WASP Software Engineering and Cloud Computing 2023 Assignment 1

Mehrdad Farahani

August 30, 2023

1 Introduce your research/area (max 400 words)

My doctoral research primarily centers on representation learning for conversational AI, which applies to various modalities. For conversational AI, representation learning is crucial since it captures and encodes the underlying structure and information. Developing more effective representations allows systems to understand human language better and generate more coherent, contextually appropriate responses.

My initial studies focused exclusively on the "Text" modality, exploring the generative models (decoder-only) employed in open-domain dialogue systems like GPT-2. However, since open-domain dialogues are not limited to specific topics or domains, they pose a more complex and challenging task to the systems. This exploration involved an examination of the impact of multi-objectives on decoder-only models and different components of conversation, such as personas and utterances, to identify that multi-objectives can enhance the model's ability to comprehend the information exchanged between conversational participants and yield more consistent, contextually or personality appropriate responses. However, despite the potential benefit of these multi-objectives for improving models' ability to generate coherent responses, they continue to struggle with generating fabricated responses, a phenomenon known as hallucination.

To address this concern, the second phase of my research examined a range of possible solutions and architectures designed to alleviate it, focusing on the precise impact they would have on representations. I am considering Retrieval-Augmented Generation (RAG) [1] models to accomplish this, particularly in open-domain dialogue systems. The RAG model combines the benefits of generative and retrieval models. Retrieval-based models find and return suitable responses from a predefined dataset, while generative models generate text tokens. The RAG model utilizes both approaches to enhance the quality and relevance of generated responses.

2 Two principles/ideas/concepts/techniques from Robert's lectures

2.1 AI/ML Software Testing

Engineers work with explicitly programmed, deterministic systems as part of traditional software testing. However, AI and machine learning (ML) models introduce new layers of complexity. Contrary to conventional software, where code paths are systematically explored and validated, AI/ML models are probabilistic and data-driven, making them difficult to test. Models such as these learn from data, and their behavior is often difficult to predict without considering that data. Consequently, AI/ML software testing involves verifying that the software works as expected and validating that the underlying models provide reliable, accurate, and safe results across various inputs and conditions.

Our research focuses on representation learning in conversational AI, particularly in generating more effective, coherent, and contextually appropriate dialogues that cannot be proven accurate

without robust testing methodologies, so AI/ML software testing is highly relevant to our research.

According to the research theme, models may generate fabricated or incorrect responses, resulting in hallucinations. Finding the root cause of these hallucinations can be extremely difficult. A prime example of the importance of AI/ML software testing techniques can be found here. Several techniques may be used to examine how small changes in input parameters or model parameters affect the output, such as sensitivity analysis, ablation studies, or adversarial testing. By conducting such a test, it might be possible to identify which aspects of the representation learning may contribute to hallucinations.

For example, we integrate Retrieval-Augmented Generation (RAG) models to combat hallucination. The AI/ML testing approach could systematically evaluate how well the RAG model retrieves relevant information under various conditions and its robustness against hallucinations compared to decoder-only models. This way, we can provide empirical evidence for the research findings and solutions to the hallucination problem.

2.2 AI/ML Deployment and Challenges

Building and testing AI/ML models is already complex, but putting them into real-world use is entirely new. We must consider many things, such as how fast the system will run when users grow and how we will protect people's personal information. There is no doubt that speed is an essential factor, especially when it comes to chatbots and other tools that need to respond almost immediately. Let us simplify things by focusing on two key points. First, scalability, or how well the system can handle more users. Then, latency is all about keeping response times snappy no matter what.

Now, consider RAG models designed to pull from massive databases to make conversations more natural and accurate. To provide a more insightful answer, imagine the model has access to an extensive database of universal knowledge. That could take time, which is a problem, especially in open-domain dialogue systems where users are expecting quick responses. It is, therefore, a real technical challenge to strike the right balance between accurate answers and fast response times.

3 Two principles/ideas/concepts/techniques from one of the three guest lectures

3.1 AI/ML Generalization and AutoML from Fabio Calefato

When it comes to machine learning, the ability of a model to perform well on data that it has not seen before is known as generalization. This is especially important in fields like conversational AI, where the model will encounter many user inputs not part of the training set. AutoML aims to make machine learning more accessible by automating some parts of the process, improving efficiency, and speeding up research efforts.

A relevant question related to generalization is whether a system trained in one language, such as English, can still perform well in other languages while maintaining conversational quality. This would require considering both linguistic and cultural differences and could involve extending the multi-objective framework currently being researched.

AutoML could be an excellent option for our team or users who are not experts in the field. It makes it easier for them to customize or extend our conversational AI models. Additionally, it can enhance the efficiency of the model development process by automating the search for the most optimal hyperparameters and architectures, particularly for research with multiple objectives and modes. Suppose we want to swiftly adapt our dialogue system to different languages or special-

ized domains. In that case, AutoML can automate the tedious parameter tuning process for each new setting, quickly recognizing the best configurations.

4 Two Topics

4.1 Continuous Deployment

Continuous Deployment (CD) is an extension of continuous integration to ensure that we can quickly release new changes to our customers sustainably. This approach ensures that software can be automatically released to production, resulting in many daily production deployments. This method has been widely adopted in traditional software engineering but presents unique challenges when applied to AI/ML projects. In contrast with conventional software, which remains static until updated, machine learning models may degrade over time as data distributions change. This requires continuous model performance monitoring. AI/ML Continuous Deployment must account for this by automatically updating the model or triggering alerts for manual intervention. In the course of updating models, it is essential to ensure that they still work well with other components of the system. To accomplish this, new validation techniques are required to ensure the deployed models perform as they should. On the other hand, due to rapid deployment cycles, changes can be implemented quickly, making it crucial to ensure that new model versions comply with all ethical and regulatory requirements.

Areas of opportunity regarding our research:

- **Validation and testing (hallucination metric):** With a continuous deployment system equipped with robust monitoring tools, it would be possible to detect instances in which hallucinations increase after deploying a new version, triggering further investigation or reverting to a previous release.
- **Boosting experimentation:** It could facilitate rapid testing; as a result, different model architectures or multi-objective settings could be tested, allowing us to examine differences in conversational quality, contextual relevance, and other metrics more quickly.

4.2 Regulations and Compliance

Regulatory frameworks for AI/ML are in place to tackle a range of ethical, legal, and social issues. Regulatory bodies are examining these technologies to ensure they comply with established criteria, such as data privacy, system security, and accountability. Laws like Europe's GDPR and California's CCPA set rigorous guidelines for gathering, storing, and utilizing data. In specific industries like healthcare and finance, there is a particular emphasis on the algorithms being both transparent and easy to understand. Additionally, maintaining logs of data handling, model training, and decision processes could be mandated for legal or compliance purposes.

Areas of opportunity regarding our research:

- **Ethical Dialogues [2]:** The aim of our work in dialogue-based AI is to create systems that produce replies that are not just contextually relevant, but also considerate of moral and cultural nuances. This would make our models more readily compliant with emerging regulations around algorithmic fairness.
- **Preserving Data Privacy in Representation Learning [3]:** We can explore methods that either anonymize the data or protect its privacy in some other way when used in dialogue models. This would make our work compliant with data protection laws such as GDPR.

References

[1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2020.

[2] Hao Sun, Zhexin Zhang, Fei Mi, Yasheng Wang, W. Liu, Jianwei Cui, Bin Wang, Qun Liu, and Minlie Huang. Moraldial: A framework to train and evaluate moral dialogue systems via moral discussions. In *Annual Meeting of the Association for Computational Linguistics*, 2022.

[3] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Xiaodong Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *USENIX Security Symposium*, 2020.