

Automaták és Formális Nyelvek

VII. Előadás

Döntési algoritmusok véges automatákon

Legyen A egy véges determinisztikus vagy nondeterminisztikus automata, akkor véges lépésben eldönthető, hogy egy $w \in \Sigma^*$ szó hozzátartozik-e az $\mathcal{L}(A)$ nyelvéhez. Ugyancsak véges lépésben eldönthető, hogy $\mathcal{L}(A)$ véges-e vagy végtelen. Igaz ugyanis az alábbi:

Tétel. (i) *Egy véges n -állapotú DFA által elfogadott L nyelv akkor és csak akkor nem üres, ha van n -nél rövidebb szó amit elfogad.*

(ii) *L pontosan akkor végtelen, ha van olyan w elfogadott szó, aminek a hosszára $n \leq |w| < 2n$ teljesül.*

Következmény Két véges determinisztikus automata, A_1 és A_2 esetén véges lépésben eldönthető, hogy ekvivalensek-e, azaz $\mathcal{L}(A_1) = \mathcal{L}(A_2)$. Döntési algoritmusokról és ezek bonyolultságáról a félév végén fogunk beszélni.

1. Környezetfüggetlen nyelvek

1. Tekintsük a következő témát:

<Mondat> $\rightarrow$ <Alany> <Állítmány>
<Alany> $\rightarrow$ <Melléknév> <Főnév>
<főnév> $\rightarrow$ fiú
<melléknév> $\rightarrow$ kicsi
<állítmány> $\rightarrow$ alszik

a következő mondatot állíthatjuk elő
a sémával "kicsi fiú alszik"

Környezetfüggetlen nyelvtan (CFG: Context free grammar)

- Úgy tekintjük, hogy ez egy olyan $G = (V, T, P, S)$ "négyes", ahol

- V - a változók halmaza

- T - az ún. terminálok halmaza

- P - az ún. produkciós szabályok nem üres halmaza

- $S \in V$ start szimbólum

Megjegyzés $T \neq \emptyset$,
 $\bar{V}$ és T végesek
és $\bar{V} \cap T = \emptyset$.

Példák:

1, $G=(V,T,P,S)$, $V = \{S\}$, $T = \{a,b\}$

$P \begin{cases} S \rightarrow aSb \\ S \rightarrow ab \end{cases}$ Egy deriváció egy terminálokból álló sorozat, vagyis egy szó

előállítás: $\underbrace{S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow a^3Sb^3 \Rightarrow \dots a^n b^n S a^{n+1} b^{n+1}}_{n+1 \text{ lépés}}$

Egy $w \in T^*$ szó előállításának derivációja: A P-beli szabályok felhasználásával S-ből eljutunk w-be.

A G nyelvtan nyelve: a G segítségével előállított szavak összessége, jele $\mathcal{L}(G)$.

$$2, G=(V,T,P,S) \quad V=\{S,A,B\}, \quad T=\{a,b\}$$

$$P \begin{cases} S \rightarrow aB & B \rightarrow b \\ S \rightarrow bA & B \rightarrow bS \\ A \rightarrow a & B \rightarrow aBB \\ A \rightarrow aS \\ A \rightarrow bAA \end{cases}$$

Igazolni fogjuk hogy $\mathcal{L}(G) = \{w \in (a+b)^* \mid w \text{ ugyanannyi } a\text{-t és } b\text{-t tartalmaz.}\}$

Jelölések

- 1, A,B,C,D,... S változókat jelentenek
- 2, a,b,c,d,... terminálokat jelentenek
- 3, u,v,w,x,y terminálokból képzett sorozatok, szavak
- 4, $\alpha, \beta, \gamma, \dots$ olyan sorozatok, amelyeket $V \cup T$ elemeivel képzünk: mondatok Pl:aSSb
- 5, X,Y,Z betűket tartjuk fenn arra a célra, hogy ismeretlen jelöljenek függetlenül attól, hogy az változók vagy terminál. $\xRightarrow{*}$ jelentése: egy, több vagy 0 " $\Rightarrow$ " lépés után

$$\begin{array}{l} A \rightarrow \alpha_1 \\ A \rightarrow \alpha_2 \\ \vdots \\ A \rightarrow \alpha_n \end{array} \quad \text{szabályokat röviden } A \rightarrow \alpha_1 | \alpha_2 | \dots | \alpha_n \text{ konvencióval fogjuk}$$

jelölni. Például: a 2-es nyelvhez a következő produkciós szabályok tartoznak

$$\begin{array}{l} S \rightarrow aB | bA \\ A \rightarrow a | aS | bAA \\ B \rightarrow b | bS | aBB \end{array}$$

Az $\mathcal{L}(G)$ nyelvet most így írhatjuk fel: $\mathcal{L}(G) = \{w \in T^* \mid S \xRightarrow{*} w\}$

- A G_1 és G_2 környezetfüggetlen nyelvtanokat ekvivalensnek nevezzük, ha $\mathcal{L}(G_1) = \mathcal{L}(G_2)$

- Az $L \subseteq T^*$ nyelvet környezetfüggetlen nyelvnek nevezzük, ha előállítható egy $G=(V,T,P,S)$ környezetfüggetlen nyelvtannal.

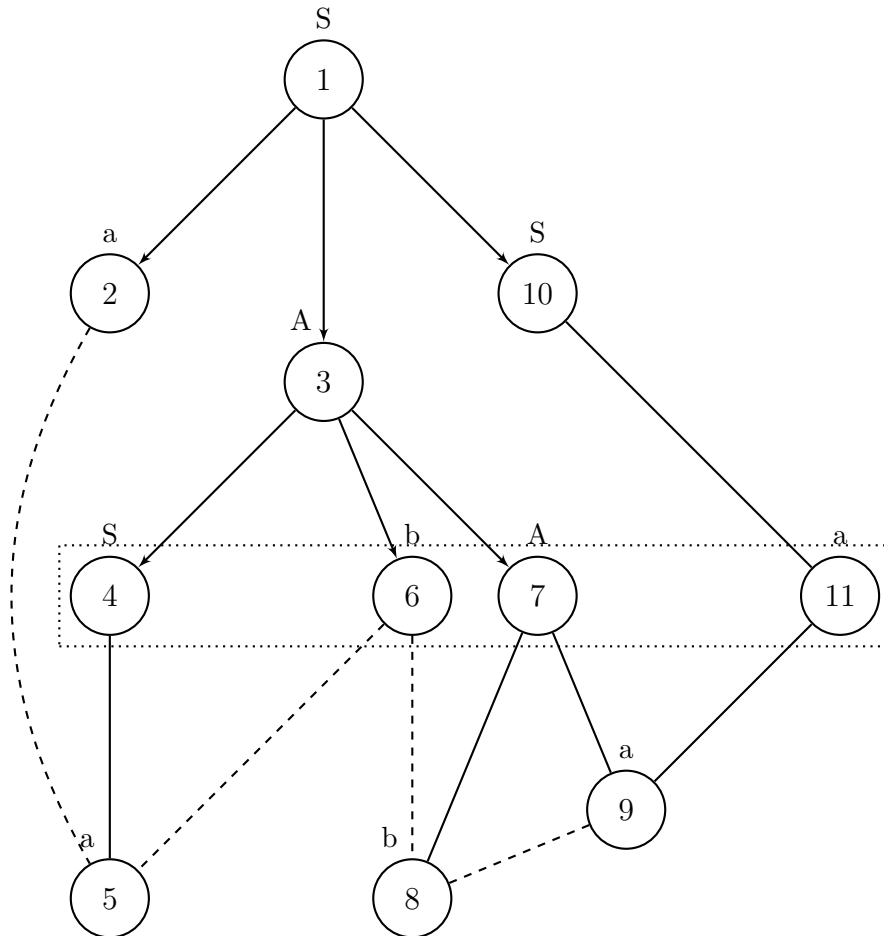
2. Derivációs fák

Tekintsük a következő környezetfüggetlen nyelvtant: $G = (\{S, A\}, \{a, b\}, P, S)$

$$P \begin{cases} S \rightarrow aAS|a \\ A \rightarrow SbA|SS|ba \end{cases} \quad \text{Állítsuk elő } G - \text{segítségével a } w=aabbaa \text{ szót:}$$

$$S \Rightarrow aAS \Rightarrow aAa \Rightarrow aSbAa \Rightarrow aabAa \Rightarrow aabbaa$$

Ezt a következőképpen ábrázolhatjuk:



A pontos számozása a következőképpen írható fel:

- balról jobbra $4 < 8$ közös örököshöz, a 3-hoz képest a 4 balra van
- felülről lefelé $3 < 5$
- amíg lehetséges mindig lefelé haladunk növekvő számozással, ha már nem lehetséges akkor jobbra lépünk.

Egy mondat, pl $SbAa$ - a fa egy szintjének felel meg

- ezek a szintek egymásba metszhetnek, pl mint aAa is
- az ágak végződését, a fa legalsó szintjén a fa aljának nevezzük - itt meg kell kapnunk az előállítandó szót $w=aabbaa$ - ez a fa "termése" vagy produktuma. w - előállításakor legalább egy fa kell létezzen.

Tétel: Legyen $G = (V, T, P, S)$ egy környezetfüggetlen nyelvtan és $\alpha \in V \cup T$ egy mondat. G -ben pontosan lehetséges a $S \xRightarrow{*} \alpha$ deriváció ha létezik olyan S gyökerű derivációs fa aminek az egyik szintjén éppen az α mondat található.

Könnyen belátható, hogy bármely $w \in T^*$ szót a G nyelvtan segítségével ugyanazt a derivációs fát felhasználva kétféleképpen is előállíthatjuk:

- i) A behelyettesítésekkel balról-jobbra haladva
- ii) A behelyettesítésekkel jobbról-balra haladva.

Például az előző G nyelvtan esetén a $w = aabaaa$ szót:

i) balról-jobbra haladva: $S \Rightarrow aAS \Rightarrow aSbAS \Rightarrow aabAS \Rightarrow aabSSS \Rightarrow aabaSS \Rightarrow aabaaS \Rightarrow aabaaa$

ii) jobbról-balra haladva: $S \Rightarrow aAS \Rightarrow aAa \Rightarrow aSbAa \Rightarrow aSbSSa \Rightarrow aSbSaa \Rightarrow aSbaaa \Rightarrow aabaaa$

Mindkét derivációt megkaphatjuk felhasználva az előbbi derivációs fát, csupán az a különbség, hogy az i) esetben mindig a fa baloldalára lépünk 1 szinttel lejjebb, a ii) esetben pedig mindig a fa jobboldalára lépünk egy szinttel lejjebb. Láthatjuk, hogy általában egy derivációs fa segítségével kettőnél több előállítás is lehetséges.

Amennyiben egy $w \in T^*$ szót balról jobbra haladva (illetve jobbról balra haladva) többféleképpen is előállíthatunk akkor ez azt jelenti, hogy a w szót több derivációs fa segítségével is előállíthatjuk. Ha az $\mathcal{L}(G)$ nyelv tartalmaz ilyen tulajdonságú w szót akkor a nyelvet és a nyelvtant is kétértelműnek (ambiguousnak) nevezzük. Pl. ilyen nyelv az $L = \{a^n b^n c^m d^m \mid n, m \geq 1\} \cup \{a^n b^m c^n d^m \mid n, m \geq 1\}$.

Magyarázat: Az $\{a^n b^n c^m d^m \mid n, m \geq 1\}$ nyelv környezetfüggetlen, így előállíthatjuk egy G_1 k.f nyelvtannal. Hasonlóan a $\{a^n b^m c^n d^m \mid n, m \geq 1\}$ nyelvet is előállíthatjuk egy G_2 k.f nyelvtannal. Az $a^n b^n c^n d^n$, $n \geq 1$ alakú szavak mindkét nyelvhez hozzátartoznak, így őket két különböző nyelvtannal, tehát mindegyiküket két különböző derivációs fával is elő lehet állítani.

3. Környezetfüggetlen nyelvek egyszerűsítése

Legyen $G = (V, T, P, S)$ egy környezetfüggetlen nyelvtan és $\mathcal{L}(G) \neq \emptyset$. Azt mondjuk, hogy az $X \in V \cup T$ szimbólum nem felesleges, ha 1) vagy egy előállított szóban szereplő terminál; 2) vagy egy olyan változó ami részt vesz egy $S \xRightarrow{*} \alpha X \beta \xRightarrow{*} w$ alakú derivációban, ahol $w \in T^*$ egy szó, α és β pedig mondatok. Ellenkező esetben X -et feleslegesnek nevezzük.

Megjegyzés: Az a tény, hogy X nem felesleges még nem jelenti azt, hogy nélkülözhetetlen valamely szó előállításához.

Tétel1: Minden nem üres L környezetfüggetlen nyelv előállítható egy olyan környezetfüggetlen nyelvtannal amely nem tartalmaz felesleges szimbólumokat.

Bizonyítás. Mivel L k. független, létezik egy olyan $G = (V, T, P, S)$ k. független nyelvtan amire $\mathcal{L}(G) = L$. G -ből a felesleges szimbólumokat két lépésben fogjuk kiküszöbölni.

1. Lépés. Először töröljük azokat a terminálokat amelyek egyetlen $w \in L$ szóban sem szerepelnek. Töröljük azokat az $A \in V$ változókat is amelyek esetén nem létezik valamilyen $A \xRightarrow{*} w$ deriváció (amelyek nem állítanak elő szót, tehát „meddőek”) és minden olyan P -beli szabályt amely valamilyen törölt terminált vagy változót tartalmaz.

Így egy $G' = (V', T', P', S)$ nyelvtant kapunk. Mivel $\mathcal{L}(G) \neq \emptyset$, legalább egy $S \xRightarrow{*} w$ alakú deriváció biztos létezik, vagyis S -t nem törölhetük, ezért $S \in V'$. Így $V' \neq \emptyset$. Vegyük észre hogy a törölt szimbólumok egyetlen w szó előállításában sem vehetnek részt. (A terminálok esetén ez nyilvánvaló, a törölt változók esetén pedig ha egy ilyen átalakításban résztvennének, akkor nekik maguknak is egy terminálsorozattá, vagyis egy w szóvá kellene átalakulni, ami nem lehetséges.) Ez azt jelenti, hogy ezek a szimbólumok feleslegesnek minősülnek, másrészt azt, hogy minden $w \in L$ szó megkapható a felhasználásuk nélkül is. Ezért $\mathcal{L}(G') = \mathcal{L}(G) = L \neq \emptyset$. Ez pedig csak úgy lehetséges ha $T' \neq \emptyset$ és $P' \neq \emptyset$. Tehát G' valóban egy k. független nyelvtan.

2. Lépés. G' -ből töröljük mindazokat a változókat és az őket tartalmazó P' -beli szabályokat, amelyek S -ből nem érhetők el, vagyis azokat az $A \in V$ elemeket amelyek nem vesznek részt egyetlen $S \xRightarrow{*} \alpha A \beta$ alakú derivációban sem. Így egy $G'' = (V'', T', P'', S)$ nyelvtant kapunk. Mivel $S \xRightarrow{*} S$, ezért S -t nem töröljük, így $S \in V''$ és $V'' \neq \emptyset$. Azt is tudjuk, hogy $T' \neq \emptyset$. Vegyük észre, hogy a törölt szimbólumok (és az őket tartalmazó szabályok) egyetlen $S \xRightarrow{*} \alpha A \beta \xRightarrow{*} w$ alakú derivációban sem vehetnek részt - tehát feleslegesek és egyetlen w szó előállításában sem szerepelhetnek. Így nélkülük is minden $w \in \mathcal{L}(G')$ szót elő tudunk állítani, ami azt jelenti hogy $P'' \neq \emptyset$, így G'' egy környezetfüggetlen nyelvtan, mi több, $\mathcal{L}(G'') = \mathcal{L}(G')$.

Összegezve azt kapjuk hogy $\mathcal{L}(G'') = \mathcal{L}(G') = L$, tehát G'' is az L nyelvet állítja elő. Most már elegendő csupán annyit igazolnunk hogy $G'' = (V'', T', P'', S)$ nem tartalmaz felesleges szimbólumokat.

Azt már megbeszéltük hogy T' nem tartalmaz felesleges terminálokat. Lássuk be hogy V'' sem tartalmazhat felesleges változót. Valóban, legyen $X \in V''$ egy tetszőleges változó. Akkor V'' szerkesztése alapján létezik egy $S \xRightarrow{*} \alpha X \beta$ deriváció, ahol $\alpha, \beta \in (T' \cup V')^*$, azaz

$$\alpha = X_1 X_2 \dots X_k, \quad X_i \in (T' \cup V'), \quad 1 \leq i \leq k \text{ és}$$

$$\beta = Y_1 Y_2 \dots Y_m, \quad Y_j \in (T' \cup V'), \quad 1 \leq j \leq m \text{ alakú.}$$

Mivel $X \in V'' \subseteq V'$, ezért V' szerkesztése szerint kell létezzen egy $X \xRightarrow{*} v \in (T')^*$ deriváció is G' -ben, így G -ben is. Mivel minden $X_i \in V'$ és $Y_m \in V'$ esetben ezek a változók T' termináljaiból álló szavaká alakíthatók, ezért léteznek az $X_i \xRightarrow{*} u_i \in (T')^*$, $1 \leq i \leq k$ és az $Y_j \xRightarrow{*} w_j \in (T')^*$, $1 \leq j \leq m$ derivációk is. Így véges lépésben azt kapjuk, hogy $\alpha X \beta \xRightarrow{*} u_1 u_2 \dots u_k v w_1 w_2 \dots w_m := w \in (T')^*$. (Hiszen $u_1 u_2 \dots u_k v w_1 w_2 \dots w_m$ maga is egy T' elemeiből képzett szó). Tehát G -ben létezik egy

$$S \xRightarrow{*} \alpha X \beta \xRightarrow{*} w \in T^*$$

deriváció, ami azt jelenti hogy X nem felesleges változó. Tehát G'' nem tartalmaz felesleges szimbólumokat. $\square$

Tétel2: Ha L egy környezetfüggetlen nyelv és $L \setminus \{\epsilon\} \neq \emptyset$, akkor $L \setminus \{\epsilon\}$ nyelv előállítható egy olyan $G = (V, T, P, S)$ környezetfüggetlen nyelvtannal, ami nem tartalmaz

- 1) felesleges szimbólumokat;
- 2) $A \Rightarrow B$ alakú szabályokat;
- 3) $A \Rightarrow \epsilon$ alakú szabályokat.

