RNDr. Drahomíra „johanka" Spoustová: Kombinované statisticko-pravidlové metody značkování češtiny (Formální popis českých vět a otázky jeho implementace)

Oponentský posudek disertační práce

Předložená disertační práce představuje další příspěvek k problému analýzy českého textu na úrovni tvaroslovné (morfologické). Synteticky navazuje na některé předchozí práce v oblasti morfologického značkování českých vět, speciálně pak na práce z oblasti statistického značkování češtiny Jana Hajiče, Barbory Vidové Hladké, Pavla Krbce a Jana Raaba (dř. Votrubce) a na práci Pavla Květoně v oblasti disambiguačních pravidel. Jádrem práce je důkladný výzkum možností kombinací několika automatických nástrojů (taggerů statistických i pravidlových). Autorka se zabývala i dalšími úzce souvisejícími oblastmi, jako je valence adjektiv (kterou potřebovala přímo uplatnit v jádru své disertační práce), a použití obdobných metod na syntaktickou analýzu češtiny jako jeden z možných směrů dalšího rozvoje vytvořených metod.

Práce je psána česky, je členěna do 5 kapitol a obsahuje anglické shrnutí a seznam literatury. Po krátkém úvodu, ve kterém autorka shrnuje cíl a obsah práce, popisuje v první kapitole současný stav dané problematiky, tj. použité taggery, systém na použití lingvistických pravidel (jazyk LanGR a jeho interpret), použitá data a v následujících experimentech použité evaluační metriky. Ve druhé kapitole popisuje práci na povrchově-valenčním slovníku českých deverbativních adjektiv, který je použit v pravidlovém systému a který vyžadují některá pravidla. Tento slovník dosud v češtině neexistoval, proto autorka sama tento slovník vytvořila konverzí existujícího povrchově-valenčního slovníku sloves. Třetí kapitola je popisem vlastní autorčiny práce na jádru disertace. Autorka popisuje základní linii postupné modifikace kombinačních experimentů a prezentuje jejich výsledky vyhodnocené podle standardních metrik. Autorka zde konstatuje a dokládá na experimentálních výsledcích, že nalezla takovou kombinaci metod a jejich spojení, která dává historicky dosud nejlepší výsledky českého taggingu. Na konci kapitoly se zmiňuje i o možnostech dalšího posunu úspěšnosti těchto metod. Čtvrtá kapitola pak popisuje další experimenty vzhledem k parsingu (syntaktické analýze českých vět), které však podle vlastního konstatování autorky (na základě provedených a popsaných experimentů) nepřinesly oproti současným statistickým parserům žádoucí zlepšení. V páté kapitole autorka shrnuje dosažené výsledky.

Hodnocení:

Především je nutno konstatovat, že autorka splnila cíl, tj. zvýšení úspěšnosti českého taggingu. V experimentálních pracích uvedeného typu se přitom tento cíl podaří splnit v méně než polovině případů, navíc rozdíl úspěšností mezi předchozím nejlepším taggerem (J. Raab-Votrubec, projekt Morče) a jejím výsledkem je výrazný; autorka dokládá, že se jedná o relativní redukci chyby 11.48%, avšak uvážíme-li, že nulová chyba není při dané metodice tvorby testovacích korpusů a jejich evaluace možná, jedná se o redukci chyby z praktického hlediska nepochybně ještě větší. Zajímavé a prakticky důležité jsou i další výsledky autorčiny práce: poukázala na to, že i na první pohled „zvláštní" kombinace, které by intuitivně neměly mít naději na úspěch, zlepšení přinášejí. Prokázala ovšem také, že pravidla v jejich současné podobě – byť doznala podstatných změn oproti předchozím kombinačním experimentům z roku 2001 – nepřinášejí výrazné zlepšení, avšak při absolutních procentech úspěšnosti, kterých se nyní v tomto problému dosahuje, je i malé zlepšení pro některé účely přínosné

(např. pro značkování ČNK), zvláště pokud nezáleží tolik na rychlosti zpracování. Důležité a pozitivní je i to, že autorka popsala i kombinační experimenty, které jsou méně časově náročné, a přitom zhoršení je minimální (např. vypuštěním pravidlové části). Vedlejším, avšak nezanedbatelným přínosem práce jsou i další výsledky: volně přístupné modifikace valenčního slovníku ve formě seznamu povrchových valenčních rámců deverbativních adjektiv jako zdroje pro další lingvistické experimenty, a popis experimentů v oblasti parsingu (byť z hlediska úspěšnosti parsingu „nevydařených" – je alespoň zřejmé, kterým směrem se dále nemá postupovat).
Z hlediska obsahu je jen škoda, že podobně autorka nepopsala i velké množství experimentů v oblasti taggingu, které nutně musela udělat, aby došla nakonec k výsledkům úspěšnějším, než kterých dosahují práce předchozí; to platí, i když tyto experimenty nakonec k cíli nevedly. V závěru třetí kapitoly autorka naznačuje další možnosti vylepšení (resp. možných experimentů, které by vylepšení přinesly, avšak bylo by nutné je provést, aby bylo možno toto vylepšení měřit a zjistit tak, zda vede k ještě lepším výsledkům). Chybí však důkladnější (podrobnější) obsahová analýza chyb, která se již ve „výšinách" úspěšnosti, ve které se nyní začínáme pohybovat, nedá vynechat – prosté kombinování dalších a dalších metod a známých algoritmů „naslepo" už totiž nemůže vést k podstatnému zlepšení, je třeba se i z lingvistického hlediska zamyslet nad příčinou zbývajících chyb a hledat cíleně metody nebo jejich modifikace, které povedou k dalším podobně (absolutně) úspěšným pracím. V tomto tedy autorka svým následovníkům práci neulehčila. Rovněž bych uvítal, kdyby autorka v závěru práce uvedla i řadu nápadů, o kterých vím, že je v průběhu práce měla, a které z prostých časových důvodů už nerealizovala – i tím by snad mohla pomoci v dalším rozvoji českého a obecně flexivního taggingu, byť tyto další možné kombinace experimentálně neověřovala.
Z formálního hlediska nemám námitek, práce je psaná logicky a čte se plynule. Některá motta u úseků práce jsou snad „drsnější" povahy, bylo by však překvapivé spíše to, kdyby podobné věci – vzhledem k tomu, jak autorku v komunitě kolegové znají – v práci chyběly. Seznam literatury je dostatečný a reflektuje výchozí situaci i prameny, které autorka v práci použila. Z práce je zřejmý vlastní přínos autorky a jsou identifikovány a správně citovány použité prameny a software. Evaluace je provedena standardně a pokud jsem mohl práci sledovat v jejím průběhu, autorka striktně dodržovala ve světě obvyklá „pravidla hry" z hlediska používání testovacích dat - její výsledky tak snesou i v tomto smyslu přísná mezinárodní kritéria a bude je možné publikovat v zahraničí (což by si zasloužily, mj, i proto, že práce je psaná česky a pro zahraniční zájemce tedy jazykově nepřístupná).

Autorka prokázala, že je schopná samostatného vědeckého myšlení a samostatné vědecké práce a doporučuji tedy, aby jí po obhajobě byl udělen titul Ph.D. v oboru MFF UK I-3 „Matematická lingvistika".

Praha, 3.6.2007