The Chomsky-Schützenberger Theorem

Deepali Nemade and Nikhil Panwar

Department of Computer Science and Automation,
Indian Institute of Science, Bangalore.

24th November, 2011

Homomorphism

A *Homomorphism* is a map $h : \Gamma^* \to \Sigma^*$ such that for all $x, y \in \Gamma^*$

$$h(xy) = h(x)h(y)$$
$$h(\epsilon) = \epsilon$$

Context Free Grammar

A *Context Free Grammar* (CFG) is a quadruple

$$G = (N, \Sigma, P, S)$$

Context Free Grammar

A *Context Free Grammar* (CFG) is a quadruple

$$G = (N, \Sigma, P, S)$$

where,

Context Free Grammar

A *Context Free Grammar* (CFG) is a quadruple

$$G = (N, \Sigma, P, S)$$

where,

N is a finite set of Non terminal Symbols,

Context Free Grammar

A *Context Free Grammar* (CFG) is a quadruple

$$G = (N, \Sigma, P, S)$$

where,

N is a finite set of Non terminal Symbols,
Σ is a finite set of Terminal Symbols,

Context Free Grammar

A *Context Free Grammar* (CFG) is a quadruple

$$G = (N, \Sigma, P, S)$$

where,

N is a finite set of Non terminal Symbols,
Σ is a finite set of Terminal Symbols,
P is a finite subset of $N \times (N \cup \Sigma)^*$ (set of productions)

Context Free Grammar

A *Context Free Grammar* (CFG) is a quadruple

$$G = (N, \Sigma, P, S)$$

where,

N is a finite set of Non terminal Symbols,
Σ is a finite set of Terminal Symbols,
P is a finite subset of $N \times (N \cup \Sigma)^*$ (set of productions)
$S \in N$ is the start symbol

Chomsky Normal Form

A CFG is in "Chomsky normal form(CNF)" if all productions are of the form

Chomsky Normal Form

A CFG is in "Chomsky normal form(CNF)" if all productions are of the form

$$A \to BC$$

Chomsky Normal Form

A CFG is in "Chomsky normal form(CNF)" if all productions are of the form

$$A \to BC$$

or

Chomsky Normal Form

A CFG is in "Chomsky normal form(CNF)" if all productions are of the form

$$A \rightarrow BC$$

or

$$A \rightarrow a$$

Chomsky Normal Form

A CFG is in "Chomsky normal form(CNF)" if all productions are of the form

$$A \rightarrow BC$$
or
$$A \rightarrow a$$

where,

$A, B, C \in N$ and $a \in \Sigma$

$PAREN_n$

$PAREN_n$ denote the language consisting of all balanced strings of parentheses of n distinct types.

$PAREN_n$

$PAREN_n$ denote the language consisting of all balanced strings of parentheses of n distinct types.

This language is generated by the grammar-

$$S \rightarrow [^1 S]^1 | [^2 S]^2 | \ldots\ldots\ldots\ldots | [^n S]^n | \varepsilon$$

$PAREN_n$

$PAREN_n$ denote the language consisting of all balanced strings of parentheses of n distinct types.

This language is generated by the grammar-

$$S \rightarrow [^1 S]^1 | [^2 S]^2 | .... | [^n S]^n | \varepsilon$$

The languages $PAREN_n$ are called *Dyck Languages* in the literature.

$PAREN_n$

$PAREN_n$ denote the language consisting of all balanced strings of parentheses of n distinct types.

This language is generated by the grammar-

$$S \rightarrow [^1 S]^1 | [^2 S]^2 | \ldots\ldots\ldots\ldots | [^n S]^n | \varepsilon$$

The languages $PAREN_n$ are called *Dyck Languages* in the literature.

Example: $[^1 [^2]^2]^1 [^3]^3$

Theorem

The Chomsky-Schützenberger Theorem

Every context-free language is a homomorphic image of the intersection of a parenthesis language and a regular set. In other words, for every CFL A, there is an $n >= 0$, a regular set R, and a homomorphism h such that

$$A = h(PAREN_n \cap R)$$

Proof

Let $G = (N, \Sigma, P, S)$ be an arbitrary CFG in Chomsky Normal Form (CNF).
Let the productions in P are denoted by $\pi, \rho, \sigma \ldots$

Proof

Let $G = (N, \Sigma, P, S)$ be an arbitrary CFG in Chomsky Normal Form (CNF).
Let the productions in P are denoted by $\pi, \rho, \sigma \ldots$

For $\pi \in P$, define π'

$$
\begin{array}{ll}
A \rightarrow [^1_\pi B]^1_\pi [^2_\pi C]^2_\pi & \text{if } \pi = A \rightarrow BC, \\
A \rightarrow [^1_\pi]^1_\pi [^2_\pi]^2_\pi & \text{if } \pi = A \rightarrow a
\end{array}
$$

Proof

Let $G = (N, \Sigma, P, S)$ be an arbitrary CFG in Chomsky Normal Form (CNF).
Let the productions in P are denoted by $\pi, \rho, \sigma \ldots$

For $\pi \in P$, define $\pi^{'}$

$$
\begin{array}{ll}
A \to [^1_\pi B]^1_\pi [^2_\pi C]^2_\pi & \text{if } \pi = A \to BC, \\
A \to [^1_\pi]^1_\pi [^2_\pi]^2_\pi & \text{if } \pi = A \to a
\end{array}
$$

and define the grammar $G^{'} = (N, \Sigma, P^{'}, S)$ with

$$
\begin{array}{l}
\Gamma = \{[^1_\pi,]^1_\pi, [^2_\pi,]^2_\pi | \pi \in P\} \\
P^{'} = \{\pi^{'} | \pi \in P\}
\end{array}
$$

Proof continued...

$$L(G') \subseteq PAREN_{\Gamma}$$

Proof continued...

$L(G') \subseteq PAREN_\Gamma$

Properties satisfied by strings in $L(G')$ that are not satisfied by strings in $PAREN_\Gamma$ in general:

Proof continued...

$L(G') \subseteq PAREN_\Gamma$
Properties satisfied by strings in $L(G')$ that are not satisfied by strings in $PAREN_\Gamma$ in general:

- **Property 1 :** Every $]^1_\pi$ is immediately followed by a $[^2_\pi$.

Proof continued...

$L(G^{'}) \subseteq PAREN_{\Gamma}$

Properties satisfied by strings in $L(G^{'})$ that are not satisfied by strings in $PAREN_{\Gamma}$ in general:

- **Property 1 :** Every $]_{\pi}^{1}$ is immediately followed by a $[_{\pi}^{2}$.
- **Property 2 :** No $]_{\pi}^{2}$ is immediately followed by a left parenthesis.

Proof continued...

$L(G^{'}) \subseteq PAREN_\Gamma$
Properties satisfied by strings in $L(G^{'})$ that are not satisfied by strings in $PAREN_\Gamma$ in general:

- **Property 1 :** Every $]^1_\pi$ is immediately followed by a $[^2_\pi$.
- **Property 2 :** No $]^2_\pi$ is immediately followed by a left parenthesis.
- **Property 3 :** If $\pi = A \rightarrow BC$, then every $[^1_\pi$ is immediately followed by $[^1_\rho$ for some $\rho \in P$ with left hand side B, and every $[^2_\pi$ is immediately followed by $[^1_\sigma$ for some $\sigma \in P$ with left-hand side C.

Proof continued...

$L(G^{'}) \subseteq PAREN_{\Gamma}$

Properties satisfied by strings in $L(G^{'})$ that are not satisfied by strings in $PAREN_{\Gamma}$ in general:

- **Property 1 :** Every $]^1_\pi$ is immediately followed by a $[^2_\pi$.
- **Property 2 :** No $]^2_\pi$ is immediately followed by a left parenthesis.
- **Property 3 :** If $\pi = A \rightarrow BC$, then every $[^1_\pi$ is immediately followed by $[^1_\rho$ for some $\rho \in P$ with left hand side B, and every $[^2_\pi$ is immediately followed by $[^1_\sigma$ for some $\sigma \in P$ with left-hand side C.
- **Property 4 :** If $\pi = A \rightarrow a$, then every $[^1_\pi$ is immediately followed by $]^1_\pi$ and every $[^2_\pi$ is immediately followed by $]^2_\pi$.

Proof continued...

In addition, all strings x such that $A \xrightarrow[G']{*} x$ satisfy the property

- **Property** (v_A)**:** The string x begins with $[^1_\pi$ for some $\pi \in P$ with left-hand side A.

Proof continued...

In addition, all strings x such that $A \xrightarrow[G']{*} x$ satisfy the property

- **Property** (v_A)**:** The string x begins with $[^1_\pi$ for some $\pi \in P$ with left-hand side A.

Now we can define a regular expression that satisfies all the above properties as:

$$R_A = \{\ \mathrm{x} \in \Gamma^* \mid \mathrm{x} \text{ satisfies } \textit{property}\,1 \text{ through } (v_A)\ \}$$

Proof continued...

Lemma

$$A \xrightarrow[G']{*} x \iff x \in (PAREN_\Gamma \cap R_A)$$

Proof Continued...

Proof of Lemma:

Prove $\Rightarrow$: $A \xrightarrow[G']{*} x \Rightarrow x \in (PAREN_\Gamma \cap R_A)$

Proof Continued...

Proof of Lemma:

Prove $\Rightarrow$: $A \xrightarrow[G']{*} x \Rightarrow x \in (PAREN_\Gamma \cap R_A)$

Applying induction on the length of derivation:

Proof Continued...

Proof of Lemma:

Prove $\Rightarrow$: $A \xrightarrow[G']{*} x \Rightarrow x \in (PAREN_\Gamma \cap R_A)$

Applying induction on the length of derivation:

Basis: n=1

$$A \rightarrow [^1_\pi B]^1_\pi [^2_\pi C]^2_\pi$$
$$A \rightarrow [^1_\pi]^1_\pi [^2_\pi]^2_\pi$$

Since RHS satisfies all properties so it is true for $n = 1$.

Proof Continued...

Proof of Lemma:

Prove $\Rightarrow$: $A \xrightarrow[G']{*} x \Rightarrow x \in (PAREN_\Gamma \cap R_A)$

Applying induction on the length of derivation:

Basis: n=1

$$A \rightarrow [^1_\pi B]^1_\pi [^2_\pi C]^2_\pi$$
$$A \rightarrow [^1_\pi]^1_\pi [^2_\pi]^2_\pi$$

Since RHS satisfies all properties so it is true for $n = 1$.

Induction Hypothesis: Let $A \xrightarrow[G']{*} \alpha$
where, α is a sentential form of length n that satisfies all properties.

Proof Continued...

Proof of Lemma:

Prove $\Rightarrow$: $A \xrightarrow[G']{*} x \Rightarrow x \in (PAREN_\Gamma \cap R_A)$

Applying induction on the length of derivation:

Basis: n=1

$$A \rightarrow [^1_\pi B]^1_\pi [^2_\pi C]^2_\pi$$
$$A \rightarrow [^1_\pi]^1_\pi [^2_\pi]^2_\pi$$

Since RHS satisfies all properties so it is true for $n = 1$.

Induction Hypothesis: Let $A \xrightarrow[G']{*} \alpha$

where, α is a sentential form of length n that satisfies all properties.

Induction Step: Proving it for $n + 1$ length of derivation

$$A \xrightarrow[G']{1} [^1_\pi B]^1_\pi [^2_\pi C]^2_\pi \xrightarrow[G']{*} \alpha$$

Proof Continued...

Prove $\Leftarrow$: $x \in (PAREN_\Gamma \cap R_A) \Rightarrow A \xrightarrow[G']{*} x$

Proof Continued...

Prove $\Leftarrow$: $x \in (PAREN_\Gamma \cap R_A) \Rightarrow A \xrightarrow[G']{*} x$

Applying induction on the length of x:

Proof Continued...

Prove $\Leftarrow$: $x \in (PAREN_{\Gamma} \cap R_A) \Rightarrow A \xrightarrow[G']{*} x$

Applying induction on the length of x:

It follows from properties that x is a string of balanced parentheses of the form

$$x = [^1_\pi y]^1_\pi [^2_\pi z]^2_\pi$$

for some $y, z \in \Gamma^*$ and π with left hand side A.

Proof continued...

If $\pi = A \rightarrow BC$, then

Proof continued...

If $\pi = A \rightarrow BC$, then

From property 3, y satisfies (v_B) and z satisfies (v_C).
Also y and z are balanced.

Thus $y \in PAREN_\Gamma \cap R_B$ and $z \in PAREN_\Gamma \cap R_c$.

Proof continued...

If $\pi = A \rightarrow BC$, then

From property 3, y satisfies (v_B) and z satisfies (v_C).
Also y and z are balanced.

Thus $y \in PAREN_\Gamma \cap R_B$ and $z \in PAREN_\Gamma \cap R_c$.

By induction hypothesis, $B \xrightarrow[G']{*} y$ and $C \xrightarrow[G']{*} z$ therefore,

$$A \xrightarrow[G']{1} [^1_\pi B]^1_\pi [^2_\pi C]^2_\pi \xrightarrow[G']{*} [^1_\pi x]^1_\pi [^2_\pi y]^2_\pi = x$$

Proof continued...

