Formális nyelvek - 4.

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

Aktív, elérhető nemterminálisok

Definíció

A környezetfüggetlen grammatika egy nemterminálisát **aktívnak** nevezzük, ha levezethető belőle terminális szó; egyébként **inaktívnak** vagy **nem aktívnak** mondjuk.

A környezetfüggetlen grammatika egy nemterminálisát **elérhetőnek** nevezzük, ha legalább egy olyan mondatformában előfordul, amely a kezdőszimbólumból levezethető; egyébként **nem elérhetőnek** mondjuk.

Hasznos/nem hasznos nemterminálisok

Definíció

Egy nemterminálist **hasznosnak** nevezünk, ha aktív és elérhető, egyébként **nem hasznos**.

(Egy nemterminális nem hasznos, ha vagy inaktív, vagy nem elérhető, vagy mindkét tulajdonság teljesül rá.)

Redukált grammatika

Definíció

Egy környezetfüggetlen grammatika redukált, ha minden nemterminálisa aktív és elérhető.

Redukált grammatika - folytatás

Tétel

Minden környezetfüggetlen grammatikához meg tudunk konstruálni egy vele ekvivalens redukált környezetfüggetlen grammatikát.

Megjegyzés:

A nem elérhető és a nem aktív nemterminálisok, valamint azok a szabályok, amelyekben előfordulnak, meghatározhatók és eliminálhatók anélkül, hogy a generált nyelv megváltozna.

Bizonyításvázlat:

Legyen $G = (N, T, P, S)$ egy környezetfüggetlen grammatika. Tekintsük az alábbi halmazokat:

$$A_1 = \{X \mid X \to u \in P, u \in T^*\},$$

$$A_{i+1} = A_i \cup \{X \mid X \to w \in P, w \in (T \cup A_i)^*\},\ i = 1, 2, \ldots$$

Nyilvánvaló, hogy az A_i, $i = 1, 2, \ldots$, halmazok nemcsökkenő hierarchiát alkotnak a tartalmazásra nézve. Így létezik olyan k szám, hogy $A_k = A_l$ teljesül minden $l \geq k$-ra. Ekkor A_k a G grammatika aktív nemterminálisainak halmaza.

Bizonyításvázlat - folytatás:

Ezután tekintsük az

$$R_1 = \{S\},$$

$$R_{i+1} = R_i \cup \{Y \mid X \to uYw \in P,\, X \in R_i, u, w \in (N \cup T)^*\},\ i = 1, 2, \dots$$

halmazokat.

Az R_i, $i = 1, 2, \dots,$ halmazok a tartalmazásra nézve nemcsökkenő hierarchiát alkotnak. Így létezik olyan m szám, hogy $R_m = R_l$ minden $l \geq m$ esetben. Könnyen látható, hogy az R_m halmaz G elérhető nemterminálisainak halmaza.

Az A_k és az R_m halmazok kiszámolása után eliminálunk minden olyan nemterminálist, amely nem eleme az $A_k \cap R_m$ halmaznak együtt azokkal a szabályokkal amelyekben előfordulnak. A fenti procedúrát megismételjük mindaddig, amíg egy redukált grammatikát nem kapunk.

Környezetfüggetlen grammatikák - levezetési fa

A **környezetfüggetlen grammatikák** levezetéseit ún. **levezetési fákkal** is jellemezhetjük.

A levezetési fa a szó előállításának lehetőségeiről ad információkat.

A **levezetési fa** egy **irányított gráf**, amely speciális tulajdonságoknak tesz eleget.

Levezetési fa - folytatás

Legyen V véges nemüres halmaz, amelynek elemeit **csúcsoknak** nevezzük.

Az **élek** E halmaza csúcsok rendezett párjaiból álló halmaz, azaz, $E \subseteq V \times V$.

Minden $e = (n_1, n_2)$ élre $s(e) = n_1$ az él kiindulási csúcsa és $t(e) = n_2$ a végcsúcsa.

