Semantic Parsing Grammar Format

Tomáš Lebeda[1]

1 Úvod

Jedním z často používaných přístupů k sémantické analýze textu, která je součástí metod zpracování přirozené řeči, je parsování vstupního textu pomocí formálních gramatik.

Tato práce představuje nový formát bezkontextových formálních gramatik, pojmenovaný jako *Semantic Parsing Grammar Format (SPGF)*, založený na existujícím standardu Speech Recognition Grammar Specification (SRGS). Hlavním cílem je poskytnout rozšířenou funkčnost parsování a ergonomičtější práci se samotnými gramatikami. Součástí práce je také implementovaný přidružený parser, který kromě samotného parsování textu a tvorby derivačních stromů zajišťuje také analýzu a validaci gramatik. Klíčovými inovacemi SPGF jsou možnost definovat více vstupních bodů v gramatikách a flexibilní kritéria pro úspěšné parsování, které umožňují částečné shody vstupního textu s gramatikou.

2 Návrh a implementace

Důvodem pro tvorbu SPGF byla absence některých funkcionalit v existující implementaci (součást platformy SpeechCloud) pro úlohu sémantické korespondence obrazu a řeči. Konkrétně se jednalo o absenci speciálního pravidla `$GARBAGE`, nevhodné návratové datové struktury a omezené parsovací strategie.

Prvním rozšířením je možnost definovat více vstupních bodů v každé gramatice, což je praktické například pro situace, kdy není předem známo, jaká z daných struktur bude odpovídat vstupnímu textu, nebo když může vstup vyhovovat více různým strukturám zároveň.

Dalším rozdílem je kritérium pro úspěšné parsování. SPGF nevyžaduje, aby byl text využit v derivačním stromu celý, od začátku do konce. Pokud tedy kořenovému pravidlu vyhovuje pouze první část vstupního textu, SPGF bude takový derivační strom považovat za úspěšný, zatímco SRGS nikoli. SPGF však definuje i speciální pravidla, pomocí kterých lze vynutit chování shodné s SRGS standardem.

Dalším rozdílem je práce s tagy, které lze v SPGF přiřadit pouze za *elementy* (tokeny, reference na pravidla nebo sekvencím). To výrazně zjednodušuje strojové zpracování a s ním spojené parsování.

Významným rozšířením oproti SRGS je práce s opakováním elementů a parsovacími strategiemi. V rámci definice opakování elementů byla přidána možnost specifikovat strategii parsování pro jednotlivá pravidla či dokonce elementy. To umožňuje kombinovat různé strategie v rámci jednoho derivačního stromu. Jedná se o způsob řešení nejednoznačných situací, které mohou vznikat v místě opakování elementů nebo alternativních expanzí pravidel. Pro SPGF

[1] student navazujícího studijního programu Kybernetika a řídící technika, obor Kybernetika, specializace Umělá inteligence a automatizace, e-mail: lebedat@students.zcu.cz

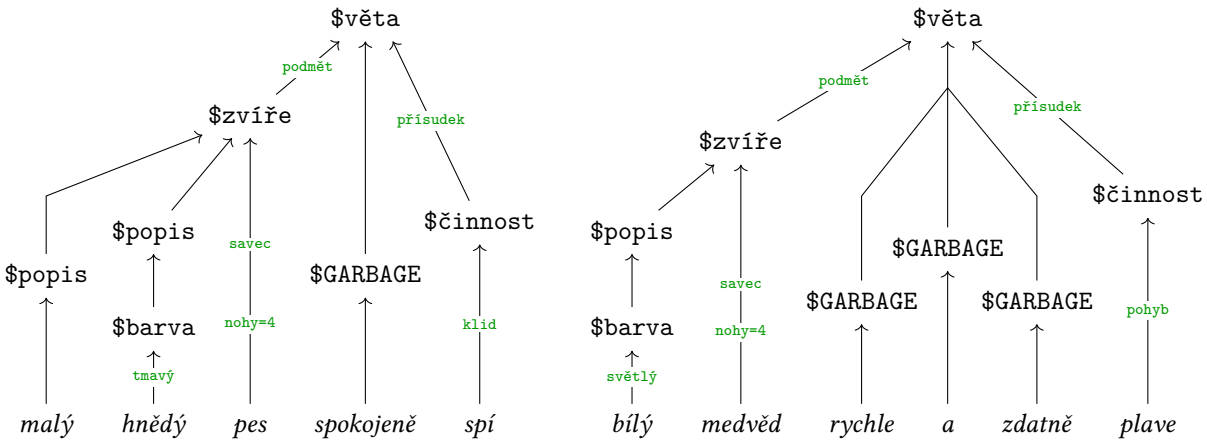
byly implementované tři strategie: *greedy*, *lazy* a *thorough*. Strategie *greedy* je výchozí chování pro SRGS a při parsování dává přednost opakování stejného elementu před postupem na další. *Lazy* strategie přistupuje k parsování opačně a při parsování upřednostňuje postup na další element. Třetí (nejsložitější) je strategie *thorough*, jejíž princip spočívá v tom, že v místě nejednoznačnosti se řešení rozdělí na více paralelních větví a ty se dále řeší nezávisle. Tento přístup je rekurzivní, takže vzniká stromová struktura možných derivačních stromů a na konci jsou vrácené všechny úspěšné možnosti. Díky tomu je možné získat všechna možná řešení a zachytit tak vazby a struktury, které by jinak nebylo možné získat.

3 Výsledky

Výsledkem je úspěšně vytvořený formát gramatik včetně parseru a validátoru, který umožňuje efektivní a univerzální způsob, jak parsovat vstupní text do derivačních stromů pro různé aplikace a nasazení. Příklad jednoduché gramatiky je na Výpisu 1, k němu jsou na Obrázku 1 znázorněné výsledné derivační stromy pro dva různé vstupní texty „*malý hnědý pes spokojeně spí*“ a „*bílý medvěd rychle a zdatně plave*“.

```
public $věta = $zvíře {podmět} $GARBAGE<L:*> $činnost {přísudek};
$zvíře = $popis<0-3> ((pes | medvěd) {nohy=4} |
         delfín {nohy=0}) {savec} | kapr {nohy=0} {ryba};
$popis = $barva | velký | malý | chlupatý;
$barva = (hnědý | černý) {tmavý} | bílý {světlý};
$činnost = spí {klid} | žere | plave {pohyb};
```

Výpis 1: Ukázka SPGF gramatiky



Obrázek 1: Ukázka dvou různých derivačních stromů

Literatura

Psutka, J., Müller, L., Matoušek, J., Radová, V. (2006). *Mluvíme s počítačem česky.* Prague: Academia.

W3C. (2004). „Speech Recognition Grammar Specification Version 1.0.“ [online] Available at: https://www.w3.org/TR/speech-grammar/