Michael Zhao
CS 4395.001
8 April 2023

ACL Paper Summary: "Is GPT-3 Text Indistinguishable from Human Text? SCARECROW: A Framework for Scrutinizing Machine Text"

In "Is GPT-3 Text Indistinguishable from Human Text? SCARECROW: A Framework for Scrutinizing Machine Text," Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi* (Paul G. Allen School of Computer Science & Engineering, University of Washington and the Allen Institute for AI) propose a framework that determines the range and severity of errors in machine-generated text. SCARECROW covers ten different categories of errors, from redundancy to commonsense errors. These errors are detected through crowdsourced human annotations and later expanded on through classification models. This system was used to collect over 41,000 error spans and quantified the differences between GPT-3 generated text and human-generated text.

In a previous work, "All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text", Clark et. al. showed that annotators are no longer able to reliably detect the differences between GPT-3 generated text and human-written text.

The types of errors can be grouped into three main categories. The first consists of the language-related errors. These errors include grammar and usage (simply incorrect or improper placement of words), off-prompt (unrelated or contradicts the prompt), redundant (repetition lexically or semantically), self-contradiction (when a phrase contradicts something it generated previously), and incoherent (something confusing but doesn't fit the aforementioned categories). Generally, these are the easiest to detect as they make the sentence sound awkward or completely incomprehensible to the annotator. The second large category of errors are the factual errors. These errors include bad math (most commonly found in conversions), encyclopedic (facts that are wrong), and commonsense (violates "basic understanding of the world"). These errors may be harder to detect as not all annotators possess the mathematical or encyclopedic knowledge to catch errors of this type. The last category of errors is something that the authors named "reader issues". The two errors in this category are "Needs Google" (Requires search to verify claim) and "Technical Jargon" (which requires more expertise to understand). This category of errors is hardest to catch and can be quite subjective. Additionally,

many of the findings show that humans tend to create equal or higher amounts of these errors, so this category of errors are discarded for some metrics used in the paper.

The study uses increasingly larger sizes of models, from GPT-2 Small to GPT-2 XL to GPT-3, as well as Grover, which is trained in-domain. The main measurement that is used in the paper is the span coverage, which indicates the average proportion of tokens that are covered by annotations of a specific error type. One of the key findings lies in the way errors are decreased by scaling models up. In the case of Encyclopedic, Commonsense, and Incoherent errors, the size of the model directly correlates with the decrease of these errors seen in machine-generated text. Humans also tend to make the least amount of these type of errors. On the other hand, for Off-Prompt, Bad Math, and Grammar/Usage errors, scaling only benefitted to an extent but then plateaued off after GPT-2 XL. Self-Contradiction and Redundant errors were harder to measure as models like GTP-2 Small were quite incoherent and thus not given much chance for self-contradiction. Redundancy was also hard to find and varied based on how many other errors were counted. In the last category of errors, the control (human-written text) showed the highest number of “Needs Google” and “Technical Jargon” errors. This makes sense as generally these types of errors aren’t technically wrong but still hinders the readability of a piece of text.

The SCARECROW methodology is composed of a few steps. First, the text is generated by feeding a one-sentence human-written prompt into a model to generate a one-paragraph output. The text is generated or can be written by a human (control), but annotators do not know which one did which. The length of output text (80-145 tokens) is picked to “balance expressiveness with scope”. It’s hard to evaluate the capabilities on shorter segments of text, and too long of text leads to errors of other types, which the paper leaves for future work. The annotators are told to select the smallest span of text that has an error. Once selected, the annotator would then label it with the type of error, the severity, and reasoning. The collection of data was facilitated through Amazon Mechanical Turk (AMT), which is a system that allows for crowdsourcing of data. Each worker is paid $40 for a training program and test, which they must pass with a 90 or higher to continue annotating. Workers are then given $3.50 per paragraph annotated. This methodology is continued for 13k annotations of 1.3k paragraphs, resulting in over 41k final spans.

The final research that the team behind the paper did was whether it was possible for machines to detect and classify these types of errors. This task was given to a model as “span classification”. Using a pre-trained model (RoBERTa-large), they trained the model for 15 epochs with AdamW and a learning rate of 10^-6. The results show that humans have a much

better precision than models for almost every category except Common-Sense. However, humans tend to have a much lower recall rate. A single human judge can be considered a “high precision, low recall judge”. The models were quite successful at identifying “Needs Google” with a relatively high precision (above 0.6) and almost perfect recall. The models also were able to predict “Grammar/Usage”, “Incoherent”, and “Redundant” errors with decently high recall at the cost of poor precision.

As of today, they have received 30 citations. I believe this paper is important as it lays the groundwork for the study of GPT-generated text and how good it actually is. It provides a quantitative framework for this type of evaluation, which is hard to find these days. Researchers of these models can also use this study as a groundwork for improving their models, especially as GPT-4 is already publicly available in beta.

* Noah A. Smith has received 40379 citations, followed by Yejin Choi, who has received 28793 citations.