POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	**Named Entity Recognition in Czech**
Jméno autora:	**Radek Štulc**
Typ práce:	bakalářská
Fakulta/ústav:	Fakulta elektrotechnická (FEL)
Katedra/ústav:	Katedra kybernetiky
Oponent práce:	Ing. Radek Mařík, CSc.
Pracoviště oponenta práce:	Katedra telekomunikační techniky

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání **průměrně náročné**

Hodnocení náročnosti zadání závěrečné práce.

Jednoduchá verze splnění zadání patří mezi základní úlohy zpracování přirozeného jazyka (NLP). Zaměření se na český jazyk je sice specialitou, která byla ještě 5 lety před velkou výzvou, nicméně dnes již existují jazykové modely zaměřené na češtinu a i NER datové sady v češtině.

Splnění zadání **splněno s menšími výhradami**

Posuďte, zda předložená závěrečná práce splňuje zadání. V komentáři případně uveďte body zadání, které nebyly zcela splněny, nebo zda je práce oproti zadání rozšířena. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.

Jednotlivé body zadání jsou všechny alespoň z části předloženou prací pokryty. Ačkoliv NER je stále velmi aktivní oblast NLP, práce uvádí pouze 2 reference pro NER v angličtině a 1 referenci pro češtinu, všechny z let 2019-2021. Chybí širší přehled SOTA technik posledních 3 let, který by pokryl například i techniky využívající velké jazykové metody (LLM), vzhledem k tomu, že NLP metody se v posledních 3 letech rapidně vyvíjejí se změnami pozorovatelnými i v rámci jednoho či dvou měsíců. Práce obsahuje porovnání starších jazykových modelů, ale architektury různých řešení pro NER porovnány explicitně nejsou.

Zvolený postup řešení **správný**

Posuďte, zda student zvolil správný postup nebo metody řešení.

Předložená práce prezentuje náhledy do řešení, které se váže k SOTA metodám z roku 2021. Student věnoval poměrně velkou část své práce konverzi datové sady do vhodného formátu, ale v práci zřejmě úplně chybí explicitní popis jeho řešení NER, ačkoliv si jistou představu o architektuře čtenář může udělat z formátů vstupních dat, zmínek v sekci 4.2 věnované parametrům trénovací smyčky. Některé kroky obsahují deklarace, že se něco nezdařilo realizovat, např. v sekci 4.2.2., avšak bez bližšího popisu tohoto selhání. Opomenu-li, že v práci chybí adekvátní rešerše NER metod a popis testované architektury řešení, pak zbývající kroky prezentovaného postupu smysl dávají a zkoušené řešení považuji za rozumné na úrovni bakalářské práce.

Odborná úroveň **C - dobře**

Posuďte úroveň odbornosti závěrečné práce, využití znalostí získaných studiem a z odborné literatury, využití podkladů a dat získaných z praxe.

Principy metod NLP jsou poměrně velmi povrchně popsány přímo v úvodu práce. Dále jsou v práci uvedeny základní znaky jazykových modelu let 2019-2021 s větším zaměřením na podporu českého jazyka. Student si s nadhledem všímá kritických konfiguračních parametrů jejich trénování a použití. Podle některých nepřesných vyjádření, např. pro „attention mechanism“ na str. 3, „benefits“ na str. 20, a dalších vágních výrazů v sekcích 4.1.2 a 4.2, lze usuzovat na to, že student pochopil hlavní principy technik NER, ale chybí mu detailnější znalosti potřebné k přesným úvahám. Očividně se vyhýbá definicím použité terminologie, např. v celé práci nejsou definovány termíny jako „named entity“, na str. 24 nerozlišuje mezi „multi-class“ a „multi-label klasifikací“ a k nim příslušným aktivačním funkcím a optimalizačním kritériím. Vzhledem k tomu, že chybí přesné vymezení vstupů jednotlivých bloků výpočetního postupu, lze při prvním čtení stěží odhadovat, co autor někdy zamýšlí, např. výrazem „the number of labels is doubled“, str. 20. Není zřejmé, kdy se co přesně mění, např. v tabulce 4.2 mají „tokeny“ té samé entity jiné „labels“ (15 vs. 16). Popis některých kroků je v některých případech zřejmě kauzálně

opačně, než jak odpovídá realitě, viz „The cross-entropy loss was replaced with BCEWithLogitsLoss3 and to fit this loss, I need to one-hot encode my labels.“ na str. 22.
Práce obsahuje i řadu spekulací, které by zřejmě bylo vhodně přesunout do kapitoly s diskusí. Nicméně je potřeba zdůraznit, že se jedná o bakalářskou práci studenta, který si problematiku NLP musel dostudovat sám.

Formální a jazyková úroveň, rozsah práce **C - dobře**

Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku.

Práce je napsána v anglickém jazyku, který se až na několik ne-anglicky vyjádřených vět dají poměrně bez problémů číst. Např. interpretaci vět popisující Perl skripty v sekci 4.1.1, či věty „All versions were organized in their files“ na str. 19 lze pouze tušit. Chybějící členy či množná čísla zde uvádět nebudu. Rovněž student ne vždy korektně používá budoucí či minulý čas. Obvykle anglické věty nezačínají předložkami, např. 4.1.3. Rozsahem a typografií je práce jinak v pořádku.

Výběr zdrojů, korektnost citací **C - dobře**

Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení závěrečné práce. Charakterizujte výběr pramenů. Posuďte, zda student využil všechny relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.

Práce obsahuje korektní citace na více než 30 referencí. Škoda, že nejsou více zaměřené na problematiku NER.

Další komentáře a hodnocení

Vyjádřete se k úrovni dosažených hlavních výsledků závěrečné práce, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, publikačním výstupům, experimentální zručnosti apod.

I přes řadu mých negativních poznámek k práci, dosažená přesnost výsledků je velmi zajímavá. Trochu zarážející je nesoulad mezi přesností publikovanou v článcích a tou, kterou dosáhl student pro model Czert-B (tab. 4.4) .

III. CELKOVÉ HODNOCENÍ, OTÁZKY K OBHAJOBĚ, NÁVRH KLASIFIKACE

Shrňte aspekty závěrečné práce, které nejvíce ovlivnily Vaše celkové hodnocení. Uveďte případné otázky, které by měl student zodpovědět při obhajobě závěrečné práce před komisí.

Práce studenta poskytuje velmi zajímavé výsledky pro NER úlohu v českém jazyku. Nicméně, srovnání možných architektur řešení a popis použitého postupu má k dokonalosti daleko. Navrhuji následující otázky do diskuse:

1) Jak si student představuje použít Heavisideovu funkci pro tzv. „multi-label classification“? Jak při použití takové funkce probíhá automatické derivování při trénování neuronové sítě?
2) Jak student zajistil použití těch samých hodnotících funkcí při přebírání výsledků z publikací?
3) Mohl by při prezentaci předložit schéma implementovaného výpočetního postupu?

Předloženou závěrečnou práci hodnotím klasifikačním stupněm **C - dobře.**

Datum: 3.6.2024 Podpis: