CYK algoritmus*

1 Chomsky-normálforma, ismétlés

Definíció 1.1. Chomsky-normálforma Egy $G = (N, T, P, S)$ környezetfüggetlen grammatikát Chomsky-normálformájúnak nevezünk, ha minden szabálya a következő alakú: $A \to a$ vagy $A \to BC$, ahol $a \in T, B, C \in N$.

Tétel 1.1. Minden ε-mentes környezetfüggetlen grammatikához megkonstruálható egy vele ekvivalens Chomsky-normálformájú környezetfüggetlen grammatika.

1. feladat: Határozzuk meg a következő grammatika Chomsky-nomrálformáját! $G = (\{A, B, S\}, \{a, +, *, (,)\}, P, S)$, P:

$S \to A|A + S$

$A \to B|B * A$

$B \to a|(S)$

Megoldás:

- ε-mentesítés: ✓
- Álterminálisok, hosszredukció:

  $S \to A|A\underline{X_1S}$

  $A \to B|B\underline{X_2A}$

  $B \to a|X_3\underline{SX_4}$

  $X_1 \to +$

  $X_2 \to *$

*A jegyzet Dr Csuhaj-Varjú Erzsébet, Dr Tichler Krisztián, Nagy Sára, Veszprémi Anna anyagai alapján készült

$X_3 \rightarrow ($

$X_4 \rightarrow)$

$S \rightarrow A|AZ_1$

$A \rightarrow B|BZ_2$

$B \rightarrow a|X_3Z_3$

$Z_1 \rightarrow X_1S$

$Z_2 \rightarrow X_2A$

$Z_3 \rightarrow SX_4$

$X_1 \rightarrow +$

$X_2 \rightarrow *$

$X_3 \rightarrow ($

$X_4 \rightarrow)$

- Lánctalanítás: láncszabályok: $S \rightarrow A, A \rightarrow B$

$S \rightarrow BZ_2|a|X_3Z_3|AZ_1$

$A \rightarrow a|X_3Z_3|BZ_2$

$B \rightarrow a|X_3Z_3$

$Z_1 \rightarrow X_1S$

$Z_2 \rightarrow X_2A$

$Z_3 \rightarrow SX_4$

$X_1 \rightarrow +$

$X_2 \rightarrow *$

$X_3 \rightarrow ($

$X_4 \rightarrow)$

2 CYK algoritmus

A CYK algoritmus (Cocke–Younger–Kasami) egy alulról felfelé elemzést végrehajtó algoritmus, amellyel eldönthető a tartalmazás problémája adott környezetfüggetlen grammatika és szó esetén, azaz adott G grammatika és $u = (u_1 \cdots u_n) \in T^*$ szó esetén eldönti, hogy $u \in^? L(G)$. A grammatikának Chomsky-normálformában kell lennie. Az algoritmus során egy alsóháromszög mátrixot töltünk ki. A mátrix celláit alulról felfelé, oszlopait balról jobbra számozzuk. Az egyes cellák nemterminális szimbólumokat tartalmaznak.

Az algoritmus lényegében részszavak levezethetőségét vizsgálja, első lépésben az egy szimbólumból álló résszszavakat, legutolsó lépésben a teljes szót. A mátrix (i,j)-dik cellájában egy olyan nemterminális szerepel, melyből levezethető az $u_{j,j+i-1}$ résszó, amennyiben ilyen van, egyéként nem szerepel semmi, azaz olyan A nemterminálisok kerülnek melyekre, $A \to BC \in P$ és B szerepel az (k,j) cellában, C a $(i-k, j+k)$ cellában, valamely $k \in [1..i-1]$. Ez lényegében azt jelenti, hogy B levezeti $u_j \cdots u_k$-t, C pedig $u_{j+k+1} \cdots u_{j+i-1}$ részszót. Ha a mátrix $(n,1)$ indexű (bal felső) cellájában megjelenik a start szimbólum, akkor a szó eleme a grammatika által generált nyelvnek.

A CYK algoritmus futási ideje $O(n^3)$. Az első sor kitöltése n lépés, a második $n-1$ a harmadik $2*(n-2)$, ... , az n. $(n-1)*1$.

2. feladat: Adott a következő CNF grammatika és az $u = aabbcc$ szó, szemléltessük a CYK algoritmus működését!

$S \to AB|BC$

$A \to XA|a$

$X \to a$

$C \to YC|c$

$Y \to c$

$B \to UV|VW$

$U \to XX$

$W \to YY$

$V \to ZZ$

$Z \to b$

A bemenetként kapott szó hossza 6, így először készítünk egy 6×6-os táblázatot. Az első sor kitöltésekor 1 hosszúságú résszavak levezetését vizsgáljuk, azaz olyan $A \to x$ alakú szabályokat keresünk, ahol $A \in N, x \in T$, tehát olyan szabályokat, melyek közvelenül terminális szimbólumot vezetnek le.

	1	2	3	4	5	6
6						
5						
4						
3						
2						
1	A, X	A, X	Z	Z	C, Y	C, Y
	a	a	b	b	c	c

A második sor kitöltésekor a 2 hosszúságú résszavak levezethetőségét vizsgáljuk. Itt már $A \to BC$ alakú szabályokkal foglalkozunk. Nézzük meg például, hogyan kaphatjuk meg a 2. sor 3. cellájának értékét. A fenti képleteknek megfelelően a mátrix $(2, 3)$ cellájában olyan nemterminális(ok) szerepel(nek), melyekből levezethető az eredeti szavunk 3. karakterétől kezdődő 2 hosszúságú résszszava, azaz $u_{3,4}$. Ekkor tehát olyan $A \to BC$ szabályokat keresünk, melyre B a mátrix k. sorának j. C pedig a mátrix $i - k$. sorának $j + k$ cellájában szerepel (a sorszám mindig a részszó hosszát, az oszlopszám a résszó első karakterének indexét adja meg). Konkrét értékeket behelyettesítve (figyelembe véve, hogy k most csak 1 lehet) a mátrix $(1, 2)$ és $(2 - 1, 2 + 1)$ celláit vesszük figyelembe. Mindkét cellában Z nemterminális szerepel, tehát olyan szabályt keresünk, melynek jobboldala ZZ, találhatunk is egy ilyen szabályt: $V \to ZZ$, ennek megfelelően a mátrix $(2, 3)$ cellájába V nemterminális kerül.

	1	2	3	4	5	6
6						
5						
4						
3						
2	A, U	∅	V	∅	C, W	
1	A, X	A, X	Z	Z	C, Y	C, Y
	a	a	b	b	c	c

A harmadik sor kitöltésekor 3 hosszúságú résszavakat vizsgálunk. Most nézzük meg, hogyan tölthetjük ki, a 3. sor 2. celláját! Ekkor tehát a 2. karaktertől kezdődő, 3 hosszúságú részszó levezethetőségét vizsgáljuk ($u_{2,4}$). A képletek alpján (k, j) és $(i - k, j + k)$ cellákat kell figyelnünk, ahol $k \in [1..2]$. Ez azt jelenti, hogy a következő esetek lesznek:

- Ha $k = 1$: $(1,2)$ és $(2,3)$ (zöld cellák), ekkor az $(1,2)$ cellában lévő nemterminális levezeti $u_{2,2}$-t (a 2. karaktert), a $(2,3)$-ban lévő pedig $u_{3,4}$-et. Olyan szabályokat keresünk tehát, melynek jobboldala AV vagy XV, ilyet azonban nem találunk, úgyhogy egyelőre a $(3,2)$ cella üresen marad.
- Ha $k = 2$: $(2,2)$ és $(1,3)$ (sárga cellák), mivel $(2,2)$ egy üres cella, biztosan nem lesz találtunk most sem.

A fenti két eset alapján a $(3,2)$ cella üres marad.

	1	2	3	4	5	6
6						
5						
4						
3	∅	∅	∅	∅		
2	A, U	∅	V	∅	C, W	
1	A, X	A, X	Z	Z	C, Y	C, Y
	a	a	b	b	c	c

A többi sor kitöltése hasonlóan.

	1	2	3	4	5	6
6	S					
5	S	S				
4	B	∅	B			
3	∅	∅	∅	∅		
2	A, U	∅	V	∅	C, W	
1	A, X	A, X	Z	Z	C, Y	C, Y
	a	a	b	b	c	c

A $(6,1)$ cella képviseli, az 1. karaktertől induló 6 hosszúságú részszót, ami maga az u szó. Mivel ebben a cellában végül megjelenik a startszimbólum $u \in L(G)$

3. feladat: Adott a 1. feladat CNF grammatikája és az $(a + a) * a$ szó, szemléltessük a CYK algoritmus működését!

	1	2	3	4	5	6	7
7	S						
6	$\emptyset$	$\emptyset$					
5	S, A, B	$\emptyset$	$\emptyset$				
4	$\emptyset$	Z_3	$\emptyset$	$\emptyset$			
3	$\emptyset$	S	$\emptyset$	$\emptyset$	$\emptyset$		
2	$\emptyset$	$\emptyset$	Z_1	Z_3	$\emptyset$	Z_2	
1	X_3	S, A, B	X_1	S, A, B	X_4	X_2	S, A, B
	(	a	+	a	)	*	a

2.1 CYK algoritmus - szintaxis fa

A CYK algoritmust szintaxis fa építésre is használhatjuk. Ehhez megszámozzuk a grammatika szabályait. Ha a mátrix (i, j) cellájában A nemterminális szerepel, amely az $A \to BC$ szabály baloldala, akkor mellé írjuk a szabály sorszámát, továbbá azt a k számot, amelyre B a mátrix j. oszlopának k. sorában, C pedig a $(j + k)$. oszlop $(i - k)$. sorában szerepel.

Az algoritmus futtatása után a mátrix $(1, 1)$ cellájából indulva az indexek segítségével rekurzívan lépkedve a szintaxis fa felépíthető. Amennyiben egy cellában több nemterminális található, úgy a fa többféleképp is felépíthető, a szó nem egyertélemű́en áll elő a grammatikából.

Például a következő grammatika és az $u = aaaab$ szó esetén:

Először lejátsszuk az algoritmust, kiegészítve a fentiekkel:

$S \to AB^1$

$A \to AA^2|a^3$

$B \to b^4$

	1	2	3	4	5
5	$S^{4,1}$				
4	$A^{1,2}$	$S^{3,1}$			
3	$A^{1,2}$	$A^{1,2}$	$S^{2,1}$		
2	$A^{1,2}$	$A^{1,2}$	$A^{1,2}$	$S^{1,1}$	
1	$A^{0,3}$	$A^{0,3}$	$A^{0,3}$	$A^{0,3}$	$B^{0,4}$
	a	a	a	a	b

Induljunk ki az $(5, 1)$ cellából, itt azt láthatjuk, hogy az algorimtus futása során $k = 4$ volt, továbbá az 1. szabályt alkalmaztuk ($S \to AB$). Ez azt jelenti, hogy a fa gyökere nyílván S lesz, és ennek két gyereke A, valamint B. A kérdés az, hogy a két gyereket a mátrix mely celláiban keressük. Mivel tudjuk, hogy $k = 4$ volt, ebből következik, hogy az első gyerek a mátrix $(4, 1)$ cellájában lesz, a második pedig $(5 - 4 = 1, 1 + 4 = 5)$ cellájában (zöld háttér) lesz. A fa építését ugyanígy ezeken a cellákon keresztül folytatjuk.

Az előálló szintaxisfa, a csúcsokat az adott nemterminális mátrixbeli pozíciója alapján $(sor, oszlop)$ formában címkézzük.