Evaluating Systems that Generate Content

Ian Soboroff
NIST

ABSTRACT

The astounding emergence of ChatGPT and other AI systems that generate content, and their apparently incredible performance, are an inspiration to the research community. The performance of these LLMs is so impressive it is widely supposed that we can use them to measure their own effectiveness! We have had evaluation methods for generated content, including question answering, summarization, and translation, and in this talk I dust them off and present both a historical view and how we might approach those methods today. tl;dr, we have a lot of work to do.