NLU Lab: Paper Reading (14 Feb 2024)

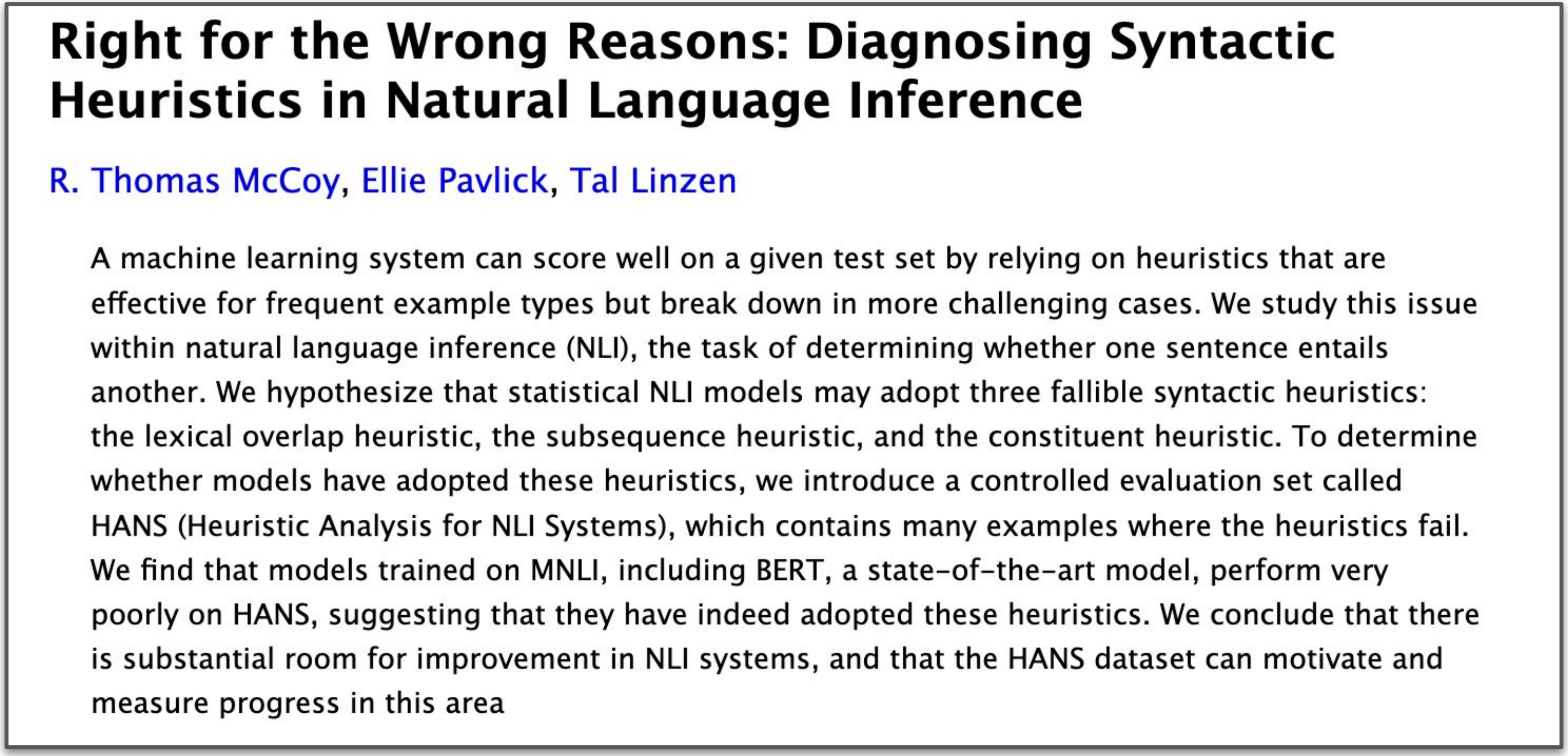
Today:

- What is a (NLP) paper?
- Why are they written & read
- How to read them (effectively)

Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference
(McCoy et al. 2019)

arXiv: 1902.01007

What can you learn from title + abstract?



Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference

R. Thomas McCoy, Ellie Pavlick, Tal Linzen

A machine learning system can score well on a given test set by relying on heuristics that are effective for frequent example types but break down in more challenging cases. We study this issue within natural language inference (NLI), the task of determining whether one sentence entails another. We hypothesize that statistical NLI models may adopt three fallible syntactic heuristics: the lexical overlap heuristic, the subsequence heuristic, and the constituent heuristic. To determine whether models have adopted these heuristics, we introduce a controlled evaluation set called HANS (Heuristic Analysis for NLI Systems), which contains many examples where the heuristics fail. We find that models trained on MNLI, including BERT, a state-of-the-art model, perform very poorly on HANS, suggesting that they have indeed adopted these heuristics. We conclude that there is substantial room for improvement in NLI systems, and that the HANS dataset can motivate and measure progress in this area

Kinds of NLP Papers

- Theory!

    “Prove transformers cannot learn to multiply arbitrary sequences in S_5”

- Empiricism!

    “Models rely on shallow heuristics to solve NLI tasks”

- Task solving!

    “Behold: the Transformer”

Empirical tests need benchmarks!

“Benchmark” = way to test on common ground

Good:

“Modifying model X by doing Y improves performance on Z”

Bad:

“Sentiment classification models work better on English than on Icelandic”

Why is the good example good?

Why is the bad example bad?

Sections & their purposes

From McCoy et al. (2019):

1. Introduction
2. Syntactic Heuristics
3. Dataset Construction
4. Experimental Setup
5. Results
6. Discussion
7. Augmenting training data with HANS-like examples*
8. Related Work
9. Conclusion

Where do they state their hypothesis?

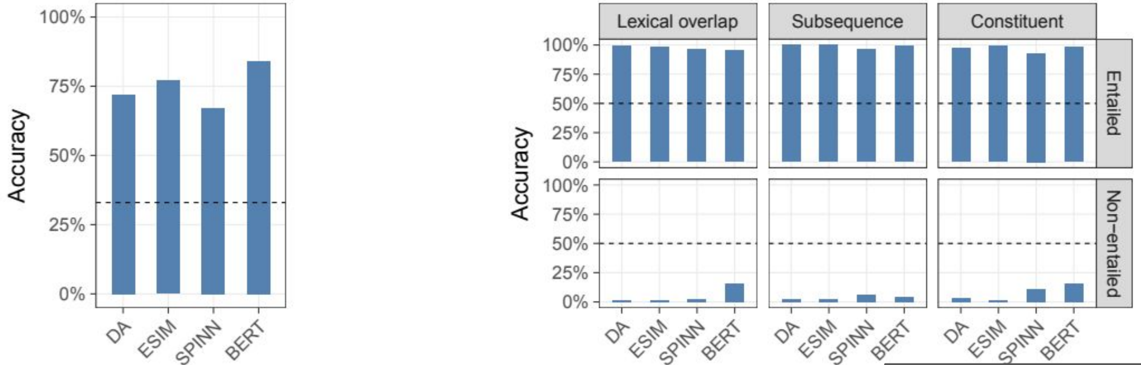
Why are (3) and (4) different sections?

What's going on with (7)?

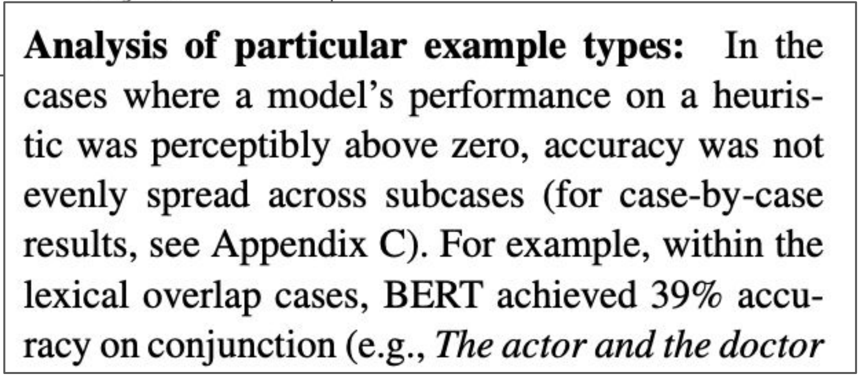
What's missing?

What is their hypothesis?

Quantitative & qualitative explanations



(a) (b)



Analysis of particular example types: In the cases where a model's performance on a heuristic was perceptibly above zero, accuracy was not evenly spread across subcases (for case-by-case results, see Appendix C). For example, within the lexical overlap cases, BERT achieved 39% accuracy on conjunction (e.g., *The actor and the doctor*

Meta Questions

- Why is this paper relevant?
- Who is it written for?
- Why are the authors writing it?
- Why are ***you*** reading it?