Élek egy $e_0, e_1, \ldots, e_k$ sorozatát az $s(e_0)$-ból kiinduló $t(e_k)$-ig vezető k hosszúságú **irányított útnak** nevezzük, ha $s(e_{i+1}) = t(e_i)$, ahol $i = 0, 1, \ldots, k-1$.

Levezetési fa - folytatás

A (V, E) rendezett párt **irányított fának** nevezzük, ha van olyan $r \in V$ **csúcs**, amelyre teljesülnek a következők:

1. Az E halmaz egyetlen élének végcsúcsa sem azonos r-rel.
2. Minden r-től különböző csúcshoz V-ben létezik egy r-ből kiinduló irányított út.

Az r **csúcsot a fa gyökerének** nevezzük. Minden fának egyetlen gyökere van.

A fa minden csúcsa gyökere a fa valamely **részfájának**.

Az n **csúcs leszármazottjainak azokat az** n' **csúcsokat nevezzük**, amelyekre $(n, n') \in E$ (azaz, (n, n') él).

Azokat a csúcsokat, amelyeknek nincs leszármazottjuk, **levélnek** mondjuk.

Levezetési fa

A környezetfüggetlen grammatikák levezetéseit fákkal is leírhatjuk.

A $G = (N, T, P, S)$ grammatika feletti egy **levezetési fának** nevezünk egy fát, ha teljesülnek a következők:

- A levezetési fa **gyökerének címkéje** S.
- **Minden további csúcs címkéje** $(N \cup T \cup \{\varepsilon\})$ **valamely eleme**.
- Ha egy **csúcs címkéje** X és **leszármazottjainak címkéi balról jobbra olvasva rendre** $X_1, \ldots X_m$, $m \geq 1$, **akkor** $X \to X_1 \ldots X_m \in P$**.**
- Minden **levél címkéje** a $T \cup \{\varepsilon\}$ halmaz valamely eleme, és ha egy csúcsnak leszármazottja ε, akkor ennek a csúcsnak ez az **egyetlen leszármazottja** van.

A levezetési fa levelei címkéinek sorozata a **levezetési fa** határa.

Levezetési fa - folytatás

Minden, a G grammatikában történő **levezetéshez** hozzá tudunk rendelni egy **levezetési fát**.

A levezetési fa **nem minden esetben** adja meg a levezetés során alkalmazott **szabályok sorrendjét**.

Két levezetés lényegében azonos, ha csak a **szabályok alkalmazásának sorrendjében** különbözik, azaz, **ugyanahhoz a levezetési fához tartozik**.

Legbaloldalibb levezetés

Egy környezetfüggetlen grammatika feletti **levezetési fa** egy **egyetlen** legbaloldalibb levezetést határoz meg.

A **legbaloldalibb levezetés** során minden levezetési lépésben arra a nemterminálisra kell szabályt alkalmazni, amely a levezetési lépéshez tartozó mondatformában balról a legelső.

(Példa: Ha $u_1 A_1 u_2 A_2 \ldots A_n u_{n+1}$, $u_i \in T^*$, $1 \leq i \leq n+1$, $A_j \in N$, $1 \leq j \leq n$, mondatforma a $G = (N, T, P, S)$ grammatikában, akkor legbaloldalibb levezetés esetén az A_1 nemterminálist kell helyettesíteni.)

A környezetfüggetlen grammatika által generált nyelv üres volta

Tétel

Minden környezetfüggetlen grammatikáról eldönthető, hogy az általa generált nyelv az üres nyelv-e vagy sem.

Bizonyításvázlat:

Legyen $G = (N, T, P, S)$ környezetfüggetlen grammatika. Az általánosság megszorítása nélkül feltehetjük, hogy G ε-mentes.

Legyen n a G nemterminális szimbólumainak száma.

Tegyük fel, hogy létezik egy $S \Longrightarrow_G^* u$ levezetés G-ben, ahol $u \in T^*$.

Tekintsük az ezen levezetéshez tartozó levezetési fát.

Ha a leghosszabb út hossza ebben a levezetési fában nagyobb, mint n, akkor van olyan v szó $L(G)$-ben, amely levezetési fájában a leghosszabb út hossza nem nagyobb, mint n.

Ez nyilvánvaló, hiszen ha az út hossza nagyobb, mint n, akkor legalább egy nemterminális szimbólum legalább kétszer fordul elő ezen az úton.

Bizonyításvázlat - folytatás:

Tekintsünk két azonos címkéjű csúcsot az úton és helyettesítsük a fa gyökeréhez közelebb eső csúcshoz tartozó részfát a másik csúcshoz tartozó részfával. Akkor továbbra is terminális szót kapunk.

Megismételve ezt az eljárást annyiszor, ahányszor szükséges, addig csökkenthetjük az út hosszát, amíg legfeljebb n hosszúságú utat kapunk.

Ebből következően, ha $L(G)$ nem üres, akkor léteznie kell egy szónak a nyelvben, amelyhez tartozó levezetési fában a leghosszabb út nem hosszabb, mint n.

Minthogy azokat a levezetési fákat, amelyekre az igaz, meg tudjuk határozni, a kérdést el tudjuk dönteni.

Korollárium

Az, hogy egy $G = (N, T, P, S)$ környezetfüggetlen grammatika A nemterminálisa inaktív-e, eldönthető.

A probléma ekvivalens azzal a problémával, hogy a $G_A = (N, T, P, A)$ grammatika által generált nyelv üres-e. Ha $L(G_A)$ üres, akkor A nem aktív.

Korollárium

Az, hogy egy $G = (N, T, P, S)$ környezetfüggetlen grammatika A nemterminálisa elérhető-e, eldönthető.

Bizonyításvázlat:

Távolítsunk el a P szabályhalmazból minden szabályt, amelynek A van a baloldalán. Jelöljük az így kapott szabályhalmazt P_1-gyel. Tekintsük a

$$G_A^\varepsilon = ((N - \{A\} \cup T, \{A\}, P_1 \cup \{X \to \varepsilon \mid X \in (N - \{A\} \cup T)\}, S)$$

grammatikát.

Ha A nem elérhető, akkor $L(G_A^\varepsilon) = \{\varepsilon\}$, máskülönben tartalmaznia kell egy olyan szót, amelyben A legalább egyszer előfordul. Megkonstruáljuk a G' környezetfüggetlen grammatikát, amelyre $L(G') = L(G_A^\varepsilon) - \{\varepsilon\}$. Minthogy el tudjuk dönteni, hogy $L(G')$ üres-e vagy sem, el tudjuk dönteni A elérhetőségét is G-ben.

Bar-Hillel vagy pumpáló lemma

Tétel

Minden L környezetfüggetlen nyelvhez meg tudunk adni két p és q természetes számot úgy, hogy minden olyan szó L-ben, amely hosszabb, mint p

$$uxwyz$$

alakú, ahol $|xwy| \leq q$, $xy \neq \varepsilon$, továbbá minden

$$ux^iwy^iz$$

alakú szó is benne van az L nyelvben minden $i \geq 0$ egész számra $(u, x, w, y, z \in T^*)$.

Bizonyításvázlat:

Legyen L ε-mentes nyelv, amelyet a $G = (N, T, P, S)$ Chomsky-normálformájú grammatika generál.

Tegyük fel, hogy G nemterminálisainak száma n és legyen $p = 2^n$, valamint legyen $q = 2^{n+1}$.

Ha $|\beta| > p$ valamely β szóra L-ben, akkor a β levezetési fájában a leghosszabb út hossza nagyobb, mint n. (A fa bináris fa, azaz minden csúcsából legfeljebb két csúcs származik és, ha a bináris fában a leghosszabb út hossza k, akkor a leveleinek száma legfeljebb 2^k.)

Tekintsük az utolsó $n + 1$ élét a leghosszabb útnak. Akkor lennie kell egy A nemterminálisnak, amely ezen az úton legalább kétszer előfordul.

