Zadání úlohy do projektu z předmětu IPP 2011/2012

Zbyněk Křivka a Dušan Kolář
E-mail: {krivka, kolar}@fit.vutbr.cz, {54 114 1313, 54 114 1238}

DKA: Determinizace konečného automatu

Zodpovědný cvičící: Zbyněk Křivka (krivka@fit.vutbr.cz)

1 Detailní zadání úlohy

Vytvořte skript pro zpracování a determinizaci konečného automatu. Skript bude zpracovávat textový zápis zadaného konečného automatu a případně generovat ekvivalentní deterministický konečný automat přesně podle algoritmu, který negeneruje nedostupné stavy (viz čtvrtá přednáška předmětu Formální jazyky a překladače (IFJ), snímek 24/44).

1.1 Formát vstupu

Komentáře do konce řádku začínají znakem #. Bílé znaky jako konec řádku (`\n` i `\r`), mezera či tabulátor jsou ignorovány (až na později definované případy).

Stav je reprezentován identifikátorem jazyka C, který nezačíná ani nekončí podtržítkem. Vstupní symbol je reprezentovaný libovolným znakem, který může, ale nemusí, být v apostrofech. U stavů i vstupních symbolů záleží na velikosti písmen[1]. Znaky, které musí být v apostrofech, jsou všechny metaznaky popisu automatu a všechny bílé znaky, tj. `'('`, `')'`, `'{'`, `'}'`, `''''`, `'-'`, `'>'`, `','`, `'.'`, `'|'`, `'#'`, `' '`, znak konce řádku v apostrofech a znak tabulátoru v apostrofech. Apostrof musí být navíc uvnitř apostrofů zdvojený. Prázdná dvojice apostrofů `''` reprezentuje prázdný řetězec[2].

Celý konečný automat je zapisován podobnou notací jako ve formálních jazycích (IFJ). Celý konečný automat je pětice uzavřená do kulatých závorek. Každá komponenta kromě komponenty určující startující stav je dále uzavřena ve složených závorkách a od ostatních komponent oddělena čárkou. Jednotlivé prvky množin reprezentujících komponenty (opět kromě startujícího stavu) jsou odděleny také čárkou.

Nejprve je definována konečná množina stavů, následuje neprázdná vstupní abeceda, poté definice množiny pravidel, dále určení startujícího stavu a nakonec množina koncových stavů. Množina pravidel je popsána seznamem pravidel. Každé pravidlo je zapsáno ve tvaru: $p\,a$`->`q, kde p je *výchozí stav*, a je *čtený vstupní symbol* (může být i vynechán nebo nahrazen reprezentací prázdného řetězce), následuje dvojznak pomlčka s většítkem reprezentující šipku (tento dvojznak nesmí být rozdělen jiným znakem) a poslední část pravidla q určuje *cílový stav*. Pokud není a prázdné nebo uvedeno v apostrofech, tak musí být p a a odděleno alespoň jedním bílým znakem.

Příklad vstupního zápisu konečného automatu:

```
# Příklad konečného automatu ve vstupním formátu úlohy DKA
({s, p,q,r ,nonTermState, # stav neukončující i nedostupný
fin1}, # stav bude označen jako koncový
{a, '''', '{', ')', 'b', č ,b }, {s č->p,
q'a' ->   r, q'{'-> r, r a->r, p''''->q ,
r->s # takto vypadá pravidlo, které nečte vstup; lze i takto: r '' -> s
```

[1] tj. definice automatu je case-sensitive.

[2] V IFJ byl prázdný řetězec značen řeckým písmenem ε. V projektu z IPP tomu tak nebude.

```
},
# následuje komponenta definující startující stav
r
, {fin1, s, r} ) # koncové stavy a ukončení definice automatu
  # zde může následovat libovolný počet bílých znaků nebo komentářů
```

1.1.1 Kontrola správnosti vstupu

Vstupní automat nesplňující popsaná lexikální a syntaktická pravidla ukončí skript s chybou a návratovou hodnotou `40`. Pokud je porušena neprázdnost vstupní abecedy nebo nejsou první a druhá komponenta disjunktní nebo startující stav není v množině stavů nebo množina koncových stavů není podmnožinou množiny stavů, tak skript ukončete se sémantickou chybou a návratovou hodnotou `41`. Při opakování stejných pravidel/stavů/symbolů v rámci jedné komponenty jsou tato ve výsledné množině pouze jednou (tj. množiny na výstupu nejsou multimnožiny). Kombinace více chyb nebude testována.

