Propiedades de los Lenguajes Libres de Contexto

2 de marzo de 2017

Contenido

Introducción

- *Simplificación de CFG's*. Esto facilita la vida, porque podemos decir que si un lenguaje es CF, entonces tiene una gramática de una forma especial.
- *Lema de Pumping para CFL's*. Similar al caso regular. No lo vamos a cubrir.
- *Propiedades de cerradura*. Algunas, pero no todas, de las propiedades de cerradura de los lenguajes regulares se cumplen en los CFL's.
- *Propiedades de decisión*. Podemos probar la membresía y vacío, pero por ejemplo, la equivalencia de CFL's no es decidible.

Forma Normal de Chomsky (CNF)

Queremos mostrar que todo CFL (sin ϵ) se genera por una CFG donde todas las producciones son de la forma:
$A \rightarrow BC$ o $A \rightarrow a$
Donde A, B, y C son variables, y a es un terminal. A esto se le conoce como CNF, y para llegar a ella debemos:

1. Eliminar símbolos no-útiles, aquellos que no aparecen en ninguna derivación $S \stackrel{*}{\Rightarrow} w$, para el símbolo de inicio S y la cadena de símbolos terminales w.
2. Eliminar las producciones-ϵ, es decir, producciones de la forma $A \rightarrow \epsilon$.
3. Eliminar producciones unitarias, es decir, producciones de la forma $A \rightarrow B$, donde A y B son variables.

Eliminando símbolos no-útiles

- Un símbolo X es *útil* para una gramática $G = (V, T, P, S)$, si hay una derivación: $S \stackrel{*}{\Rightarrow}_G \alpha X \beta \stackrel{*}{\Rightarrow}_G w$ para una cadena terminal w. A los símbolos que no son útiles se les denomina *inútiles*.
- Un símbolo X es *generador* si $X \stackrel{*}{\Rightarrow}_G w$, para alguna $w \in T^*$.
- Un símbolo X es *alcanzable* si $S \stackrel{*}{\Rightarrow}_G \alpha X \beta$, para algún $\alpha, \beta \subseteq (V \cup T)^*$.
- Un símbolo útil es generador y alcanzable.
- Cabe notar que si eliminamos a los símbolos no-generadores primero, y luego a los no-alcanzables, nos quedamos únicamente con símbolos útiles.

Ejemplo

- Sea G: $S \to AB|a, A \to b$.
- S y A son generadores, B no lo es. Si eliminamos B tenemos que eliminar $S \to AB$, dejando la gramática $S \to a, A \to b$. Ahora sólo S es alcanzable.
- Eliminando A y b nos deja con $S \to a$. Con el lenguaje $\{a\}$.
- El orden importa, de otra manera (para este ejemplo), si eliminamos primero los símbolos no-alcanzables, nos damos cuenta de que todos los símbolos son alcanzables.
- A partir de: $S \to AB|a, A \to b$. Después eliminamos B como no-generador, y nos quedamos con $S \to a, A \to b$, que todavía contiene símbolos inútiles.

Teorema

Sea $G = (V, T, P, S)$ una CFG tal que $L(G) \neq \emptyset$. Sea $G_1 = (V_1, T_1, P_1, S)$ la gramática obtenida:

1. Eliminando todos los símbolos no-generadores y las producciones en las que ocurren. Sea la nueva gramática $G_2 = (V_2, T_2, P_2, S)$.
2. Eliminando de G_2 todos los símbolos no-alcanzables y las producciones en que ocurren.

G_1 no tiene símbolos inútiles, y $L(G_1) = L(G)$.

Cálculo de Símbolos Generadores y Alcanzables

- Necesitamos algoritmos para calcular los símbolos generadores y alcanzables de $G = (V, T, P, S)$.
- Los símbolos generadores $g(G)$ se calculan con el siguiente algoritmo de cerradura:
  1. **Base:** Todo símbolo de T es generador, se genera a sí mismo.
  2. **Inducción:** Suponemos que tenemos la producción $A \to \alpha$, y cada símbolo de α es generador. Entonces A es generador (esto incluye $\alpha = \epsilon$, las reglas que tienen a ϵ en el cuerpo son generadoras).

