LAP-lunttilappu v. 1.5 PK 2015-01-09

Kurssin tueksi — EI tenttiin mukaan!

Joukko-opin peruskäsitteitä

Joukko on kokoelma alkioita; $\{\}$, $\{1, 2, 3\}$, $\{\text{kissa}, \text{koira}\}$, $\mathbb{N} = \{0, 1, 2, \ldots\}$, $2\mathbb{N} = \{x \in \mathbb{N} \mid x \text{ on parillinen}\}$. Sitä että a on *joukon A alkio* eli *kuuluu joukkoon A* merkitään $a \in A$; $1 \in \{1, 2, 3\}$, $4 \in 2\mathbb{N}$, $4 \notin \{1, 2, 3\}$, $5 \notin 2\mathbb{N}$, kana $\notin \{\text{kissa}, \text{koira}\}$.

Tyhjässä joukossa $\emptyset$ ei ole yhtään alkiota: $\emptyset = \{\}$.

Joukko A on *joukon B osajoukko*, merk. $A \subseteq B$, jos jokainen sen alkio on myös joukon B alkio; $\{1\} \subseteq \{1, 2, 3\}$, $\{1, 2, 3\} \subseteq \mathbb{N}$, $2\mathbb{N} \subseteq \mathbb{N}$; $\emptyset \subseteq A$ ja $A \subseteq A$ jokaisella joukolla A.

Joukkojen A ja B *yhdiste* $A \cup B$ koostuu alkioista, jotka kuuluvat joukkoon A tai joukkoon B (tai molempiin): $A \cup B = \{x \mid x \in A \text{ tai } x \in B\}$; $\{1, 2, 3\} \cup \{2, 4\} = \{1, 2, 3, 4\}$.

Joukkojen A ja B *leikkaus* $A \cap B$ koostuu alkioista, jotka kuuluvat kumpaankin joukkoon A ja B: $A \cap B = \{x \mid x \in A \text{ ja } x \in B\}$; $\{1, 2, 3\} \cap \{2, 4\} = \{2\}$.

Joukkojen A ja B *erotus* $A \setminus B$ koostuu alkioista, jotka kuuluvat joukkoon A mutta eivät kuulu joukkoon B: $A \setminus B = \{x \in A \mid x \notin B\}$; $\{1, 2, 3\} \setminus \{2, 4\} = \{1, 3\}$.

Joukkojen A ja B *karteesinen tulo* $A \times B$ on niiden alkioparien joukko: $A \times B = \{(a, b) \mid a \in A, b \in B\}$; $\{1, 2, 3\} \times \{2, 4\} = \{(1, 2), (1, 4), (2, 2), (2, 4), (3, 2), (3, 4)\}$.

Funktio f joukosta A joukkoon B on sääntö, merk. $f : A \to B$, joka liittää jokaiseen $a \in A$ yksikäsitteisen $f(a) \in B$.

Joukon A *potenssijoukko* $\mathcal{P}(A)$ on joukon A osajoukkojen kokoelma: $\mathcal{P}(A) = \{X \mid X \subseteq A\}$; $\mathcal{P}(\{1, 2, 3\}) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$.

Merkkijonot ja formaalikielet

Aakkosto on epätyhjä ja äärellinen joukko $\Sigma = \{a_1, a_2, \ldots, a_\sigma\}$ *merkkejä eli symboleja*; Esim. binääriaakkosto $= \{0, 1\}$, DNA-aakkosto $= \{C, A, T, G\}$, ASCII, UNICODE.

Merkkijono eli *sana* on järjestetty jono symboleja. Sanan $w = a_1 a_2 \cdots a_n$ *pituus* $|w|$ on sen sen merkkien lukumäärä n. *Tyhjä merkkijono* ε ei sisällä yhtään merkkiä; $|\varepsilon| = 0, |0| = 1, |\text{kissa}| = 5$.

Aakkoston Σ kaikkien merkkijonojen joukkoa merkitään Σ^*; $\{0, 1\}^* = \{\varepsilon, 0, 1, 00, 01, 10, 11, 000, 001, \ldots\}$. $\Sigma^+ = \Sigma^* \setminus \{\varepsilon\}$.

Merkkijonojen $v = a_1 a_2 \cdots a_m$ *ja* $w = b_1 b_2 \cdots b_n$ *katenaatio* vw muodostuu niiden peräkkäin asetetetuista merkeistä: $vw = a_1 a_2 \cdots a_m b_1 b_2 \cdots b_n$; jos $v = \text{JÄ}$ ja $w = \text{BÄ}$, niin $vw = \text{JÄBÄ}$; $\varepsilon w = w \varepsilon = w$ kaikilla $w \in \Sigma^*$.

Sanan i-kertainen toisto: $w^i = ww \cdots w$ (i kertaa); $w^0 = \varepsilon$ ja $w^1 = w$; Jos $w = \text{HE}$, niin $w^3 = \text{HEHEHE}$.

(Aakkoston Σ) merkkijonojen joukkoja ($L \subseteq \Sigma^*$) kutsutaan *(aakkoston Σ) kieliksi*; esim. $\emptyset$, Σ^*.

Kielten A ja B katenaatio AB muodostuu sanoista, joiden alkuosa voidaan valita joukosta A ja loppuosa joukosta B: $AB = \{xy \mid x \in A, y \in B\}$; $\{\text{J}, \text{JÄ}\}\{\varepsilon, \text{BÄ}, \text{ÄBÄ}\} = \{\text{J}, \text{JBÄ}, \text{JÄBÄ}, \text{JÄ}, \text{JÄÄBÄ}\}$; Kaikilla kielillä A pätee $A\{\varepsilon\} = \{\varepsilon\}A = A$ ja $A\emptyset = \emptyset A = \emptyset$.

Kielen A sulkeuma A^* koostuu sanoista, jotka voidaan muodostaa katenoimalla nolla tai useampia sen sanoja: $A^* = \{w_1 w_2 \cdots w_k \mid k \geq 0, w_i \in A \text{ jokaisella } i = 1, 2, \ldots, k\}$; $\{ab, ba\}^* = \{\varepsilon, ab, ba, abab, abba, baab, baba, ababab, \ldots\}$; $\emptyset^* = \{\varepsilon\}^* = \{\varepsilon\}$; muiden kielten sulkeuma on ääretön.

Säännölliset kielet ja lausekkeet, äärelliset automaatit

Aakkoston Σ *säännöllinen lauseke* on muotoa x missä $x \in \{\emptyset, \varepsilon\} \cup \Sigma$, tai (EF), $(E \cup F)$ tai E^*, missä E ja F ovat säännöllisiä lausekkeita; Esim. $\emptyset$, ε, a, $((ab^*) \cup c^*)$, $((a \cup b)^* c)$.

Säännöllisen lausekkeen E kuvaama kieli $L(E) \subseteq \Sigma^*$ määritellään induktiivisesti: $L(\emptyset) = \emptyset$, $L(x) = \{x\}$ kun $x \in \Sigma \cup \{\varepsilon\}$, $L((EF)) = L(E)L(F)$, $L((E \cup F)) = L(E) \cup L(F)$ ja $L(E^*) = L(E)^*$; $L(\varepsilon) = \{\varepsilon\}$, $L(a) = \{a\}$, $L((ab^*) \cup c^*)) = \{a, ab, abb, \ldots, \varepsilon, c, cc, ccc, \ldots\}$, $L(((a \cup b)^* c)) = \{c, ac, bc, aac, abc, bac, bbc, aaac, \ldots\}$.

Määritelmä: Kieli on *säännöllinen* joss se voidaan kuvata säännöllisellä lausekkeella.

Äärellinen automaatti (FA) $M = (Q, \Sigma, \delta, q_0, F)$ on viisikko, jossa Q on äärellinen joukko *tiloja*, Σ on (syöte)aakkosto, δ *siirtymäfunktio*, $q_0 \in Q$ *alkutila* ja $F \subseteq Q$ joukko (hyväksyviä) *lopputiloja*. Jos siirtymäfunktio liittää jokaiseen tilaan $q \in Q$ ja merkkiin $a \in \Sigma$ yksikäsitteisen kohdetilan $\delta(q, a) \in$