If $\pi = A \to a$, then

Proof continued...

If $\pi = A \rightarrow a$, then

From property 4, $y = z = \epsilon$, and

$$A \rightarrow [^1_\pi]^1_\pi [^2_\pi]^2_\pi = x$$

Proof continued...

If $\pi = A \to a$, then

From property 4, $y = z = \epsilon$, and

$$A \to [^1_\pi]^1_\pi [^2_\pi]^2_\pi = x$$

It follows from Lemma that $L(G') = PAREN_\Gamma \cap R_S$.

Proof continued...

Applying Homomorphism

Define homomorphism $h : \Gamma^* \to \Sigma^*$ as follows:

Proof continued...

Applying Homomorphism

Define homomorphism $h : \Gamma^* \to \Sigma^*$ as follows:

For π of the form $A \to BC$, take

$$h([^1_\pi) = h(]^1_\pi) = h([^2_\pi) = h(]^2_\pi) = \epsilon,$$

Proof continued...

Applying Homomorphism
Define homomorphism $h : \Gamma^* \to \Sigma^*$ as follows:

For π of the form $A \to BC$, take

$$h([^1_\pi) = h(]^1_\pi) = h([^2_\pi) = h(]^2_\pi) = \epsilon,$$

For π of the form $A \to a$, take

$$h(]^1_\pi) = h([^2_\pi) = h(]^2_\pi) = \epsilon,$$
$$h([^1_\pi) = a$$

Proof continued...

Applying Homomorphism

Define homomorphism $h : \Gamma^* \to \Sigma^*$ as follows:

For π of the form $A \to BC$, take

$$h([^1_\pi) = h(]^1_\pi) = h([^2_\pi) = h(]^2_\pi) = \epsilon,$$

For π of the form $A \to a$, take

$$h(]^1_\pi) = h([^2_\pi) = h(]^2_\pi) = \epsilon,$$
$$h([^1_\pi) = a$$

Applying h to the production π of P' gives the production π of P thus $L(G) = h(L(G')) = h(PAREN_\Gamma \cap R_S)$.

Proof continued...

Applying Homomorphism

Define homomorphism $h : \Gamma^* \rightarrow \Sigma^*$ as follows:

For π of the form $A \rightarrow BC$, take

$$h([^1_\pi) = h(]^1_\pi) = h([^2_\pi) = h(]^2_\pi) = \epsilon,$$

For π of the form $A \rightarrow a$, take

$$h(]^1_\pi) = h([^2_\pi) = h(]^2_\pi) = \epsilon,$$
$$h([^1_\pi) = a$$

Applying h to the production π of P' gives the production π of P thus $L(G) = h(L(G')) = h(PAREN_\Gamma \cap R_S)$.

This completes the proof of the Chomsky-Schützenberger theorem.

Example

Apply the theorem on $\{\ a^n b^n \mid n >= 0\ \} = h(L(G^{'}) \cap R)$

Example continued...

Let $G = (N, \Sigma, P, S)$ be the CFG corresponding to our CFL $\{a^n b^n | n \;= 0\}$ where,

$$S \to aSb$$
$$S \to \epsilon$$

Converting it to CNF we get

$$\begin{array}{ll} \pi & S \to AC \\ \sigma & C \to SB \\ \rho & A \to a \\ \lambda & B \to b \\ \gamma & S \to \epsilon \end{array}$$

Example continued...

Now define grammar $G' = (N, \Gamma, P', S)$ with

$$\Gamma = \{[^1_\pi,]^1_\pi, [^2_\pi,]^2_\pi | \pi \in P\} \ ,$$
$$P' = \{\pi' | \pi \in P\} \ .$$

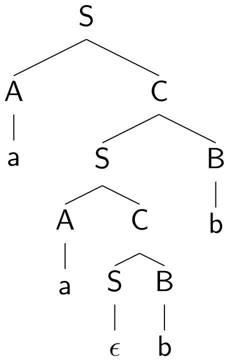
where production P' are,

π	$S \rightarrow [^1_\pi A]^1_\pi [^2_\pi C]^2_\pi$
σ	$C \rightarrow [^1_\sigma S]^1_\sigma [^2_\sigma B]^2_\sigma$
ρ	$A \rightarrow [^1_\rho]^1_\rho$
λ	$B \rightarrow [^1_\lambda]^1_\lambda$
γ	$S \rightarrow [^1_\gamma]^1_\gamma$

Example continued...

Consider a string x generated by grammar G: $x = aabb$

The parse tree generated by G is



THANK YOU