COMP 550: READING ASSIGNMENT 3

JAN TIEGGES NOVEMBER 17, 2023

The paper presents the Winograd Schema Challenge (WSC), a test designed to assess a machines intelligence. In the WSC, a machine is presented with a sentence containing an ambiguous pronoun or possessive adjective and is asked to choose the correct referent for it.

The WSC is designed to be Google-proof, i.e. there should be no obvious statistical test over text corpora that reliably disambiguates the correct answer. This is important because it ensures testing a machine's ability to understand language and context, rather than simply relying on statistical patterns in large datasets. By focusing on language understanding rather than a specific application, WSC encourages the development of more flexible and generalizing language models. However, a disadvantage of this is the difficulty to evaluate the performance in a real-world setting.

The authors compare the WSC against the Turing test and Recognizing Textual Entailment (RTE). While all these tests are designed to assess the ability of machines to demonstrate human-like intelligence, they differ fundamentally in a number of aspects. The Turing test is about the ability of a machine to hold a conversation with a human in natural language, while RTE focuses on the ability to determine whether one sentence entails another. The scope of the Turing test is naturally more general than that of the other two, as it is not limited to a specific area of language comprehension. The Turing test requires human judges to evaluate the performance of the machine, while the RTE and Winograd Schema challenge can be scored automatically. In addition, the Turing test requires a large data set of human conversations, while RTE and WSC require a data set of sentence pairs and ambiguous sentences respectively, thus making them much easier to implement.

The authors strongly outline the weaknesses of the Turing test, which does not really assess the ability of a machine to be intelligent, but rather the ability to imitate a human being. They impose new requirements on tests for machine intelligence, including the need for background knowledge, google-proofness, and the possibility of automatic evaluation. The WSC presented here is a step in the right direction, as it requires thinking about the meaning of words and sentences in context.

One limitation of the paper is that they do not back up many of their claims with actual experiments. Only a few examples of WSC and its solution are given, but the performance of existing models on this task is not evaluated. Therefore, it is difficult to determine how effective WSC is in practice. The paper provides no formal proof that there is no statistical test over text corpora that reliably disambiguates the correct answer.

The fact that large language models (LLMs) can solve the Winograd Schema Challenge (WSC) with high accuracy does not necessarily mean that they can "think in the full-bodied sense that we usually reserve for people". While LLMs have demonstrated impressive language comprehension abilities, they are still fundamentally different from human cognition. LLMs problems such as hallucinations show that they are ultimately just statistical models that learn patterns in large data sets. On the other hand, their inner workings is still a very unexplored research field. In the end, the WSC may not be as Google-proof as originally thought, but it was still a good step towards a better test for the intelligence of machines.