible value of "m" for a *non-empty* context-free language?

Today

- Grammar transformations.
- Chomsky normal form.
- The pumping lemma for context-free languages.

Next lecture

- Closure properties.
- Algorithms.

- Deadline for the next quiz: 2019-03-04, 10:00.
- Deadline for the fifth assignment: 2019-03-03, 23:59.