Referát č. 9 Chomského normální forma bezkontextových gramatik

Ivo Kostecký, kos0148

Při převodu bezkontextové gramatiky do Chomského normální formy se po odstranění pravidel typu A → ε provádí krok, který odstraňuje pravidla typu X → Y (neterminál se přepisuje na neterminál). (Pak tedy dostaneme jen pravidla typu A → α, kde α je terminál nebo |α| ≥ 2, přičemž generovaný jazyk se nezměnil.)

Na vhodném příkladu **ilustrujte algoritmus** odstraňující ona pravidla X → Y **a ukažte jeho korektnost**.

Chomského normální forma

Chomského normální forma je tvar či specifický způsob zápisu formální gramatiky.

Definice

Je přímo v zadání, jinými slovy: Bezkontextová gramatika je v Chomského normální formě, jestliže každé její pravidlo je tvaru X → Y Z nebo X → a, kde „X, Y, Z" označují neterminální symboly a „a" terminální symbol. *Tedy na levé straně je jeden neterminál, na pravé terminál nebo maximálně 2 další neterminály.*

Tvar CHNF

- *A* → *BC* nebo
- *A* → α nebo
- *S* → ε *(je povoleno, pokud gramatika generuje prázdný řetězec a zároveň se S nevyskytuje na pravé straně žádného pravidla)*

kde *A*, *B* a *C* jsou neterminály, α je terminál, *S* je startovní neterminál a ε je prázdný řetězec, přičemž *B* ani *C* nemohou být startovacím neterminálem.

Algoritmus

1. odstranění ϵ-pravidel;
   a. *nahrazení ϵ těmi terminály a neterminály, které by se tam mohly objevit a přidat je za stávající pravidla. Je třeba to provést důkladně a zahrnou všechny možné vzniknuvší kombinace!*
2. odstranění jednoduchých pravidel;
   a. *samostatné neterminály v jednom znaku nahradit jejich přepisem*
3. odstranění terminálů mimo $A \to a$;
   a. *zavést substituci novým neterminálem za všechny terminály, které jsou součástí neterminálů, kromě osamocených terminálů nebo skupiny terminálů.*
4. pohlcování neterminálních symbolů.
   a. *Všechny neterminály, které jsou delší než 2 znaky se musí zkrátit zavedením dalších substitucí. Ponechá se první znak a zbývající se substituují dalším neterminálem.*
   b. *Výsledkem je sada pravidel ve tvaru X → Y Z, přičemž pro zjednodušení mohou být zapsána i za sebou se znakem „nebo |".*

Příklady

Transformace bezkontextových gramatik

Příklad 1: Převeďte následující bezkontextovou gramatiku do Chomského normální formy.

$$\begin{array}{lcl} S & \to & AbB \mid d \\ A & \to & CAb \mid B \\ B & \to & cSd \mid \epsilon \\ C & \to & a \mid xd \end{array}$$

1. Odstraníme ϵ-pravidla

$$\begin{array}{lcl} S & \to & AbB \mid d \mid Ab \mid bB \mid b \\ A & \to & CAb \mid B \mid Cb \\ B & \to & cSd \\ C & \to & a \mid xd \end{array}$$

2. Odstraníme jednoduchá pravidla

$$\begin{array}{lcl} S & \to & AbB \mid d \mid Ab \mid bB \mid b \\ A & \to & CAb \mid Cb \mid cSd \\ B & \to & cSd \\ C & \to & a \mid xd \end{array}$$

3. Odstraníme terminální symboly mimo pravidel typu $A \to a$

$$\begin{array}{lcl} S & \to & ADB \mid d \mid AD \mid DB \mid b \\ A & \to & CAD \mid CD \mid ESF \\ B & \to & ESF \\ C & \to & a \mid GF \\ D & \to & b \\ E & \to & c \\ F & \to & d \\ G & \to & x \end{array}$$

4. Zkrátíme pravidla pohlcením neterminálních symbolů

$$\begin{array}{lcl} S & \to & AH \mid d \mid AD \mid DB \mid b \\ A & \to & CJ \mid CD \mid EK \\ B & \to & EK \\ C & \to & a \mid GF \\ D & \to & b \\ E & \to & c \\ F & \to & d \\ G & \to & x \\ H & \to & DB \\ J & \to & AD \\ K & \to & SF \end{array}$$

Příklad převodu na Chomského NF

Původní gramatika

$A \to B \mid C$
$B \to 0B1 \mid 01$
$C \to D \mid E$
$D \to 1D0 \mid 1$
$E \to 0E \mid 0$

Po odstranění $X \to Y$

$A \to 0B1 \mid 01 \mid 1D0 \mid 1 \mid 0E \mid 0$
$B \to 0B1 \mid 01$
$\cancel{C \to 1D0 \mid 1 \mid 0E \mid 0}$
$D \to 1D0 \mid 1$
$E \to 0E \mid 0$

Po nahrazení terminálů

$A \to NBJ \mid NJ \mid JDN \mid 1 \mid NE \mid 0$
$B \to NBJ \mid NJ$
$D \to JDN \mid 1$
$E \to NE \mid 0$
$N \to 0$
$J \to 1$

Chomského normální forma

$A \to NA_1 \mid NJ \mid JA_2 \mid 1 \mid NE \mid 0$
$A_1 \to BJ$
$A_2 \to DN$
$B \to NB_1 \mid NJ$
$B_1 \to BJ$
$D \to JD_1 \mid 1$
$D_1 \to DN$
$E \to NE \mid 0$
$N \to 0$
$J \to 1$

Automaty a gramatiky, Roman Barták

(další příklady na zvláštním listu)

Použití

Pro pochopení formy je třeba mít představu o tom, čeho forma to je.

Bezkontextové gramatiky

S bezkontextovými jazyky a gramatikami se hojně setkáváme při syntaktické analýze textu, třeba zdrojových kódů programovacích jazyků. Slovo „bezkontextová“ v názvu gramatiky znamená, že na levé straně každého pravidla stojí jeden neterminál bez sousedních symbolů (tedy není v „kontextu“).

G = ({VÝROK, PŘÍVLASTEK, PODMĚT, PŘÍSUDEK}, {malý, velký, pes, kocour, běží, leží, spí}, VÝROK, $P_{\text{dále}}$)

1. **VÝROK → PŘÍVLASTEK PODMĚT PŘÍSUDEK**
2. **PŘÍVLASTEK →** `malý`
3. **PŘÍVLASTEK →** `velký`
4. **PODMĚT →** `pes`
5. **PODMĚT →** `kocour`
6. **PŘÍSUDEK →** `běží`
7. **PŘÍSUDEK →** `leží`
8. **PŘÍSUDEK →** `spí`

Nyní zkusíme vygenerovat nějaké výroky. Začneme u počátečního symbolu a postupně aplikejeme některé z možných pravidel:

1. **VÝROK**
2. **PŘÍVLASTEK PODMĚT PŘÍSUDEK** (použité pravidlo: 1)
3. `malý` **PODMĚT PŘÍSUDEK** (použité pravidlo: 2)
4. `malý` `pes` **PŘÍSUDEK** (použité pravidlo: 4)
5. `malý` `pes` `spí` (použité pravidlo: 8)

Význam

Všechny bezkontextové jazyky lze generovat bezkontextovými gramatikami v jistém speciálním tvaru. Derivační stromy, příslušné k těmto gramatikám, pak mají jednotný tvar. Tato skutečnost usnadňuje cestu k závěrům v některých úvahách o bezkontextových jazycích. Příkladem je Chomského nebo Greibachova normalní forma.

Převedení zápisu gramatiky do CHNF vede k vytvoření jednoznačného derivačního stromu. Gramatika G je jednoznačná, jestliže každé slovo z L(G) má právě jedno levé odvození (právě jeden derivační strom), jinak je víceznačná a jako takovou ji není možné vždy transformovat na ekvivalentní jednoznačnou gramatiku.

Hlavní výhodou v Chomského normální formě je že, každé odvození řetězce s n písmeny má přesně 2n - 1 kroků, je tedy možné určit, zda je řetězec v jazyku všech derivací ve známém počtu operací.

Derivační strom

Derivační strom je v reprezentant syntaktické struktury slovního řetězce podle formální gramatiky. V derivačním stromě jsou vnitřní uzly označeny jako neterminály gramatiky, zatímco koncové uzly jsou označeny jako terminály. Derivační stromy mohou být využity pro generování nebo analýzu vět v přirozeném jazyce, stejně tak jako během zpracování počítačových jazyků.

