Az Earley-féle elemzésről

Legelőször tennék egy olyan jellegű megjegyzést, hogy az Earley elemző működik nem egyértelmű nyelvtanokkal (nyelvekkel) is. Az elemzés során $\{A \to \alpha_1 \cdot \alpha_2, i\}$ alakú elemeket rakosgatunk be csoportokba a nemsokára bemutatásra kerülő szabályok alapján. Az elemek jelentése a következőképpen írható le: ha egy $\{A \to \alpha_1 \cdot \alpha_2, i\}$ alakú elem van a k. csoportban, akkor az elemezni kívánt jelsorozat $(i+1)$. és k. karaktere által határolt részszót a pont előtti rész (α_1) generálni tudja, méghozzá úgy, hogy menet közben éppen a nyelvtan $A \to \alpha_1\alpha_2$ szabályát használjuk.

Az Earley elemző lényegében top–down elemzést hajt végre, méghozzá úgy, hogy megpróbál minden lehetséges módon ballevezetést generálni. Magyarul, ha éppen egy nemterminális szimbólumnál tartunk, akkor az összes, az adott nemterminálissal rendelkező baloldalú szabályt kipróbálja, hátha abba az irányba is tud menni egy levezetés.

Tudom, hogy a jegyzetben egy kissé másképp van leírva, de a levezetést egy $\{\cdot S, 0\}$ elemből fogom indítani, hogy az összes lépést algoritmusszerűen, és teljesen konzisztens módon tegyük meg. (Akinek nem tetszik ez a megoldás, az tekintse úgy, mintha a nyelvtanba bevennénk egy új, $S' \to S$ szabályt, és ebből indítanánk próbálkozásainkat.)

Három szabály szerint tehetünk be elemet egy csoportba.

1. Ha a k–1. csoportban van $\{A \to \alpha_1 \cdot \mathtt{a}\alpha_2, i\}$ alakú elem, és az elemzendő szó k. karaktere $\mathtt{a}$ (magyarul a k–1. csoportban van olyan elem, ahol a pont után az elemzendő szó k. karaktere szerepel), akkor a k. csoportba be lehet venni egy $\{A \to \alpha_1\mathtt{a} \cdot \alpha_2, i\}$ elemet. (Ezeket hívják a csoport alapító tagjainak.)

   Ezen szabály alkalmazását az ábrákon dupla vonalas nyíllal jelöljük.

   $$\{A \to \alpha_1 \cdot \mathtt{a}\alpha_2, i\} \Longrightarrow \{A \to \alpha_1\mathtt{a} \cdot \alpha_2, i\}$$

   Ha a k–1. csoportban ugyan van olyan elem, ahol a pont után terminális szimbólum szerepel, de az nem egyezik meg az elemzendő szó k. karakterével, akkor értelemszerűen nem tehetünk be új elemet ezen szabály segítségével.

   Nézzük meg, hogy mit is jelent ezen szabály alkalmazása egy elem generálásakor! A k–1. csoportban olyan elemet keresünk, ahol a pont olyan terminális előtt áll, amely megegyezik az elemzendő szó k. karakterével. Ha találunk ilyet, akkor elkészíthetjük a k. csoport első elemét (elemeit), mégpedig úgy, hogy a ponttal az adott terminálist átugorjuk. Magyarul itt egyfajta ellenőrzés történik, ami azt vizsgálja, hogy a mondatjelölt k. karaktere generálható-e.

2. Ha a k. csoportban van $\{A \to \alpha_1 \cdot B\alpha_2, i\}$ alakú elem, és a nyelvtanunk tartalmaz $B \to \gamma$ alakú szabályt, akkor a k. csoportba bevesszük a $\{B \to \cdot\gamma, k\}$ alakú elemet.

   A szabály alkalmazását az ábrákon pontozott vonalas nyíllal jelöljük.

   $$\{A \to \alpha_1 \cdot B\alpha_2, i\} \dashrightarrow \{B \to \cdot\gamma, k\}$$

A szabály alkalmazása előtt a pont egy nemterminális előtt áll. Ez lényegében azt jelenti, hogy egy válaszútnál vagyunk, mivel el kéne tudnunk dönteni, hogy merre megy helyes levezetés. A problémát egy huszáros vágással oldjuk meg, méghozzá úgy, hogy minden irányt bejárunk. Ennek fényében minden olyan szabály miatt új elemet veszünk be a k. csoportba, melynek bal oldalán a már a csoportban levő elem pont utáni nemterminálisa szerepel. Az újonnan bevett elemek indexe (vessző utáni szám) pedig értelemszerűen k lesz, mivel az elemben jelen levő szabály segítségével a $k{+}1$. karakter lehet az első, amit generálni tudunk.

