Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Michelle Elizabeth
Název práce	Conversational Agents for Task-Oriented Dialogue
Rok odevzdání	2024
Studijní program	Informatika **Studijní obor** Language Technologies and Computational Linguistics

Autor posudku	David Mareček **Role** oponent
Pracoviště	Institute of Formal and Applied Linguistics

Text posudku:

The goal of the thesis is to employ Large Language Models in Task Oriended Dialogues (e.g. searching for a restaurant, a train, booking a hotel, etc.) using the ReAct prompting strategy, which is based on generating thoughts, using tools, observing their outputs and repeat these steps until being able to answer the question. The evaluation is done using a simulated user agent.

Structure of the thesis:

A short introduction with goals and motivations is followed by two chapters about theoretical background. The first chapter describes Task oriented Dialogue systems and respective evaluation metrics. The second chapter shows the theory of Large Language Models and prompting strategies including the ReAact strategy.

Chapter 3 present the experimental setting. It describes the evaluation mechanisms, datasets and tools used and the methodology of the system agent (how the LLMs are prompted to use the ReAct strategy) and simulated user agent (LLMs compared to existing rule-based user simulator) Chapter 4 presents the results and show both the quantitative and qualitative analysis including various examples where the system failed. Chapter 5 discusses the achieved results, compare them to other approaches and outline some future work. The final conclusion summarises the work done.

Evaluation:

The thesis is written in very good English, it is well structured and it is rather short (44 pages of text). The first two background chapters about Task oriented Dialogue systems and Large Language Models are very well written (just one note: some abbreviations in text are different from abbreviations in Figure 1.1). However, in the third and fourth chapter describing the experiments and results, I would expect more details. For example, since the core of the thesis is prompting GPT models to behave according to the ReAct scenario, I would like to see an example of a full

prompt and what and how many examples were used and also whether different prompt variants were tried (see Question 1).

I appreciated the examples in the qualitative analysis, they really helped me to understand the method and the problems, however, many examples were not clearly described (see Questions 2, 3, 4).

The results comparison with other approaches is sufficient. I am missing more justifications on what are the main sources of errors. There is a claim in conclusions "LLMs might just imitating the examples", but no concrete example was shown in the thesis. Also regarding the claim "Difficulty in understanding user requests leads to repeated utterances from the user", in some examples in the thesis I saw it was the user agent who was generating unnecessary and confusing utterances and who evidently did not understand the LLM (see Question 3). It is a pity that user agent based on LLM did not work. I believe that with the LLM-based user agent or with the real human user, the system would work much better.

But overall, I like the thesis very much. Fitting a general LLM into such precisely structured framework must be very interesting but difficult at the same time. Since I am not an expert in this field, it is hard for me judge how much work was done to make all the experiments run. It uses existing data, frameworks and evaluation tools. I can imagine that writing a couple of prompts and incorporating it into an existing framework may be done in two days, but I can also imagine that analysing and connecting all the needed components may take months.

Questions:

1. Figure 3.2: How many examples were used in the few-shot setting? How many different tools/slots they covered? What is the difference between {tools} and [{tool_names}]? Have you experimented with different prompts or you tried only the one presented? And have you experimented with different numbers of shots?

2. Figure 4.2: The system found just one police station. Does it mean than there is only one police station in the database? Or the system always return only one record?

3. Figure 4.4: We see the hallucinated slot "time: after 14:15" but it is not shown in the figure how it was generated after the first user request. Does these hallucinations occur when the slot was not present in the examples in prompt? Or is it independent?

4. Figure 4.8: I do not see much difference, the result is the same by both the systems. Only GPT3.5 in the first db_query badly specifies the leaveAt. But it is correct in the second time. In both cases, the system forgot the specified time. What is the output of db_query if there is a mistake in the input?

5. Aren't the rule-based systems better because the user agent is also rule-based? The user utterances are often very unnatural and sometimes do not follow the previous response.

6. How much coding was necessary to perform the thesis experiments? Was it mainly the LLM prompting or is there a lot of hidden coding work in connecting the components?

Conclusion:

Overall, I rate the work as very solid, perhaps a less extensive, but I fully recommend it for defense.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 3. 9. 2024

Podpis: