Aggregating analyses in surveys and toolkits

Each analysis paper asks a very specific question.

How do we ask, *what does the field currently know about BERT?*

Answer: meta-studies compiling results

Analysis Methods in Neural Language Processing: A Survey

Yonatan Belinkov[12] and **James Glass**[1]

[1]MIT Computer Science and Artificial Intelligence Laboratory
[2]Harvard School of Engineering and Applied Sciences
Cambridge, MA, USA
{belinkov, glass}@mit.edu

A Primer in BERTology: What we know about how BERT works

Anna Rogers, Olga Kovaleva, Anna Rumshisky
Department of Computer Science, University of Massachusetts Lowell
Lowell, MA 01854
{arogers, okovalev, arum}@cs.uml.edu

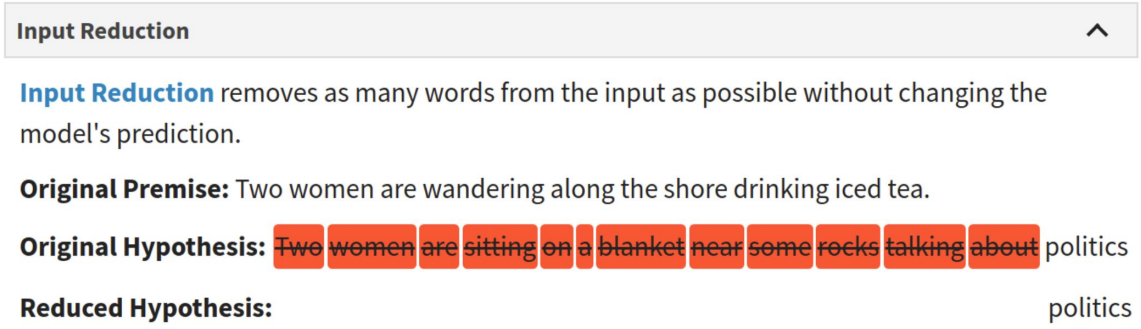
Aggregating analyses in surveys and toolkits

How do we ask, *what can I easily find out about **my** model?*

Answer: interpretability toolkits!

AllenNLP Interpret:
A Framework for Explaining Predictions of NLP Models

Eric Wallace[1] **Jens Tuyls**[2] **Junlin Wang**[2] **Sanjay Subramanian**[1]
Matt Gardner[1] **Sameer Singh**[2]
[1]Allen Institute for Artificial Intelligence [2]University of California, Irvine
ericw@allenai.org, sameer@uci.edu



Takeaways

Neural models are complex, fascinating objects that we don't currently understand, but we're making strides to understand them better!

A wide variety of analysis methods have been developed, for:

- Understanding a model's behavior on specific phenomena
- Understanding what a model learns about a topic or task
- Understanding what seemingly innocuous input changes make a model fail
- Many other things, with more coming every day!

These methods can be integrated into your future NLP projects!