Oponentský posudek doktorské disertační práce

Mgr. Pavel Pecina: Lexical Association Measures: Collocation Extraction

Předložená disertační práce se zabývá mírami lexikální asociace (*lexical association measures*) a jejich využitím pro automatickou extrakci kolokací.
Hlavní přínos spočívá v podrobné analýze asociativních měr a jejich možných kombinací do komplexnějších modelů.

Obsah práce

Práce se skládá ze 7 kapitol. V úvodní kapitole autor představuje problematiku, formuluje úlohu práce, uvádí možné využití výsledků a zmiňuje předchozí práce, na které navazuje.
Následuje rozsáhlá kapitola věnovaná teoreteickým základům, je zde z různých hledisek posuzován pojem *kolokace* a definován úkol automatické extrakce kolokací. Ten je pak demonstrován na příkladech. Autor podrobně diskutuje přístupy uvedené v ostatních pracech zabývajících se týmž tématem. Autor také uvádí postup, který sám zvolil pro řešení vytyčeného úkolu.
Třetí kapitola popisuje zvolené asociativní míry použité v práci, autor představuje 82 měr, které rozděluje do skupin podle toho jestli jsou založeny na měření statistické asociace nebo na analýze kontextu.
Čtvrtá kapitola popisuje datové zdroje, které byly použity v experimentech. Autor pracoval se třemi korpusy - dvěma českými (Pražský závislostní korpus a Český národní korpus) a jedním švédským (korpus Parole). V případě Pražského závislostního korpusu provedl autor dva typy předzpracování - při jednom použil závislostní parsing, ve druhém se spolehl pouze na povrchový slovosled. Pro korpus Parole je definován modifikovaný úkol - extrakce *support verb construction*.
V páté kapitole autor podrobně evaluuje výsledky dosažené na zmíněných korpusech vůči ručně anotovaným datům. Pro porovnávání výsledků navrhuje a používá metriku *Average Precision*. Jsou uvedeny a vzájemně porovnány úspěšnosti jednotlivých asociačních měr. Na jednotlivých korpusech je pořadí jejich výkonosti různé, což autor v práci komentuje. Zde bych autorovi lehce vytkl, že je v práci uvedené malé množství příkladů získaných kolokací, které by čtenáři umožnili lepší představu o úspěšnosti metod.
Následující kapitola jde za hranice evaluace "po jednotlivých metrikách" a velmi zajímavě popisuje možnosti jejich vzájemné kombinace pomocí metod strojového učení z ručně anotovaných dat (supervised metodami). Autor aplikoval několik aktuálně používaných a úspěšných metod, např. logistickou regresi, Support Vector Machines a neuronové sítě. Dále je zmíněna redukce modelu, ve které se při minimální ztrátě kvality výrazně zjednoduší výpočetní model.
V závěrečné kapitole autor shrnuje dosažené výsledky.

Celá práce je psána anglicky, velmi srozumitelně a správně a je doplněna dvěma přílohami a seznamem literatury.

Přínos práce

Autor, svým vzděláním informatik, prokazuje hlubokou znalost statistických metod. Zvolené téma extrakce signifikatních kolokací je ve výpočení lingvistice velmi aktuální. V práci je dosaženo významných a originálních výsledků, které mohou mít dobré využití v řadě praktických aplikacích.
Práce také dává komplexní přehled evaluačních metrik využitelný i při řešení jiných problémů.
Speciálně bych ocenil důkladnou analýzu výsledků a jejich další zlepšování - autor zde přichází s velmi zajímavými způsoby dalšího vyladění zdánlivě vyřešené úlohy.
Ke kvalitě práce značně přispívá podrobná statistická analýza výsledků včetně např. testů signifikance rozdílů jednotlivých metod a visualizace výsledků. Statistická evaluace je ve výpočetní ligvistice běžná, ale autorova preciznost v tomto směru ještě více sbližuje aplikovanou linvistiku se statistickou.

Dotazy k obhajobě

Pro Pražský závislostní korpus dosahují míry používající pouze povrchový slovosled lepších výsledků než míry používající informace ze závislostního parsingu. To je pro mě překvapivé, zejména proto, že jsou zde kolokace definované jako syntakticky souvislé dvojice. Měl by pro to autor nějaké vysvětlení? Případně by mě zajímalo, jestli by bylo možné nějakým způsobem tyto informace pro účely rozhodování zkombinovat?
V kapitole 5 byly kolokace určovány unsupervised metodami. Zajímalo by mě, zda výrazné zlepšení při použití kombinace metrik, tak jak je popsáno v kapitole 6, přičítá autor právě zapojením informací pocházejících z ruční evaluace?
Je možné, že další dotazy ještě vyplynou v průběhu osobní prezentace a následné diskuse.

Závěr

Předkládaná disertační práce prokazuje hluboký vhled do problematiky a autor prokázal, že je schopen samostatně vědecky pracovat a řešit složité problémy. Doporučuji, aby byla práce přijata jako disertační a aby Matematicko-fyzikální fakulta Univerzity Karlovy v Praze udělila Pavlu Pecinovi titul Ph.D.

Praha 20. 8. 2008
RNDr. Jiří Semecký Ph.D.
Google
Krakow, Poland