3. Ha a k. csoportban van $\{B \to \gamma\cdot, j\}$ alakú, és a j. csoportban van $\{A \to \alpha_1 \cdot B\alpha_2, i\}$ alakú elem, akkor a k. csoportba bevesszük a $\{A \to \alpha_1 B \cdot \alpha_2, i\}$ elemet.

   Egy megjegyzés: a fenti állításban ha az első feltétel teljesül (van olyan elem, ahol a pont a szabály jobboldalának végén helyezkedik el), akkor a második is automatikusan teljesülni fog.

   A szabály alkalmazását a következőképpen fogjuk ábrázolni:

$$\begin{tikzcd} & \{B \to \gamma\cdot, j\} \arrow[d, dash, dashed] & \\ \{A \to \alpha_1 \cdot B\alpha_2, i\} \arrow[r] & {} & \{A \to \alpha_1 B \cdot \alpha_2, i\} \end{tikzcd}$$

   Itt a pont az elem szabály részének végén szerepel, ami azt jelenti, hogy a $j{+}1$. karaktertől a k. karakterig terjedő részszót a B nemterminálisból indulóan lehet generálni. Annak idején a j. csoportba azért vettük be a $\{B \to \cdot\gamma, j\}$ elemet, mert úgy véltük, hogy a szabályt tudjuk használni a mondatjelölt levezetése során. A mostani szituációban kiderült, hogy (részben) igazunk volt.

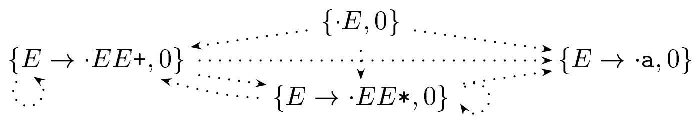
   A j. csoportban levő $\{A \to \alpha_1 \cdot B\alpha_2, i\}$ elem szerint α_1 generálja a mondatjelölt $i{+}1$. és j. karakter közti részét. Most azt a tanulságot szűrtük le, hogy B tudja generálni a $j{+}1$. és a k. karakter közti részt. Mindent egybevéve azt mondhatjuk, hogy $\alpha_1 B$ képes elkészíteni a jelsorozat $i{+}1$. és k. karakter közti részét. Erre utal a k. csoportba újonnan bevett $\{A \to \alpha_1 B \cdot \alpha_2, i\}$ elem.

Az elemekben szereplő pont $(\cdot)$ szerepéről azt is lehet mondani, hogy azt jelöli, hogy hol tartunk éppen aktuálisan a szó elemzésével. Igyekszem az itt elmondottakat egy példával is illusztrálni. A példa legyen a posztfix aritmetikát generáló nyelvtan, míg az elemzendő mondat az `aa+a*`.

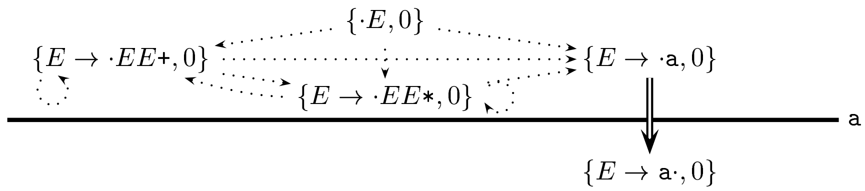
Kezdetben az $\{\cdot E, 0\}$ elemből indulunk ki. A pont után az E nemterminális szimbólum áll, tehát használhatjuk a második szabályt, és ezáltal generálhatunk három új elemet a 0. csoportba.

$$\begin{tikzcd} & \{\cdot E, 0\} \arrow[dl, dotted] \arrow[d, dotted] \arrow[dr, dotted] & \\ \{E \to \cdot EE\texttt{+}, 0\} & \{E \to \cdot EE\texttt{*}, 0\} & \{E \to \cdot\texttt{a}, 0\} \end{tikzcd}$$

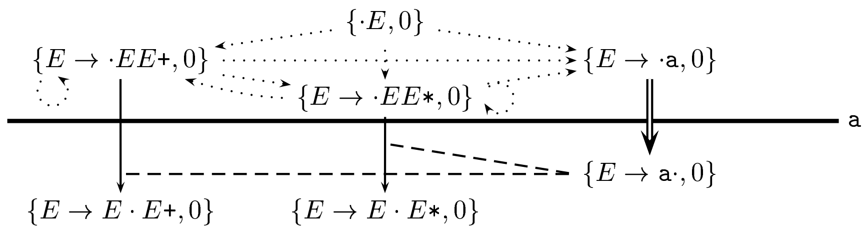
Az újonnan bekerült elemek közül kettőben a szabály részben a pont után nemterminális szimbólum (E) áll. Tehát ismét alkalmazhatjuk a második szabályt néhányszor. A szabály alkalmazása ugyan új elemet nem hoz be a csoportba, viszont néhány új nyíllal gazdagodik az ábra.



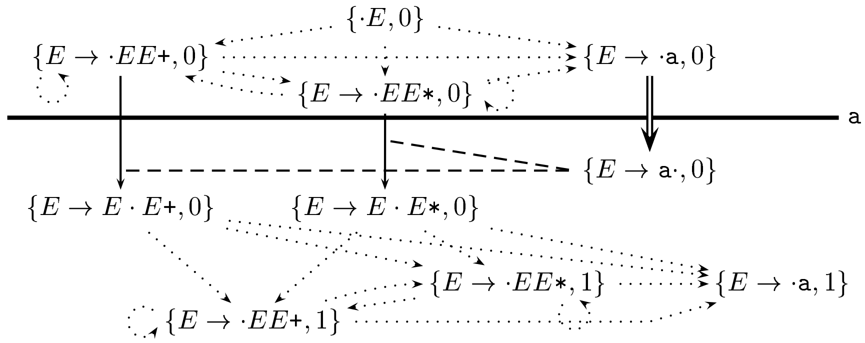
Miután minden lehetséges kombinációt végignéztünk, befejezhetjük a 0. csoporttal való foglalatoskodást. Az 1. csoport első elemét az első szabály segítségével kapjuk, méghozzá úgy, hogy az $\{E \to \cdot\texttt{a}, 0\}$ elemre alkalmazzuk.



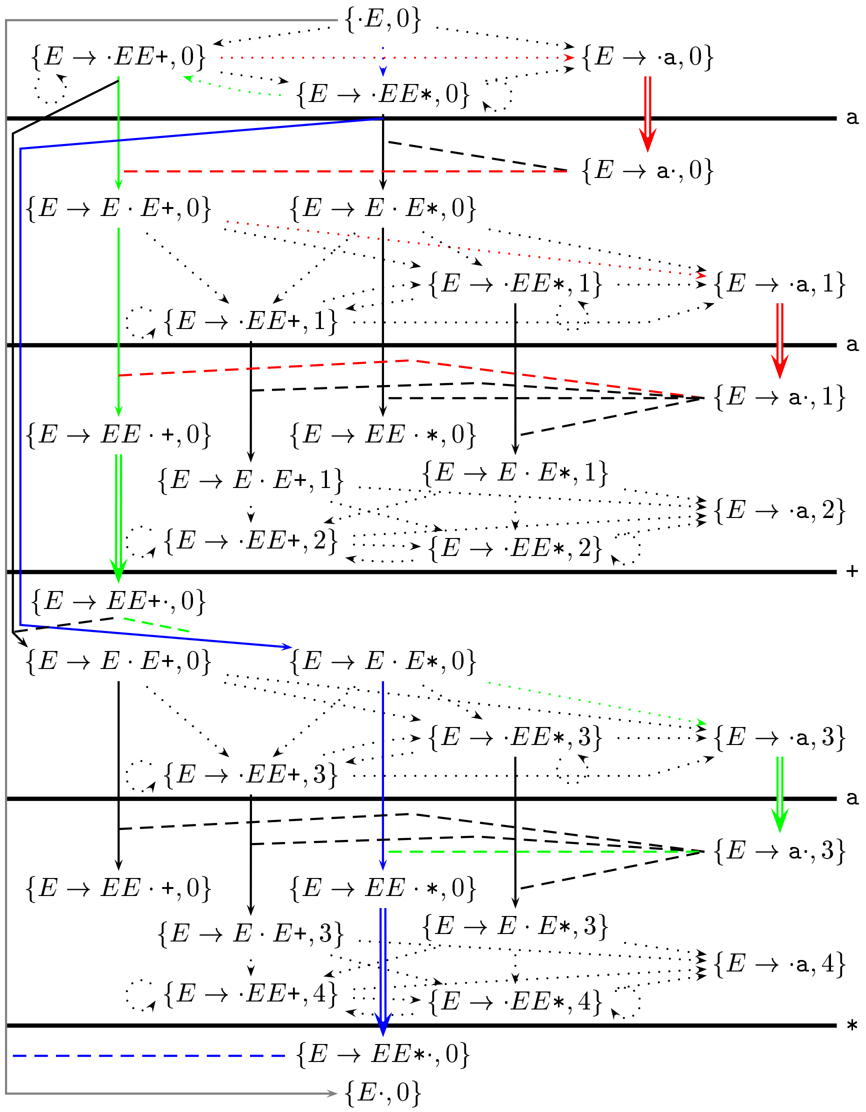
Az új elemnél a pont a jobboldal végére került. Az index azt jelzi, hogy a 0. csoportban kell kutakodni olyan elemek után, melyekben a pont után az E nemterminális szimbólum áll. Két elemet is találunk, ahol a ponttal átléphetjük az E nemterminálist.



De ekkor ismét nemterminális szimbólum áll a pont után az újonnan bekerülő elemekben, tehát a második szabály alkalmazható néhányszor. Ennek eredményeképpen az alábbi szituációhoz jutunk.



Eddigi gondolatmenetünket folytatva megkaphatjuk a végeredményt.



Ilyenkor jön jól egy színes lézernyomtató. ;-)

Felmerülhet a kérdés, hogy a csoportok elkészítésekor miért nem elégedtünk meg csupán az elemek jelölésével, illetve hogy miért használunk színeket az ábrán. Annak a megállapításához, hogy a jelsorozat eleme-e a nyelvtan által generált nyelvnek, tökéletesen elegendő a fekete–fehér ábra és az elemek nyilvántartása. Viszont a levezetés előállításakor jól jön a plusz információ.

Az utolsó, ötödik csoportban az $\{E\cdot, 0\}$ elem jelzi, hogy a vizsgált jelsorozat mondata a nyelvnek. Mivel a 0. csoportbeli $\{\cdot E, 0\}$ elemből (egyszeres) folytonos nyíl megy az előbb említett ötödik csoportbeli elemhez, ezért azt az útszakaszt részletesebben meg kell néznünk.

1. Ez a gyakorlatban azt jelenti, hogy meg kell keresnünk azt az elemet (jelen esetben $\{E \to EE{*}\cdot, 0\}$), amelyből szaggatott vonal vezet a már említett folytonoshoz.

2. Meg kell keresni azt az útvonalat, amelyen az előbbihez hasonlító elemek találhatók, annyi különbséggel, hogy az utóbbiaknál a pontot folyamatosan balrább és balrább tuszkoljuk. Ezt az útvonalat az ábrán kék színnel jelöltem. Az útvonal kötelező jelleggel tartalmaz az elején egy pontozott, a végén pedig egy szaggatott vonalat. A köztes vonaldarab pedig egyszeres folytonos vonal, ha nemterminálist léptünk át az adott lépésben, és dupla folytonos, ha terminálist.

3. Az előbb kijelölt útvonal mentén található egyszeres folytonos vonalaknál további elemzést kell végezni hasonlóan az itt leírakhoz. Ezt kb. úgy lehet felfogni, mintha rekurzív hívást hajtanánk végre minden egyszeres folytonos vonalnál. A rekurzív hívás mélységét jelölik a különböző színek. A rekurzív hívásoknál történik tényleges nyelvtani szabály alkalmazás, tehát a levezetés követéséhez szükséges szabálysorszámok kiírását például a pontozott vonalaknál lehet elvégezni.