Q eli on muotoa $\delta : Q \times \Sigma \to Q$, kyseessä on *deterministinen automaatti* (DFA). Jos jokaiseen tilaan q ja merkkiin a liittyy joukko vaihtoehtoisia kohdetiloja $\delta(q, a) \subseteq Q$ eli siirtymäfunktio on muotoa $\delta : Q \times \Sigma \to \mathcal{P}(Q)$, kyseessä on *epädeterministinen automaatti* (NFA). Jos siirtymäfunktio sallii tilan vaihtamisen myös lukematta syötemerkkiä eli se on muotoa $\delta : Q \times (\Sigma \cup \{\varepsilon\}) \to \mathcal{P}(Q)$, kyseessä on *$\varepsilon$-automaatti* ($\varepsilon$-NFA). Jos automaatti pääsee siirtymäfunktionsa mukaisesti alkutilastaan q_0 johonkin lopputilaan $q \in F$ lukien syötteen $w \in \Sigma^*$ kaikki merkit, automaatti *hyväksyy* sanan w. *Automaatin M tunnistama kieli* $L(M) \subseteq \Sigma^*$ koostuu niistä merkkijonoista, jotka automaatti hyväksyy; $w \in L(M) \sim$ automaatti M hyväksyy sanan w, ja $w \notin L(M) \sim$ automaatti M *hylkää* sanan w.

Tulos: Jokaisen NFA:n tai ε-NFA:n M tunnistama kieli voidaan tunnistaa myös deterministisellä automaatilla, joka syntyy n.s. *osajoukkokonstruktiolla*:

Tilan $q \in Q$ ε-sulkeuma $\mathcal{E}(q) \subseteq Q$ sisältää tilan q ja siitä ε-siirtymin saavutettavat tilat: $q \in \mathcal{E}(q)$ ja $p \in \mathcal{E}(q) \Rightarrow \delta(p, \varepsilon) \subseteq \mathcal{E}(q)$. Yleistetään *$\varepsilon$-sulkeuma* ja *siirtymäfunktion δ arvo* merkillä $a \in \Sigma$ *tilajoukoille* $Q' = \{q_1, \ldots, q_n\}$: $\mathcal{E}(Q') = \mathcal{E}(q_1) \cup \cdots \cup \mathcal{E}(q_n)$ ja $\delta(Q', a) = \delta(q_1, a) \cup \cdots \cup \delta(q_n, a)$.

ε-NFA $M = (Q, \Sigma, \delta, q_0, F) \Rightarrow$ **DFA** $\hat{M} = (\hat{Q}, \Sigma, \hat{\delta}, \hat{q_0}, \hat{F})$: $\hat{Q} \leftarrow \{\hat{q_0} \leftarrow \mathcal{E}(q_0)\}$; **while** $\hat{\delta}(\hat{q}, \cdot)$ puuttuu joltain $\hat{q} \in \hat{Q}$, muodosta se: **for each** $a \in \Sigma$: $\hat{\delta}(\hat{q}, a) \leftarrow \mathcal{E}(\delta(\hat{q}, a))$ ja $\hat{Q} \leftarrow \hat{Q} \cup \{\mathcal{E}(\delta(\hat{q}, a))\}$; Aseta $\hat{F} \leftarrow \{\hat{q} \in \hat{Q} \mid \hat{q} \cap F \neq \emptyset\}$.[1]

Tulos: Jokaisen säännöllisen lausekkeen E kuvaama kieli voidaan tunnistaa äärellisellä automaatilla. Kielen $L(E)$ tunnistava ε-NFA M_E syntyy lausekkeesta E esim. "Thompson-konstruktiolla".

Tulos: Jokaisen äärellisen automaatin M tunnistama kieli voidaan kuvata säännöllisellä lausekkeella. Perustelu: Automaatista M voidaan muodostaa kieltä $L(M)$ kuvaava lauseke (muuntamalla M kaksitilaiseksi *lausekeautomaatiksi*).

Säännöllisten kielten sulkeumaominaisuudet ja pumppauslemma

Tulos: Jos A ja B ovat aakkoston Σ säännöllisiä kieliä, niin myös AB, $A \cup B$, A^*, $\overline{A} = \Sigma^* \setminus A$, $A \setminus B$, $A \cap B$ sekä $A^R = \{a_1 a_2 \cdots a_n \mid a_n a_{n-1} \cdots a_1 \in A, a_i \in \Sigma\}$ ovat säännöllisiä kieliä. Perustelut kieliä kuvaavia lausekkeita tai tunnistavia automaatteja muokkaamalla.

