Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	František Trebuňa		
Název práce	Persona-Aware Chatbot Response Generation		
Rok odevzdání	2024		
Studijní program	Informatika	**Studijní obor**	Umělá inteligence

Autor posudku	Ondřej Dušek	**Role**	vedoucí
Pracoviště	Ústav formální a aplikované lingvistiky		

Text posudku:

Téma práce Tématem diplomové práce Františka Trebuni je generování odpovědí chatbota pro chitchat (zdvořilostní společenskou konverzaci) s konzistentní osobností. To představuje důležitý problém už od vzniku prvních chatbotů založených na generativních neuronových modelech (dříve LSTM, tj. cca od r. 2015): neuronové chatboty obecně mívají velké problémy s nekonzistencemi v odpovědích a na opakované nebo synonymní otázky si často protiřečí. Přitom konzistence odpovědí a vytváření jednotlié osobnosti je pro chitchat jeden z hlavních požadavků, které by každý chatbot měl splňovat. Tato situace taky vedla k vytvoření benchmarku PersonaChat a jeho varianty ConvAI2, která představuje zároveň nejdůležitější trénovací datovou sadu zaměřenou specificky na tento problém. Jedná se o přepisy chatů dvou osob, z nichž každá má předepsanou "personu", tj. krátkou biografii se základními vlastnostmi (rodina, povolání, koníčky apod.), kterými se během konverzace musí řídit. V podstatě výsledné generování odpovědí odpovídá dnes populárnímu přístupu retrieval-augmented generation, tj. generování na základě informací dodaných ve vstupním kontextu modelu.

Autor v rámci svojí diplomové práce staví nové modely pro chitchat na benchmarku PersonaChat/ConvAI2. Podle zadání práce problém řeší finetunovaním "menších" předtrénovaných jazykových modelů (zejména GPT-2), úpravami jejich architektur i vylepšením trénovacích režimů. Nad rámec původního plánu všechny tyto experimenty porovnává i s posledními velkými jazykovými modely (LLM), konkrétně GPT-3.5.

Shrnutí obsahu Autor nejprve provedl podrobnou analýzu samotného datasetu ConvAI2 a pak ve svých experimentech začal s finetunováním modelů GPT-2, které v době začátku experimentů (2022) představovaly nejrozšířenější jazykový model používaný pro podobné účely. K základnímu modelu postupně přidává jednotlivá vylepšení:

- Vyladění hyperparametrů a použití beam searche, včetně jeho sofistikovanějších variant (diverse beam search, beam search with sampling)
- Řazení několika variant výstupu (ranking) podle modelu natrénovaného s pomocí multi-task training pro generování i výběr následující odpovědi
- Specifické trénování modelu pro řazení variant odpovědí, vč. přístupu LambdaRank
- Přímé trénování modelu podle preferencí (direct preference optimization), což je nový přístup vyvinutý a používaný hlavně pro trénování LLM.

K tomu práce obsahuje i promptování modelu GPT-3.5, včetně několika kroků prompt engineeringu, specificky zaměřeného na vylepšení automatických metrik na datasetu ConvAI2.

Pro evaluaci autor používá zejména F1 skóre na slovech vzhledem k referenční odpovědi, což je hlavní metrika benchmarku ConvAI2 od jeho vzniku. Tu doplňuje o další vhodné automatické metriky zaměřené na jiné kvality (různorodost, plynulost). Zároveň ale přidává manuální evaluaci, kde se zaměřuje na párové porovnání (výběr preferované odpovědi, s možností nerozhodných výsledků).

Výsledky práce ukazují, že ač autor dosáhl svými vylepšeními menších modelů nejlepších dosavadních výsledků vzhledem k F1 skóre, podle manuální evaluace vychází jednoznačně nejlépe promptování GPT-3.5. To ukazuje na limity datasetu ConvAI2 a metriky F1, které autor částečné ukázal již v úvodní analýze. GPT-3.5 sice dopadá podle F1 vyrazně hůř, ale vzhledem k manuálním evaluacím je preferován i oproti referenčním odpovědím ze samotného datasetu. Dalo by se říct, že GPT-3.5 v podstatě problém s konzistencí odpovědí řeší zcela.

Práce tak jednak posouvá nejlepší dosažené výsledky na daném benchmarku a zároveň si je plně vědoma nedostatků tohoto benchmarku. Přitom prezentuje neotřelé přístupy pro generování odpovědí v chatbotech – zejména použití LambdaRank dopadá velmi dobře, podobně i přístup direct preference optimization, který je (aspoň pokud vím) tímto způsobem u menších modelů použit vůbec poprvé.

Struktura textu Text diplomové práce je rozdělen na pět číslovaných kapitol, plus nečíslovaný úvod a závěr. Krátký úvod motivuje práci a uvádí do problému v ní řešeného. 1. kapitola velmi obšírně popisuje celý potřebný teoretický základ pro experimenty – neuronové reprezentace, jazykové modely i řazení odpovědí. Kapitola 2 obsahuje popis datasetu ConvAI2, včetně autorovy úvodní analýzy a popisu zjištených problémů. 3. kapitola popisuje metody evaluace používané v experimentech. Kapitola 4 pak shrnuje nejdůležitější předchozí práce na benchmarku ConvAI2, které inspirovaly autorovy vlastní experimenty. 5. kapitola pak představuje hlavní popis všech experimentů v práci, včetně evaluace a finální diskuse. Závěr krátce shrnuje všechny výsledky, vztahuje je k otázkám kladeným v úvodu a představuje nápady na potenciální rozšíření do budoucna.

Text práce je psán velmi kvalitní, dobře srozumitelnou angličtinou. Struktura práce je dobře rozvržená, jednotlivé části na sebe přirozeně navazují a text se dobře čte.

Průběh prací Autor na experimentech souvisejících s tímto tématem pracoval cca od r. 2021 s několika přerušeními, cca od listopadu 2023 se velmi intenzivně věnoval experimentům popsaným ve finální verzi textu. Mezitím vznikl i vedlejší produkt práce, knihovna VisuaLLM pro vizualizace generování z jazykových modelů, kterou autor prezentoval jako článek na prestižní konferenci INLG 2023 v Praze. Veškeré experimenty i psaní textu jsme s autorem probírali velmi intenzivně na pravidelných online schůzkách. Během celého procesu autor ukázal schopnost výborné práce s literaturou i s daty, implementace složitých modelů i korektní evaluace; bylo vidět skutečné zaměření na kvalitu výsledku, kritické myšlení a invence při řešení problémů.

Celý text práce jsme společně podrobně prošli, některé části autor přepracovával a procházeli jsme je i vícekrát. Přitom jsme vyřešili veškeré mé připomínky, takže s výsledným textem jsem naprosto spokojen jak po obsahové, tak po formální stránce.

Celkové hodnocení Celkově práci jednoznačně doporučuji k obhajobě; nemám žádné dotazy.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 1. 9. 2024

Podpis: