TL;DR: Mining commonsense knowledge from corpora suffers from reporting bias: over-representing the rare at the expense of the trivial [1]. We study to what extent pre-trained language models overcome this issue. We find that while their generalization capacity allows them to better estimate the plausibility of frequent but unspoken actions, outcomes, and properties, they also tend to overestimate that of the very rare, amplifying the bias that already exists in their training corpus.

I. Actions and Events:

[1]: Discrepancy between corpus occurrences and real-world frequency: trivial actions (breathing) are rarely mentioned, while noteworthy events (murdering) are discussed disproportionately.

Are language models better at estimating frequencies?
• Estimate real-world and corpus frequency of various actions
• Compute LM scores of “The person is ____” as a proxy for frequency

LMs provide a worse estimate of action frequency due to overestimating very rare actions.

References:

2. Events Outcomes:

[1]: an event outcome is more likely to be mentioned in text if it’s not certain. Are language models better at estimating event outcomes?
  $P_{LM}[cause] \leq P_{LM}[effect]$ (So I As a result, I As one would expect) (effect)

LMs predict both expected outcomes as well as sensational and unlikely outcomes.

The man turned on the faucet. As a result,

- the water in the bathtub began to flow
- the man’s blood was sprayed everywhere

[1]: extract typical outcomes from constructions that indicate a speaker’s expectation about the world was not met.

The man turned on the faucet, yet, for some reason, water did not flow from the spout.

- Zero-shot LM-based model for COPA with disconfirmed expectations typically performed worse.
  $P_{LM}[cause] \leq P_{LM}[effect]$ (yet but I however, surprisingly I for some reason I somehow?) [negated effect])

- This is likely due to the insensitivity of LMs to negation [3].

3. Properties:

[1]: people are more likely to state unusual properties of a concept (blue pencil) than usual ones (yellow pencil).
[4]: LMs learn associations between concepts and their properties indirectly.

LMs tend to over-generalize and often confuse mutually exclusive values.

• Example: predicting masked colors in sentences from Wikipedia:

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Corpus</th>
<th>BERT-L</th>
<th>RoBERTa-L</th>
<th>BERT-L+FT</th>
<th>RoBERTa-L+FT</th>
</tr>
</thead>
<tbody>
<tr>
<td>The ___ banana is tasty.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The ___ apple is sweet.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The ___ cat is cute.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The ___ dove is beautiful.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The ___ cow eats grass.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The ___ dog runs in the park.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Further training on corpus “ground truth” substantially improves performance.