Oponentský posudek doktorské disertační práce

Mgr. Jiří Semecký: Verb Valency Frames Disambiguation

Oponentka: Markéta Lopatková

V předkládané disertační práci se autor zabývá lingvistickou úlohou přiřazování slovníkových položek / valenčních rámců jednotlivým výskytům sloves, tedy úlohou spadající do problematiky 'word sense disambiguation' (WSD). Využívá přitom různé metody strojového učení. Svou úlohu řeší jako klasifikační úlohu, tedy úlohu rozřadit dané objekty (výskyty slovesa) do jedné z předem definovaných tříd.

Problematika WSD je jednou z klíčových úloh při automatickém zpracování přirozeného jazyka (NLP, natural language processing), se kterou se nezbytně střetne autor jakékoliv pokročilejší aplikace, jmenujme například strojový překlad či dobývání informací. Automatické přiřazování slovníkových položek jednotlivým výskytům sloves, které tvoří základ syntaktické struktury věty, umožňuje využít existující bohaté jazykové zdroje pro zpracování jazyka. Předkládaná disertační práce přináší hluboký vhled do této problematiky, která v takové šíři a hloubce nebyla zatím pro češtinu řešena (a nejsou mi známy ani srovnatelné výsledky pro jiné jazyky).

Obsah a členění práce

Práce sestává z osmi kapitol – po krátké úvodní kapitole, ve které autor stručně představuje řešenou problematiku a formuluje svůj úkol, se v druhé kapitole věnuje metodám strojového učení (ML, machine learning), které v dalších částech využívá pro klasifikaci slovesných výskytů. Tato kapitola dává srozumitelný přehled hlavních metod ML, popisuje tzv. Naïve Bayes Classifier, metodu rozhodovacích stromů (a vyvozování pravidel z implementace rozhodovacích stromů), metodu Support Vector Machine a metodu maximální entropie.

Krátká třetí kapitola je věnována vztahu mezi významy sloves a jejich jednotlivými valenčními rámci. Autor konstatuje, že v jím užívaných slovnících jsou valenční rámce dobrou aproximací slovesných významů.

Ve čtvrté kapitole jsou představeny datové zdroje, na kterých je práce postavena – jsou to valenční slovníky VALLEX a PDT-VALLEX a k nim příslušející korpusy, VALEVAL a Pražský závislostní korpus (PDT).

Pátá kapitola zmiňuje další lexikální zdroje, které v práci většinou využity nejsou, nicméně by mohly být využity pro podobné experimenty (BRIEF, WordNet, FrameNet, PropBank), a krátce charakterizuje některé práce z oblasti WSD, a to jak pro češtinu, tak pro jiné jazyky.

Šestá a sedmá kapitola tvoří jádro celé práce. V šesté kapitole jsou popsány a diskutovány jednotlivé typy rysů, které jsou využity při všech metodách ML. Jsou rozčleněny do pěti skupin – čistě morfologické rysy, rysy založené na syntaktické struktuře věty, rysy životnosti, rysy pro idiomy a rysy využívající nejvyšší úrovně ontologie WordNetu.

V sedmé kapitole autor velmi podrobně rozebírá a vyhodnocuje provedené experimenty. Porovnává výsledky jednotlivých metod ML i vliv výběru rysů, diskutuje význam jednotlivých měr úspěšnosti.

Práce má dvoustránkový závěr.

Dále práce obsahuje seznamy obrázků a tabulek a odpovídající seznam citované literatury. Je doplněna přílohou se seznamem všech použitých rysů a přílohou s vybranými rozhodovacími stromy generovanými algoritmem C5 z PDT.

Hodnocení práce

Autor, svým vzděláním informatik, prokazuje, že si během doktorského studia osvojil velmi dobré lingvistické základy. V práci se projevuje zejména porozumění problémům týkajících se valence sloves, a to zvláště při návrhu lingvisticky relevantních rysů pro ML.

Práce dokumentuje velké množství experimentů s různými metodami strojového učení – autor prezentuje výsledky šesti metod ML, kde každou metodu natrénoval pro 20 různých kombinací rysů a na dvou typech dat, a to pro 109 sloves ('base lemmas') z korpusu VALEVAL a pro všechna slovesa z PDT (z téměř 5 tisíc 'base lemmas'), která se vyskytla v trénovacích datech.

Zejména rozsáhlá sedmá kapitola (str. 83-116), věnovaná zhodnocení a diskusi výsledků provedených experimentů, svědčí o hlubokém pochopení problematiky. Ráda bych zde vyzdvihla zvláště dva aspekty. Prvním z nich je rozbor rysů, které nejvíce přispěly k úspěšnosti rozhodovacích stromů. Nepřekvapuje, že nejlépe se uplatnily rysy založené na syntaktické struktuře, neboť valenční slovníky mají popisovat právě syntaktické chování sloves. Tyto rysy – společně s rysy morfologickými – se také vyskytovaly v nejvyšších 'patrech' rozhodovacích stromů. Zajímavé ovšem je, že v některých případech se jako relevantní uplatnily rysy, které ve slovnících vůbec zachyceny nejsou (například rozhodovací strom pro sloveso *dělit* z PDT se rozhoduje podle předložky *podle* závisející na slovese, která typicky uvozuje volné doplnění CRIT, a jako taková v PDT-VALLEXu uvedena není).

Druhým velmi zajímavým aspektem, který bych ráda zdůraznila, je podrobný rozbor evaluačních metrik ('recall', 'precission', 'coverage' a 'accuracy') a stanovení, co v konkrétní úloze znamenají. Autor ukazuje, že uváděná čísla pro 'accuracy' se týkají úlohy přiřazování valenčních rámců všem (tektogramatickým) slovesům, pro která byl natrénován klasifikátor. To ale typicky není úloha, kterou chceme řešit – typicky chceme přiřadit valenční rámce (a tedy rozlišit význam) pro všechna slovesa ve zpracovávaném textu, bez ohledu na to, na jakých datech byly klasifikátory natrénovány. Autor proto dopočítává úspěšnost ('precission', 'recall' i 'accuracy') i pro takto formulovanou úlohu. Je přitom zřejmé, že přesně vystihuje potřeby automatického zpracování textu, a přitom dobře rozumí způsobu, jak se specifika konkrétních úloh promítají do problematiky vyhodnocování experimentů.

Připomínky, dotazy a návrhy

a) Obsah práce

Autor navrhl řadu rysů založených na syntaktické struktuře věty, tedy na morfologických charakteristikách slov v určitém syntaktickém vztahu ke zpracovávanému slovesu (konkrétně dětí a rodičů uzlu reprezentujícího sloveso). Zajímalo by mě, zda přitom bral v úvahu způsob zachycování koordinace a apozice v PDT, a tedy i v parserech na datech PDT založených. Konkrétně zda při určování, zda jsou splněny jednotlivé navržené rysy, zkoumal pouze děti a rodiče uzlu reprezentujícího sloveso, či zda 'přeskakoval' uzly pro koordinaci a apozici.

Druhá připomínka se týká rysu životnosti, 'animacy' – nesouhlasím s interpretací autora, že takto navržený soubor rysů zachycuje životnost. Konkrétně u zájmen, oddíl 6.5.2, první

tabulka, se navrhuje považovat za životná ta zájmena, která mají v morfologickém tagu hodnotu druhé pozice (detailní slovní druh):

5 – zájmeno "on" ve tvarech po předložce (tj. "n-": "něj", "něho", ...),
6 – reflexívní zájmeno "se" v dlouhých tvarech ("sebe", "sobě", "sebou"),
9 – vztažné zájmeno "jenž", "již", ... po předložce ("n-": "něhož", "níž", ...),
H – krátké tvary osobních zájmen ("mě", "mi", "ti", "mu", ...),
P – osobní zájmena (vč. tvaru "tys").

Tato zájmena však nezastupují pouze životná jména, např. ve větě *Dej ty stoly k sobě* bude mít tvar *sobě* tag P6-X3----------, přestože reflexivní zájmeno zde zastupuje neživotné substantivum *stul*.

Valenční teorie, kterou využívají oba slovníky, nepracuje s životností u prvních dvou aktantů sloves (na rozdíl např. od FrameNetu, který je v práci též zmiňován), lze proto předpokládat, že rys životnosti bude mít jen okrajové využití.

Třetí poznámka se týká možného rozšíření práce. V experimentech se pracuje s rysy pro idiomatické významy sloves ve VALLEXu – v tomto směru by bylo možné využít též funktory DPHR (závislá část frazému), u PDT-VALLEXu též CPHR (část slovesně jmenného výrazu).

Poslední poznámka se týká statistické významnosti uváděných výsledků. Autor ve svých nejlepších experimentech nesporně dosáhl vysoké úspěšnosti – 'accuracy' kolem 80% oproti 'baseline' 68,27% (resp. přes 77% oproti 60,74%) u VALEVALu a přes 78% oproti 71,98% u PDT. Některé metody, příp. kombinace rysů se však liší jen o desetiny, či dokonce o setiny procenta – jaký rozdíl ve výsledcích je signifikantní a kde může jít pouze o statistickou chybu?

b) Formální zpracování a drobné nepřesnosti a překlepy

Po formální stránce je práce přehledně a logicky členěna. Vzhledem k velkému rozsahu a zajímavosti popisovaných experimentů i k celkové obsahové úrovni práce lze jen litovat, že se autor nevyvaroval občasných ne zcela jasných formulací, drobných chyb obsahových i jazykových a překlepů. Jde například o následující:

- str. 40, pod výčtem ... shoda podmětu s přísudkem,
- str. 45, lexeme ... ve FGD lexém zahrnuje celou dvojici vidového páru, ve VALLEXu 1.5 i PDT-VALLEXu mají dokonavé i nedokonavé sloveso vždy vlastní slovníkové heslo,
- str. 50, 2. charakteristika VALLEXu ... slovesa pro zpracování byla vybírána podle frekvence v ČNK,
- str. 63, 1. odst. ... nejasná formulace,
- str. 65, fig. 6.1 ... nesouhlasí popiska horizontální osy,
- str. 69 a dál ... subordinated verb,
- str. 71 ... tag u *si* má být P7-X3----------,
- str. 79 ... animacy (i dál),
- str. 90, odstavec popisující oraculum ... nejasná formulace

Tyto i některé další drobné nedostatky dostal autor k dispozici ve formě korektorských oprav.

Shrnutí a závěr

Předložená práce dokumentuje velmi rozsáhlou práci s lingvistickými zdroji. Návrh lingvisticky relevantních rysů pro strojové učení a široká diskuse získaných výsledků

prokazují hluboké pochopení zkoumané problematiky. Problematika WSD je vysoce aktuální a výsledky autora dokumentované v předložené práci jsou originální a jednoznačně přínosné pro další aplikace v oblasti NLP. Autor prokazuje schopnost samostatné vědecké práce a originální přístup k řešeným problémům.

Práce splňuje všechny požadavky kladené na disertační práci, navrhuji tedy přijmout ji jako práci požadovanou pro udělení titulu Ph.D.

Praha, 12.8.2007

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