ČAJ – sémantická analýza

Pavel Rychlý

září 1995

Třetí částí systému Čaj je sémantická analýza české věty. Jejím cílem je z věty získat podstatu (sémantiku), tj. o čem se v ní jedná, a zaznamenat ji v nějakém kódu pro další použití. Výsledkem by měl být kód, ze kterého je již odfiltrována formální struktura věty, tvary jednotlivých slov a použitá slova vůbec. Pokud tedy na vstup dáme dvě věty, které jsou významově shodné, ale používají jiných jazykových konstrukcí (jiné pořadí slov ve větě, synonyma, atd.), výstupem by měl být v obou případech stejný výsledný kód. Příkladem mohou být věty:
Po lese chodí muž, nesoucí stroj.
Pán nese v lese stroj.

1 Základní pojmy

Pro reprezentaci sémantiky je použito tzv. *sémů*. Toto označení i princip pochází z prací dr. V. Smetáčka, *Sémantický analyzátor - SÉMAN*. Nejdříve tedy zavedeme nejdůležitější pojmy:

sém vyjadřuje nějaký společný rozlišovací znak skutečnosti, nějaký rys, pomocí kterého se jednotlivá slova rozlišují. Sémy jsou označovány dvojmístnými nebo trojmístnými znakovými kódy.

sémový operátor umožňuje modifikaci obsahu jednoho sému (např. aplikace sémového operátoru Y znamená nedostatek vlastnosti vyjádřené sémem). Operátory jsou označovány jedním písmenem.

pomocný pořadatel vyjadřuje sém se specifickým významem, jako jsou čísla, vlastní jména, atd.

sémový výraz je sém se všemi operátory, popř. včetně pořadatelů

sémový ekvivalent soubor sémových výrazů popisující sémantiku dané lexikální jednotky

Jednou ze základních výhod tohoto přístupu je, že umožňuje poměrně snadno pracovat s různě podrobnými slovy. Sémový ekvivalent slova, které je méně podrobné, obecnější, je totiž podmnožinou sémového ekvivalentu slova podrobnějšího, konkrétního. Např. slovo *muž* může být vyjádřeno sémovým ekvivalentem (lidský mužský), slovu *chlapec* může odpovídat (lidský mužský mladý) a stejný sémový ekvivalent bychom mohli dostat pro spojení *mladý muž*.

Dr. Smetáček se pokusil vytvořit úplný slovník sémů, použitelný pro veškeré druhy textů. Slovník dnes obsahuje asi 650 sémů, asi 14 sémových operátorů a 8 pomocných pořadatelů. S pomocí tohoto souboru se podle dr. Smetáčka popíše slovní zásoba celého jazyka. Podle mého názoru ovšem tento popis nemůže být jednoznačný. Dále neřeší uspokojivě problém homonym. V jeho slovníku *slovo — sémový ekvivalent* jsou pro homonyma spojeny všechny významy do jednoho sémového ekvivalentu, což neodpovídá podstatě věci.

2 Rozšíření systému sémů

V naší práci jsme se pokusili implementovat podobný model v jazyce Prolog. Základní princip byl zachován, navíc byl rozšířen o hierarchickou strukturu sémů, která je pro zachycení jednoznačné sémantiky potřebná.

Homonyma jsou řešena pomocí základního prostředku Prologu – backtrackingu. Ve slovníku *slovo — sémový ekvivalent* je slovo uvedeno pro každý svůj význam, tj. např. slovo se třemi významy je ve slovníku uvedeno třikrát. Při běhu programu se nejdříve uplatní první výskyt, pokud neprojde úspěšně sémantickou kontrolou, uplatní se další výskyt, dokud se nenajde sémantický ekvivalent, který vyhovuje všem kontrolám.

Sémantická kontrola je prováděna po převodu slov na jejich sémové ekvivalenty. Je řešena pomocí slovníku, který obsahuje vždy dvojici sémových ekvivalentů (skupinu sémů) a příznak, určující v jakém vzájemném vztahu se mohou vyskytovat. Jsou zde vyjádřeny tzv. pádové rámce sloves. První sémový ekvivalent odpovídá slovesu, druhý kontextu tohoto slovesa v závislosti na daném příznaku. V systému Čaj bylo v predikátu `ss` (sémantický slovník) použito následujících šest různých příznaků:

p – podmět,

o – objektiv,

a – adresát,

l – lokativ,

t – temporativ,

m – modativ.

Potom např. `ss(p, pohyb, bytost).` znamená, že se slovesem vyjadřujícím nějaký pohyb (jeho sémový ekvivalent obsahuje sém "pohyb") se jako podmět musí vázat nějaká bytost (slovo se sémantickým ekvivalentem obsahujícím "bytost"). Nebo `ss(o, působit, objekt).` znamená, že se slovesem vyjadřujícím působení (akci) se ve 4. pádě (objektiv) pojí slovo s významem objektu. Použitý princip by šel jednoduše rozšířit na další sémantické kontroly, jako je spojení přídavné jméno - podstatné jméno, aby bylo zajištěno nepřijmutí např. spojení *usměvavý balón*, *mladá židle* apod.

3 Použití v programu

Práce s programem vypadá následovně:
Uživatel zadá větu a počítač ji zpracuje. Před vlastním zpracováním je věta zkontrolována. Pokud není správně, vypíše se hlášení o syntaktické resp. sémantické chybě, podle toho, zda byla věta odmítnuta již při syntaktické kontrole, nebo až při kontrole sémantické. Jestliže je správně, provede se podle typu věty další zpracování.

Pro větu oznamovací je význam věty uložen do databáze. Pokud se v dané větě pojednává o podmětu již dříve známém, jsou dodatečné informace přidány do databáze k danému podmětu. V případě, že ve větě není žádná nová informace, databáze se nezmění. Pokud se ve větě objeví zcela nový podmět, je v databázi vytvořen nový záznam o tomto podmětu a jeho dalších vlastnostech.

Tázací věta může být pouze větou zjišťovací a je potom zpracována jako dotaz, na který program odpoví `Ano` nebo `Ne`.

Věta rozkazovací slouží k provádění pomocných funkcí a neslouží k přímé práci s uloženými znalostmi. Musí být v některém specielním tvaru, a podle rozkazu se provede daná akce. Rozpoznávané akce jsou:

skonči! – ukončení práce programu a návrat do systému,

vypiš slova! – vypsání databáze známých slov,

vypiš znalosti! – vypsání databáze zatím naučených znalostí,

zapomeň! – zapomenutí naučených znalostí (počáteční stav databáze).

Příklad běhu programu:

```
?- go.
|: xxxxxxxx.
Syntaktická chyba!
|: pes štěkala.
Syntaktická chyba!
|: pes štěká.
|: pes nese v lese hrad.
Sémantická chyba!
|: pes nese v lese maso.
|: nese pes maso?
Ano
|: štěká v lese pes?
Ne
|: chlapec nese stroj.
|: nese pán stroj?
Ano
|: chodí muž?
Ano
|: vypiš znalosti!
bytost zvire :
   pohyb pusobit([objekt]) - lok([misto]) obj([objekt,potrava])
   stekat
bytost osoba mlady :
   pohyb pusobit([objekt]) - obj([objekt,uzivat])
|: pán chodí.
To už znám.
```

```
|: konči!

Yes
```

4 Současná omezení a možnosti rozšíření

Systém Čaj neřeší:

- časové souvislosti, tj. např. sémantika slov dříve, později, ...
- logická analýza věty, negace na různých úrovních (slovo, věta). Např. souvětí „*Když přišel domů, pustil si televizi.*“ nebo „*Dostal pětku, protože neuměl slovíčka.*“ můžeme brát jako „*Přišel domů. Pustil si televizi.*“ resp. „*Dostal pětku. Neuměl slovíčka.*“ Logická analýza nám tedy dává dodatečné informace, které v jednoduchých systémech můžeme zanedbat.
- slovní spojení. Některá slova (zvláště v odborné terminologii) mají ve spojení jiný význam než složení významů jednotlivých slov.
- v současné implementaci je potřeba všechny tvary explicitně uvést ve slovníku *slovo — sémový ekvivalent*. Pro některé tvary by se výsledný sémantický ekvivalent měl skládat ze sémantického ekvivalentu nějakého základu a sémantického ekvivalentu modifikátoru (množné číslo, zdrobnělina, některé přípony, předpony, odvozeniny, zvratná slovesa apod.)

Dále je nutné si uvědomit, že sémový slovník (soubor všech sémů se svými „významy“) je nutné volit specificky podle dané aplikace.