# Hodnocení vedoucího diplomové práce

**Autor práce:**

Bc. Adam Mištera

**Název práce:**

Mezi-jazyčné transformace sémantických prostorů

**Obsah práce:**

Práce se zabývá projekcemi sémantických prostorů za účelem vytvoření sdíleného jazykově nezávislého prostoru. Sémantické prostory se za poslední dvě dekády projevily jako velmi užitečný nástroj při zpracování přirozeného jazyka (NLP). Vektorová reprezentace významu (bod v sémantickém prostoru) pomáhá bojovat s problémem řídkosti dat tím, že zavádí míru podobnosti slov a frází.

Výzkum v NLP oblasti se v posledních letech soustředí na reprezentace významu, které jsou nezávislé na jazyku. Jedním z lákavých směrů, pro dosažení tohoto cíle, začíná vytvořením sémantických prostorů separátně pro každý jazyk zvlášť a následuje projekcí těchto prostorů do jednoho univerzálního. Právě nezávislost trénování těchto prostorů je klíčovou výhodou, protože lze kdykoliv přidat další jazyk.

Diplomová práce zkoumá lineární a nelineární transformace sémantických prostorů. Kvalita vícejazyčných sémantických prostorů je měřena na datasetech slovních podobností, slovních analogií a strojovém překladu, a to na šesti jazycích z třech různých jazykových rodin. Mimo jiné je měřena tzv. *hubness*, jakožto zástupce problémů vznikajících v prostorech s vysokou dimenzí (občas označováno jako *prokletí dimenzionality*).

**Kvalita řešení a dosažené výsledky:**

Student se v první řadě soustředil na trénování lineárních transformací dle nejlepších známých postupů. Následně zkoumal použití neuronových sítí pro zavedení nelinearity s cílem překonat lineární transformace.

Při dodržení správného postupu pro trénování obyčejných lineární transformací se tento úkol však projevil jako velmi obtížný. Nelinearita sice dosahovala podobných výsledků na datasetu slovních podobností, současně ale vedla k nevhodnému pokřivení prostoru, k výrazné navýšení hubness a k výraznému poklesu úspěšností na slovních analogií či strojovém překladu. Tento problém byl následně odstraněn využitím hinge loss s negativním samplováním pro cenovou funkci, která přímo optimalizuje hubness. Výsledný prostor konzistentně překonal všechny lineární transformace na všech testovacích úlohách a na všech jazycích.

V následujících experimentech se diplomant věnoval zavedení lokální informace do mezi-jazyčných transformací. Sémantický prostor byl nejdříve rozdělen do segmentů pomocí shlukování. V různých segmentech byly následně použity jiné transformace optimalizované pro daný segment. Shlukování sice v průměru nepřekonalo neuronovou síť s hinge loss, překonalo ale všechny lineární transformace a dosáhlo nejlepších výsledků v kategorii strojového překladu.

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 pro redukci dimenzionality a neuronových sítí.

Hinge loss je sice v teorii mezi-jazyčných transformací známá, pouze ale ve spojení s lineárními transformacemi. Využití shlukování pro zavedení lokální informace při transformacích, také dosud nebylo publikováno. Dle mého názoru tvoří diplomová práce solidní základ pro následující vědeckou činnost.

**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 60 stran vlastního textu (80 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 podpůrného kódu je napsána v jazyce Java. Pro trénování neuronových sítí je použit Python. Odevzdaný kód je funkční a dle přiloženého návodu lze všechno s trochou snahy spustit a změřit. Autor dodržuje zažité konvence pro psaní kódu. 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 7.6.2021

Ing. Tomáš Brychcín, Ph.D.