Mgr. Pavel Pecina:
Lexical Association Measures (Collocation Extraction)
Disertační práce

(vyjádření školitele)

Předložená anglicky psaná disertační práce se zabývá binárními lexikálními asociačními mírami a jejich kombinacemi z hlediska posouzení, zda dané (dvouslovné) spojení je tzv. kolokací. Tyto míry vybírá z několika ve světě používaných variant a také je z toho hlediska rigorózně empiricky vyhodnocuje (včetně podrobných testů signifikace) na čtyřech různých korpusech češtiny a švédštiny, a popisuje i (soutěžní) zpracování téže problematiky na dalších datech pro němčinu a češtinu.

Práce je rozdělena do sedmi kapitol. V úvodu autor popisuje základní termíny, motivaci, cíle a stanovená omezení pro vlastní disertační. Ve druhé kapitole se velmi podrobně zabývá pojmy z dané oblasti a ukazuje, že je třeba především najít vhodnou „definici" pro kolokace, aby práce měla solidní základy. Autor takovou definici vybral v sekci 2.1 a dále ji používá. V další části kapitoly 2 pak popisuje způsob tvorby evaluačních dat, která pro tento účel nebyla k dispozici a autor tedy musel tato data specifikovat a zorganizovat jejich anotaci. V kapitole 3 pak shrnuje známé míry statistické asociace používané pro (binární) lexikální asociaci (celkem 82 takových měr). V další kapitole popisuje použitá data (tj. průběh a výsledek manuální anotace pro evaluaci veškerých experimentů použitých v práci (s výjimkou soutěžních dat pro němčinu a češtinu použitých v experimentech popsaných v příloze A). Autor vytvořil(a) pomoci anotátorů celkem 4 korpusy kolokací: závislostní na základě PDT, povrchový na základě PDT a ČNK, a švédský na základě korpusu PAROLE (v tomto případě verbonominální konstrukce/kolokace). Pátá kapitola se věnuje popisu a vyhodnocení experimentů s uvedenými korpusy a mírami. Autor rovněž (vzhledem k relativně velké inherentní volnosti při vyhodnocování těchto měr) popisuje několik postupů při volbě měření a testů statistické významnosti (a jejich vizualizaci). V šesté kapitole pak popisuje navazující experimenty s kombinací asociativních měr a ukazuje, že tyto kombinace výsledky podstatně zlepší, a navrhuje a ověřuje zde i metodu minimalizace počtu kombinovaných metod při jen malé ztrátě přesnosti. V poslední číslované kapitole autor věcně shrnuje dosažené výsledky. V příloze A pak popisuje, jak jím dosažené výsledky byly úspěšně využity na soutěžních datech MWE 2008 Shared Task (publikováno dříve) na datech pro němčinu a češtinu. V příloze B pak systematicky ukazuje výsledky všech experimentů v tabulkách a grafech, včetně těch, které nebyly přímo uvedeny v textu práce.

Hodnocení:

Práce je psána velmi dobrou angličtinou. Je psána zcela jasně a na vysoké odborné úrovni. Jakkoli jsou jádrem práce kap. 4, 5 a 6, z hlediska zpracování vlastní disertační práce je třeba vysoko hodnotit text popisující výchozí podmínky v úvodu a dále pak úvodní odstavce k jednotlivých kapitolám, ať už z věcného nebo jazykového hlediska (a i z hlediska úplnosti a kvality referencí a citované literatury). Tyto části mají nejvyšší možnou knižní úroveň (a jakož i celá práce, stálo by za to o takovém typu publikace uvažovat, neboť takto shrnující a komplexní publikace o lexikálních asociativních mírách s jednotným hodnotícím přístupem neexistuje).

Věcně je třeba ocenit především velké množství asociativních měr použitých v práci, jdoucí vysoko nad obvyklé porovnání několika měr v řádu jednotek (připomeňme, že bylo použito 82 měr ve vlastním textu a 55 měr pro soutěž MWE 2008). Chybí mi zde pouze zmínka o tom, proč autor některé míry ze svých dřívějších publikací (kde jich bylo použito kulatých 100) vynechal.

Trvalým výsledkem této práce pak jsou matematické a algoritmické postupy, kterými je možno pro jednotlivé úlohy z oblasti NLP dojít k výběru vhodných lexikálních asociativních měr nebo (je-li to potřebné kvůli požadavku vyšší přesnosti) jejich efektivně stanoveným kombinacím. Dalším velmi užitečným výsledkem jsou pochopitelně i evaluační data, která autor se svými anotátory vytvořil. O kvalitě těchto dat svědčí mj. i to, že byla použita i v celosvětové soutěži. Je třeba se rovněž zmínit i o tom, že autor měl přijaté referáty o předběžných výsledcích této práce na nejprestižnějších světových konferencích tohoto oboru.

Práce je formálně zcela v pořádku, seznam citované literatury je velmi obsáhlý a zejména vzhledem k tématu úplný.

Předloženou disertaci Mgr. Pavla Peciny tedy hodnotím velmi vysoko a považuji za práci, která splňuje nároky kladené na disertační práci; doporučuji ji k obhajobě.

Praha, 22.8.2008