Lexical Generalization Improves with Model Size And Longer Training

Elron Bandel, Yoav Goldberg, Yanai Elazar

Larger language models trained for **longer** are **better,**
But the **dev set performance does not reflect it!**

1. Approach

- How do models asses the semantic relatedness of sentences with similar words?
- We analyze the adoption and avoidance of **lexical overlap heuristics** in models of different sizes and in different training phases.

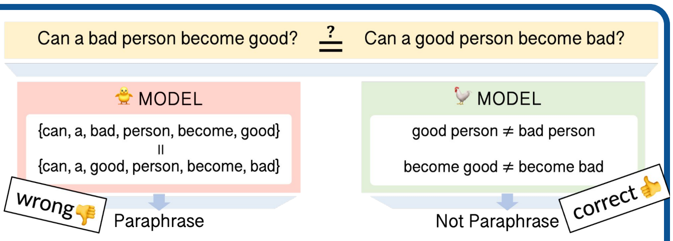

Can a bad person become good? ≟ Can a good person become bad?
🐥 MODEL
{can, a, bad, person, become, good}
=
{can, a, good, person, become, bad}
wrong👎
Paraphrase
🐓 MODEL
good person ≠ bad person
become good ≠ become bad
Not Paraphrase
correct👍

Figure 1: The lexical overlap heuristic - if two sentences have similar words they must be semantically similar! (wrong)

2. Measuring Heuristic Use

✔ = Performance on test-set **consistent** with heuristic

✘ = Performance on test-set **inconsistent** with heuristic

HEUR = ✔ − ✘

Higher HEUR values indicate high use of the lexical overlap heuristic

3. Test-sets

New Dataset!

3.1 Text-pair Classification

Classification test-sets that control the lexical overlap heuristic

PAWS (Paraphrase Identification)

HANS (NLI)

But they are also:

- Synthetic
- Similar to each other

3.2 ALSQA: a High Lexical-overlap Reading Comprehension Test

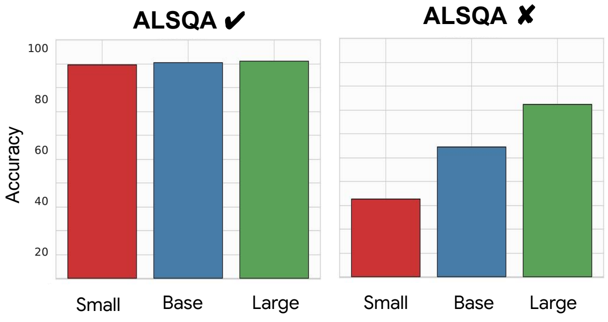
We asked crowdworkers to rewrite SQuAD2.0 questions to questions with high lexical overlap with the context passage.

- **Answerable questions** are consistent with the lexical overlap heuristic ✔
- **Unanswerable questions** are not consistent with the lexical overlap heuristic ✘

Passage	Question	Answerable
...compacts like the 1974 Mustang I were a prelude to the DOT "downsize" revision of vehicle categories. By 1977, GMś full - sized cars reflected the crisis. By 1979, virtually all " full - size " American cars had shrunk, featuring smaller engines and smaller outside dimensions.	By which year did full sized American cars shrink to be smaller ?	A ✓
	What vehicle category did Chrysler change to in 1977 ?	NA ✗

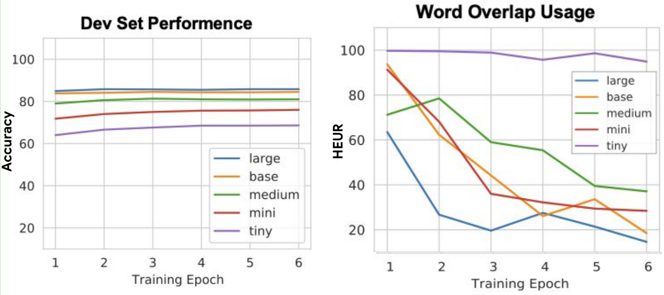
Conclusion I: Pretrain **Larger** Models

While models of different sizes perform equally on the subset consistent with the heuristic (✔), in the inconsistent subset (✘) **larger models generalize better and are less likely to adopt the heuristic**.



Conclusion II: Finetune Them For **Longer**

When training models **longer,** they tend to **abandon the use of lexical overlap heuristics**.



... while maintaining similar performance on the standard test

Remaining questions:

What other heuristics models employ for prediction?

How the size of the model and fine-tuning duration affect the abandonment of lexical heuristics?