Megjegyzés: A környezetfüggetlen nyelvtanok egyszerűsítése során nem az a cél, hogy a nyelvtan ne tartalmazzon redundanciát és így minnél kisebb és tömörebb legyen, hanem az hogy minnél áttekinthetőbb legyen a szabályrendszer. Ez két úgynevezett normálforma segítségével valósítható meg.

Chomsky féle normálforma

Tétel3: Minden L környezetfüggetlen nyelv amire $\epsilon \notin L$ előállítható egy olyan G környezetfüggetlen nyelvtannal amely csak $A \Rightarrow BC$ és $A \Rightarrow a$ alakú produkciós szabályokat tartalmaz.

Greibach féle normálalak

Tétel3: Minden L környezetfüggetlen nyelv amire $\epsilon \notin L$ előállítható egy olyan G környezetfüggetlen nyelvtannal, amely csak $A \Rightarrow a\alpha$ alakú produkciós szabályokat tartalmaz. ($a \in T, \alpha \in (T \cup V)^*$).

Például $G_1 = (V, T, P, S)$ P $\left\{ \begin{array}{l} S \rightarrow AS | BS | c \\ A \rightarrow a \\ B \rightarrow b \end{array} \right.$

egy Chomsky féle normálalakban írt nyelvtan.

$$G_2 = (S, A, B, a, b, P, S) \text{ P } \begin{cases} S \rightarrow bA|aB \\ A \rightarrow bAA|aS|a \\ B \rightarrow aBB|bS|b \end{cases}$$

egy Greibach féle normálalakban írt nyelvtan.

Feladat1 Írjuk át a G_2 nyelvtant Chomsky féle normálformába.

Mivel több produkciós szabály is van amelyik nem felel meg a Chomsky féle normálforma definíciójának, pl. bAA vagy bA szabályok sem felelnek meg, ezért helyettesítéseket végzünk. Minden olyan szabályt helyettesítünk egy másikkal amely a definíciónak nem felel meg (pl. $D_A \rightarrow AA$).

$$\begin{aligned} S &\rightarrow C_b A | C_a B \\ A &\rightarrow C_b D_A | C_a S | a \\ B &\rightarrow C_a D_B | C_b S | b \\ C_a &\rightarrow a \\ C_b &\rightarrow b \\ D_A &\rightarrow AA \\ D_B &\rightarrow BB \end{aligned}$$

Feladat2 Adott az alábbi környezetfüggetlen nyelvtan

$$\begin{aligned} S &\rightarrow aAS|SC|DE|a \\ A &\rightarrow SbA|SS|ba \\ B &\rightarrow aB|bS|b \\ F &\rightarrow BB \\ G &\rightarrow AF \end{aligned}$$

- Kiküszöböljük ki a nyelvtanból a felesleges szimbólumokat;
- Írjuk fel az egyszerűsített nyelvtant Greibach féle normálalakban.

Megoldás:

1. lépésben kiküszöböljük azokat a terminálokat amelyek nem vezetnek sehova, látható, hogy C, D, E teljesen feleslegesek, így kapjuk, hogy

$$\begin{aligned} S &\rightarrow aAS|a \\ A &\rightarrow SbA|SS|ba \\ B &\rightarrow aB|bS|b \\ F &\rightarrow BB \\ G &\rightarrow AF \end{aligned}$$

2.lépésben kiküszöböljük a láncszabályokat, azaz

$$\begin{aligned} S &\rightarrow aAS|a \\ A &\rightarrow SbA|SS|ba \\ B &\rightarrow aB|bS|b \end{aligned}$$

b) Greibach féle normálalaknál minden produkciós szabály kisbetűvel kell kezdődjön, azaz két esetben ($A \rightarrow SbA|SS$) is más szabályokkal kell helyettesítenünk

$$S \rightarrow aAS|a$$

$$A \rightarrow aASbA|abA|aS|aASS|ba$$

$$B \rightarrow aB|bS|b$$