

Západočeská univerzita v Plzni, Fakulta aplikovaných věd, Katedra kybernetiky

Posudek na dizertační práci

Ing. Jan Hoidekr: Metody redukce OOV ve statistických jazykových modelech založených na třídách

Předložená práce, jejímž autorem je Ing. Jan Hoidekr, se věnuje problematice redukce OOV ve statistických jazykových modelech pro rozpoznávání řeči, konkrétně rozpoznávání řeči pro diktování soudních rozhodnutí. To je téma velmi aktuální a s rostoucím počtem soudních procesů a růstem ceny pracovní síly nejen v ČR nabývá na významu. Výsledky práce Ing. Hoidekra jsou navíc obecnější a použitelné v dalších oblastech.

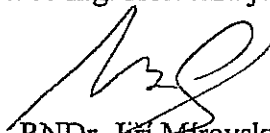
Práce je rozdělena do osmi kapitol. Po úvodu autor v kapitolách 2 a 3 blíže specifikuje cíle práce a v kapitolách 4 a 5 se zabývá již existujícími postupy v daném oboru. V šesté kapitole autor navrhuje řešení vytyčených cílů a v kapitole sedmé se věnuje experimentům a jejich výsledkům. Osmá kapitola je krátký závěr.

Po formální a obsahové stránce (pomímám teď stránku jazykovou) je předložená práce víceméně v pořádku. Obsahuje všechny potřebné kapitoly a formální náležitosti. Zvolený postup řešení vychází z autorovy dřívější práce a obohacuje ji o netriviální pravděpodobnostní model vycházející z řady dostupných datových zdrojů, které autor dobře využívá. Cíl práce je splněn, jak ukazují výsledky experimentů - rozpoznávání OOV slov se pomocí použitých metod zlepšilo a zároveň nedošlo k výraznému zhoršení celkového rozpoznávání. Mám jen několik snad menších výhrad: V některých vzorcích, zvláště z počátku kapitoly 4, jsou drobné faktické či formátovací chyby; jsou to naštěstí vzorce dobře známé. Kapitoly 4 a 5, popisující současný stav v daném oboru, by si zasloužily mnohem více odkazů. Literatura na konci práce je rozsáhlá, jen místy neúplná (chybí název sborníku či jméno konference). Totéž platí pro seznam autorových publikací - je dostatečně rozsáhlý a potvrzuje, že autor svoji práci publikoval na mezinárodní úrovni, ale chybějí v něm důležité části bibliografických údajů. Zajímavý by byl, ale chybí, rozbor chyb, které vytvořený systém dělá oproti systému původnímu.

Největší výhrady mám k jazykové úrovni předložené práce. Obsahuje velké množství chyb, které naznačují, že její konečná písemná podoba vznikala narychlo a že vysázenou podobu práce autor snad ani nečetl. Autor soustavně chybí v psaní čárek, místy v gramatice, v textu jsou přehlédnutá opakující se či naopak chybějící slova, překlipy, zapomenuté kusy vět. Naštěstí to nebrání srozumitelnosti obsahu textu, což dokazuje, že jako prostředek dorozumívání je čeština systémem velmi robustním. Z toho důvodu nepovažuji jazykovou úroveň práce pro otázku její dizertability za fatální.

Na základě výše uvedeného konstatuji, že Ing. Jan Hoidekr prokázal schopnost samostatně vědecky pracovat. Zvolené téma patří v oboru informačních technologií k aktuálním a potřebným a věřím, že výsledky práce se uplatní jako základ pro další vědecký výzkum i přímo v praxi. Mohu tedy prohlásit, že práce Ing. Hoidekra je dizertabilní a doporučuji ji k obhajobě.

V Praze, 27. listopadu 2012



RNDr. Jiří Mírovský, Ph.D.
Ústav formální a aplikované lingvistiky
Univerzita Karlova v Praze, MFF

POSUDEK NA DISERTAČNÍ PRÁCI

Téma práce: Metody redukce OOV ve statistických jazykových modelech založených na třídách.

Doktorand: Ing. Jan HOIDEKR

Posudek vypracoval: Doc. Ing. Petr POLLÁK, CSc.
ČVUT FEL K13131, Technická 2, 166 27 Praha 6

Disertační práce ing. Hoidekra se věnuje doplnění jazykového modelu v rozpoznávací spojitě řeči o slova, která nejsou zahrnuta v modelu obecném získaném z trénovacího korpusu. Autor se úžeji zaměřuje na doplnění místopisných názvů v jazykových modelech pro přepis soudních rozhodnutí. To je nezbytné pro zvýšení přesnosti přepisu do textové podoby bez nutnosti ruční korekce špatně rozpoznaných různých místopisných názvů. Realizovaný výzkum je stále aktuální, ačkoliv je obecně problematika jazykového modelování dlouhodobě řešena v mezinárodním i národním měřítku a i pro jazykové modelování češtiny byly již dosaženy významné výsledky. Na tyto výsledky autor ve své práci navazuje.

Je nutné konstatovat, že předložená práce je spíše stručnější, obsahuje pouze 73 stran vlastního čistého textu. Přes tuto skutečnost popisuje originální metodiku řešení daného problému a cíle práce, které si autor stanovil, považují za disertabilní. Hlavní přínosy předložené práce vidím v následujících bodech.

- V první řadě autor naplňuje první cíl, podat přehled používaných metod jazykového modelování a metod doplňování slovníku rozpoznávání. Toto je obsahem kapitol 4 a 5. Autor předkládá souhrn používaných technik jazykového modelování. Popis je místy až příliš detailní a s některými popisovanými technikami ve své práci autor nakonec vůbec nepracuje (např. s jazykovými modely na bázi neuronových sítí), jejich podrobný popis je tedy trochu nadbytečný. Na druhou stranu mi zde ovšem chyběly konkrétní zmínky o publikovaných výsledcích dosažených citovanými autory a zejména pak ke konci práce jejich srovnání s výsledky dosaženými autorem samotným.
- Vlastním přínosem autora je metoda úpravy jazykového modelu umožňující vložení nových slov do jazykového modelu. Autor se zaměřuje na jazykové modely pro automatický přepis soudních rozhodnutí a navrhnul proceduru rozšíření jazykového modelu o třídy reprezentující místopisné názvy, konkrétně města a ulice, včetně určení pravděpodobností výskytu jednotlivých názvů v definovaných třídách. Tato informace byla získána z několika veřejně dostupných rejstříků a pro zvolenou metodiku určení jednotlivých pravděpodobností z těchto zdrojů byla potvrzena korelovanost s četnostmi výskytu v dostupném korpusu soudních rozhodnutí. Z větší části tohoto korpusu byl získán i přímo trénovaný jazykový model obsahující skutečné výskyty místopisných názvů v dostupných trénovacích datech. Ačkoliv je řešení zaměřeno na uvedenou úzkou tematickou oblast přepisů soudních rozhodnutí, navržená metodika je zobecnitelná a jistě použitelná i pro jiné tematické oblasti. Trochu širší diskuse na toto téma by byla určitě velmi zajímavá a je škoda, že v práci chybí.
- Významným přínosem je dále zpracování a označování dostupných textů pro vytvoření jazykových modelů a definování zvolených tříd, podobně také předzpracování dat z veřejných rejstříků do podoby použitelné pro výše diskutovanou tvorbu a doplnění jazykového modelu. Zde možná opět trochu chyběla diskuse, jak některé tyto kroky automatizovat, zejména s ohledem na v práci zmiňovaný požadavek aktualizace místopisných názvů.
- Úspěšné doplnění jazykových modelů bylo potvrzeno v experimentech, provedených na testovacích korpusech. Vytvořené modely dosahovaly nízké úrovně OOVr i perplexity. Kvalita

těchto modelů byla také potvrzena vlastním rozpoznáváním realizovaných na vytvořených nahrávkách soudních rozhodnutí z vybraných testovacích sad. Bylo dosaženo vysoké přesnosti rozpoznávání (ACC 84%), pro testovací sadu obsahující více místopisných názvů dosažená přesnost překonala výsledky získané s přímo tvořeným jazykovým modelem (ACC 80% oproti 78%).

Ačkoliv se autor nezabýval do detailu akustickým modelováním, pro úplnou představu o dosažené výsledné přesnosti přepisu čtených rozhodnutí, mi zde chyběla informace, zda byly testovací audio nahrávky vytvořeny jedním či více mluvčími, a v popisu rozpoznávače pak informace, zda byly použity přímo natrénované akustické modely nezávislé na mluvčím, či zda byly akustické modely adaptovány, jak bývá v podobných úlohách obvyklé.

Po formální stránce je předložená práce na akceptovatelné úrovni. Text je psaný srozumitelnou češtinou, grafická úprava je dobrá. Při podrobnějším čtení jsem narazil na některé formální prohřešky. Vyskytují se zde občas gramatické chyby, je možné nalézt problémy s interpunkcí, text je místy moc heslovitý, jsou zde nepřesné a nekompletní citace (zejména ty vlastní - bez dalšího dohledávání není často zřejmé, kde přesně byla daná práce publikována), častý je i výskyt jednopísmenných předložek na konci řádku, a rušivý je také nesystematický tok textu a matematických vztahů, které netvoří dohromady standardní věty. Pečlivější finální edice by práci jistě velmi prospěla. Rovněž není vhodné psát disertační práci v 1. osobě množného čísla, jakkoliv je zřejmé a pochopitelné, že mnohá komplexní řešení (jako třeba konstrukce rozpoznávače spojitě řeči) musí být výsledkem kolektivní práce.

K práci bych měl také připomínku k rozsahu jednotlivých kapitol. Část věnovaná popisu známých technik je spíše rozsáhlejší, zatímco popisu vlastní realizované práce, vlastních modifikací tvorby jazykového modelu i popisu experimentů je věnováno prostoru méně. Zejména v disertační práci bych očekával poměr opačný. V rešeršní části pak autor odkazuje celkem 40 publikací včetně vlastních prací, což představuje spíše menší rozsah. Je ovšem nutné na druhou stranu konstatovat, že významné práce, zejména pak ty pro český jazyk, jsou zde citovány.

Přes výše zmíněné připomínky není o originálním přínosu autora v dané oblasti pochyb, jakkoliv je spíše menšího rozsahu. Originalita a přínos dosažených výsledků je potvrzena publikační činností autora. Autor publikoval výsledky své práce jako spoluautor na mezinárodních workshopech a konferencích. Ačkoliv zde sice chybí příspěvky na nejprestižnějších konferencích řady Interspeech či ICASSP, stejně jako časopisecké publikace, autorovi publikace jsou většinou kvalitní a jsou zaindexované v mezinárodních databázích.

Na základě všech výše uvedených skutečností lze tedy na závěr konstatovat, že předložená práce přináší originální výsledky, a proto práci **doporučuji** k obhajobě za účelem získání vědecké hodnosti doktora na Západočeské univerzitě v Plzni.



V Praze dne 7. prosince 2012