Chomskyn hierarkia ja yhteysherkät kieliopit

Laskennan teorian opintopiiri

Tuomas Hakoniemi

21. helmikuuta 2014

Käsittelen tässä laskennan teorian opintopiirin harjoitustyössäni *muodollisten kielioppien* Chomskyn hierarkiaa, sekä hieman tarkemmin erästä hierarkian tasoa, *yhteysherkkiä kielioppeja*. Hierarkia on nelitasoinen. Esittelen työssä hierarkian tasot sekä niitä vastaavat automaatit. Tarkoituksena on antaa lyhyt yleiskatsaus, joten monet todistuksista sivuutetaan, sillä etenkin todistukset kielioppien ja automaattien ekvivalensseista ovat hyvin pitkiä. Kaikki todistukset ovat samankaltaisia. Ensinäkin kutakin kielioppia kohden konstruoidaan automaatti, joka simuloi kieliopin lauseenmuodostusta. Toisaalta kutakin automaattia kohden konstruoidaan kielioppi, joka simuloi automaatin toimintaa. Todistukset on löydettävissä Hopcroftin ja Ullmanin kirjasta [1].

1 Muodollinen kielioppi

Määritellään ensin, mitä tarkoitetaan muodollisella kieliopilla.

Määritelmä 1. *Muodollinen kielioppi* G on nelikko

$$G = (V, \Sigma, P, S),$$

missä

1. V on äärellinen epätyhjä joukko *välikkeitä*;
2. Σ on äärellinen epätyhjä joukko *päätemerkkejä*;
3. $P \subset (V \cup \Sigma)^* V (V \cup \Sigma)^* \times (V \cup \Sigma)^*$ on äärellinen epätyhjä joukko *produktiosääntöjä* tai *produktiota*;
4. $S \in V$ on *aloitusmerkki*.

5. $V \cap \Sigma = \emptyset$

Produktiolle $(\alpha, \beta) \in P$ käytetään merkintää $\alpha \to \beta$.

Kukin muodollinen kielioppi määrittelee (tai tuottaa) muodollisen kielen. Muodollisen kieliopin määrittämän kielen merkkijonoja ovat ne pelkistä päätemerkeistä koostuvat merkkijonot, jotka on muodostettavissa soveltamalla äärellisen monta kertaa produktiosääntöjä lähtien aloitusmerkistä. Produktiota voidaan soveltaa merkkijonoon, jos produktion etujäsen on merkkijonon osajono. Tällöin soveltamisella tarkoitetaan kyseisen osajonon korvaamista produktion takajäsenellä. Seuraavat kaksi määritelmää formalisoivat tämän.

Määritelmä 2. Olkoon $\alpha, \beta, \gamma, \delta, \omega \in (V \cup \Sigma)^*$, $A \in V$ ja $\beta A \gamma \to \omega$ produktio. Sanotaan, että $\alpha\beta A\gamma\delta$ *tuottaa suoraan* merkkijonon $\alpha\omega\delta$, merkitään $\alpha\beta A\gamma\delta \Rightarrow \alpha\omega\delta$. Sanotaan, että merkkijono α *tuottaa* merkkijonon β, merkitään $\alpha \overset{*}{\Rightarrow} \beta$, jos on olemassa äärellinen jono merkkijonoja $(\delta_i)_{i<n}$ s.e.

$$\alpha \Rightarrow \delta_0, \delta_0 \Rightarrow \delta_1, \ldots, \delta_{n-1} \Rightarrow \beta$$

Määritelmä 3. Muodollisen kieliopin G *tuottama kieli* $L(G)$ määritellään seuraavasti:

$$L(G) = \{\omega \in \Sigma^* \colon S \overset{*}{\Rightarrow} \omega\}$$

2 Säännölliset kieliopit

Pienin Chomskyn hierarkian luokista on *säännöllisten kielioppien* luokka. Säännöllisiä kielioppeja on kahta muotoa, *oikealle* ja *vasemalle lineaariset* kieliopit.

Määritelmä 4. Muodollinen kielioppi G on *oikealle lineaarinen*, jos sen kaikki produktiot ovat muotoa

$$A \to \omega B \text{ tai } A \to \omega,$$

missä $A, B \in V$ ja $\omega \in \Sigma^*$.

Vastaavasti G on *vasemmalle lineaarinen*, jos sen kaikki produktiot ovat muotoa

$$A \to B\omega \text{ tai } A \to \omega.$$

Muodollinen kielioppi G on *säännöllinen*, jos se on oikealle tai vasemmalle lineaarinen. Säännöllisen kieliopin tuottamaa kieltä kutsutaan *säännölliseksi* kieleksi.

Laskennan mallien kurssilla säännölliseksi kieleksi kutsuttiin kieltä, jonka jokin äärellinen automaatti tunnistaa. Kyseinen määritelmä ja juuri annettu määritelmä ovat yhtäpitävät.

Lause 5. *Olkoon G säännöllinen kielioppi. Tällöin on olemassa äärellinen automaatti, joka tunnistaa kielen $L(G)$*

Lause 6. *Olkoon L jonkin äärellisen automaatin tunnistama kieli. Tällöin on olemassa sekä oikealle lineaarinen kielioppi G että vasemmalle lineaarinen kielioppi G' s.e. $L(G) = L = L(G')$.*

3 Yhteydettömät kieliopit

Chomskyn hierarkian seuraava taso on *yhteydettömien kielioppien* luokka.

Määritelmä 7. Muodollinen kielioppi G on *yhteydetön* jos sen kaikki produktiot ovat muotoa

$$A \to \omega,$$

missä $A \in V$ ja $\omega \in (V \cup \Sigma)^*$.

Yhteydettömän kieliopin tuottamaa kieltä kutsutaan *yhteydettömäksi* kieleksi.

Selvästi jokainen säännöllinen kielioppi on myös yhteydetön. Käänteinen ei kuitenkaan päde. Esimerkiksi kielelle $L = \{a^n b^n : n \in \mathbb{N}\}$ on olemassa yhteydetön kielioppi, mutta ei säännollistä kielioppia.

Yhteydettömille kieliopeille on olemassa seuraava normaalimuoto.

Lause 8. *Olkoon L yhteydetön kieli, joka ei sisällä tyhjää merkkijonoa. Tällöin on olemassa yhteydetön kielioppi G, jonka kaikki produktiot ovat muotoa*

$$A \to BC \text{ tai } A \to a,$$

missä $A, B, C \in V$ ja $a \in \Sigma$, s.e. $L = L(G)$.

Sanotaan, että kielioppi G on Chomskyn normaalimuodossa*. Jokainen yhteydetön kielioppi on siis* ekvivalentti *Chomskyn normaalimuodossa olevan kieliopin kanssa.*

Seuraava lause on todistettu Laskennan mallien kurssilla.

Lause 9. *Kieli L on yhteydetön jos ja vain jos jokin epädeterministinen pinoautomaatti tunnistaa kielen.*

4 Yhteysherkät kieliopit

Chomskyn hierarkian toiseksi laajin luokka on *yhteysherkkien kielioppien* luokka.

Määritelmä 10. Muodollinen kielioppi on yhteysherkkä, jos sen kaikilla produktioilla

$$\alpha \to \beta$$

pätee $|\alpha| \leq |\beta|$, missä $|\alpha|$ tarkoittaa merkkijonon α pituutta. Yhteysherkän kieliopin tuottamaa kieltä kutsutaan *yhteyherkäksi kieleksi.*

Chomskyn normaalimuodon olemassaolosta nähdään suoraan, että jokainen yhteydetön kieli, joka ei sisällä tyhjää merkkijonoa, on myös yhteysherkkä. Ylläolevalla määritelmällä yhteysherkkä kieli ei kuitenkaan voi sisältää tyhjää merkkijonoa. Tällöin yhteydettömät kielet eivät olisi yhteysherkkien kielien osajoukko. Tämä puute voidaan korjata sallimalla tyhjän merkkijonon ϵ lisääminen yhteysherkkään kieleen. Yhteysherkkää kielioppia muutetaan tällöin lisäämällä uusi alkumerkki S_0 ja sille produktiot $S_0 \to \epsilon|S$, missä S on alkuperäinen alkumerkki (samanlainen muutos voidaan tehdä Chomskyn normaalimuodon esitykseen). Nyt yhteydettömät kielet ovat yhteysherkkien kielien osajoukko. Seuraava esimerkki osoittaa, että sisältyvyys on aito.

Esimerkki 11. Kieli $L = \{a^n b^n c^n \colon n \geq 1\}$ on yhteysherkkä, mutta tunnetusti L ei ole yhteydetön. Seuraava yhteysherkkä kielioppi tuottaa kielen L.

1. $S \to aSBC$
2. $S \to aBC$
3. $CB \to BC$
4. $aB \to ab$
5. $bB \to bb$
6. $bC \to bc$
7. $cC \to cc$

Nimitystä yhteysherkkä kielioppi selittää seuraava normaalimuoto yhteysherkille kieliopeille. Normaalimuodossa α ja β muodostavat yhteyden tai kontekstin, jossa produktiota voi soveltaa.

Lause 12. *Olkoon L yhteysherkkä kieli, joka ei sisällä tyhjää merkkijonoa. Tällöin on olemassa muodollinen kielioppi G, jonka kaikki produktiot ovat muotoa*

$$\alpha A\beta \to \alpha\gamma\beta,$$

missä $\alpha, \beta, \gamma \in (V \cup \Sigma)^*$, $\gamma \neq \epsilon$ ja $A \in V$, s.e. $L = L(G)$.

4.1 Lineaarisesti rajoitettu automaatti

Yllä on esitetty säännöllisille ja yhteydettömille kielille tutut niitä vastaavat automaatit. Seuraavaksi esitellään yhteydettömien kielien karakterisaatio automaattien avulla.

Määritelmä 13. *Lineaarisesti rajoitettu automaatti* on epädeterministinen Turingin kone, joka täyttää seuraavat kaksi ehtoa.

1. Syöteaakkosto sisältää kaksi erikoissymbolia, vasemman reunamerkin X_L sekä oikean reunamerkin X_R, $X_L \neq X_R$.

2. Lukupää ei voi liikkua X_L:n vasemmalle puolen tai X_R:n oikealle puolen eikä kone voi kirjoittaa mitään kyseisten merkkien paikalle.

Laskennan alussa nauhan sisältö on $X_L\omega X_R$, missä ω on syöte.

Lineaarisesti rajoitettu automaatti on siis Turingin kone, jonka käytettävissä oleva nauha on rajoitettu nauhan osalle, joka laskennan alussa sisältää syötteen sekä reunamerkit. Muuttamalla määritelmää siten, että käytettävissä olevan nauhan osan pituus olisi syötteen pituuden suhteen lineaarinen, saisimme mallin jonka laskennallinen kyky on yhtäläinen yllä annetun määritelmän kanssa. Tämä selittää nimen lineaarisesti rajoitettu automaatti.

Yhteysherkkiä ovat täsmälleen ne kielet, jotka voidaan tunnistaa lineaarisesti rajoitetulla automaatilla. Intuitiivisesti lineaarisesti rajoitettu automaatti tunnistaa yhteysherkän kielen, sillä tarkistaakseen kuuluuko jokin merkkijono kieleen simuloimalla kielen kielioppia, automaatti ei tarvitse nauhaa enempää kuin merkkijonon pituuden verran. Tämä johtuu yhteysherkän kielen kieliopin produktioiden muodosta - ei ole tilannetta, jossa kieleen kuuluva merkkijono voitaisiin johtaa vain sitä pidemmän merkkijono avulla.

Lause 14. *Kieli L, joka ei sisällä tyhjää merkkijonoa, on yhteysherkkä jos ja vain jos jokin lineaarisesti rajoitettu automaatti tunnistaa kielen.*

Yhteysherkkien kielien joukko on jo hyvin iso. Yllä on annettu yksinkertaiset esimerkit kielistä, jotka eivät ole säännöllisiä tai yhteydettömiä. Tällaisia simppeleitä esimerkkejä ei-yhteysherkistä kielistä ei ole. Kaikki esimerkit ei-yhteysherkistä kielistä perustuvat jonkunlaiselle diagonaaliargumentille. Seuraavaksi osoitamme, että jokainen yhteysherkkä kieli on rekursiivinen, ja että on olemassa myös rekursiivinen kieli joka ei ole yhteysherkkä.

Lause 15. *Jokainen yhteysherkkä kieli on rekursiivinen.*

Todistus. Olkoon L yhteysherkkä kieli ja LBA lineaarisesti rajoitettu automaatti, joka tunnistaa kielen L. Käytämme hyväksi tietoa, että lineaarisesti rajoitetulla automaatilla voi käytettävissä olevan nauhan äärellisen pituuden takia olla vain äärellisen monta eri konfiguraatiota. Olkoon k erilaisten konfiguraatioden lukumäärä. Nyt voidaan muodostaa kaikilla syötteillä pysähtyvä Turingin kone M, joka tunnistaa kielen L, siten että M simuloi LBA:n laskentaa maksimissaan $k+1$:n askeleen verran. M pysähtyy ja hyväksyy/hylkää, jos LBA pysähtyy ja hyväksyy/hylkää, mutta jos laskenta on jatkunut $k+1$ askeleen verran M pysähtyy ja hylkää. Tällöin tiedetään, että jokin konfiguraatio on toistunut ja LBA on loopissa.

□

Sen todistamiseksi, että yhteysherkät kielet ovat rekursiivisten kielten aito osajoukko, tarvitaan seuraava aputulos.

Lemma 16. *Olkoon $(M_i)_{i\in\mathbb{N}}$ numeroituva perhe kaikilla syötteillä pysähtyviä Turingin koneita. Tällöin on olemassa rekursiivinen kieli L, s.e $L \neq L(M_i)$ kaikilla $i \in \mathbb{N}$.*

Todistus. Määritellään kieli $L \subset (0 \cup 1)^*, \epsilon \notin L$ seuraavasti. Kaikilla $\omega \in (0\cup1)^*$, $\omega \in L$ jos ja vain jos M_i ei hyväksy syötettä ω, missä ω on i:n binääriesitys. L on rekursiivinen, sillä, annettuna epätyhjä merkkijono $\omega \in (0\cup1)^*$, voimme muodostaa koneen M_i, ja tarkistaa hyväksyykö se merkkijonon. Nyt kaikilla $M_i, i \in \mathbb{N}$, $L \neq L(M_i)$. □

Korollaari 17. *On olemassa rekursiivinen kieli, joka ei ole yhteysherkkä.*

Todistus. Tämä seuraa nyt siitä, että voimme numeroida yhteysherkät kieliopit, joiden päätemerkkien joukko on $\{0, 1\}$. □

5 Rajoittamattomat kieliopit

Chomskyn hierarkian laajin luokka on *rajoittamattomien kielioppien* luokka.

Määritelmä 18. Rajoittamaton kielioppi on muodollinen kielioppi, jonka produktiot voivat olla mitä muotoa tahansa.

Seuraavat kaksi lausetta osoittavat, että rajoittamattomat kieliopit määrävät täsmälleen rekursiivisesti numeroituvien kielien joukon.

Lause 19. *Olkoon G rajoittamaton kielioppi. Tällöin on olemassa Turingin kone M s.e. $L(G) = L(M)$.*

Lause 20. *Olkoon M Turingin kone. Tällöin on olemassa rajoittamaton kielioppi G s.e. $L(M) = L(G)$.*

Viitteet

[1] John E. Hopcroft, Jeffrey D. Ullman: Introduction to Automata Theory, Languages and Computation. Addison-Wesley 1979

[2] Michael Sipser: Introduction to the Theory of Computation. Thomson Course Technology, 2006