Ejemplo

- Sea G: $S \to AB|a, A \to b$
- Entonces, primero $g(G) = \{a, b\}$.
- Como $S \to a$ ponemos a S en $g(G)$, y porque $A \to b$ añadimos también a A, y eso es todo, el conjunto de símbolos generadores es $\{a, b, A, S\}$.

Teorema

- El algoritmo anterior encuentra todos y sólo los símbolos generadores de G.
- El conjunto de símbolos alcanzables $r(G)$ de $G = (V, T, P, S)$ se calcula con el siguiente algoritmo de cerradura:
  1. *Base*: $r(G) = \{S\}$, S es alcanzable.
  2. *Inducción*: Si la variable $A \in r(G)$ y $A \to \alpha \in P$ entonces se añaden todos los símbolos de α a $r(G)$.

Ejemplo

- Sea $G : S \to AB|a, A \to b$
- Entonces, primero $r(G) = \{S\}$.
- Con base en la primera producción añadimos $\{A, B, a\}$ a $r(G)$.
- Con base en la segunda producción añadimos $\{b\}$ a $r(G)$ y eso es todo.

Teorema: El algoritmo anterior encuentra todos y solo los símbolos alcanzables de G.

Eliminando Producciones-ϵ

- Aunque las producciones ϵ son convenientes, no son esenciales. Si L es CF, entonces $L - \{\epsilon\}$ tiene una CFG sin producciones ϵ.
- La estrategia consiste en descubrir cuáles variables son *nulificables*.
- Se dice que la variable A es *nulificable* si $A \stackrel{*}{\Rightarrow} \epsilon$.
- Sea A nulificable, entonces en todas las producciones en donde A aparece en el cuerpo, digamos $B \rightarrow CAD$, creamos dos versiones de la producción, una sin A, $B \rightarrow CD$ y otra con A, $B \rightarrow CAD$. Si utilizamos la producción con A no permitimos que A derive a ϵ.

Algortimo

- El siguiente algoritmo calcula $n(G)$, el conjunto de símbolos nulificables de una gramática $G = (V, T, P, S)$ como sigue:
  1. *Base*: $n(G) = \{A : A \rightarrow \epsilon \in P\}$
  2. *Inducción*: Si $\{C_1, C_2, \ldots, C_k\} \subseteq n(G)$ y $A \rightarrow C_1, C_2, \ldots, C_k \in P$, entonces $n(G) = n(G) \cup \{A\}$.
  3. Nota, cada C_i debe ser una variable para ser nulificable, entonces se consideran sólo las producciones con cuerpos conformados de variables.

Teorema

- En cualquier gramática G, los únicos símbolos nulificables son las variables encontradas por el algoritmo anterior.
- Una vez que conocemos los símbolos nulificables, podemos transformar G en G_1 como sigue:
  1. Para cada $A \rightarrow X_1 X_2 \dots X_k \in P$ con $m \leq k$ símbolos nulificables, reemplazar por 2^m reglas, una con cada sub-lista de los símbolos nulificables ausentes.
  2. Excepción: Si $m = k$ no añadimos la regla donde borramos todos los m símbolos nulificables.
  3. Borrar todas las reglas de la forma $A \rightarrow \epsilon$.

Ejemplo

- Considere la siguiente gramática:
  $S \rightarrow AB$
  $A \rightarrow aAA|\epsilon$
  $B \rightarrow bBB|\epsilon$
- A y B son nulificables porque tienen a ϵ en el cuerpo de una de sus producciones.
- S también es nulificable, porque $S \rightarrow AB$ tiene puros símbolos nulificables.
- Ahora para construir las nuevas producciones sin ϵ, consideremos la primera: $S \rightarrow AB$.
- De aqui construimos las producciones con y sin los símbolos nulificables (y sin eliminar todas):
  $S \rightarrow AB|A|B$.

Ejemplo (cont.)

- Para $A \rightarrow aAA$, hacemos algo parecido y nos queda: $A \rightarrow aAA|aA|aA|a$, que como hay dos iguales podemos elimiar una.
- Finalmente para B es parecido, por lo que la gramática final queda como:
  $S \rightarrow AB|A|B$
  $A \rightarrow aAA|aA|a$
  $B \rightarrow bBB|bB|b$
- La gramática anterior no cambia el lenguaje, excepto que ϵ ya no está presente.

Teorema

- Si la gramática G_1 se construye a partir de G con la construcción anterior para eliminar producciones ϵ, entonces $L(G_1) = L(G) - \{\epsilon\}$.

Eliminando Producciones Unitarias

- $A \rightarrow B$ es una producción unitaria, cuando A y B son variables. Las producciones unitarias se pueden eliminar.
- Veamos la gramática:
  $I \rightarrow a|b|Ia|Ib|I0|I1$
  $F \rightarrow I|(E)$
  $T \rightarrow F|T * F$
  $E \rightarrow T|E + T$
  tiene las producciones unitarias $E \rightarrow T$, $T \rightarrow F$, y $F \rightarrow I$.
- Podemos expandir T en la producción $E \rightarrow T$ y obtener: $E \rightarrow F|T * F$.
- Expandiendo $E \rightarrow F$ nos da: $E \rightarrow I|(E)$.
- Finalmente expandemos $E \rightarrow I$ y obtenemos:
  $E \rightarrow a|b|Ia|Ib|I0|I1|(E)|T * F|E + T$

Eliminando Producciones Unitarias

- El método de expansión trabaja siempre y cuando no haya ciclos en las reglas, por ejemplo en: $A \rightarrow B, B \rightarrow C, C \rightarrow A$.
- Para calcular $u(G)$, el conjunto de todos los pares unitarios de $G = (V, T, P, S)$ utilizamos el siguiente algoritmo de cerradura.

Algoritmo

- *Base*: (A, A) es un par unitario para cualquier variable A. Esto es, $A \stackrel{*}{\Rightarrow} A$ en cero pasos.
  $u(G) = \{(A, A) : A \in V\}$.
- *Inducción*: Suponemos que $(A, B) \in u(G)$ y que $B \to C \in P$ donde C es una variable. Entonces añadimos (A, C) a $u(G)$.

Teorema: El algoritmo anterior encuentra todos y solo los pares unitarios de una CFG G.

Algoritmo

Para eliminar producciones unitarias, procedemos de la siguiente manera. Dada $G = (V, T, P, S)$, podemos construir $G_1 = (V, T, P_1, S)$:

1. Encontrando todos los pares unitarios de G.
2. Para cada par unitario (A, B), añadimos a P_1 todas las producciones $A \to \alpha$, donde $B \to \alpha$ es una producción no unitaria en P.
3. Note que es posible tener $A = B$; de esta manera, P_1 contiene todas las producciones unitarias en P.
4. $P_1 = \{A \to \alpha : \alpha \notin V, B \to \alpha \in P, (A, B) \in u(G)\}$

Ejemplo

A partir de la gramática:

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

Creamos un nuevo conjunto de producciones usando el primer elemento del par como cabeza y todos los cuerpos no unitarios del segundo elemento del par como cuerpos de las producciones:

Ejemplo (cont.)

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

Ejemplo (cont.)

Eliminamos las producciones unitarias. La gramática resultante es equivalente a la original.

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

$T \rightarrow T * F|(E)|a|b|Ia|Ib|I0|I1$

$F \rightarrow (E)|a|b|Ia|Ib|I0|I1$

$I \rightarrow a|b|Ia|Ib|I0|I1$

Resumen

Para “limpiar” una gramática podemos:

1. Eliminar producciones-ϵ
2. Eliminar producciones unitarias
3. Eliminar símbolos inútiles

en este orden.

Forma Normal de Chomsky, CNF

- Ahora se mostrará que cada CFL no vacío sin ϵ tiene una gramática G sin símbolos inútiles, de tal manera que cada producción tenga la forma: $A \to BC$, donde $\{A, B, C\} \subseteq T$, o $A \to \alpha$, donde $A \in V$, y $\alpha \in T$.
- Para lograr esto, iniciamos con alguna gramática para el CFL, y:
  1. "Limpiamos la gramática".
  2. Hacemos que todos los cuerpos de longitud 2 o más consistan solo de variables.
  3. Dividimos los cuerpos de longitud 3 o más en una cascada de producciones con cuerpos de dos variables.

