Posudek bakalářské práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Václav Hrouda
Název práce	Automatizace generování popisů produktů pomocí neuronových jazykových modelů
Rok odevzdání	2024
Studijní program	Informatika
Specializace	Databáze a web

Autor posudku	Ing. Zdeněk Kasner	Vedoucí
Pracoviště	Ústav formální a aplikované lingvistiky	

K celé práci

	lepší	OK	horší	nevyhovuje
Obtížnost zadání		X		
Splnění zadání	X	X		
Rozsah práce *… textová i implementační část, zohlednění náročnosti*		X		

Práce se zabývá automatickým generováním textových popisů produktů ze strukturovaných detailů o produktu. Tento přístup není běžně používaný v praxi a začíná být zajímavým až s příchodem jazykových modelů založených na architektuře Transformer, které dokážou generovat přirozeně znějící text na základě zadaného vstupu. Problémem modelů ale nadále zůstává přesnost výstupů a výpočetní náročnost. Cílem práce bylo tedy prozkoumat možnosti tohoto přístupu a vhodnost jeho nasazení. Specifikem této práce navíc je, že se pracuje s daty v češtině, na kterou existuje oproti angličtině daleko méně vhodných trénovacích datových sad a předtrénovaných modelů.

Původním záměrem bylo zaměřit se výhradně na přístupy spojené s dotrénováním menších modelů předtrénovaných pro češtinu. Pro tento přístup hrála dostupnost většího množství trénovacích dat z reálného e-shopu a zároveň neexistence větších jazykových modelů se schopností generovat český text. S příchodem velkých jazykových modelů jsme se se studentem rozhodli porovnat ještě dva další přístupy: využití volně dostupného modelu Mistral v kombinaci s překladovým modelem a využití komerčního modelu GPT-3.5 (ChatGPT).

Jádrem práce zůstává dotrénování jazykového modelu pro češtinu. Student získal přístup na univerzitní cluster AIC, kde s pomocí grafických karet provedl řadu experimentů s modelem `gpt2-small-czech-cs`. Pozitivně hodnotím, že student si pro napojení na trénovací infrastukturu připravil webovou aplikaci, která pokryla vše od monitorování experimentů až po závěrečné vyhodnocení výstupů z modelů. Student také zajistil větší množství trénovacích dat (~50k příkladů) z reálného e-shopu a jejich předzpracování pro dotrénování modelu. Více času mohlo být naopak věnováno hledání vhodných hyperparametrů modelů, případně vyzkoušení alternativních menších modelů.

Student dále získal skrz API přístup k velkému jazykovému modelu Mistral a překladovému modelu CUBBITT (obojí běžícímu lokálně na infrastruktuře LINDAT), a k modelu GPT-3.5 přes OpenAI API. Tyto modely již netrénoval, pouze využil instrukce a vstupní data k vygenerování výstupů. K vyhodnocení kvality výstupů student použil automatické metriky založené na lexikální podobnosti v kombinaci s hodnocením pomocí lidských anotátorů. Z vyhodnocení vyplývá, že výstupy z velkých jazykových modelů (i bez využití trénovacích dat) jsou mnohem kvalitnější, než výstupy z `gpt2-small-czech-cs`. Ani jejich výstupy ale nejsou bezchybné a proti nasazení jazykových modelů v praxi navíc nadále mluví jejich výpočetní náročnost, případně nutnost platit za komerční API.

(pokračování na další straně)

Se spoluprácí se studentem jsem spokojený. Problematický byl pouze časový horizont, ve kterém práce vznikala: zatímco zadání vzniklo před rokem, převážná část práce začla vznikat tři měsíce zpátky, což vedlo k menším kompromisům ohledně experimentů a lehce nevyvážené textové části práce. Student ovšem na projektu pracoval intenzivně a prezentoval pokrok každý týden na osobních schůzkách. Zadání považuji celkově za splněné: v práci sice chybí menší vícejazyčné modely a rozsáhlejší diskuze vhodnosti přístupů, nad rámec původního zadání ale práce obsahuje porovnání dotrénovaných modelů s velkými jazykovými modely, které považuji v kontextu nedávného vývoje v oblasti generování přirozeného jazyka za cennější.

Textová část práce

	lepší	OK	horší	nevyhovuje
Formální úprava *… jazyková úroveň, typografická úroveň, citace*	X	X		
Struktura textu *… kontext, cíle, analýza, návrh, vyhodnocení, úroveň detailu*		X	X	
Analýza		X		
Vývojová dokumentace		X		
Uživatelská dokumentace		X		

Práce je psaná v češtině a její hlavní část má 55 stran. Jazykově i stylisticky je práce až na drobné detaily v pořádku. Text doplňují diagramy, screenshoty aplikace, grafy s výsledky a ukázky kódu. Výsledky by ovšem mohly být prezentovány v kompaktnější a přehlednější podobě – grafy zabírají vícero samostatných stran a nejsou k dispozici v podobě tabulek přímo u textu.

V práci jsou obsaženy všechny potřebné sekce od teorie, přes návrh experimentů, jejich implementaci a vyhodnocení. Obsah sekcí je ovšem nevyvážený: zatímco teorie neuronových modelů je zpracována důkladně, konkrétnímu zaměření práce (generování popisků produktů a českým modelům) je věnováno méně pozornosti. Velkou část návrhu experimentů tvoří popis webové aplikace a přezpracování dat, zato pro experimenty je popsán pouze jeden konkrétní běh. Stejně tak je stručnější i diskuze výsledků experimentů (chybí např. konkrétní ukázky výstupů) a jejich dopadů na zkoumané téma. Práce je ovšem v rámci možností kompletní a obsahuje (v kombinaci s kódem) dostatek informací pro replikaci experimentů.

Implementační část práce

	lepší	OK	horší	nevyhovuje
Kvalita návrhu *… architektura, struktury a algoritmy, použité technologie*	X			
Kvalita zpracování *… jmenné konvence, formátování, komentáře, testování*	X			
Stabilita implementace	X			

Student se implementační části práce věnoval důkladně a nad rámec povinností. Kód je obsažený v příloze práce. Základem je webová aplikace založená na webovém frameworku Flask, kterou student vytvořil a používal během práce. Aplikace umožňuje spravovat přidělené prostředky na výpočetním clusteru, předzpracovávat data a ukládat je do SQL databáze, spouštět a monitorovat trénování modelu pomocí knihovny `transformers` a evaluovat modely pomocí automatických metrik a lidských anotátorů. V externí příloze jsou dále přiložené výstupy z modelů, výstupy evaluačních metrik, schéma databáze, a dotrénovaný model použitý pro experimenty. Implementační část je z mého pohledu kompletní a nemám k ní výhrady.

Celkové hodnocení Výborně (-)

Práci navrhuji na zvláštní ocenění Ne

Datum 24.1.2024 Podpis