Vytváření derivačních stromů ze specificky formátované gramatiky vede na jednoznačné derivační stromy, což je velký výhoda pro porovnávání bezkontextových gramatik mezi sebou.

Důkaz

Nejprve věta 12.2.2 o odstranitelnosti ε přechodů: Ke každé bezkontextové gramatice G lze sestrojit ekvivalentní nevypouštějící gramatiku G‘ tž. L(G‘) = L(G) − {ε}. *Důkaz ve skriptech str. 62.*

Věta 12.3.2 *Ke každé bezkontextové gramatice* G *lze sestrojit gramatiku* G' *v CHNF tž.* $L(G') = L(G) - \{\varepsilon\}$.

Důkaz: Podle věty 12.2.2 můžeme rovnou předpokládat, že G je nevypouštějící (jinak ji do této formy převedeme). Ukážeme, jak postupnou transformací G zkonstruujeme ekvivalentní gramatiku G' v CHNF.

Nejprve z gramatiky G odstraníme pravidla typu $X \rightarrow Y$:

Pro každý neterminál A zkonstruujeme množinu $\mathcal{D}_A = \{B \mid A \Rightarrow^* B\}$; pak pro každé pravidlo $B \rightarrow \alpha$, kde $B \in \mathcal{D}_A$ a α není rovno jednomu neterminálu, přidáme pravidlo $A \rightarrow \alpha$. Nakonec odstraníme všechna pravidla typu $X \rightarrow Y$. Snadno lze ověřit, že generovaný jazyk zůstává zachován.

Dále pro každý terminál a zavedeme nový neterminál A_a a přidáme pravidlo $A_a \rightarrow a$. Pak na pravé straně každého pravidla $X \rightarrow \alpha$, kde $|\alpha| \geq 2$, nahradíme každý výskyt terminálu a neterminálem A_a.

Je zřejmé, že generovaný jazyk zůstává stále zachován; jediná pravidla porušující podmínku CHNF mohou být typu $X \rightarrow Y_1Y_2 \dots Y_n$, kde $n \geq 3$ (Y_i jsou samozřejmě neterminály).

Každé pravidlo uvedeného typu lze ovšem nahradit soustavou $X \rightarrow Y_1Z_1$, $Z_1 \rightarrow Y_2Z_2, \dots, Z_{n-3} \rightarrow Y_{n-2}Z_{n-2}$, $Z_{n-2} \rightarrow Y_{n-1}Y_n$, kde $Z_1, Z_2, \dots, Z_{n-2}$ jsou nově přidané neterminály.

Opět je snadné se přesvědčit, že generovaný jazyk se nezmění a uvedenými změnami vzniklá gramatika G' je požadovanou gramatikou v CHNF. $\square$

Jednoduše si lze představit důkaz jako problém grafu o 2 uzlech nad sebou. Pokud horní a spodní uzel rozdělím a vložím mězi ně uzel další, pak se zcela logicky nemění význam grafu, pouze je v něm více uzlů. Přesně stejný případ vytváří gramatika v Chomského normální formě, protože nemění význam a pouze dochází k přidání pravidel se specifickými vlastnostmi.

Co se tím říká? Nedochází k porušení správnosti gramatiky, neboť důkaz odstranění ε přechodů se provádí výše, to je první zásah do konstrukce gramatiky: ve stručnosti se pouze přepisují symboly na nová umístění, ale jde jen komplikaci zápisu. Dalším porušením integrity by mohlo být přepisování pravidel neterminály, ale lze jistě souhlasit, že zavedením substituce nedochází ke změně významu původního výrazu a právě tato úprava se v průběhu převodu do CHNF odehrává. Intuitivně lze vyčíst, že v rámci algoritmu nedochází k modifikaci samotné gramatiky.

Zdroje:

- https://cs.wikipedia.org/wiki/Chomského_normální_forma
- http://www.cs.vsb.cz/jancar/TEORET-INF/teoret-inf.pdf
- http://www.cs.vsb.cz/kot/down/ti2007/TI-text-20070410.pdf
- http://www1.osu.cz/home/habibal/kurzy/ytzi2.pdf
- http://voho.cz/wiki/formalni-gramatika/
- http://people.cs.clemson.edu/~goddard/texts/theoryOfComputation/9a.pdf