Pumppauslemma: Jokaisella säännöllisellä kielellä L on *pumppauspituus* $p \in \mathbb{N}$: jokainen vähintään sen pituinen kielen L sana s voidaan jakaa osiin $s = xyz$, missä (1) $y \neq \varepsilon$, (2) $|xy| \leq p$ ja (3) $xy^iz \in L$ kaikilla $i \in \mathbb{N}$. Käyttö: Osoitetaan kieli L ei-säännölliseksi valitsemalla jokin pumppauslemman mukainen sana $s \in L$ josta voidaan perustella, että se ei täytä lemman ehtoja. Esim. jos kieli $L = \{a^n b^n \mid n \in \mathbb{N}\}$ olisi säännöllinen, sillä olisi jokin pumppauspituus $p \in \mathbb{N}$. Nyt $s = a^p b^p \in L$ on pumppauslemman mukainen sana, jonka ehtojen (1) ja (2) mukaisissa jaoissa $s = xyz$ "pumppaustermi" $y = a^k$ jollain $k > 0$, joten ehto (3) ei toteudu; siksi kieli L ei voi olla säännöllinen.

Kontekstittomat kieliopit ja kielet

Kontekstiton kielioppi (CFG) $G = (V, \Sigma, P, S)$ on nelikko, jossa V on kieliopin aakkosto, $\Sigma \subseteq V$ on *päätesymbolien* aakkosto ja $S \in N$ on *lähtösymboli*, missä $N = V \setminus \Sigma$ on *välikesymbolien* aakkosto. P on joukko *sääntöjä* eli *produktioita*, jotka ovat muotoa $A \to \alpha$, missä $A \in N$ ja $\alpha \in V^*$. Kieliopilla voidaan *johtaa* jonosta $\alpha \in V^*$ jono $\beta \in V^*$, merk. $\alpha \Rightarrow^* \beta$, jos jonon α voi muuttaa jonoksi β korvaamalla välikesymboleja niiden sääntöjen oikeilla puolilla. Jos $S \Rightarrow^* \alpha$, niin α on kieliopin *lausejohdos*, ja jos lisäksi $\alpha \in \Sigma^*$, se on *kieliopin (tuottama) lause*. *Kieliopin tuottama* tai *kuvaama kieli* $L(G)$ koostuu kieliopin tuottamista lauseista: $L(G) = \{w \in \Sigma^* \mid S \Rightarrow^* w\}$.

Määritelmä: Kieli on *kontekstiton* joss joku kontekstiton kielioppi tuottaa sen.

Tulos: Kontekstittomat kielet on säännöllisten kielten aito laajennus. Perustelu: Jokainen säännöllinen kieli voidaan kuvata jo *lineaarisella kieliopilla*. Toisaalta esim. kielioppi, jonka produktiot ovat $S \to aSb$ ja $S \to \varepsilon$ tuottaa ei-säännöllisen kielen $\{a^n b^n \mid n \geq 0\}$.

Sulkeumaominaisuuksia: Jos A ja B ovat kontekstittomia kieliä, niin myös AB, $A \cup B$, A^* ja A^R ovat kontekstittomia; $A \cap B$ ja $\overline{A}$ eivät välttämättä ole kontekstittomia.

Pinoautomaatti (PDA) $M = (Q, \Sigma, \Gamma, \delta, q_0, F)$ on kuusikko, jossa Q on äärellinen joukko tiloja, Σ on (syöte)aakkosto, Γ on *pinoaakkosto*, δ siirtymäfunktio, $q_0 \in Q$ alkutila ja $F \subseteq Q$ joukko lopputiloja. Pinoautomaatin siirtymäfunktio δ :

[1] $X \leftarrow Y$ tarkoittaa sijoitusta "X saa arvon Y".

$Q \times (\Sigma \cup \{\varepsilon\}) \times (\Gamma \cup \{\varepsilon\}) \to \mathcal{P}(Q \times (\Gamma \cup \{\varepsilon\})$ voi riippua tilan ja syötemerkin lisäksi *pinon* huipulla olevasta merkistä, ja kohdetilaan siirtymisen lisäksi voi lisätä/poistaa/vaihtaa pinon huipulla olevan merkin. Syötteen hyväksymisen ja kielen tunnistamisen määritelmä on muuten sama kuin äärellisillä automaateilla.

Tulos: Jokainen kontekstiton kieli voidaan tunnistaa (epädeterministisellä) PDA:lla, ja jokaisen PDA:n tunnistama kieli on kontekstiton.

Kieliopin $G = (V, \Sigma, P, S)$ mukaisen lauseen $w \in L(G)$ *jäsennyspuu* on järjestetty puu, jonka (1) juuri on S, (2) lehtien katenointi muodostaa jonon w ja (3) jokaisella sisäsolmulla $A \in N$ on lapsisolmuina $x_1, x_2, \ldots, x_k$ vain jos $A \to x_1 x_2 \cdots x_k \in P$. Kielioppi on *moniselitteinen* jos se sallii jollekin syötteelle vaihtoehtoisia jäsennyspuita.

LL(1)-kielioppi mahdollistaa kielen osittavan (top-down), vasemmalta oikealle etenevän jäsentämisen siten, että sovellettavat produktiot määräytyvät yhden päätesymbolin kurkistuksella.

Jonoille $\alpha \in V^*$ määritelty joukko $\textsc{first}(\alpha) \subseteq (\Sigma \cup \{\varepsilon\})$ koostuu päätesymboleista, jotka voivat aloittaa jonosta α johdettavissa olevan jonon; jos $\alpha \Rightarrow^* \varepsilon$, myös $\varepsilon \in \textsc{first}(\alpha)$. Välikkeille A määritelty joukko $\textsc{follow}(A) \subseteq \Sigma$ koostuu päätesymboleista, jotka voivat esiintyä kieliopin lausejohdoksissa heti välikkeen A oikealla puolella.

Formaali LL(1)-ehto: Jos $A \to \alpha_1 \mid \alpha_2 \mid \cdots \mid \alpha_k$ ovat välikesymbolin A vaihtoehtoiset säännöt, niin (1) $\textsc{first}(\alpha_i) \cap \textsc{first}(\alpha_j) = \emptyset$ kun $i \neq j$ ja (2) jos $A \Rightarrow^* \varepsilon$ niin $\textsc{follow}(A) \cap \textsc{first}(\alpha_i) = \emptyset$ jokaisella ei-nollautuvalla α_i eli kun $\alpha_i \not\Rightarrow^* \varepsilon$.

Kieliopin muokkaaminen LL(1)-muotoon (ei ole aina mahdollista!):

Sääntöjen yhteisten alkuosien poisto: Korvaa (esim.) säännöt $A \to \alpha\beta_1 \mid \alpha\beta_2$, missä $\alpha \neq \varepsilon$ on niiden pisin yhteinen alkuosa, säännöillä $A \to \alpha A'$ ja $A' \to \beta_1 \mid \beta_2$.

Välittömän vasemman rekursion poisto: Korvaa (esim.) produktiot $A \to Aa \mid Ab \mid c \mid d$ produktioilla $A \to cA' \mid dA'$ ja $A' \to aA' \mid bA' \mid \varepsilon$.

Yleinen vasemman rekursion poisto: Käsittele välikkeet jossain järjestyksessä $A_1, \ldots, A_n$: (i) Jos välikkeellä A_i on muotoa $A_i \to B\alpha$ olevia sääntöjä, joissa B on aiemmin käsitelty välike, korvaa ne säännöillä jotka muodostuvat laventamalla B sen säännöillä $B \to \beta_1 \mid \cdots \mid \beta_k$; (ii) Poista välikkeen A_i mahdollinen välitön vasen rekursio kuten yllä.

Rekursiivisesti etenevä LL(1)-jäsennys: Laske etukäteen sääntöjen $\textsc{first}$-joukot. Tee kieliopin kullekin symbolille oma jäsennysproseduuri. Välikesymbolin jäsennysproseduuri valitsee sovellettavan säännön viimeksi luetun päätesymbolin ja sääntöjen $\textsc{first}$-joukkojen perusteella. Päätesymbolin jäsennysproseduuri tarkistaa, että viimeksi luettu päätesymboli on oikea ja lukee seuraavan päätesymbolin.