

Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce Jaroslav Šafář
Název práce Practical Neural Dialogue Management Using Pretrained Language Models
Rok odevzdání 2023
Studijní program Informatika **Studijní obor** Umělá inteligence

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

Text posudku:

Téma práce Diplomová práce Jaroslava Šafáře se zabývá problémem řízení dialogu v dialogových systémech. Autor se specificky soustředí na problémy sledování dialogového stavu (state tracking), tj. jak se vyvíjí preference uživatele v dialogu, a výběr akce systému (dialogovou strategii, dialogue policy/action selection), tj. určení, na co se systém má doptat, či zda má poskytnout odpověď z databáze. Motivací práce je snaha jít trochu jiným směrem než současný trend, kdy předtrénované generativní modely se pokoušejí zastávat všechny funkce dialogového systému (end-to-end), včetně generování slovo po slově, což vede jednak k horší kontrolovatelnosti výsledných systémů, jednak vyžaduje velké množství trénovacích dat, aby systémy pracovaly spolehlivě. Autor proto provádí výběr systému explicitně a tedy kontrolovatelně, a generování odpovědi ponechává na externím systému, jímž mohou být např. 100% spolehlivé šablony. Výsledné řešení navíc testuje na omezenějších trénovacích množinách.

Shrnutí obsahu Přestože práce míří jiným směrem než end-to-end dialogové systémy, autor shodně s nimi a v souladu s posledními trendy pro svoje systémy používá předtrénované jazykové modely (Flan-T5 a RoBERTa), přičemž věnuje velkou pozornost linearizaci strukturovaných dat na jejich vstupu. Své systémy pro sledování stavu a výběr akce autor testuje na datasetu MultiWOZ 2.2, který je dnes standardním benchmarkem pro dialogové systémy, a přidává varianty s pouhými 30% a 50% trénovacích dat. Oba typy systémů jsou evaluovány pomocí běžných automatických metrik a autor dále přidává vlastní detailní analýzu. Modely pro sledování stavu dosahují ve všech konfiguracích výborných výsledků, lepších než předchozí publikované práce. Modely pro výběr akcí vykazují ve všech konfiguracích sice relativně nízká skóre, nicméně zde je přímé porovnání s evaluačními daty příliš striktní, protože predikované akce jsou často akceptovatelné, i když přesně neodpovídají akcím zaznamenaným v datech. Tyto modely nicméně mají i tak problémy s komplexnějšími dialogovými situacemi a pro praktické použití zřejmě budou vyžadovat další úpravy.

Struktura textu Práce obsahuje úvod a šest číslovaných kapitol. Úvod nastiňuje v krátkosti motivaci práce a představuje obsah všech kapitol. První kapitola detailně vysvětluje potřebné teoretické koncepty v oblasti neuronových sítí a předtrénovaných jazykových modelů, kapitola 2 obsahuje úvod do problému dialogového managementu a shrnutí předchozích prací v oblasti. Kapitola 3 představuje autorův vlastní přístup k problému a soustředí se zejména na linearizaci dat pro použití v předtrénovaných modelech. Čtvrtá kapitola popisuje průběh experimentů, použitý dataset MultiWOZ 2.2 a použité evaluační metriky. Kapitola 5 představuje a diskutuje výsledky experimentů a kapitola 6 podává krátké závěrečné shrnutí. Text je logicky rozdělen a psán velmi dobře srozumitelnou angličtinou.

Průběh prací Autor na tématu dialogového managementu začal pracovat už v r. 2021, naše spolupráce pak probíhala v různé intenzitě celou dobu až do odevzdání (samozřejmě směrem k odevzdání se zintenzivnila). Autor provedl experimenty i s několika jinými architekturami, které neměly tak dobré výsledky a nakonec nebyly pro finální modely použity a do práce se nedostaly. Veškeré experimenty jsme průběžně konzultovali, text práce jsme před odevzdáním taktéž podrobně probrali a autor zapracoval většinu mých připomínek. Spolupráce s autorem na diplomové práci probíhala bez problémů, nemám k ní žádné výhrady.

Celkové hodnocení Výsledek splňuje všechny běžné nároky kladené na diplomovou práci. Některá původně zamýšlená rozšíření (integrace obou typů modelů, detailnější evaluace) bohužel vzhledem k nedostatku času nebyla realizována a evaluační část je tak omezenější, než bych si přál, i tak ale práce splňuje definované zadání a představuje validní model pro sledování stavu i výběr akce v dialogu. Model pro sledování stavu navíc dosahuje velmi vysokého skóre, zřejmě díky velmi dobře zvolenému základnímu předtrénovanému modelu Flan-T5, ale dílem jistě také správnou volbou linearizace vstupních dat. Celkově práci doporučuji k obhajobě; nemám žádné dotazy.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 28. 8. 2023

Podpis: