Review of Thesis Proposal

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University, Prague

Thesis Title:	Semantic Accuracy in Natural Language Generation
Candidate:	Mgr. Patrícia Schmidtová
Supervisor:	Mgr. et Mgr. Ondřej Dušek, Ph.D.
Reviewer:	Ing. Alexandr Rosen, Ph.D. Charles University, Faculty of Arts, Institute of the Czech National Corpus

Topic

The proposal is concerned with the issue of reliability of LLMs. Their sloppy treatment of facts, the main complaint of many users, has earned this behaviour the label of hallucination. This specific flaw in semantic accuracy has replaced other concerns about NLG, such as fluency, and its choice as the main topic of the proposal is very timely and perfectly justifiable. There are two research questions to be answered in the proposed thesis (§1.1). The first is about evaluation of the generated output (*How can we determine if a generated text is semantically accurate given its source data?*), while the second, more ambitious question, is about why the models hallucinate and what can be done about it (*Which internal or external mechanisms affect the semantic accuracy of an LLM's output and how can we manipulate them to achieve better accuracy?*).

Content

The *Introduction* sets the stage by nicely presenting the issue of semantic accuracy and hallucination (*If a human does not know the answer to a question, the socially acceptable behavior is to say 'I do not know' instead of making up a plausibly sounding lie.*). Arguments for restricting semantic accuracy first to faithfulness and then to hallucination are sound, and so is the choice of NLG tasks: data-to-text generation and summarization, without access to data outside the model. Still somewhat an introductory part, §2 on *Theoretical Foundations* (*necessary for proper understanding of this thesis* – §1.4) describes the transformer architecture and (more briefly) the technique of probing hidden layers.

The most extensive §3 (*Evaluating Semantic Accuracy*) deals with RQ1. Besides mapping the field by discussing methods and pros / cons of human and automatic evaluation, §3.2–§3.5 overview experience gained from the author's previous and ongoing work, including her Master's thesis, and (in §3.6) RQ1-related plans for the thesis. The plans include exploring evaluation methods used in MT, adopting LLMs as judges, and using methods based on source-target alignments. Based on the experience, crowd-sourced evaluation will be compared to experts' judgments. Also, the preferred scenario assumes that LLMs will provide feedback on experiments while humans will evaluate reported results.

The shorter §4 on RQ2 (*Improving and Understanding Semantic Accuracy in NLG*) has a structure similar to §3, starting with an overview of previous attempts to improve semantic accuracy by using external information or explain why LLMs are so sensitive to specific prompts, followed by presenting interpretability, implemented as probing, as the preferred direction of research in *our future work*. However, there is no past work by the author reported in this section and just one topic in the *Ongoing Work* §4.3: the effect of grammatical errors in the prompts. The *Future Work* §4.4 suggests two topics: decreasing the distance between source and target representations (following up on the alignment-based evaluation), and exploring the effects of prompt wording, with the ideal (not necessarily achievable) outcome suggested as a model more robust to variations in the prompt. In the final two paragraphs *Challenges and Limitations* the author curbs unrealistic expectations: *We do not expect we will be able to solve the issue of hallucination within this thesis*, suggesting moreover that the issue may not have a solution at all. Instead, there is a more modest aim: to understand *some of the faithfulness-related processes taking place inside the black box*. Finally, the rapid progress in the research of interpretability is admitted as the reason behind the less elaborate plans in this section.

Conclusion (§5) summarizes the goals: (i) to advance the understanding why LLMs produce semantic errors by using interpretability techniques and to *outline ways* to make the models more consistent and reliable by using the knowledge (RQ2), and (ii) to improve the art of evaluation of LLMs by *mapping the current state of research and encouraging the adoption of good practices*, adapting promising MT metrics to NLG tasks, assessing LLMs as evaluators, and measuring the *our contribution by a combination of automatic metrics and carefully designed human evaluation*. The proposal is concluded by an extensive list of references (6 pages).

Comments

The section on theoretical foundations (§2) is in fact a dense overview of how language models are built (§2.1) and how their hidden layers can be examined (§2.2). So, it's more about methodologies and techniques. I'd expect that it would approach the RQs theoretically, e.g. by comparing LLMs to humans, reviewing assumptions about deviations from faithfulness and if the deviations can be traced inside the model, if models can be honest and admit they don't know, also suggesting references on those topics. Alternatively, the section could just be renamed to better fit the idea that it should help the reader understand the following content.

There are quite a few topics listed in §3 as ongoing or future work, probably too many to fit into the thesis together with an appropriate description and results, and to be presented in the proposal with more details. Maybe they could stay but treated and presented as components of a single coherent methodology? Or maybe some of them could be side-tracked.

I believe RQ2 to be more challenging than RQ1, yet §4 is much shorter. Some reasons are understandable: the rapid developments in the field, or the realistic assessment of chances to resolve the hallucination issue. However, the plans here do not match the expectations raised in the RQ. From that perspective they can give only partial answers. Moreover, the author's experience and ongoing work concerning this topic seems to be also partial, especially in comparison with $3. Maybe RQ2 can be modified so that it better matches the plan? Alternatively, the topic of evaluation could be made even more prominent, although I'd hate to see the potentially very interesting understanding/improving part be dropped altogether.

The author should be careful to clearly distinguish her contribution, the core of the thesis, from collective work.

Question

§4.4.: We remain skeptical about whether the issue [of hallucination] can be fully solved (Kalai and Vempala, 2024).

- Any arguments? The reference does not seem to substantiate the statement.

Summary

The thesis proposal proves admirable competence and experience of the author in the topic, especially in the Introduction and in the section on evaluation. However, the reviewer is worried that there are too many topics planned, some of them and out of balance and not specific enough (although it is reasonable to leave some space for new developments in the field). The suggestions above are very tentantive, also because the reviewer trusts the author to find her way. Overall, the proposal promises an excellent, interesting and timely thesis.

Praha, 21 May 2024

Signature: