Formális nyelvek - 3.

Csuhaj Varjú Erzsébet

Algoritmusok és Alkalmazásaik Tanszék
Informatikai Kar
Eötvös Loránd Tudományegyetem
H-1117 Budapest
Pázmány Péter sétány 1/c

E-mail: csuhaj@inf.elte.hu

Környezetfüggetlen grammatikák normálformái

Grammatikai transzformációkkal nyert grammatikák,

- melyek bizonyos szintaktikai feltételeknek vagy **tulajdonságoknak** tesznek eleget,
- **általában valamilyen szempontból egyszerűbbek, mint az eredeti grammatikák**,
- de **ugyanazon típusba** tartoznak,
- és **ugyanazt a nyelvet generálják**.

Tétel (ε-mentesítés)

Minden $G = (N, T, P, S)$ **környezetfüggetlen grammatikához** meg tudunk konstruálni egy **vele ekvivalens** $G' = (N', T, P', S')$ **környezetfüggetlen grammatikát** úgy, hogy

- G' minden szabályának jobboldala **nemüres szó**,
- **kivéve** azt az esetet, ha az **üres szó benne van a** G **által generált nyelvben**,
- mely esetben $S' \to \varepsilon$ az **egyetlen** olyan szabály, melynek jobboldala az üres szó és ekkor S' **nem fordul elő** a G' egyetlen szabályának jobboldalán sem.

Bizonyításvázlat:

Tekintsük a $G = (N, T, P, S)$ környezetfüggetlen grammatikát. Ha P nem tartalmaz $X \to \varepsilon$ alakú szabályt, akkor $G' = G$.

Tegyük fel, hogy P-ben van $X \to \varepsilon$ alakú szabály. Definiáljuk az $U_i \subseteq N$ halmazokat a következőképpen:

$$U_1 = \{X \mid X \to \varepsilon \in P\},$$

$$U_{i+1} = U_i \cup \{X \mid X \to u \in P \text{ és } u \in U_i^*\},\ i \geq 1.$$

Nyilvánvaló, hogy az U_i sorozat, $i = 1, 2, \ldots,$ a tartalmazásra nézve hierarchiát alkot és van olyan k index, hogy $U_k = U_{k+1}$ és így $U_k = U_j$ minden $j \geq k$-ra.

Legyen $U = U_k$.

Ekkor azonnal látható, hogy $X \Longrightarrow_G^* \varepsilon$ akkor és csak akkor, ha $X \in U$.

(Vagyis, $\varepsilon \in L(G)$ akkor és csak akkor, ha $S \in U$.)

Bizonyításvázlat - folytatás:

Ezután megkonstruáljuk a P_1 szabályhalmazt a következőképpen:

Minden olyan $X \to u$ szabály benne van P_1-ben, amelyre $u \neq \varepsilon$ és van olyan $v \in (N \cup T)^*$ sztring, hogy $X \to v \in P$ és u-t v-ből úgy kapjuk meg, hogy U-beli nemterminálisok valahány, azaz nulla vagy több előfordulását elhagyjuk v-ből. P_1-ben nincs más szabály.

(**Példa**: Legyen $A, B \in U$ és $C \notin U$, akkor az $S \to ACAB$ szabályból a következő szabályokat képezzük: $S \to ACAB$, $S \to CAB$, $S \to ACB$, $S \to ACA$, $S \to CB$, $S \to CA$, $S \to AC$, $S \to C$.)

Legyen $G_1 = (N, T, P_1, S)$. Ekkor látható, hogy $L(G_1) \subseteq L(G) - \{\varepsilon\}$, hiszen minden $X \to u$ szabály alkalmazása megfelel az $X \to v$ szabály alkalmazásának, amelyet valahány $Z \Longrightarrow_G^* \varepsilon$ levezetés alkalmazásával kombinálunk, ahol $Z \in U$ és Z előfordul v-ben.

Megfordítva, ha $S \Longrightarrow_G^* u$ és $u \neq \varepsilon$, akkor $S \Longrightarrow_{G_1}^* u$, hiszen az $X \to \varepsilon$ típusú szabályok alkalmazása elkerülhető P_1 megfelelő szabályának alkalmazásával.

Bizonyításvázlat - folytatás:

A fentiek alapján $L(G_1) = L(G) - \{\varepsilon\}$. Ha $\varepsilon \notin L(G)$, akkor $G' = G_1$.

Ha $\varepsilon \in L(G)$, akkor vesszük a $G' = (N \cup \{S'\}, T, P_1 \cup \{S' \to \varepsilon, S' \to S\}, S')$ grammatikát, amely az $L(G)$ nyelvet generálja.

ε-mentes grammatika

Definíció

A G grammatikát ε-**mentesnek** nevezzük, ha egyetlen szabályának jobboldala sem az üres szó.

Tétel

Minden környezetfüggetlen G grammatikához meg tudunk konstruálni egy G' ε-mentes környezetfüggetlen grammatikát, amelyre $L(G') = L(G) - \{\varepsilon\}$ teljesül.

Az állítás közvetlen következménye a megelőző állításnak, az előző bizonyításban konstruált G_1 grammatika pontosan ilyen.

Környezetfüggetlen grammatikák Chomsky-normálformája

Definíció

A $G = (N, T, P, S)$ környezetfüggetlen grammatikát **Chomsky-normálformájúnak** mondjuk, ha minden egyes szabálya vagy

1. $X \to a$, ahol $X \in N$, $a \in T$, vagy

2. $X \to YZ$, ahol $X, Y, Z \in N$ alakú.

Chomsky-normálforma - folytatás:

Tétel:

Minden ε-mentes $G = (N, T, P, S)$ környezetfüggetlen grammatikához meg tudunk konstruálni egy vele **ekvivalens** $G' = (N', T, P', S)$ **Chomsky-normálformájú** környezetfüggetlen grammatikát.

Bizonyításvázlat:

Ha minden szabály $X \to a$ vagy $X \to YZ$ alakú, ahol $a \in T$, $X, Y, Z \in N$, akkor $G' = G$. Ha nem, akkor a következőképpen járunk el.

Az általánosság megszorítása nélkül feltehetjük, hogy a P szabályhalmaz elemei terminális szimbólumokat csak az $X \to a$, $a \in T$ alakú szabályokban tartalmaznak (egy megelőző normálforma tételhez hasonlóan járunk el). Ebben az esetben minden további szabály $X \to u$ alakú, ahol $u \in N^+$.

3) Ekkor minden

$$X \to Y_1 Y_2 \dots Y_k,\ k \geq 3$$

alakú szabályt helyettesítünk egy

$$\begin{aligned} X &\to Y_1 Z_1, \\ Z_1 &\to Y_2 Z_2, \\ &\dots, \\ Z_{k-2} &\to Y_{k-1} Y_k, \end{aligned}$$

szabályhalmazzal, ahol $Z_1, \dots, Z_{k-2}$ új, a szabályhoz bevezetett nemterminálisok.

Bizonyításvázlat - folytatás:

Így egy $G_1 = (N', T, P_1, S)$ grammatikát kapunk, ahol P_1 olyan szabályokból áll, amelyek alakja az alábbi három típus valamelyike:

1. $X \to a$, $X \in N'$, $a \in T$,
2. $X \to Y$, $X, Y \in N'$,
3. $X \to YZ$, $X, Y, Z \in N'$.

Az N' halmaz N elemeiből, valamint azokból az új nemterminálisokból áll, amelyeket külön-külön bevezettünk P azon szabályaihoz, amelyek jobboldalának hosszát csökkentettük.

Bizonyításvázlat - folytatás:

Könnyű belátni, hogy a $L(G_1) = L(G)$.

Legyen $S = u_1 \Longrightarrow u_2 \ldots \Longrightarrow^* u_m = u$ levezetés G-ben, ahol $u \in T^*$ és $u_j \in (N \cup T)^*$, $1 \leq j \leq m-1$.

Akkor az $u_i \Longrightarrow u_{i+1}$, $1 \leq i \leq m-1$, közvetlen levezetési lépésben P valamely szabályát alkalmazzuk.

Ha a szabály $X \to a$, vagy $X \to Y$, vagy $X \to YZ$ alakú, akkor a közvetlen levezetési lépés megfelel egy közvetlen levezetési lépésnek G_1-ben.

Ha valamely $X \to Y_1Y_2 \ldots Y_k,\ k \geq 3$ alakú szabályt alkalmaztuk, akkor a megelőzőek alapján létezik G_1-ben egy $u_i \Longrightarrow^* u_{i+1}$ levezetés, amely megfelel a szabály alkalmazásának. Így $L(G) \subseteq L(G_1)$.

$L(G_1) \subseteq L(G)$ is fennáll, mivel minden $X \to Y_1Y_2 \ldots Y_k,\ k \geq 3$ alakú szabályhoz új, szabályonként különböző nemterminálisokat vezettünk be, vagyis nem fordulhat elő olyan eset, hogy a szabályok alkalmazása közben olyan mondatforma is megjelenik, amely nem vezet $L(G)$-beli szóra.

Vagyis, $L(G_1) = L(G)$.

Bizonyításvázlat - folytatás:

A továbbiakban az $X \to Y$ **alakú** szabályokat, ahol X és Y nemterminálisok, és amelyeket **láncszabályoknak nevezünk**, elimináljuk a szabályhalmazból.

Ezen célból minden egyes $X \in N'$ nemterminálisra definiáljuk az $U_i(X)$ halmazokat a következőképpen:

$$U_1(X) = \{X\},$$

$$U_{i+1}(X) = U_i(X) \cup \{Y \mid Y \to Z \in P_1,\ Z \in U_i(X)\},\ i = 1, 2, \ldots.$$

Nyilvánvaló, hogy van olyan k természetes szám, hogy $U_k(X) = U_{k+1}(X)$, és így $U_k(X) = U_l(X)$ teljesül minden l-re, ahol $l \geq k$.

Legyen $U_k(X) = U(X)$.

Látható, hogy $Y \Longrightarrow^* X$ akkor és csak akkor, ha $Y \in U(X)$.

Bizonyításvázlat - folytatás:

Definiáljuk P'-t a következőképpen:

1. $X \to a \in P'$ akkor és csak akkor, ha van olyan $A \in N'$, ahol $X \in U(A)$ és $A \to a \in P_1$,

2. $X \to YZ \in P'$ akkor és csak akkor, ha van olyan $A \in N'$, ahol $X \in U(A)$ és $A \to YZ \in P_1$.

További szabály nincs P'-ben.

Látható, hogy $X \to a \in P'$ akkor és csak akkor, ha $X \Longrightarrow^*_{G_1} a$ és $X \to YZ \in P'$ akkor és csak akkor, ha $X \Longrightarrow^*_{G_1} A \Longrightarrow^*_{G_1} YZ$ teljesül valamely A-ra.

Ezek alapján megmutatható, hogyha egy terminális szó generálható a G_1 grammatikával, akkor generálható a G' grammatikával is, és a fordított állítás is fennáll. Vagyis, $L(G) = L(G')$.

Következmények:

Tétel: (szóprobléma)

Minden $G = (N, T, P, S)$ környezetfüggetlen grammatika esetében eldönthető, hogy egy tetszőleges $u \in T^*$ szó benne van-e G grammatika által generált nyelvben vagy sem.

Bizonyításvázlat:

Mivel az előzőek alapján az, hogy az üres szó benne van-e a G grammatika által generált nyelvben eldönthető, elég az $u \neq \varepsilon$ esetre szorítkozni. Az általánosság megszorítása nélkül feltehetjük, hogy G Chomsky-normálformájú. Ekkor az u szó $k = 2|u| - 1$ lépésben levezethető G-ben ($|u|$ az u szó hosszát jelöli). (Lásd a G szabályainak alakját). Minthogy a G grammatikában a legfeljebb k lépésben levezethető szavak véges halmazt alkotnak, ezért el tudjuk dönteni, hogy u előfordul-e ebben a halmazban vagy sem.

Következmények:

Korollárium:

Minden $G = (N, T, P, S)$ környezetfüggetlen grammatika és minden $L \subseteq T^*$ véges nyelv esetében eldönthető, hogy igazak-e a következő állítások: $L \subseteq L(G)$, valamint $L \cap L(G) = \emptyset$.