

Posudek diplomové práce

Karel Vandas:
**Automatické určování sémantických preferencí pro slovesná valenční
doplnění**
(Posudek vedoucí práce)

Obsah práce

Autor se ve své práci zabývá automatickým určováním sémantických preferencí pro slovesná valenční doplnění. Pracoval přitom s řadou jazykových zdrojů, zejména s valenčním slovníkem VALLEX a korpusem Valeval, s korpusem CzEng a dále s českou větví lexikální databáze WordNet (ČWN), přičemž se detailně seznámil s jejich datovými formáty a s automatickými nástroji použitými pro jejich přípravu. Dále využil své zkušenosti s vybranými nástroji pro tzv. klastrovou analýzu.

Po úvodní motivační kapitole (kap. 1) autor představuje koncept valence (kap. 2), základní principy klastrové analýzy (kap. 3) a data, se kterými pracuje (kap. 4). Jádro práce tvoří kapitoly 5 a 6, které jsou věnovány návrhu experimentů a podrobnému vyhodnocení výsledků pro dvě vybraná slovesa. Práce končí krátkou diskusí výsledků a závěrem. Dále práce obsahuje (kromě seznamu obrázků, tabulek a použité literatury) uživatelskou dokumentaci popisující instalaci vyvinutého nástroje (příloha A), příklady datových formátů (příloha B) a CD s vyvinutým nástrojem, použitými daty a dokumentací.

Práce je psána anglicky.

Hodnocení

Cílem práce bylo navržení systému, který pro vzorek sloves určí (na základě jejich výskytu ve velkém korpusu) sémantickou charakteristiku jednotlivých valenčních doplnění.

Systém pracuje v několika krocích: (i) Pro vybraná slovesa jsou nejprve ze zvolených korpusů (Valeval a CzEng) získány a automaticky zpracovány příkladové věty. (ii) Na základě vytvořené reprezentace se systém pokouší tyto věty rozřadit do klastrů. (iii) V rámci klastrů se zjišťují sémantické preference valenčních doplnění, a to na základě hyperonym pro lexikální obsazení jednotlivých doplnění (tzv. abstrakce).

Autor se rozhodl práci doplnit webovým rozhraním, které umožňuje měnit nastavení experimentů a sledovat jejich výstupy.

K vlastnímu návrhu experimentů mám jeden základní dotaz – klastrování vět (bod (ii) výše) nedává rozumné výsledky (typicky vzniká mnoho triviálních klastrů obsahujících jedinou příkladovou větu a jeden klaster s většinou příkladových vět, a to i v případě, že systému "napovíme" správný počet klastrů). Proč tedy klastrování předchází vyhledávání hyperonym? Logickým krokem, který by mohl pomoci vyřešit problém řídkosti dat, se zdá právě práce se "zobecněnými" daty (tedy s jejich abstrakcí).

Další okruh připomínek se týká textového zpracování – přestože obecná struktura práce je logická a přehledná, jednotlivé kroky experimentů nejsou popsány dostatečně jasně (např. jaká jsou vstupní/výstupní formáty dílčích kroků)?

- Zcela nejasné je, co se vlastně děje v kroku popsaném v sekci 5.1.2 Automatic Analysis (Jak se má vektorová reprezentace vět=dokumentů pomocí váhy *tf-idf* k automatické analýze z korpusu CzEng? Jak se získá vektorová reprezentace ze stromové struktury dat?)
- Který klastrovací algoritmus byl použit, 'Agglomerative Hierarchical Clustering' (AHC) nebo 'k-means' algoritmus? Sekce 5.1.3 Clustering of Verb Examples nepodává vysvětlení;

v úvodu sekce 6.2.2 autor zmiňuje porovnání výsledků k-means a AHC, sekce však zřejmě zůstala nedopsána.

- Je nejasné, jak se při hledání hyperonym určují příslušné synsety (desambiguace).
- K obtížné srozumitelnosti textu přispívá skutečnost, že řada aspektů je popisována na několika místech (např. formát dat korpusu CzEng je zmiňován na str. 21 (4.1 CzEng) a str. 24 (4.4.2), přitom jeho popis lze nalézt jen v příloze B3; navíc je autorem používaný zjednodušený formát dat označen poněkud zavádějícím způsobem jako Origin, str. 24, 75).
- Autor navíc užívá řadu termínů, které nejsou běžně užívány či jde o jeho vlastní termíny, aniž by je definoval nebo dostatečně popsal (např. 'abstraction', 'granularity' vs. 'generalization' vs. '(level of) abstraction'; 'Valeval (defined)' vs. 'Vallex data source' vs. 'Valeval examples', 'Valeval (mixed)').

Některé další drobné připomínky

- str. 14 ... Zde se obecně se definuje 'cosine similarity', což je v pořádku; nikde v práci jsem ale nenašla, s jakou normou (jakými normami) systém pracuje?
- str. 16 ... Pokud chápu správně, váha *tf-idf* se vztahuje vždy k danému dokumentu a termínu, nikoli k celému souboru dokumentů, příklad je tedy špatně.
- str. 26 ... Jedním z bodů algoritmu je normalizace, v práci není vysvětleno, co se tím míní.
- str. 33/35 ... Proč se zavádí míra 'recall' (sekce 6.2.1), potom se však používá 'accuracy' (6.3.1)?
- str. 34 ... Jaká je motivace pro volbu 'baseline'? Kolik je to pro slovo *skupina* (6.3.2)? Míry se zavádějí v intervalu $\langle 0,1 \rangle$, uvádějí se ale v procentech %
- str. 37 ... Při evaluaci slovesa *postavit* se někdy pracuje s 52 (AHC) a někdy se 42 (k-means) příklady?
- str. 53 ... Co rozumíte pod "overall accuracy for abstraction identification"?
- webová aplikace a nastavení experimentů:
 - Nenašla jsem popis rysu 'nominative/accusative'.
 - Nastavení rysů 'feature extraction weight' není příliš intuitivní.
 - Jaká je motivace pro rys 'depth' (pracujeme-li s valenčními doplněními slovesa)?
 - Termíny v aplikaci ne vždy korespondují s termíny v textu, což je velice zavádějící (např. 'analytical position' (aplikace) vs. 'tag position'/'part of speech tag' (str. 28) vs. 'morphological tag simplification' (str. 18); 'Valeval (defined)' (aplikace) vs. 'Vallex data source' (text, str. 35).

Závěr

Posuzovaná diplomová práce splnila zadání, autor při jejím zpracování postupoval samostatně a projevil vysoký zájem o problematiku. Vlastní práce ovšem poněkud trpí nejasnou strukturou experimentů (a to jak z hlediska návrhu experimentů, tak z hlediska jejich popisu). I přes výše uvedené připomínky práci doporučuji k obhajobě.

V Praze, 1.9.2012

doc. RNDr. Markéta Lopatková, Ph.D.
Ústav formální a aplikované lingvistiky MFF UK