

Posudek vedoucího diplomové práce

Klasifikátor pro sémantické vzory užívání anglických sloves

Autor: Bc. Vincent Kríž

Matematicko-fyzikální fakulta Univerzity Karlovy, Praha 2012

Hlavním tématem předložené diplomové práce je aplikace strojového učení v oblasti lexikální sémantiky anglických sloves. Diplomová práce měla za úkol analyzovat a co nejlépe využít dostupná data pro konstrukci automatického klasifikátoru sémantiky anglických sloves. Cílem práce bylo navrhnout, implementovat a empiricky evaluovat klasifikátor pro rozpoznání sémantických vzorů užívání anglických sloves klasifikovaných pomocí nové lexikografické metody *Corpus Pattern Analysis* (CPA) s využitím automatického parsingu angličtiny a metod strojového učení.

Jako datového zdroje autor využívá několika desítek (přesně 30) anglických sloves zpracovaných metodou CPA, což je důsledně korpusová, empirická metoda, která analyzuje typické vzory užívání slov v jazykovém korpusu a popisuje význam sloves pomocí kontextových preferencí definovaných jak syntakticky, tak sémanticky (pomocí specifikace sémantického typu typických kolokátů). Kombinací sémantických typů a syntaxe metoda CPA originálním způsobem konkuruje tradičním přístupům k formálnímu popisu a k automatickému rozpoznávání významu sloves. Autor práce se zaměřuje na *supervised* metody strojového učení, přičemž rozsah dostupných trénovacích dat je několik stovek (zpravidla 250-500) instancí pro každé sloveso.

Předložená práce je velmi dobře formálně zpracována a má standardní strukturu. Ačkoliv je práce poměrně rozsáhlá (cca 90 stran text + 40 stran dodatky), je zpracována velmi pečlivě. Autor se zpravidla vyjadřuje zcela jasně a přesně a téměř bez gramatických chyb. Práce nejprve obsahuje úvod do problematiky a popis slovníku *Pattern Dictionary of English Verbs* zpracovaného pomocí metody CPA (kap. 1 a 2), dále rešerši z literatury týkající se standardních metod strojového učení pro automatickou klasifikaci (kap. 3) a známých přístupů k rozlišování a automatickému rozpoznávání sémantiky sloves (kap. 4) a analýzou dostupných dat (kap. 5). Další tři kapitoly jsou již věnovány výslovně aplikaci metod strojového učení. V kap. 6 se autor věnuje výběru atributů pro strojové učení, v kap. 7 přípravě konkrétních dat pro danou úlohu a v kap. 8 jsou popsány konkrétní metody trénování, optimalizace a evaluace automatického klasifikátoru. V závěrečné kapitole jsou stručně shrnuty nejdůležitější výsledky.

Předložená práce dále obsahuje značně rozsáhlý seznam referencí na použité odborné zdroje a přílohu na CD s programy a daty, které autor vytvořil nebo při vlastní práci použil (stručný popis těchto příloh lze nalézt v dodatku E). V dodatcích jsou velmi přehledně uvedeny nejdůležitější výsledky experimentů.

Po věcné stránce nemá předložená práce žádné závažné nedostatky. Nalezl jsem pouze velmi malé množství drobných chyb nebo nejasností. Jakožto vedoucí práce jsem s předloženým výsledkem velmi spokojen a mohu říci, že ve všech částech předložená práce splňuje má očekávání.

Hlavní přínos práce je ve velkém množství pečlivě a seriózně zpracovaných experimentů včetně *feature selection* a evaluace a v tom, že autor se jako první pokusil o statistické učení na základě dat zpracovaných pomocí CPA. Mimo jiné jde také o první pokud využít Bickovy sémantické prototypy ke zlepšení klasifikace sloves pomocí sémantiky jejich argumentů.

Za zmínku také stojí fakt, že některé výsledky resp. části předložené práce byly využity pro přípravu odborných publikací (viz uvedené reference). Autor diplomové práce je spoluautorem příspěvků publikovaných (resp. k publikaci přijatých) do odborných mezinárodních konferencí EACL '2012 a EURALEX '2012.

Závěr

Předložená práce adekvátně zasahuje všechny cíle vytyčené v jejím zadání a svou formou, rozsahem, náročností a kvalitou více než stoprocentně vyhovuje požadavkům kladeným na diplomovou práci. Autor jednoznačně splnil uložené cíle. Práci hodnotím jako naprůměrně kvalitní a doporučuji ji k obhajobě. Navrhují předloženou práci uznat jako diplomovou.

V Praze, 9. května 2012

RNDr. Martin Holub, Ph.D.