Using Control Tasks To Study the Effectiveness of Linguistic and Cognitive Probing Models

Anand Bollu | Advised By: Prof. Leila Wehbe & Mariya Toneva

Overview

- Control tasks help determine whether neural language representations capture particular information of interest
- We propose and evaluate a novel construction of control tasks motivated by permutation tests to better contextualize probe selectivity

Background

What is a Probe?

- Probes are models that aim to reveal whether a language representation make certain task labels (e.g. part-of-speech) accessible

What are Control Tasks?

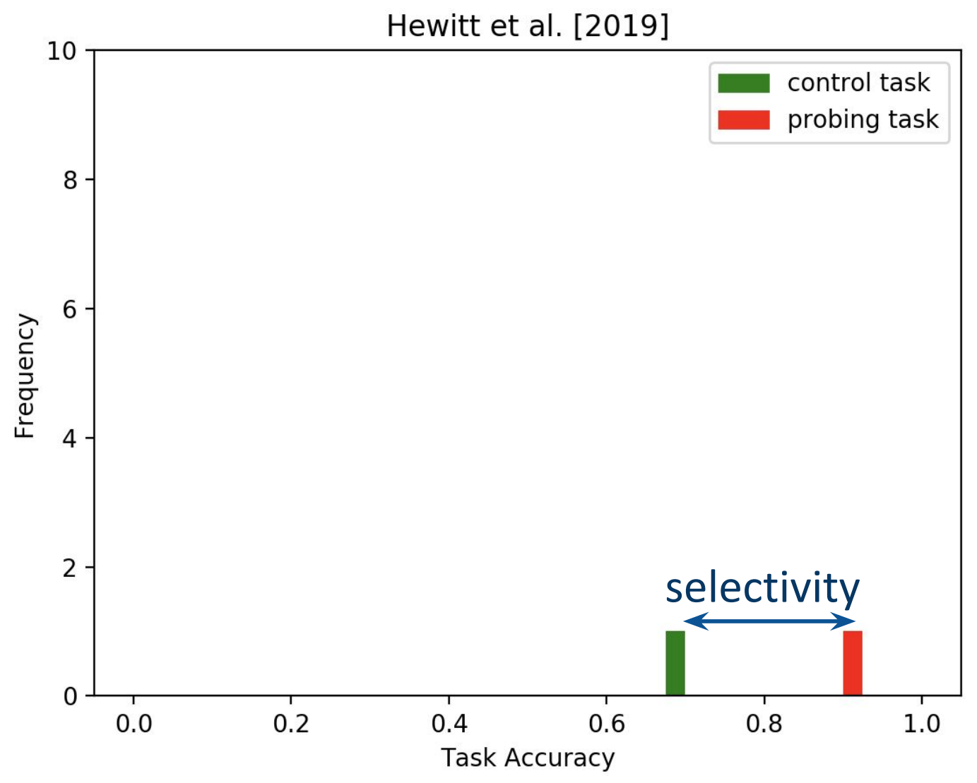
- Hewitt et al. [2019] introduced control tasks to identify whether high probing task accuracy is observed because:
  - **A.** the language representation encodes task-relevant information, or
  - **B.** the probe is expressive enough to learn the task by itself given sufficient data

Sentence	The	brown	squirrel	...
Probing Task Labels (POS)	DT	JJ	NNP	...
Control Task Labels	NNP	DT	JJ	...

Control tasks have the same input and output space as a probing task and define random behavior. They can only be learned by a probe that memorizes the mapping.

selectivity = [probing task accuracy] - [control task accuracy]

- Need a way to assess what a ***good*** selectivity threshold is

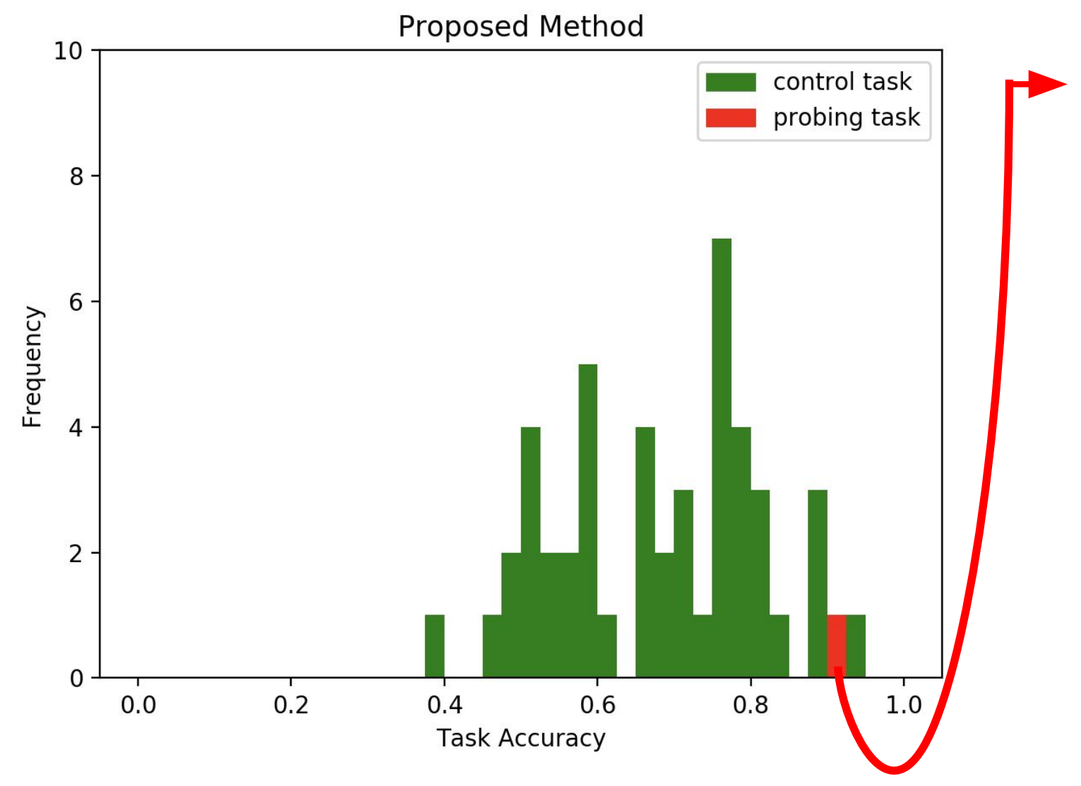


Proposed Method

- Permutation tests can be used to measure how likely it is that the probing task accuracy was obtained by chance.
- We structure our experiment in terms of null & alternative hypotheses:

