Master Thesis Review

Faculty of Mathematics and Physics, Charles University

Thesis Author	Michelle Elizabeth
Thesis Title	Conversational Agents for Task-Oriented Dialogue
Submission Year	2024
Study Program	Computer Science
Branch of Study	Language Technologies and Computational Linguistics
Review Author	Ondřej Dušek **Role** Supervisor
Department	Institute of Formal and Applied Linguistics

Review Text:

Thesis Topic The topic of Michelle Elizabeth's master's thesis is large language model (LLM) prompting for task-oriented dialogue (TOD). This is very timely nowadays as all LLM-related topics; however, it also stands in contrast to most regular uses of LLMs, as TOD is much more rigid and consistency-focused than the unstructured conversations (chat-attuned) LLMs are designed for. This task essentially shows some of today's LLM's fundamental limits, in terms of consistency and outside world access.

The thesis approaches the TOD problem using a very promising promting paradigm for LLMs called ReAct. It iterates multiple steps of "thought, action, observation", where the LLM is tasked with reflecting its next steps, taking actions by calling external API functions, and processing their results. This builds on top of the very popular chain-of-thought prompting in a way that is very apt for TOD: allowing multi-step processing, integrating outside tools, and pushing the model to plan its actions.

Contents Summary While LLM prompting may seem simple at the first sight, the overall implementation work involved in the thesis was far from trivial. The author implemented a full end-to-end setup with a prompted LLM-based TOD system and a conjoined user simulator, to be evaluated in end-to-end full dialogues on MultiWOZ, the most popular TOD benchmark of today, which features multiple interconnected tourist information domains (hotels, restaurants, trains, etc.). This is the next level compared to most papers published on TOD nowadays, which mostly tend to take shortcuts to simplify evaluation and the problem itself. First, most current works evaluate by only generating a single-turn response and comparing to gold-standard data, then referring to the next turn of the gold-standard dialogue and starting response generation again from there, regardless of whether the previous response was reasonable. The full end-to-end dialogue used in this thesis is a much more realistic and much more challenging setting. Second, most works produce delexicalized outputs (i.e., various entity names are replaced by placeholders, to be filled directly from the database using a rule). This thesis works with full entity names everywhere, which again makes the setup more realistic and challenging than most previous works.

The implementation involved both an LLM-prompted TOD system and an LLM-prompted simulated user, which also included dialogue goal tracking. However, a pre-existing, partially rule-based user simulator was used for the final experiments to improve performance, as the double LLM setup performed rather poorly and the main focus was on the LLM TOD system. The author focused on evaluation with GPT-3.5 and GPT-4, which are among the most prominent LLMs in use today. The LLM TOD system was evaluated with the user simulator on the MultiWOZ domains and compared to multiple baselines on standard metrics. The evaluation also includes some prompt variation and a focus on response generation only. The simulation-based evaluations are accompanied by a manual qualitative analysis with detailed insights on error cases and their potential causes.

While the resulting system shows some promising results, it lags behind the baselines featuring more traditional architectures. It demonstrates the complexity of the TOD task and the MultiWOZ domains and shows the limits of today's LLMs on this problem. As frequently noted elsewhere, LLMs have problems with hallucination (ungrounded outputs), which also affects tracking the dialogue state in this experiment. Furthermore, their reasoning may be inconsistent and may not stick to the instructions; LLMs seem to struggle to generalize simple examples supplied in prompts onto more complex situations within the same domains. Furthermore, LLMs seems to have a hard time role-playing: They often divulge the internal processing to the user, and the LLM-based user agent tended to switch its role to play the system (which is probably closer to the LLM training).

The results indicate that LLM-based TOD systems are viable, but require a lot of control over the LLM; some problem simplifications such as delexicalization may still be required. Furthermore, the results are more positive when evaluated with humans instead of a simulator (see below).

Text Structuring The text is structured into five numbered chapters, plus an unnumbered introduction and conclusion. The introduction provides a little historical perspective and motivates the thesis work. Chapters 1 and 2 present the theoretical background and related works for the experminents in the thesis; the former introduces task-oriented dialogue systems and their evaluation, while the latter gives an overview of the recent developments around transformer-based language models and LLMs. It also describes LLM prompting including the ReAct approach. Chapters 3 and 4 describe the experimental work conducted by the author herself. Chapter 3 starts by introducing the MultiWOZ benchmark, the overall system setup and implementational detals. Chapter 4 then includes the evaluation, both in terms of automatic metrics and a qualitative manual analysis. Chapter 5 provides a high-level discussion of the results and suggests potential areas of improvement. The formal conclusion gives a brief summary of the whole work.

The text is written in excellent English and is very easy to read. The overall structuring is very sensible, the individual chapters and sections follow each other in a natural order.

Work Progress The work on this thesis was conducted within the author's dual-masters LCT study in Prague and Lorraine, but also as part of an internship at Orange Labs, which meant that the author had three supervisors. It started in March of this year, and it continues until today. The project's time scope was not ideal with respect to Charles University's thesis submission deadlines: The author had to submit her thesis partway through her internship. However, this also means that the work did not end with the submission – up until today, the author collected over 90 dialogues per system in a large-scale in-house evaluation campaign at Orange Labs. This evaluation shows the LLM less successful in providing information than the baseline, but more highly rated by the users, i.e., more "likeable".

In her work, the author was diligent, determined and able to solve many different implementation issues successfully. She consulted her work with all three supervisors and we collectivelly agreed on the next steps – by interaction with all of us, she has essentialy been managing a very valuable collaboration. The collaboration also applies to her writing, which was similarly consulted with all three supervisors and all three of us supplied our comments on the whole draft. I am happy to say that all my major comments were successfully resolved in the final text of the thesis. I am very happy with the final text and have no reservations.

Overall evaluation I believe that the end result fulfills all the obligations for a master's thesis at Charles University. Overall, I wholeheartedly recommennd the thesis to be defended; I do not have any questions for the defense.

I recommend that the thesis be defended.

I do not nominate the thesis for a special award.

Prague, 2 September 2024

Signature: