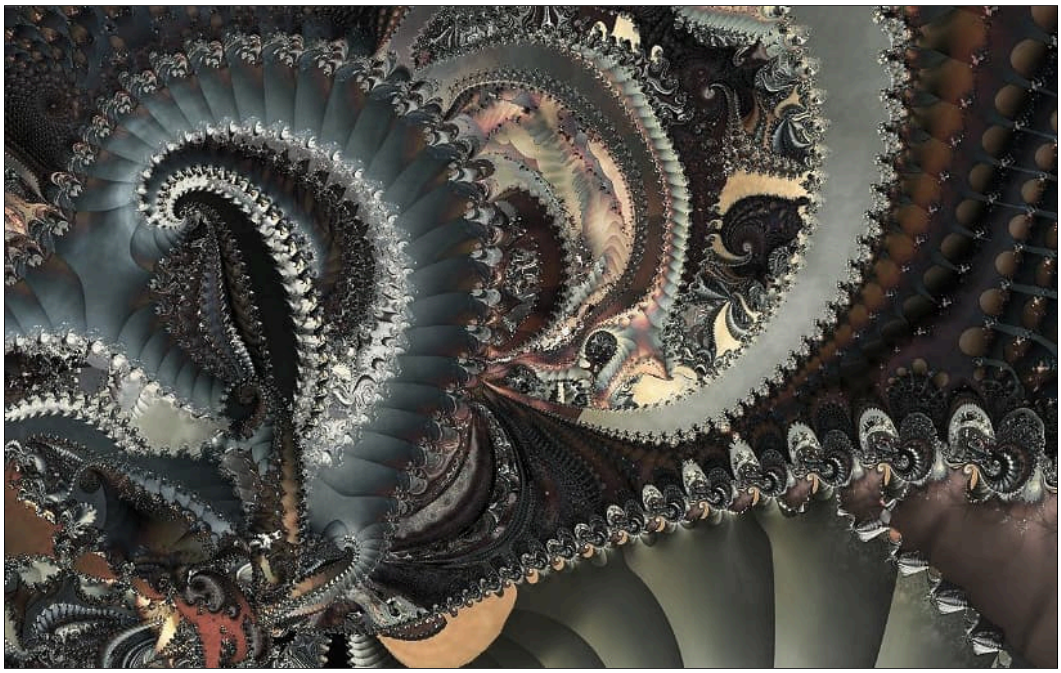
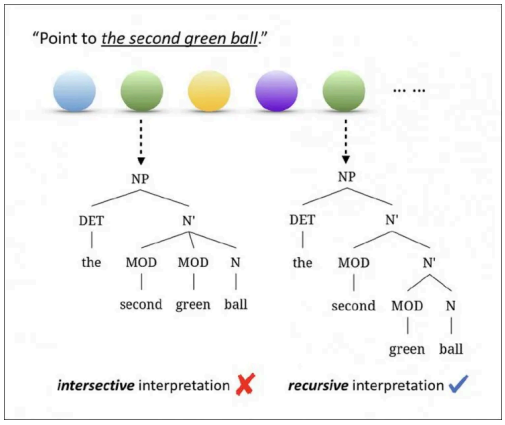
NLP Models Struggle to Understand Recursive Noun Phrases

Originally published at https://www.unite.ai/nlp-models-struggle-to-understand-recursive-noun-phrases/



Researchers from the US and China have found that none of the leading Natural Language Processing (NLP) models seem to be capable, by default, of unraveling English sentences that feature recursive noun phrases (NPs), and 'struggle' to individuate the central meaning in closely-related examples such as *My favorite new movie* and *My favorite movie* (each of which has a different meaning).



In a headline example from the paper, here is a minor puzzle that children frequently fail to unpick: the second ball is green, but the fifth *ball is the 'second green ball'.* Source: https://arxiv.org/pdf/2112.08326.pdf

The researchers set a Recursive Noun Phrase Challenge (RNPC) to several locally installed open source language generation models: OpenAI's GPT-3*, Google's BERT, and Facebook's RoBERTa and BART, finding that these state-of-the-art models only achieved 'chance' performance. They conclude†:

> *'Results show that state-of-the-art (SOTA) LMs fine-tuned on standard benchmarks of the same format all struggle on our dataset, suggesting that the target knowledge is not readily available.'*

Task	ID	Input	Gold Label	Predicted Label
Single-Premise Textual Entailment	(1a)	Premise: This is my new favorite movie. Hypothesis: This is my favorite movie.	Entailment	Entailment ✓
	(1b)	Premise: This is my favorite new movie. Hypothesis: This is my favorite movie.	Non-Entailment	Entailment ✗
Multi-Premise Textual Entailment	(2a)	Premise 1: He is a short American basketball player. Premise 2: He is a man. Hypothesis: He is an American man.	Entailment	Entailment ✓
	(2b)	Premise 1: He is a short American basketball player. Premise 2: He is a man. Hypothesis: He is a short man.	Non-Entailment	Entailment ✗
Event Plausibility Comparison	(3a)	Event 1: An animal can be harmful to people. Event 2: A dead dangerous animal can be harmful to people.	Less Plausible	Less Plausible ✓
	(3b)	Event 1: An animal can be harmful to people. Event 2: A dangerous dead animal can be harmful to people.	More Plausible	Less Plausible ✗

Minimal-pair examples in the RNPC challenge where the SOTA models made errors.

In the examples above, the models failed, for instance, to distinguish the semantic disparity between *a dead dangerous animal* (i.e. a predator that poses no threat because it is dead) and a *dangerous dead animal* (such as a dead squirrel, that may contain a harmful virus, and is a currently active threat).

(Additionally, though the paper does not touch on it, 'dead' is also frequently used as an adverb, which addresses neither case)

However, the researchers also found that additional or supplementary training that includes RNPC material can resolve the issue:

> *'Pre-trained language models with SOTA performance on NLU benchmarks have poor mastery of this knowledge, but can still learn it when exposed to small amounts of data from RNPC.'*

The researchers argue that a language model's ability to navigate recursive structures of this type is essential for downstream tasks such as language analysis, translation, and make a special case for its importance in harm detection routines:

> *'[We] consider the scenario where a user interacts with a task-oriented agent like Siri or Alexa, and the agent needs to determine whether the involved activity in the user query is potentially harmful* [i.e. to minors]. *We choose this task because many false positives come from recursive NPs.*
>
> *'For example,* how to make a homemade bomb *is obviously harmful while* how to make a homemade bath bomb *is harmless.'*

The paper is titled *Is "my favorite new movie" my favorite movie? Probing the Understanding of Recursive Noun Phrases*, and comes from five researchers at the University of Pennsylvania and one at Peking University.

Data and Method

Though prior work has studied syntactic structure of recursive NPs and the semantic categorization of modifiers, neither of these approaches is sufficient, according to the researchers, to address the challenge.

Therefore, based on the use of recursive noun phrases with two modifiers, the researchers have sought to establish whether the prerequisite knowledge exists in SOTA NLP systems (it doesn't); whether it can be taught to them (it can); what NLP systems can learn from recursive NPs; and in what ways such knowledge can benefit downstream applications.

The dataset the researchers used was created in four stages. First was the construction of a modifier lexicon containing 689 examples drawn from prior literature and novel work.

Next the researchers gathered recursive NPs from literature, existing corpora, and additions of their own invention. Textual resources included the Penn Treebank, and the Annotated Gigaword corpus.

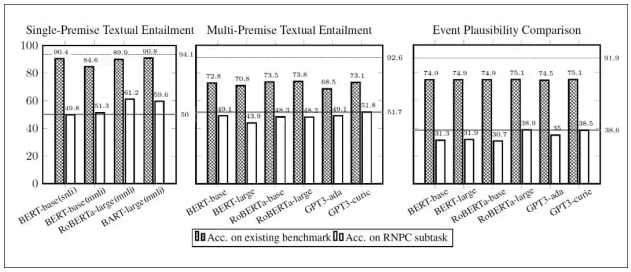
Then the team hired pre-screened college students to create examples for the three tasks that the language models would face, validating them afterwards into 8,260 valid instances.

Finally, more pre-screened college students were hired, this time via Amazon Mechanical Turk, to annotate each instance as a Human Intelligence Task (HIT), deciding disputes on a majority basis. This whittled the instances down to 4,567 examples, which were further filtered down to 3,790 more balanced instances.

The researchers adapted various existing datasets to formulate the three sections of their testing hypotheses, including MNLI, SNLI, MPE and ADEPT, training all the SOTA models themselves, with the exception of the HuggingFace model, where a checkpoint was used.

Results

The researchers found that all models 'struggle' on RNPC tasks, versus a reliable 90%+ accuracy score for humans, with the SOTA models performing at 'chance' levels (i.e. without any evidence of innate ability versus random chance in response).



Results from the researchers' tests. Here the language models are tested against their accuracy on an existing benchmark, with the central line representing equivalent human performance in the tasks.

Secondary lines of investigation indicate that these deficiencies can be compensated for at the training or fine-tuning phase of an NLP model's pipeline by specifically including knowledge of recursive noun phrases. Once this supplementary training was undertaken, the models achieved *'strong zero-shot performance on an extrinsic Harm Detection [tasks]'*.

The researchers promise to release the code for this work at https://github.com/veronica320/Recursive-NPs.

Originally published December 16th 2021 – 17th December 2021, 6:55am GMT+2: Corrected broken hyperlink.

** GPT-3 Ada, which is the fastest but not the best of the series. However, the larger 'showcase' Davinci model is not available for the fine-tuning that comprises the later phrase of the researchers' experiments.*

† *My conversion of inline citations to hyperlinks.*

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More