

Hodnocení vedoucího diplomové práce

Autor práce:

Bc. Michal Tušíl

Název práce:

Vícejazyčná sémantická podobnost textů

Obsah práce:

Práce se zabývá strojovým učením bez supervize pro porovnávání významu dvou vět napříč různými jazyky, a to bez použití přímého překladu do stejného jazyka. Diplomant zkoumá metody Distribuční sémantiky pro vytvoření sémantických prostorů, které jsou následně transformovány do sdíleného bilinguálního prostoru pomocí lineárních transformací.

Automatický odhad sémantické podobnosti je porovnáván s člověkem a je měřena Pearsonova a Spearmanova korelace. Testy jsou prováděny na datasetech z konference SemEval mezi větami v angličtině na jedné straně a španělštině, turečtině, italštině a chorvatštině na straně druhé.

Kvalita řešení a dosažené výsledky:

Hlavním přínosem práce je nový přístup kroslinguální sémantické podobnosti vět, založený na lineární transformaci vektorové reprezentace významu věty do sdíleného prostoru. Transformace je trénována na paralelních korpusech a dosahuje slibných výsledků na datasetech v různých jazycích z konference SemEval.

Téma práce, způsob řešení a množství experimentů přesahuje úroveň diplomové práce a spíše se blíží práci rigorózní. Diplomant musel proniknout do teoreticky poměrně náročné oblasti Distribuční sémantiky, strojového učení bez supervize pro reprezentaci významu věty a lineárních transformací pro bilinguální sémantické prostory.

Spolupráce s vedoucím a aktivita studenta:

Student přistupoval k řešení velmi svědomitě a aktivně. Pravidelně konzultoval svojí práci s vedoucím a dodržoval stanovené termíny. V tomto směru byla spolupráce bezproblémová.

Formální úroveň práce:

Diplomová práce se skládá z 54 stran vlastního textu (63 stran včetně úvodních stránek a referencí). Formální úroveň práce je bezproblémová. Struktura textu je dobře navržena. Matematický popis je srozumitelný a správný. Z experimentů je patrné, co a jak bylo uděláno. Autor dodržuje zažité typografické konvence. Práce je vysázena v systému LaTeX. Autor používá vektorovou grafiku. Vyjadřování v českém jazyce je odpovídající. Diplomantova práce s literaturou je na úrovni. Diplomant cituje významné konference a časopisy v dané oblasti.

Úroveň kódu:

Převážná část kódu, zaměřující se na bilinguální transformace a na sémantickou reprezentaci věty, je napsána v jazyce Java (pro trénování monolingálních sémantických prostorů diplomant používá dostupné implementace v jazyce Python). Odevzdaný kód je funkční a dle přiloženého návodu lze s trochou snahy vše spustit a změřit. Autor dodržuje konvence pro psaní kódu v jazyce Java. Kód je srozumitelný a dostatečně komentovaný. Autor používá Maven jako buildovací nástroj.

Splnění zadání:

Práce zcela splňuje zadání, a to nad rámec kritérií běžných pro diplomové práce. Navrhuji hodnocení známkou **výborně** a práci doporučuji k obhajobě.

V Plzni 1.6.2019

Ing. Tomáš Brychcín, Ph.D.