1.2 Transformace a formát výstupu

Popis algoritmu je odkázán v referencích (snímky 10 a 24). Algoritmus je kvůli testování výsledků závazný. Po načtení automatu je třeba provést nejprve odstranění případných prázdných přechodů, které nečtou žádný vstupní symbol (tzv. ε-pravidla), pomocí algoritmu ze snímku 10/44 a poté provést determinizaci algoritmem ze snímku 24/44.

Při determinizaci se provádí tzv. *slučování stavů*. Pro jednotný výsledek bude potom výsledný identifikátor pro sloučený stav definován jako spojení všech původních stavů (resp. jejich identifikátorů) pomocí znaku podtržítko v lexikografickém vzestupném pořadí (pořadí jednotlivých znaků je určeno jejich ordinální hodnotou). Například, pokud budeme při determinizaci slučovat stavy `s1`, `p2`, `P2` a `p`, tak výsledný stav bude mít identifikátor `P2_p_p2_s1`, kdy jsou jednotlivé stavy spojeny podtržítkem v lexikografickém pořadí.

1.2.1 Normální forma výstupu

Výstupní formát výsledného konečného automatu vychází ze vstupního formátu a je definován následující normální formou. Všechny komentáře a nadbytečné bílé znaky budou vypuštěny. Automat bude vypsán v úplném tvaru a s každou komponentou začínající na zvláštním řádku. Kromě množiny pravidel budou všechny komponenty právě na jednom řádku. Za každou komponentou bezprostředně následuje čárka a odřádkování[3]. Za poslední komponentou nebude čárka ale pouze odřádkování, takže uzavírací kulatá závorka bude na novém řádku. Stavy v množině stavů, resp. symboly ve vstupní abecedě, budou odděleny čárkou a jednou mezerou (za posledním prvkem nebude čárka ani mezera) a ve svých komponentách seřazeny lexikograficky vzestupně. Každý symbol vstupní abecedy bude navíc uveden v apostrofech.

Každé pravidlo z množiny pravidel bude začínat na novém řádku (tj. odřádkování bude i za otevírající levou složenou závorkou; uzavírající pravá složená závorka bude též na novém řádku). Oddělující čárka bude bezprostředně za identifikátorem cílového stavu a za ní rovnou odřádkování. Množina výsledných pravidel bude seřazena vzestupně do seznamu pravidel podle sdruženého klíče ze tří podklíčů. Primární klíč je identifikátor výchozího stavu (řazeno lexikograficky podle ordinálních hodnot znaků), sekundární je znak reprezentující vstupní symbol bez případných uvozujících apos-

[3] Odřádkování provádějte unixovým způsobem (znakem `LF`).

trofů (řazeno podle ordinální hodnoty znaku, prázdný řetězec má hodnotu 0), posledním podklíčem je identifikátor cílového stavu (řazeno analogicky jako výchozí stavy).

Vstupní symbol obsažený v pravidlech bude na výstupu vždy uveden v apostrofech (stejně tak prázdný řetězec, který bude též v apostrofech). Části pravidla budou odděleny právě jednou mezerou, jak je vidět na příkladu formátování jednoho pravidla na výstupu (včetně oddělující čárky za pravidlem):

```
stav1_stav2 'a' -> stav2_stav3,
```

Tento skript bude pracovat s těmito parametry:

- `--help` viz společné zadání všech úloh.
- `--input=filename` zadaný vstupní textový soubor v UTF-8 s popisem konečného automatu.
- `--output=filename` textový výstupní soubor (opět v UTF-8) s popisem výsledného ekvivalentního[4] konečného automatu v předepsaném formátu.
- `-e`, `--no-epsilon-rules` pro pouhé odstranění ε-pravidel vstupního konečného automatu. Parametr nelze kombinovat s parametrem `-d` (resp. `--determinization`).
- `-d`, `--determinization` provede determinizaci bez generování nedostupných stavů (viz algoritmus IFJ, 4. přednáška, snímek 24/44). Parametr nelze kombinovat s parametrem `-e` (resp. `--no-epsilon-rules`).
- `-i`, `--case-insensitive` nebude brán ohled na velikost znaků při porovnávání symbolů či stavů (tj. `a = A`, `ahoj = AhOj` nebo `A_b` = `a_B`); ve výstupu potom budou všechna velká písmena převedena na malá.

Pokud nebude uveden parametr `-e` ani `-d`, tak dojde pouze k validaci a normalizovanému výpisu načteného konečného automatu.

2 Bonusová rozšíření

- **WCH** (0,5 bod): Volba `-w`, `--white-char` ve vstupu lze oddělující čárku nahradit libovolným bílým znakem (i komentář je brán jako bílý znak).
- **RUL** (0,5 bod): Volba `-r`, `--rules-only` vstupní soubor obsahuje zkrácený vstupní zápis tj. pouze množinu pravidel automatu (nikoli ostatní komponenty).

  Ve zkráceném zápisu jsou na vstupu pouze jednotlivá pravidla (bez složených závorek ohraničujících množinu pravidel v úplném zápisu). U pravidla může navíc za cílovým stavem následovat tečka, která označuje cílový stav zároveň jako koncový (tečku není třeba opakovat u každého výskytu stejného stavu). Výchozí stav prvního pravidla je definován jako startující stav. Zkráceně zapsaný konečný automat obsahuje právě všechny stavy a symboly použité v seznamu pravidel. Všechny komentáře jsou zahazovány.

  Příklad:

  ```
  startujiciStav a -> stav2, startujiciStav -> stav2, # nějaký komentář
  stav2 a -> f., stav2 a -> startujiciStav., f'' -> f # dva stavy jsou koncové
  ```

[4] Ekvivalentní konečný automat je definován jako automat přijímající stejný jazyk.

- **STR** (1 bod): Po zadání bonusového parametru `--analyze-string="retezec"` (nelze kombinovat s `-e` nebo `-d`) provede algoritmus analýzu, zda je zadaný `retezec` řetězcem jazyka přijímaného zadaným konečným automatem. Pokud ano, vypíše na výstup samotnou číslici `1`, jinak vypíše `0`. POZOR! Návratový kód je vzhledem k validní funkčnosti skriptu v obou případech `0`. V zadaném řetězci nemusíte uvažovat výskyt znaků uvozovek a apostrofů.
- **WSA** (1,5 bodu): Po zadání bonusového parametru `--wsfa` bude vstupní konečný automat[5] transformován na dobře specifikovaný konečný automat (pomocí algoritmu ze 4. přednášky IFJ na snímku 35/44). Výsledný automat bude obsahovat maximálně jeden neukončující stav s identifikátorem `qFALSE`.

Reference:

- A. Meduna, R. Lukáš. *Přednášky předmětu Formální jazyky a překladače (IFJ): Kapitola IV. Speciální typy konečných automatů.* FIT VUT v Brně, 2011. [cit. 2012-02-09]. Dostupné z: `https://www.fit.vutbr.cz/study/courses/IFJ/private/prednesy/Ifj04-cz.pdf`

3 Specifické požadavky na dokumentaci

Popište techniku ověřování lexikální a syntaktické správnosti vstupu. V případě determinizace nepopisujte obecný algoritmus, ale specifika Vaší konkrétní implementace.

4 Poznámky k hodnocení

Výsledný automat bude nejprve porovnán na přesnou shodu pomocí nástroje `diff`. V případě neúspěchu bude s bodovou srážkou provedena normalizace jinak syntakticky správného výstupního konečného automatu a uskutečněno nové porovnání.

[5] Předpokládejte, že vstupní konečný automat nemá stav `qFALSE`.