H_0: **Probe unable to learn random mappings of inputs to outputs**

H_1: **Probe able to learn random mappings of inputs to outputs**

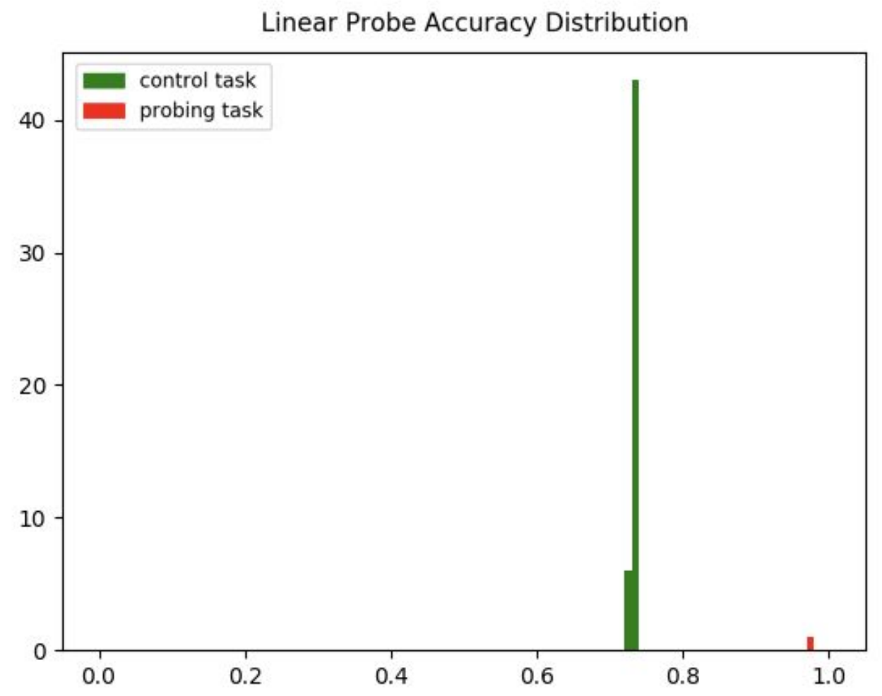


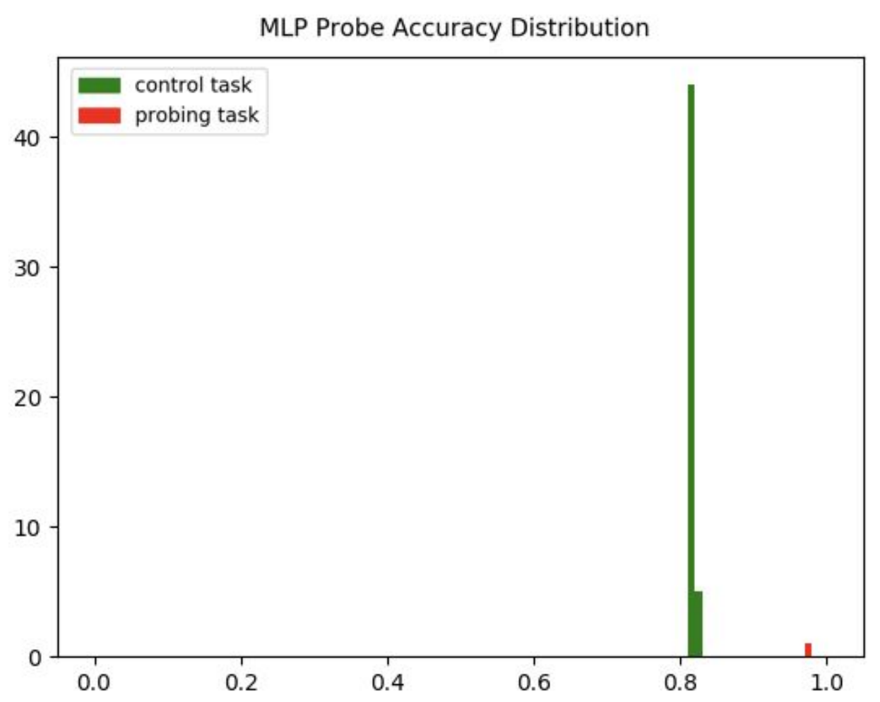
Record p-value for probing task observation

- p-value ≤ $\boldsymbol{\delta}$: Reject H_0 in favor of H_1
- p-value > $\boldsymbol{\delta}$: Fail to reject H_0

1. Record probing task accuracy
2. Construct several control tasks by permuting the labels from the original probing task
3. Record control task accuracies
4. Compute **p-value** and compare it to a significance level determined a priori ($\boldsymbol{\delta}$) to see if the probe could be expressive enough to learn the task

Results



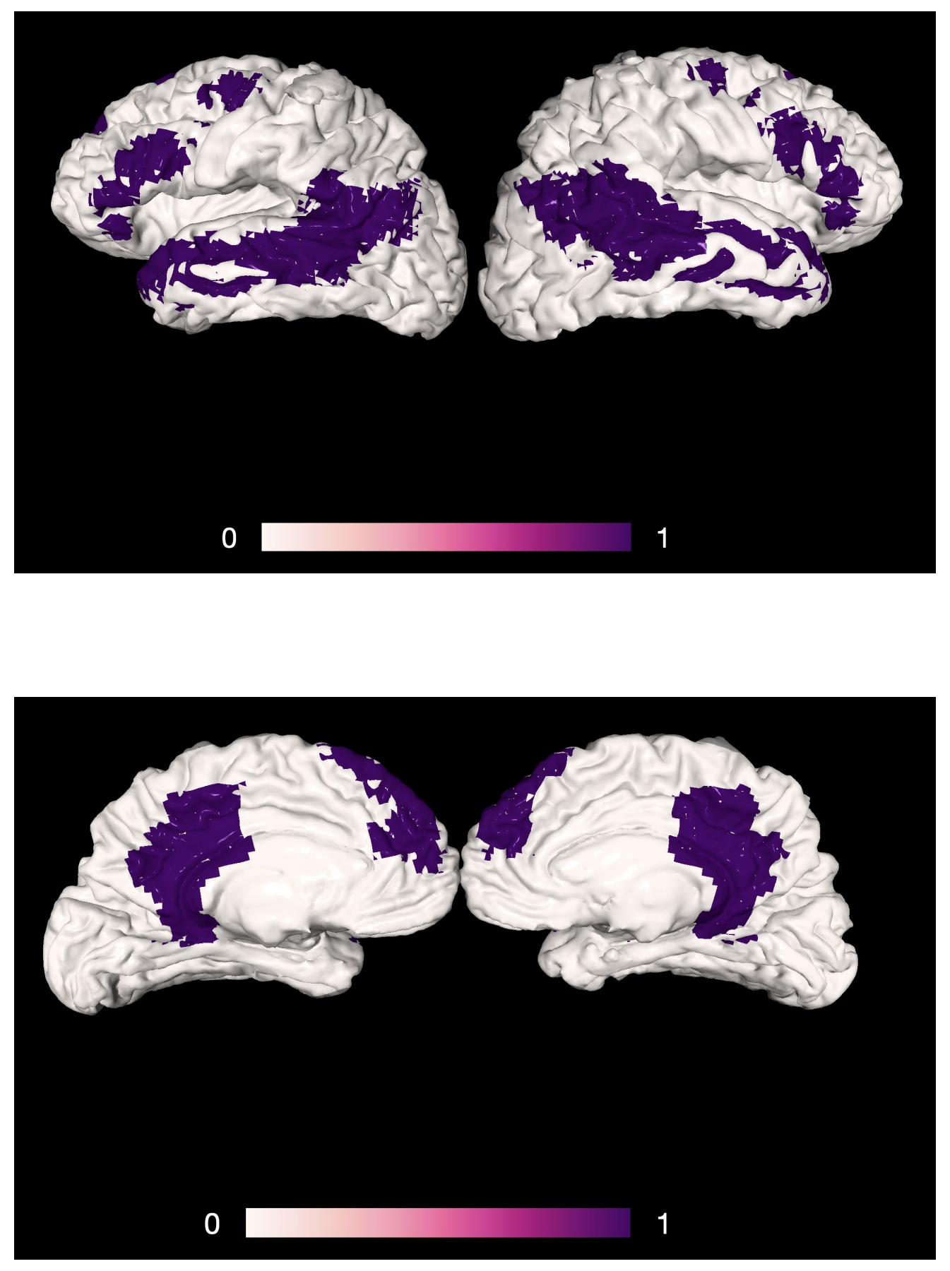


What We Found:

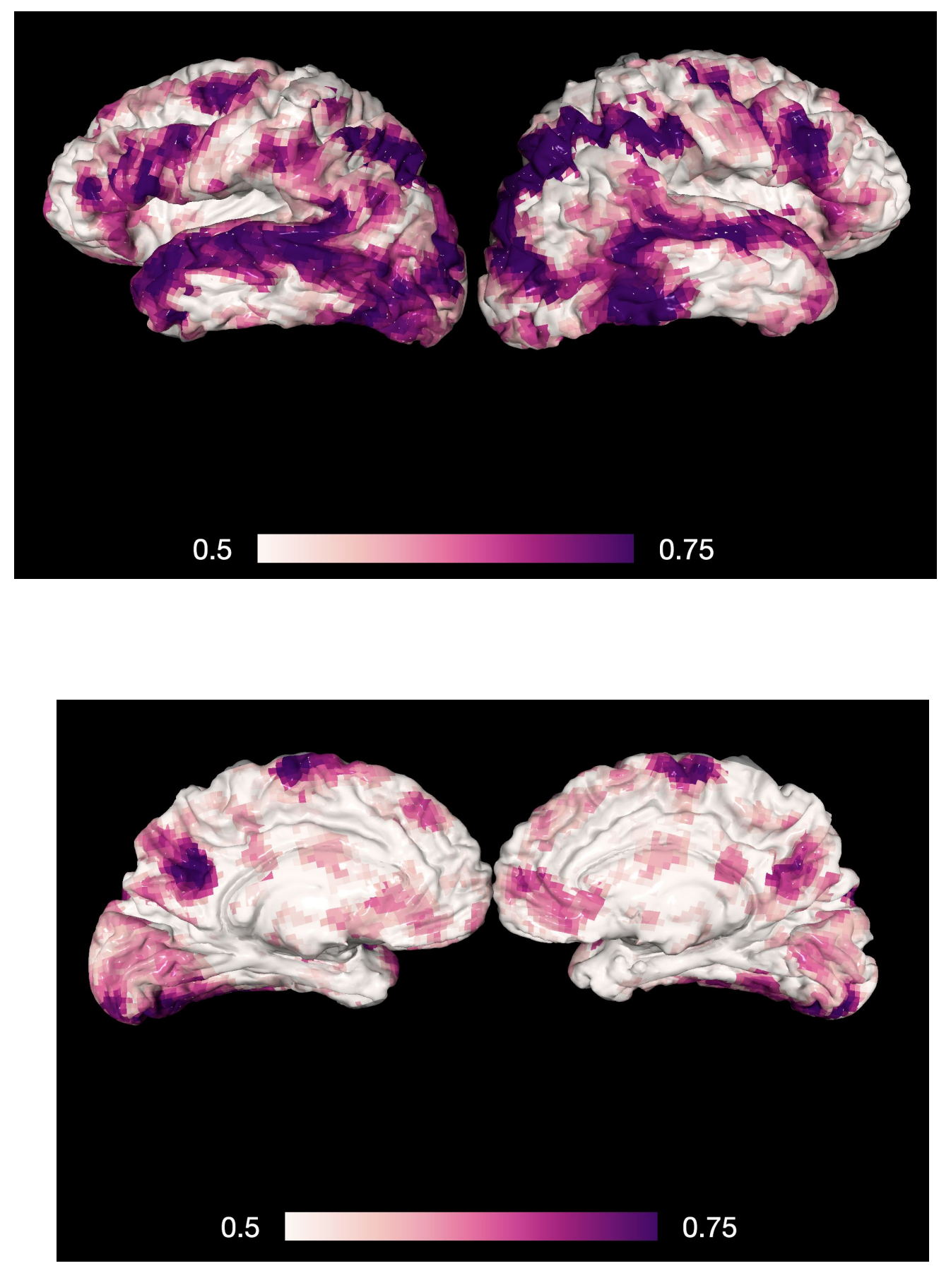
- Our approximation of the accuracy distribution possessed much lower variance than initially expected
- Permuting the label sets made it difficult to preserve the latent structure within the original labeled data

Current & Future Work

- Probes can also be used to predict information shared between neural language representations and brain recordings
- Interested in the effect of varying probe complexity on BERT-Brain probing
- Brain activity recorded with fMRI and MEG while participants read a chapter of Harry Potter (Wehbe et al. [2014a, 2014b])
- Able to probe brain recordings from the original text using linear models (Toneva et al. [2019])



Regions of Interest (highlighted in purple) are parts of the brain that are consistently activated during language processing



Linear probe voxel-by-voxel accuracies (Low accuracy in white, High accuracy in purple)

- Interested in recording MLP probe performance in this setting and comparing it to the linear probe
- Need to define how factors other than performance (e.g. selectivity, simplicity) influence probe selection

References

- John Hewitt and Percy Liang. 2019. Designing and Interpreting Probes with Control Tasks. In Proceedings of EMNLP.
- Toneva M., and Leila Wehbe. 2019. Interpreting and improving natural language processing (in machines) with natural language-processing (in the brain). Advances in Neural Information Processing Systems.
- Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., Mitchell, T., 2014a. Simultaneously uncovering the patterns of brain regions involved in different story reading Subprocesses. PLoS One 9, 1–19.
- Wehbe, L., Vaswani, A., Knight, K., Mitchell, T., 2014b. Aligning context-based statistical models of language with brain activity during reading. In: EMNLP