WASP Software Engineering Assignment: Essay

Bram Willemsen, KTH

December 2022

1 My research

The focus of my research is on the phenomenon of language grounding in (situated) dialogue, which involves studying the ways in which referring expressions produced by conversational agents relate to the entities they denote in their shared environment. In other words, we are addressing the question of how we can determine the referents of words in a conversational context. Additionally, we are also investigating the inverse problem of generating referring expressions, as effective communication requires that conversational agents be able to not only interpret the expressions used by others, but also generate their own referring expressions for use in conversation.

2 Topics

2.1 Human-Computer Interaction

From the description of my own research in Section 1 it should be fairly obvious that Human-Computer Interaction (HCI) is a topic that is highly relevant to my own work. HCI is highly interdisciplinary: although which subjects are relevant to a given HCI problem depends on the purpose and physical manifestation of the machine and the human being(s) using it, HCI brings together ideas, concepts, and techniques from various fields of study, including, but not limited to, computer vision, natural language processing, robotics, and UX design. As a result, HCI researchers and practitioners often come from a variety of backgrounds (I myself originally do not have an engineering but a humanities background, for example), and their work can be applied in a wide range of settings, from the design of consumer products and interactive media, to the development of assistive technologies and intelligent systems. HCI is a critical field that plays a central role in shaping the ways in which humans and machines interact and co-exist, and has the potential to transform the ways in which we live and work. Typically, a piece of software intended for humans to complete some task with, requires the end user to study and understand the ways in which the tool can and should be used to complete said task. There is a reason why, for certain positions, experience using certain software packages is explicitly listed on people's resumes when they apply for the job: it should indicate to the prospective employer that one is able to complete tasks with the listed tools. However, learning how to use such tools, especially when complex and when UX design was not a priority, is not trivial. In addition, it is not a given that skills using one piece of software will transfer to another: one might have to start from scratch when switching tools. Now, natural language is perhaps our most intuitive means of communication and has the potential to, to a great extent, do away with CLIs and GUIs, especially for products intended for an audience that has no expert knowledge of the system but is its intended end user. As opposed to systems using specialized software, the use of a truly conversational user interface is an option even for the lowest common denominator. Studying HCI from this perspective highlights the potential of natural language as a user interface to greatly improve the accessibility of machines for a wide range of users, even those without specialized knowledge or expertise.

2.2 Security and Privacy

The paradigm shift from the design of expert systems or otherwise more explicit coding of systems in favor of the creation of systems based on machine learning has far-reaching implications for both security and privacy. Using HCI as a lens through which we look at these issues, various questions arise. Self-driving cars, such as

those developed by Tesla, may be our best chance at reducing traffic accidents in the near future, but in their current state these autonomous vehicles are not flawless systems (if there ever is such a thing to begin with, of course). When a system is based on hand-coded rules rather than a blackbox trained on tons of data, its decision-making process is more transparent for the engineer. Without explicit efforts to go from blackbox to explainable AI, it is anyone's guess as to why certain decisions are made by a system. When dealing with high stakes situations in which countless decisions have to be made every second, how can we ensure that the blackbox makes decisions that we as humans would deem appropriate given the circumstances? How do we ensure consistency? How can we ensure that we handle cases that are out of distribution appropriately? Oftentimes, the latter is left for future work, but when a human life is at stake "edge cases" are not to be overlooked.

Privacy is also a concern when it comes to machine learning-based systems. Research has shown that it may be possible to extract the training data from large models. As language modeling moves towards larger models with billions of parameters, there are costs beyond just computational. It may be possible to extract individual training examples from these models, potentially revealing sensitive information. This is particularly problematic if the data has not been anonymized and contains personally identifiable information of unwilling participants. If these systems "fail", who should be held responsible?

2.3 Regulations and Compliance

Sticking with the example of autonomous vehicles, when an oversight in the software of a self-driving car is the cause of a fatal accident, who is at fault? The software developers, the manufacturer, or the operator of the vehicle (or, some would even add to this list, the vehicle itself)? Arguably, this question is difficult to answer even for hand-coded systems, let alone machine-learning based ones. In the event that developers could be held responsible for such unfortunate events, companies could require end users to sign waivers before they would allow the use of their product. But this is hardly a situation anyone would want to find themselves in, neither producer nor consumer. The general public as it is, is already skeptical of knowingly relinquishing control to a non-human, let alone having to sign off their rights to boot. But such dilemmas inevitably pop up when technological progress is faster than the law. Regulations cannot keep up. Related, ethical questions arise when a system would need to make a decision where as part of the outcome a fatality is inevitable. What comes to mind is the trolley problem, a thought experiment in ethics where one is asked to make a decision between two options, neither having favorable outcomes. What is the ethical choice; how does one weigh their options? Especially when each option results in the death of a person. Moreover, how do we expect machines to make ethical decisions for us when we have no clear answer ourselves as to what is the ethical decision in a given situation in the first place?

3 Future trends and directions of Software Engineering

Seeing as my work revolves around language, I follow advancements in language modeling research closely. As mentioned, over the last few years we have seen the number of parameters in language models steadily increase. What was unfathomably large in terms of size just a few years prior is now considered small. We do see that this scaling up approach has clear benefits, as indicated by constantly improving SOTA on basically all benchmarks across the board in essentially all subjects of computer science. The public has been introduced to some of the consequences of a world in which we can no longer distinguish between what is real and what has been generated by AI (think, for example, deepfakes). Image synthesis, speech synthesis, text generation: the further we push SOTA, the more likely we create a world in which we no longer know whether someone actually said something abhorrent themselves or whether it was all faked with malicious intent. These technologies have major potential to disrupt society. However, it is not all doom and gloom. These technologies can also benefit us all and have major potential to do so. As an experiment and as an indication that these technologies have the potential to be truly useful, I have instructed OpenAI's recently released conversational-esque language model ChatGPT to generate bits and pieces of this essay: I have told it to rephrase small chunks of text I had written, fed it lines of reasoning for which I told it to generate short, essay-style paragraphs, and I had it finish some of my sentences for me. I am confident that you, the reader, was not aware of this until I revealed this to you just now; you are probably wondering which parts written by the human author of this essay were augmented by

ChatGPT and which parts generated by ChatGPT were post-edited by yours truly. I take the fact that it is not immediately obvious as evidence supporting my point. Here, I have used ChatGPT as an assistive technology for next-level, tool-assisted writing, but the applications are plentiful. That is not to say that, as of right now, models like ChatGPT would be able to generate an entire essay from scratch that would be indistinguishable from a human's efforts (as of right now it took more time to use ChatGPT for this purpose than if I were to have not asked for its assistance at all). There is a reason why "prompt engineering" is currently a thing. It is, if anything, an indication that these models do still suffer from certain limitations. That being said, even though this may not be the "be all, end all" for AI, or even language modeling for that matter, the practical usefulness of these models for downstream applications is undeniable.

A particularly relevant downstream application is that of code generation. OpenAI's Codex shows the potential language models have to do away with much of manual programming. Through simple instructions in natural language, entire applications can be written from scratch. It is a step towards making ourselves obsolete, although that reality is still far away. In the meantime, such tools can speed up the process and do away with menial programming tasks that would otherwise be time-consuming.

As a proponent of open science, however, what concerns me about these developments is the fact that we are more and more at the mercy of a few big players with control over the SOTA AI behind a paywall. The irony is that these models make technology more accessible for everyone by providing a natural language user interface that non-experts can meaningfully interact with as well, but at the same time the fact that only large corporations and institutions have the means to create these models (or even host them) counteracts the democratization of this technology.