Oponentní posudek diplomové práce

Michal Auersperger:
Korektor anglické gramatiky: určité a neurčité členy

obor: Matematická lingvistika

Cíl práce

Aspirant se ve své práci zabývá tím, jak správně v anglickém textu automaticky doplňovat nebo opravovat určité a neurčité členy, což definuje jako klasifikační úlohu: (určitý, neurčitý, žádný) na nominálních frázích (ve smyslu frázové gramatiky).

Obsah práce

Práce je rozdělena do úvodu, čtyř hlavních kapitol a závěru. Po stručném úvodu s přehledem dosavadních přístupů k řešení problému následuje kapitola 1 podrobně zpracovávající jazykovou problematiku členů v angličtině. Ve druhé kapitole jsou představeny algoritmy strojového učení, logistická regrese a gradient tree boosting, které jsou v práci použity. Ve třetí kapitole "Experimental Setup" autor stručně popíše použitá data a podrobněji se věnuje návrhu příznaků (features) jak v původní literatuře, tak nových, které navrhuje sám. Ve čtvrté kapitole Evaluace přehledně představí jednotlivé provedené experimenty a jejich výsledky.

Hodnocení

Práce je napsána vynikající angličtinou, je velmi dobře strukturována a to jak ve smyslu návrhu experimentů a jejich evaluace, tak ve smyslu samotného textu. Typografické zpracování je také velmi dobré.

Přes níže uvedené dílčí nejasnosti a otázky je zřejmé, že jde o velmi kvalitní práci. Autor přesvědčivě demonstruje své porozumění jak samotnému problému umístění členů, včetně toho, že někdy neexistuje jen jedno správné řešení, a obsáhle o něm referuje z literatury. Stejně tak autor dobře popisuje modely, které použije, a opět demonstruje pečlivé studium stávajících prací. V sekci o použitých datech se dopouští některých drobných nepřesností, ale celkově, obzvláště ve výběru příznaků, opět prokazuje kompetenci. Evaluace je provedena velmi přehledně a pečlivě. Podle uvedených výsledků autorův systém překonává dosavadní publikované systémy na stejné úloze, což je v diplomové práci opravdu výborný výsledek.

Otázky:

Návrh rozšířených příznaků (features) na str. 34–35, které jsou vlastním přínosem autora, ukazuje velmi dobré porozumění jak anglickému jazyku a úloze členů, tak použité frázové gramatice. V sekci 3.1.2 se píše, že byl použit korpus BNC, jehož část psaných

textů byla rozdělěna tak, že 4/5 byly použity jako trénovací data, coiž činilo asi 6,9 mil. tokenů. To mi nesedí se známou velikostí BNC 100mil. slov, kde jen asi 10% je mluvený jazyk, viz `http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html`. Jaká data tedy přesně byla použita?

V sekci 3.2.1 na str. 32 je příznak "head form" definován jako "*lemma* of the head of the noun phrase". Opravdu se jedná o lemma a nikoliv o formu (jak u autora, tak v původním výzkumu, který zde replikuje)?

Stejně tak se píše o lemmatech u dalších příznaků, které mají v názvu "word". To je velmi neobvyklé užití slova "word" (obvykle užívaného v textu pro formu). Navíc na str. 34 se píše, že autor neužívá dalších zdrojů kromě samotného textu a parse trees. Ovšem Stanford Parser lemmatizaci, jak je obvyklé u angličtiny a PCFG, vůbec neobsahuje (viz `https://nlp.stanford.edu/software/lex-parser.shtml#Sample`), stejně jako samotná data WSJ (a celého PTB). Pro Collinsův parser (trénovaný na WSJ), domnívám se, platí totéž. Prosím o vysvětlení.

V sekci 3.2.2 Countability feature se ve zdůvodnění výběru přístupu uvádí, že varianta Nagata et al. (2005) dosahuje úspěšnosti 93%. Chybí zde ovšem srovnání s baseline jednoduššího přístupu, např. citovaným Han et al. (2006), aby bylo zřejmé, že je výběr smysluplný.

Kapitola 4 Evaluace je strukturována kompetentně a ukazuje schopnost správně strukturovat vyhodnocení experimentů i interpretovat dosažené výsledky.

Otázka k hodnotě cut-off value v sekci 4.1: Nebylo možné vyhodnotit všechny experimenty s prahovou hodnotou 3 namísto 5? Motivace k redukci počtu dimenzí v kapitole 3.3 je jasná, ale vzhledem k výsledku v tabulce 4.2 postrádám diskusi, proč nebylo možno použít hodnotu 3 a bylo nutné přijmout sníženou přesnost danou hodnotou 5. Podle tabulek 4.1–4.3 dopad této vyšší prahové hodnoty přibližně oidpovídá pozitivnímu dopadu autorových rozšířených příznaků.

V obrázku 4.2. se sice dá uhodnout, které čáry odpovídají trénovacím datům a které evaluačním, nicméně je lepší to graficky odlišit a zjednoznačnit.

V kapitole 4.4.1 ve srovnání úspěšnosti systémů na korpusu Wall Street Journal (WSJ, soustavně chybně označovaném názvem celého Penn Treebanku, který obsahuje ještě další a velmi odlišné části) a na datech z BNC postrádáme diskusi známé a značné specifičnosti jazyka WSJ, částečně dané specifickou doménou tohoto titulu. Autor na str. 57 zmiňuje, že by také bylo možno provést srovnání systémů s lidským hodnocením na původních datech WSJ, ale bohužel nezmiňuje, proč to tak neudělal a namísto toho zvolil jiná data (kde ještě navíc hraje podstatnou roli náhodný výběr z velmi rozsáhlé kolekce textů různého žánru).

Tuto sekci považuji za důležitou, protože naznačuje možné omezení celého přístupu. PCFG parsery trénované na WSJ mají v některých aspektech problémy právě s tím, že jsou trénované na tak specifických datech, tedy mají podstatně horší úspěšnost na out-of-domain datech. Pro reálné používání systému, jehož přesnost jinak působí velmi dobře (i když její interpretace je z důvodů uvedených v kapitole 4.4.1 velmi obtížná), by bylo zajímavé vědět, nakolik přesný je parsing samotný na datech mimo WSJ a zda to podle autora představuje problém. Pokud ano, jaká jsou možná řešení?

Závěr

Autor prokázal velmi dobrou erudici jak v tématu jazykovém (probematika členů v angličtině), tak v problematice klasifikačních úloh a jejich řešení. Provedl systematickou řadu experimentů, které nejen replikují publikované přístupy, ale přidávají i vlastní přínos autora, dobře je vyhodnotil a v neposlední řadě překonal ve svých experimentech dosud publikované výsledky. Předložená práce podle mě splňuje požadavky na diplomovou práci na MFF UK a doporučuji ji k obhajobě.

V Praze 5. června 2017 Pavel Straňák