Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Bc. Jan Bodnár		
Název práce	Morphological Segmentation in Czech using Word-Formation Network		
Rok odevzdání	2020		
Studijní program	Informatika	**Studijní obor**	Umělá inteligence

Autor posudku	doc. Ing. Zdeněk Žabokrtský, Ph.D.	**Role**	Vedoucí
Pracoviště	Ústav formální a aplikované lingvistiky, MFF UK		

Text posudku:

Cílem předložené práce bylo prozkoumat možnosti, jak automaticky rozsegmentovat lemmata českých slov na morfémy v situaci, kdy jsou k dispozici velmi omezená ručně značkovaná data pro tuto úlohu, ale naopak lze využít bohatou databázi slovotvorných relací.

Obsah práce: Práce se skládá z pěti kapitol. Po krátkém úvodu následuje kapitola s přehledem několika vybraných existujících přístupů k segmentaci slov a kapitola popisující základní relevantní pojmy z oblasti morfologie a z oblasti neuronových sítí. Třetí kapitola popisuju strukturu implementovaného řešení, ve kterém se kombinuje několik komponent založených na pravidlech s komponentami založenými na strojovém učení. Čtvrtá kapitola empiricky vyhodnocuje dosažené experimentální výsledky, pátá kapitola práci uzavírá. Práce je psána anglicky, včetně seznamu literatury a příloh má práce 58 stran.

Hodnocení práce:

- Silné stránky: Za hlavní přínos práce považuji velmi solidní experimentální výsledek, který pro češtinu překonává standardní baseline i dříve publikované výsledky reprezentující state-of-the-art, a to přesto, že autor neabsolvoval předměty programu Matematická lingvistika. Současná verze systému je výsledkem poměrně dlouhé řady dílčích experimentů a následných vylepšení, během kterých student získal solidní empirický i kvalitativní vhled do problematiky (jak potvrzují i pasáže s analýzou chyb). Hlavní přínost práce byl představen v publikaci na konferenci TSD2000, vygenerované segmentace budou zveřejněny v dalším vydání slovotvorné sítě DeriNet a již v současnosti je možné využít je pro nový způsob vyhledávání v Českém národním korpusu. Zajímavým vedlejším produktem práce je také nová metoda detekce pravděpodobných chyb v síti DeriNet.

- Slabé stránky: Na první pohled je bohužel patrné, že samotný text práce vznikal pod časovým tlakem a úroveň formy neodpovídá úrovni dosaženého experimentálního přínosu. Text je psaný sice srozumitelnou a poměrně plynulou, ale ne zcela bezchybnou angličtinou (ve finálním textu zůstalo například nemálo překlepů a chyb ve členech nebo interpunkci). Ještě více ovšem pokulhává sazba. Pozitivně hodnotím, že autor byl ochoten začít si osvojovat principy práce v systému LaTeX, nicméně v textu lze nalézt pasáže, které zjevně neprošly ani letmou vizuální kontrolou. Pokud jde o obsahovou stránku, některé části práce považuji za příliš stručné až zkratkovité, neumožňují tak nahlédnout komplexnost řešení (a v některých případech ani komplexnost vstupních dat). Pokud jde o použitelnost

výsledného softwaru, obecně by samozřejmě z hlediska údržby byla vhodnější "monolitičtější" architektura, u které není nutné při přetrénovávání přenastavovat parametry a hyperparametry několika různých komponent. Zde si ale dovolím připomenout, že cílem byl spíše experiment posouvající state-of-the-art, tzn. vyšší prioritu než softwarová udržitelnost dostalo hledání metod zvyšujících úspěšnost.

Otázka k obhajobě:
V tabulce 4.4 úspěšnost segmentace na celých slovech zjevně nekoreluje s úspěšností měřenou na jednotlivých morfémech nebo na hranicích morfémů. Mohu poprosit o vysvětlení?

Závěr:
Předložená práce představuje výrazný pozitivní experimentální výsledek. I přes vyjmenované výhrady k formální stránce textu práci doporučuji k obhajobě.

Domnívám se ale, že by bylo vhodné, aby autor jako budoucí doktorand všechny nedostatky pečlivě opravil a text zveřejnil znovu například jako technickou zprávu.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Datum 7. 9. 2020 **Podpis**