Forma Normal de Chomsky, CNF

- Para el paso 2, por cada terminal a que aparece en un cuerpo de longitud ≥ 2, creamos una nueva variable, A, y reemplazamos a a por A en todos los cuerpos. Después añadimos una nueva regla $A \rightarrow a$.
- Para el paso 3, por cada regla de la forma $A \rightarrow B_1 B_2 \dots B_k$, $k \geq 3$, introducimos variables nuevas $C_1, C_2, \dots C_{k-2}$, y reemplazamos la regla con:
  $A \rightarrow B_1 C_1$
  $C_1 \rightarrow B_2 C_2$
  $\dots$
  $C_{k-3} \rightarrow B_{k-2} C_{k-2}$
  $C_{k-2} \rightarrow B_{k-1} B_k$

Ejemplo

Iniciamos con la gramática (el paso 1 ya está hecho):

- $E \to E + T|T * F|(E)|a|b|Ia|Ib|I0|I1$
- $T \to T * E|(E)|a|b|Ia|Ib|I0|I1$
- $F \to (E)|a|b|Ia|Ib|I0|I1$
- $I \to a|b|Ia|Ib|I0|I1$

Para el paso 2, introducimos nuevas variables y nos quedan las siguientes reglas:

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

Ejemplo (cont.)

y al reemplazar obtenemos la gramática:

$E \rightarrow EPT|TMF|LER|a|b|IA|IB|IZ|IO$

$T \rightarrow TPE|LEL|a|b|IA|IB|IZ|IO$

$F \rightarrow LER|a|b|IA|IB|IZ|IO$

$I \rightarrow a|b|IA|IB|IZ|IO$

$A \rightarrow a, B \rightarrow b, Z \rightarrow 0, O \rightarrow 1$

$P \rightarrow +, M \rightarrow *, L \rightarrow (, R \rightarrow)$

Ejemplo (cont.)

Para el paso 3, reemplazamos:

- $E \to EPT$ por $E \to EC_1, C_1 \to PT$
- $E \to TMF, T \to TMF$ por
  $E \to TC_2, T \to TC_2, C_2 \to MF$
- $E \to LER, T \to LER, F \to LER$ por
  $E \to LC_3, T \to LC_3, F \to LC_3, C_3 \to ER$

Ejemplo (cont.)

La gramática CNF final es:

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

Ejemplo

- Encuentre una gramática equivalente para:
  $S \rightarrow AB|CA$
  $A \rightarrow a$
  $B \rightarrow BC|AB$
  $C \rightarrow aB|b$
  sin símbolos inútiles
- Con la gramática:
  $S \rightarrow ASB|\epsilon$
  $A \rightarrow aAS|a$
  $B \rightarrow SbS|A|bb$
  Eliminar: (a) producciones ϵ, (b) producciones unitarias, (c) símbolos inútiles, y (d) ponerla en CNF

Solución al 1

- A y C son generadoras por que tienen producciones con terminales. Como S tienen una producción con puros generadores (CA), entonces también en generador. B no es generador, por lo que si lo eliminamos y todas las producciones donde aparece y nos quedamos solo con los alcanzables, la gramática queda:
  $S \rightarrow CA$
  $A \rightarrow a$
  $C \rightarrow b$

Solución al 2

- Solo S es nulificable, por lo que tenemos que ponerla y quitarla en cada lugar donde ocurre:
  $S \rightarrow ASB|AB$
  $A \rightarrow aAS|aA|a$
  $B \rightarrow SbS|Sb|bS|b|A|bb$
- La única producción unitaria es: $B \rightarrow A$ por lo que la reemplazamos directamente:
  $S \rightarrow ASB|AB$
  $A \rightarrow aAS|aA|a$
  $B \rightarrow SbS|Sb|bS|b|aAS|aA|a|bb$
- A y B generan símbolos terminales, y por lo tanto también S, por lo que no hay símbolos inútiles

Solución al 2 (cont.)

- Introducir variabes y producciones $C \rightarrow a$ y $D \rightarrow b$ y ponerla en todos los cuerpos que no tienen un solo símbolo terminal
  $S \rightarrow ASB|AB$
  $A \rightarrow CAS|CA|a$
  $B \rightarrow SDS|SD|DS|b|CAS|CA|a|DD$
  $C \rightarrow a$
  $D \rightarrow b$
- Para las producciones con más de 3 símbolos se introducen nuevas variables:
  $S \rightarrow AE|AB$
  $A \rightarrow CF|CA|a$
  $B \rightarrow SG|SD|DS|b|CF|CA|a|DD$
  $C \rightarrow a$, $D \rightarrow b$
  $E \rightarrow SB$, $F \rightarrow AS$
  $G \rightarrow DS$

Propiedades de Cerradura de los CFL's

- Existen varias propiedades de los CFL, una de las más importantes es la de substitución.
- **Substitución**: Sea Σ un alfabeto y supongamos que para cada símbolo $a \in \Sigma$ definimos un lenguaje arbitrario L_a.
- Estos lenguajes definen una función s o una substitución.
- Si $w = a_1 a_2 \ldots a_n$ es una cadena en Σ^*, $s(w)$ es la concatenación de los lenguajes $s(a_1)s(a_2)\ldots s(a_n)$.

Ejemplo

- $\Sigma = \{0, 1\}, s(0) = \{a^n b^n : n \geq 1\}, s(1) = \{aa, bb\}$.
- Sea $w = 01$. Entonces $s(w) = s(0)s(1) = \{a^n b^n aa : n \geq 1\} \cup \{a^n b^{n+2} : n \geq 1\}$.
- Si $L = \{0\}^*$, entonces $s(L) = (s(0))^* = \{a^{n_1} b^{n_1} a^{n_2} b^{n^2} \ldots a^{n_k} b^{n_k} : k \geq 0, n_i \geq 1\}$

Teoremas

- **Teoremas**: Sea L un CFL sobre Σ, y s una substitución, tal que $s(a)$ sea un CFL, $\forall a \in \Sigma$. Entonces $s(L)$ es un CFL.
- **Teoremas**: Si tenemos uno o más CFL's, también son CFL el resultados de hacer: (i) unión, (ii) concatenación, (iii) Cerradura de Kleene, (iv) cerradura positiva $+$, (v) inversión, (vi) homomorfismo, (vii) homomorfismo inverso.
- **Teorema**: Si L, L_1, L_2 son CFL's, y R es lenguaje regular, entonces $L \cap R$, $L \setminus R$ son CFL y $\overline{L}$, $L_1 \setminus L_2$ no son necesariamente CFL's.

Probrar membresía en eun CFL

- Usar el algoritmo CYK (construye tabla triangular)
- En el primer renglón pone producciones tipo $A \to a$
- En el resto pone producciones que tengan una parte del prefijo de la cadena seguida de el resto de la cadena

X_{15}				
X_{14}	X_{25}			
X_{13}	X_{24}	X_{35}		
X_{12}	X_{23}	X_{34}	X_{45}	
X_{11}	X_{22}	X_{33}	X_{44}	X_{55}
a_1	a_2	a_3	a_4	a_5

Ejemplo

- Probar que la siguiente gramática genera: *baaba*
  $S \rightarrow AB|BC$
  $A \rightarrow BA|a$
  $B \rightarrow CC|b$
  $C \rightarrow AB|a$

{S,A,C}				
-	{S,A,C}			
-	{B}	{B}		
{S,A}	{B}	{S,C}	{S,A}	
{B}	{A,C}	{A,C}	{B}	{A,C}
b	*a*	*a*	*b*	*a*

Problemas no-decidibles

Los siguientes son problemas no-decidibles:

1. Es una CFG G dada ambigua?
2. Es un CFL dado inherentemente ambiguo?
3. Es la intersección de dos CFL's vacía?
4. Son dos CFL's el mismo?
5. Es un CFL dado universal (igual a Σ^*)?