How Much Do Language Models Copy From Their Training Data? Evaluating Linguistic Novelty in Text Generation Using RAVEN

R. Thomas McCoy,[1] Paul Smolensky,[2,3] Tal Linzen,[4] Jianfeng Gao,[2] and Asli Celikyilmaz[5]

[1]Princeton University, [2]Microsoft Research, [3]Johns Hopkins University, [4]New York University, [5]Meta AI

1 Overview

- **Question:** To what extent do language models (LMs) generate novel text, as opposed to copying text from their training sets?
- **Main finding:** Models show an impressive degree of novelty (albeit with occasional examples of extensive copying).
  - Thus, a reasonable default assumption is that LM-generated text is novel.
  - But when we need to be certain about novelty—e.g., when studying abstract abilities—we must explicitly check.

2 Motivation

- Why should we care if LM text is novel?
- Answer: Important for evaluating a model's abstract abilities.
- Only novel text can serve as evidence for abstraction!

Example

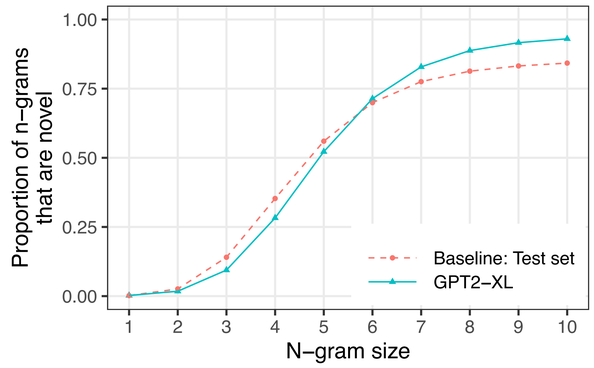
- Suppose we are evaluating if an LM captures the abstract property of *coherence*
- Situation 1: The LM's text is **coherent** and **novel**
  - → Evidence for an abstract notion of coherence.
- Situation 2: The LM's text is **coherent** but **copied**
  - → Not evidence that the LM has captured coherence.
  - → The credit for coherence belongs to the human who originally composed the text, not to the LM that copied it.

3 Approach

- Analyzed text sampled from GPT-2 using top-40 sampling.
  - GPT-2 = largest model for which training set was available.
  - Baseline: Human-generated text from GPT-2's test set.
- Checked for overlap with the training set to determine novelty of n-grams and syntactic structures.
- See paper for other models & decoding methods!

4 N-gram novelty

- **Small n-grams:** Less novel than the baseline
- **Medium & large n-grams:** More novel than the baseline

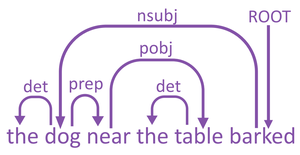


- **Supercopying:** in rare cases, models copy passages over 1,000 words long
  - Typically passages that appeared many times in training

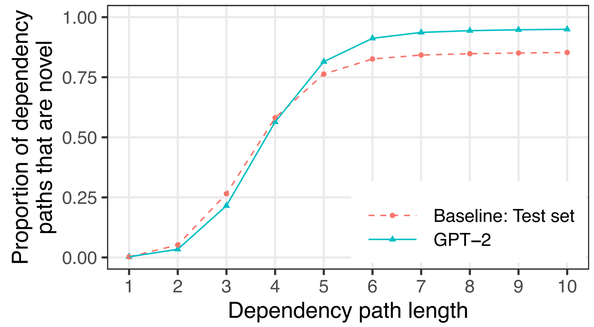
5 Syntactic novelty

- Analyzed novelty of labeled paths in a dependency tree:

Example length-3 path: [the_{det}, $table_{pobj}$, *near*]



- Similar trends as for n-grams:



6 Manual analysis of specific phenomena

- **Broad question:** Is novel LM text linguistically well-formed?
  - Evaluated with respect to morphology, syntax, and semantics.
- **Summary:** Strong performance for morphology and syntax; errors are fairly common for semantics.

	Morphology	Syntax	Semantics
Correct	0.96	0.94	0.80
Incorrect	0.02	0.01	0.11
Unclear	0.02	0.05	0.09

Examples: Morphology of novel words

- Novel plurals:
  - Correct (72/74): ***Brazilianisms, Fowleses, …***
  - Incorrect (2/74): ***1099es, SQLes***
- Some other well-formed examples:

IKEA-ness	***bagshare***
Smurfverse	***nonneotropical***
quackdom	***Disqusiquette***
Thirteenthly	***hill-elves***

Examples: Syntactic context of novel words

- GPT-2 usually places novel plurals in syntactically-appropriate contexts (e.g., with proper agreement, underlined)
  - ***FOIA-requesters*** *who think an agency has a good reason for withholding information are not always given a second opportunity to press their case.*
  - *The* ***Sarrats*** *were lucky to have her as part of their lives.*

Examples: Semantic context of novel words

- Some simple errors (red), some impressive cases (green)
  - *…adding an optional "****no-knockout****" version…so you can actually be knocked out*
  - *The concept of '****co-causation****', in which effects are thought to be caused by causes that act in parallel*

7 Conclusion

- **Summary:** LM-generated text is usually novel, both for n-grams and syntax.
  - Evidence for a range of linguistic abstractions (constituent structure, dependency structure, morphological processes…)
- **More recent models?**
  - RLHF & new prompting techniques might encourage copying.
- **Broader point:** To understand models' abilities, we must consider their training data & how they generalize beyond it.

This material is based upon work supported by the NSF GRFP. All opinions are our own.