

New test reveals AI still lacks common sense

November 18 2020, by Caitlin Dawson



Despite advances in natural language processing, state-of-the-art systems still generate sentences like "two dogs are throwing frisbees at each other." Credit: Adriana Sanchez.

Natural language processing (NLP) has taken great strides recently—but how much does AI understand of what it reads? Less than we thought, according to researchers at USC's Department of Computer Science. In a recent paper Assistant Professor Xiang Ren and Ph.D. student Yuchen Lin found that despite advances, AI still doesn't have the common sense needed to generate plausible sentences.

"Current machine text-generation models can write an article that may be convincing to many humans, but they're basically mimicking what they have seen in the training phase," said Lin. "Our goal in this paper is to study the problem of whether current state-of-the-art text-generation models can write sentences to describe natural scenarios in our everyday lives."

Understanding scenarios in daily life

Specifically, Ren and Lin tested the models' ability to reason and showed there is a large gap between current text generation models and human performance. Given a set of common nouns and verbs, state-of-the-art NLP computer models were tasked with creating believable sentences describing an everyday scenario. While the models generated grammatically correct sentences, they were often logically incoherent.

For instance, here's one example sentence generated by a state-of-the-art model using the words "dog, frisbee, throw, catch":

"Two dogs are throwing frisbees at each other."

The test is based on the assumption that coherent ideas (in this case: "a person throws a frisbee and a dog catches it,") can't be generated without a deeper awareness of common-sense concepts. In other words, common sense is more than just the correct understanding of language—it means you don't have to explain everything in a conversation. This is a fundamental challenge in the goal of developing generalizable AI—but beyond academia, it's relevant for consumers, too.

Without an understanding of language, chatbots and voice assistants built on these state-of-the-art [natural-language](#) models are vulnerable to failure. It's also crucial if robots are to become more present in human environments. After all, if you ask a robot for hot milk, you expect it to

know you want a cup of milk, not the whole carton.

"We also show that if a generation model performs better on our test, it can also benefit other applications that need commonsense reasoning, such as robotic learning," said Lin. "Robots need to understand natural scenarios in our daily life before they make reasonable actions to interact with people."

The common sense test

Common-sense reasoning, or the ability to make inferences using [basic knowledge](#) about the world—like the fact that dogs cannot throw frisbees to each other—has resisted AI researchers' efforts for decades. State-of-the-art deep-learning models can now reach around 90% accuracy, so it would seem that NLP has gotten closer to its goal.

But Ren, an expert in natural language processing and Lin, his student, needed more convincing about this statistic's accuracy. In their paper, published in the Findings of Empirical Methods in Natural Language Processing (EMNLP) conference on Nov. 16, they challenge the effectiveness of the benchmark and, therefore, the level of progress the field has actually made.


Concept-Set: a collection of objects/actions.

dog, frisbee, catch, throw



Generative Commonsense Reasoning

Expected Output: everyday scenarios covering all given concepts.

- A dog leaps to catch a thrown frisbee. [Humans]
- The dog catches the frisbee when the boy throws it.
- A man throws away his dog 's favorite frisbee expecting him to catch it in the air. 

GPT2: A dog throws a frisbee at a football player. [Machines]

UniLM: Two dogs are throwing frisbees at each other .

BART: A dog throws a frisbee and a dog catches it.


T5: dog catches a frisbee and throws it to a dog 

Figure 1: An example of the dataset of COMMONGEN. GPT-2, UniLM, BART and T5 are large pre-trained text generation models, *fine-tuned* on the proposed task.

Examples of sentences generated by state-of-the-art text generation models. Credit: from the paper: "Commongen: a constrained text generation challenge for generative commonsense reasoning."

"Humans acquire the ability to compose sentences by learning to understand and use common concepts that they recognize in their surrounding environment," said Lin.

"Acquiring this ability is regarded as a major milestone in human development. But we wanted to test if machines can really acquire such generative commonsense reasoning ability."

To evaluate different machine models, the pair developed a constrained text generation task called CommonGen, which can be used as a benchmark to test the generative common sense of machines. The researchers presented a dataset consisting of 35,141 concepts associated with 77,449 sentences. They found the even best performing model only achieved an accuracy rate of 31.6% versus 63.5% for humans.

"We were surprised that the models cannot recall the simple commonsense knowledge that 'a human throwing a frisbee' should be much more reasonable than a dog doing it," said Lin. "We find even the strongest model, called the T5, after training with a large dataset, can still make silly mistakes."

It seems, said the researchers, that previous tests have not sufficiently challenged the models on their common sense abilities, instead mimicking what they have seen in the training phase.

"Previous studies have primarily focused on discriminative common sense," said Ren. "They test machines with multi-choice questions, where the search space for the machine is small—usually four or five candidates."

For instance, a typical setting for discriminative common-sense testing is a multiple-choice question answering task, for example: "Where do adults use glue sticks?" A: classroom B: office C: desk drawer.

The answer here, of course, is "B: office." Even computers can figure this out without much trouble. In contrast, a generative setting is more open-ended, such as the CommonGen task, where a model is asked to generate a natural sentence from given concepts.

Ren explains: "With extensive [model](#) training, it is very easy to have a good performance on those tasks. Unlike those discriminative

commonsense reasoning tasks, our proposed test focuses on the generative aspect of machine common sense."

Ren and Lin hope the data set will serve as a new benchmark to benefit future research about introducing common sense to natural language generation. In fact, they even have a leaderboard depicting scores achieved by the various popular models to help other researchers determine their viability for future projects.

"Robots need to understand natural scenarios in our daily life before they make reasonable actions to interact with people," said Lin.

"By introducing common sense and other domain-specific knowledge to machines, I believe that one day we can see AI agents such as Samantha in the movie Her that generate natural responses and interact with our lives."

More information: CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning, arXiv:1911.03705 [cs.CL] arxiv.org/abs/1911.03705

inklab.usc.edu/CommonGen/

Provided by University of Southern California

Citation: New test reveals AI still lacks common sense (2020, November 18) retrieved 24 April 2024 from <https://techxplore.com/news/2020-11-reveals-ai-lacks-common.html>

| |
|---|
| This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only. |
|---|