

Poslední kapitola je věnována automatickému přiřazování patternů jednotlivým výskytům sloves, a to na základě skóre určujícího naplnění pozic jednotlivých patternů; využívá přitom statistických charakteristik získaných při určování populací sémantických typů (zde bych přivítala motivaci u vzorců pro stanovení skóre $o_t(t,w)$, tedy skóre určujícího, zda slovo w může realizovat sémantický typ t). Navržený klasifikátor dosahuje v průměru přesnosti (*accuracy*) 72,86%, což je zlepšení o velmi dobrých 11,32% oproti baseline dané přiřazením nejčastějšího patternu. Oceňuji přehlednou tabulku s výsledky pro 20 sloves z testovacího souboru i velké množství doplňujících informací (např. míra jistoty či počet přiřazených patternů s velkou mírou jistoty).

O velmi dobré úrovni výsledků popsaných v diplomové práci svědčí i skutečnost, že experiment navržený pro zjišťování mezianotátorské shody (kapitola 3) bude prezentován na mezinárodní konferenci (TSD 2010).

Práce je přehledná a logicky strukturovaná, je psána čтивě a stylisticky vyváženě. Obsahuje malé množství překlepů a gramatických chyb (jde zejména o problémy s interpunkcí – na několika místech např. chybí interpunkční čárky oddělující (vložené) vedlejší věty, např. str. 7, 26, 32, 33, 35, 42, ...) a několik dalších drobných nesrovonalostí (procenta se počítají do 100, nikoli do 1, str. 23).

Doporučuji věnovat větší pozornost citacím příslušných pramenů – autorka sice pečlivě cituje webové stránky projektů, jejichž nástroje využívá, diplomová práce (jako každá vědecká a výzkumná publikace) by měla obsahovat citace na ‘standardní’ publikace (knihy, články v časopisech a konferenčních sbornících apod.).

Závěr

Posuzovaná práce je rozsáhlá a přináší cenné výsledky, které prokazují schopnost autorky samostatně řešit zadaný problém, její vhled do zpracovávané problematiky, schopnost kreativně využívat existující lingvistická data a nástroje a navrhovat nové postupy. Též po textové stránce je práce kvalitní.

Diplomovou práci doporučuji k obhajobě.

V Praze, 30.8.2010

RNDr. Markéta Lopatková, Ph.D.

Oponentský posudek na diplomovou práci

Lenka Smejkalová Typické vzory užívání anglických sloves

Obsah práce

Autorka se zabývá velmi zajímavou problematikou vytváření a využití slovníku *Pattern Dictionary of English Verbs* (PDEV, Hanks & Pustejovsky, 2005). Popisuje současný stav slovníku PDEV, analyzuje dostupná data a navrhuje procedury pro podporu dalšího vývoje slovníku.

Po krátkém úvodu se autorka věnuje popisu slovníku PDEV, představuje tzv. patterny, tedy typické vzory užívání sloves založené na metodě korpusové analýzy vzorů užívání sloves (CPA, Corpus Pattern Analysis), a způsob jejich vytváření; dále podává krátký přehled podobných projektů pro angličtinu i pro češtinu (kapitola 2). Třetí kapitola se zaměřuje na zjištování mezianotátorské shody při vytváření PDEV. Následující kapitoly představují postupy směřující k automatickému zpracování dat a příslušné nástroje, a to nástroje pro automatickou extrakci slovesních argumentů (kapitola 4), nástroje pro odstranění víceznačnosti při přiřazování sémantických typů jednotlivým argumentům (kapitola 5) a nástroj pro automatické rozpoznávání patternů (kapitola 6).

Po krátkém závěru následují dodatky (seznam zkratek, definice patternů, příklady populací sémantických typů, dále obsah a popis programů na přiloženém CD) a seznam užité literatury.

Hodnocení

Předložená diplomová práce referuje o rozsáhlé a velmi přínosné práci autorky na projektu PDEV a je nepochybně, že cílů stanovených v zadání se autorce podařilo plně dosáhnout.

Prvním cílem práce byla analýza dostupných dat a zjištění mezianotátorské shody při anotování vzorku Pražského anglického závislostního treebanku (PEDT) a Britského národního korpusu (BNC50) pomocí patternů pro vybraná slovesa z PDEV. Autorka prokazuje dobrý vhled do problematiky a předkládá zajímavé výsledky i diskuzi neshod. V práci ovšem postrádám způsob výpočtu míry *kappa* (Cohenovo kappa pro 2 anotátory a Fleissovo kappa pro více anotátorů) – přestože jsou tyto míry považovány za standardní, jejich aplikace není triviální (chybí mi též příslušné citace).

Nejobsáhlejší kapitola je věnována automatické extrakci slovesních argumentů. Autorka využívá existujících nástrojů (zejména Stanfordský parser, Charniak-Johnsonův parser, McDonaldův parser a dále nástroj pro identifikaci argumentů z aplikace Sketch Engine), jejich výstupy ovšem s velkým porozuměním problematice upravuje pro účely extrakce argumentů (např. zpracování sponových sloves a apozice v Stanfordském parseru, kombinace McDonaldova parseru s dvěma taggery a různé přístupy pro identifikaci závislostních vztahů). Autorka též provedla rozsáhlou evaluaci řady experimentů pomocí standardních měr *precision* a *recall* na datech získaných z manuálně zpracovaných stromů PEDT.

Následující kapitola se zaměřuje na získání tzv. populace sémantických typů, tedy seznamů slov, která realizují jednotlivé sémantické typy, a na odstranění víceznačnosti vznikajících na základě alternací v patternech. Autorka poměrně stručně popisuje experimenty s filtry využívajícími statistických charakteristik (zejména PMI, *pointwise mutual information*) ručně anotovaných dat. Výsledné PR grafy by si zasloužily podrobnější diskusi. (Z oddílu 5.5 vyplývá, že v experimentech nebyl použit NE klasifikátor, přestože se na něj autorka v dalším odvolává.)