Bizonyításvázlat - folytatás:

Feleljen meg az a részfa, amely az A első előfordulásához tartozik ezen az úton (abból indul ki) az $S \Longrightarrow^* \alpha$, $\alpha \in T^*$ levezetésnek, az A második előfordulásához tartozó részfa pedig feleljen meg az $A \Longrightarrow^* w$, $w \in T^*$ levezetésnek.

Mivel A kétszer fordult elő az úton, van két olyan $x, y \in T^*$ szó, hogy $\alpha = xwy$ és $A \Longrightarrow^* xAy$.

Ezen kívül

$$S \Longrightarrow^* uAz \Longrightarrow^* uxAyz \Longrightarrow^* uxwyz = u\alpha z = \beta.$$

Az A nemterminális adott előfordulásainak pozíciójából $|\alpha| \leq 2^{n+1}$ következik. Továbbá, az $A \Longrightarrow^* xAy$ levezetéshez során legalább egy $B \rightarrow CD$ alakú szabály alkalmazása szükséges, így $|xy| \neq \varepsilon$.

Láthatjuk, hogy $S \Longrightarrow^* uwz$ és $S \Longrightarrow^* ux^iwy^iz$, $i \geq 1$, levezetések G-ben.

Következmény

Léteznek nem környezetfüggetlen mondatszerkezetű nyelvek.

Ilyen például az $L = \{a^n b^n c^n | n \geq 1\}$.

Következmény - folytatás

Megmutatjuk, hogy L nem környezetfüggetlen. Tegyük fel az ellenkezőjét. Legyen $w = a^q b^q c^q \in L$, ahol p, q a Bar-Hillel lemmának megfelelő konstansok. Nyilvánvaló, hogy $|w| > q > p$. (Lásd a lemma bizonyítását.) Ekkor bármely $u, v, x, y, z \in \{a, b, c\}^*$-ra, amelyre $w = uxvyz$, $|xvy| \leq q$, $|xy| > 0$, a lemma alapján fennáll, hogy $uvz \in L$. Viszont, mivel xy az $\{a, b, c\}$-ből legalább egy betűt nem tartalmaz, így uvz nem lehet L eleme, és így a nyelv nem teljesíti a Bar-Hillel lemma feltételeit, vagyis a nyelv nem környezetfüggetlen.

Következmény

Tétel

Eldönthető, hogy egy környezetfüggetlen grammatika végtelen nyelvet generál-e vagy sem.

Bizonyításvázlat:

Csak ε-mentes $G = (N, T, P, S)$ grammatikával foglalkozunk. Megmutatjuk, hogy $L(G)$ akkor és csak akkor végtelen, ha tartalmaz olyan β szót, hogy $p < |\beta| \leq p + q$ teljesül.

1) Ha G rendelkezik ezzel a tulajdonsággal, akkor a Bar-Hillel lemma alapján az állítás fennáll.

2) Megmutatjuk, hogy a fordított állítás is teljesül. Ha $L(G)$ végtelen, akkor tartalmaznia kell egy β szót, amelyre $p < |\beta|$. Megmutatjuk, hogy ekkor tartalmaz egy szót, amelyre $p < |\beta| \leq p + q$ fennáll.

Bizonyításvázlat - folytatás:

Tegyük fel az ellenkezőjét, azaz, azt, hogy minden β szóra, ahol $p < |\beta|$, az teljesül, hogy $p + q < |\beta|$.

De ha $p < |\beta|$, akkor β alakja $uxwyz$, ahol $uwz \in L(G)$ és $|uwz| < |\beta|$, mivel $xy \neq \varepsilon$.

Ha $p < |uwz|$, akkor a fenti érvelés megismételhető mindaddig, amíg egy $\beta' = u'x'w'y'z'$ szót kapunk, amelyre $p < |\beta'|$ és $|u'w'z'| \leq q$ teljesül.

Ekkor, mivel $|x'w'y'| \leq q$, a Bar-Hillel lemma alapján a $p < |u'x'w'y'z'| \leq p + q$ egyenlőtlenséget kapjuk, amely ellentmond feltételezésünknek.

(A $p + q$ felső korlát alapján a megfelelő levezetési fa megkereshető.)