Posudek vedoucího diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Bc. Jaroslav Šafář
Název práce	Practical neural dialogue management using pretrained language models
Rok odevzdání	2023
Studijní program	Informatika **Studijní obor** Umělá inteligence

Autor posudku	doc. RNDr. Ondřej Bojar, Ph.D. **Role** oponent
Pracoviště	Ústav formální a aplikované lingvistiky

Text posudku:

Diplomová práce Jaroslava Šafáře studuje velmi aktuální otázku, jak využít předtrénované jazykové modely v dialogových systémech pro interpretaci uživatelova záměru a sdělení. Cílem je především těžit z robustnosti těchto modelů a autor zmiňuje i známý problém tzv. halucinací, kdy model generuje fakticky nesprávné výstupy.

Práce je přehledně členěna do šesti kapitol, zabírá 87 tištěných stran včetně 27 stran bibliografie, seznamů a příloh. Po teoretickém úvodu v kapitole 1 následuje kapitola 2 s přehledem historických a moderních modelů s důrazem na aspekty podstatné pro tuto práci (sledování stavu dialogu, volba akce a end-to-end systémy). Těžištěm je třetí kapitola s vlastním návrhem vlastních modelů, čtvrtá kapitola pak tyto modely empiricky testuje a získané výsledky jsou prodiskutovány v kapitole 5. Šestá kapitola přináší shrnující závěr a výhled do budoucna.

Teoretický úvod je dostatečně přesný a podrobný, zavádí i základní koncepty hlubokého učení. Úvod do správy dialogu a existujících modelů je pro čtenáře mimo tento úzký obor přiliš hutný, ale to je s ohledem na omezený prostor možné prominout. Horší je skutečnost, že zde chybí alespoň náznak, co měří používané metriky (joint goal accuracy), a není zde ani dopředný odkaz do kapitoly 4. Také čtení ztížilo to, že termín „domain“ nebyl vysvětlen na příkladech. Z jediného příkladu „restaurant“ není zřejmé, že doména je vlastně název „formuláře“, který chceme pomocí dialogového systému vyplnit. Plně čtenář pochopí použití domén až v příkladu na konci strany 44.

Kapitola 3 popisuje vlastní návrh modelu. Autor se soustředí na přesnost popisu, včetně srozumitelného formalismu. Čtenáři by např. na straně 35 ovšem velmi pomohlo vidět i *příklad* textově reprezentovaného stavu dialogu $\mathrm{str}_S(S_t)$, kontextu C_t, a podobně i pro vícedoménové reprezentace. Příkladů stavu dialogu se čtenář dočká na straně 38, a v tu chvíli je zřejmé, že ve funkci $\mathrm{str}_S(S_t)$ jde o poměrně přímočarou linearizaci datové struktury.

Čtvrtá kapitola začíná podrobným popisem několika verzí datasetu MultiWOZ a filtrací zvolené verze pro účely vlastního experimentu tak, aby dataset obsahoval dostatek příkladů. Následuje

popis metrik, formálně přesný, ale z důvodu nedostatku příkladů opět jen velmi pracně srozumitelný.

Vyhodnocení experimentů a diskusi přináší pátá kapitola. Autor používá připravené automatické metriky a shledává, že pochopení stavu dialogu se generativním modelům daří (JGA přes 70 % oproti baseline cca 42–45 %), ale volba následující akce je velmi těžká úloha (požaduje se přesná shoda s referencí). Dosaženou přesnost cca 20 % tedy autor shledává relativně dobrým výsledkem. Kladně hodnotím alespoň minimalistické ruční srovnání výstupů systému s referenčním výstupem v sekci 5.3; bez něj by byla interpretace výsledků velmi nejistá. Ruční hodnocení by ještě lépe (1) zahrnovalo kvantitativní vyhodnocení, byť na velmi malém vzorku; zde autor možná studoval jen jeden dialog, (2) bylo přímo proloženo příklady chyb. Příklady a komentáře k nim najdeme v příloze.

Závěr práce dle očekávání jen shrnuje již řečené a ve výhledu do budoucna zejména odkazuje na potřebu pracovat s delším kontextem dialogu, např. pomocí posilovaného učení.

Příjemným obohacením je rejstřík s definicemi a propojení výskytů termínů z textu s ním. Do rejstříku by se hodilo přidat právě i termíny specifické pro dialogové systémy, mj. metriky.

Práce je psána velmi dobrou angličtinou s jen malým počtem chyb nebo překlepů. Některé z nich uvádím zde:

- Strana 8: „...which are specific information...“ by mělo být „pieces of information“.
- Strana 28: „but also a system inform memory that keeps track“. Termín „system inform memory“ je velmi nejasný.
- Obrázek 2.6 je bohužel příliš malý, čísla kroků nejsou čitelná.
- Fráze „generally unsafe for practical applications“ se s nepatrnou obměnou objevuje v prvním odstavci kap. 3 dvakrát.
- Strana 56: „modes“ místo „models“.

Otázky a komentáře:

- Jen poznamenávám, že termín „language understanding“ (strana 4) je příliš široký. Upřesnění, co tím myslíte, následuje dále a je bez problémů.
- Strana 9, bod 3 příkladu: Z formalizované podoby vypadlo heslo „restaurant“. Je to jen překlep, nebo záměr?
- Nesouhlasím s tvrzením na straně 38, že linearizované datové struktury stavu dialogu představují „a well-structured string format that *resembles the natural language that the T5 model was trained on*“. Dvojtečkové seznamy jsou v typickém textu spíše výjimka. Podobnost lexikální

pak závisí na rozumné volbě názvů pro domény („hotel–area“, „train–departure“, ale méně již „bookpeople“ nebo „intent : book_hotel“). Je zřejmě nad rámec práce zamyslet se, *proč* je T5 schopen takový strukturovaný text generovat.

- Práce podrobně popisuje inverzní funkce převádějící textovou reprezentaci do struktury. Zajímavé by bylo dozvědět se víc o jejich chování v případě, že formát není příslušným generujícím modelem přesně zachován. Počítáte nějak s takovou možností? Stála by za samostatné vyhodnocení.

- Jsou výsledky JGA uváděné v tabulce 5.1 srovnatelné s výsledky v tabulce 5.2? Tj. jde přesně o stejný test set, stejnou implementaci JGA? V ideálním případě máte k dispozici „výstupy“ konkurenčních systémů a vyhodnocení provádíte přesně stejným způsobem vy sám.

- Na str. 56 píšete: „...the generation models consistently outperform the classification models across all metrics. This can be attributed to the generation model's capacity to generate new sequences, making them more versatile and capable of handling unseen situations during training.“ Jak je to prosím s překryvem množiny tříd mezi trénovací a testovací částí dat? Ano, výstupní formát je zde sekvence, ale pokud je sekvencí ve skutečnosti jen omezený a všechny jsou známé z trénovacích dat, i sekvenční model bude de facto dělat klasifikaci. Dává na tuto otázku odpověď příloha A.1 support? Její popis není nikde dostatečně podrobný.

- Nemohu úplně souhlasit s tím, že malý rozdíl v přesnosti predikce akcí při zlatých vs. predikovaných stavech dialogů dokládá robustnost modelů (strana 56). Domnívám se, že se může jednat spíš o málo účinnou (málo rozlišující) metriku. Obecně je ACC nízké a pozorované rozdíly v ACC velmi malé, i když je velký prostor pro zlepšení.

- Souhlasím, že ACC dosahující (jen) 20 % není špatný výsledek. Bylo by ale mimořádně zajímavé vědět, jaké úrovně přesné shody dosáhnou *lidé*, kdyby řešili přesně stejnou úlohu. Dal by se udělat takový malý pokus třeba jen na 20, 50 nebo maximálně 100 příkladech?

- Píšete „The classification model is more accurate in predicting necessary actions than the generative model.“ (str. 57), přitom v tabulce 5.3 vychází ACC i F1 pro Robertu (klasifikační přístup) horší skóre než pro FLAN. Jak to tedy je?

- V příkladu v příloze A.2 je na straně 81 uveden textový kontext „I can help you with that. Do you have any special area you would like to stay? Or possibly a star request ...“ Je to prosím z referenčních dat, nebo je to minulý výstup systému, vygenerovaný z predikovaných akcí?

- V témže příkladu na straně 83 kritizujete predikci akcí: „The generative model incorrectly predicts actions to request booking day and duration, while the user already confirmed the

booking for four people.“ Této kritice nerozumím, počet osob přece neříká nic o délce a počátku pobytu. Kritiku si dle mého zaslouží naopak reference, která navrhuje zeptat se na počet osob, ale to již uživatel řekl. (Pokud jsem správně pochopil strukturu uváděných příkladů.)

Celkově jsem s diplomovou prací Jaroslava Šafáře spokojen. Na více místech by text mohl být pro čtenáře napsán srozumitelněji a některé ze závěrů jsou k diskusi. Práce však jasně dokládá schopnost korektně navrhnout, uskutečnit a přesně popsat a prodiskutovat experimenty v oblasti automatického zpracování přirozeného jazyka a doporučuji proto, aby byla přijata.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 31. 8. 2023

Podpis: