Using Roark-Hollingshead Distance to Probe BERT's Syntactic Competence

Jingcheng Niu[RH] Wenjie Lu[R] Eric Corlett[R] Gerald Penn[RH]
University of Toronto[R] Vector Institute[H]
Emails: {niu,luwenjie,ecorlett,gpenn}@cs.toronto.edu

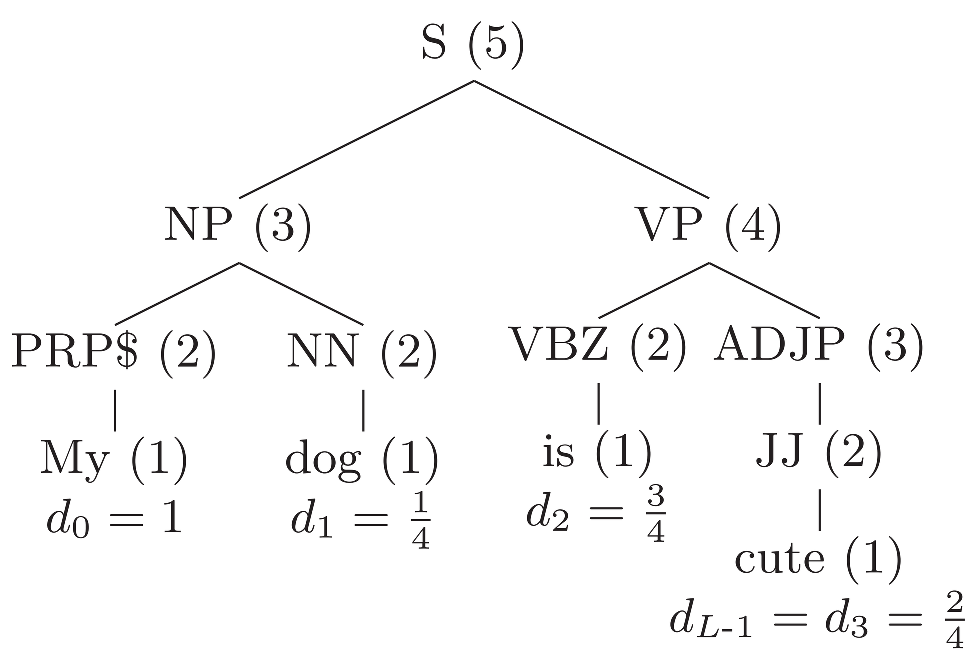
Code: https://github.com/frankniujc/rh_probe

CODE & DATA

Overview

- Probing BERT's general ability to reason about syntax is not simple.
- Performance-based probes suffer the criticism that the observed syntactic knowledge is not obtained by the LM through pretraining, but rather emerges from the probe classifier itself. Parameter-free probe (Perturbed Masking) produces unimpressive results.
- Still, we want to measure the inferential capacity of the language model itself. E.g., to induce parse trees.
- **RH Probe**: an encoder-decoder-based probing architecture with two experiments (ablation probe & attack probe). Ablation study is still a valid way to interrogate the model.
- **Finding:** BERT's word embeddings contain important syntactic information, but this information alone is **not enough** to reproduce traditional syntactic representations (e.g. phrase structure) in their entirety.

RH Syntactic Distance & the RH Conjecture



An RH distance calculation example. The heights of nodes (h) are in brackets.

$$d_i = \frac{h(t_{i-1}, t_i) - 2}{h(r) - 1}$$

$$h(t_{\text{-}1}, t_0) = h(t_{L\text{-}1}, t_L) = h(r)+1$$

$$h(u, v) = h(u \cup v), \text{otherwise.}$$

Roark and Hollingshead (2008) conjectured that RH distance is sufficient to reconstruct the structure of an entire binary constituency tree.

We proved this conjecture.

Previous Work

Probability Probing

$P(\text{gramma.}) > P(\text{ungramma.})$?

- Probability is not a particularly good reflection of syntactic well-formedness.
- It also does not reflect the modern pretrain/finetune usage of language models.

Performance-based Probing

Hewitt and Liang (2019): "When a probe achieves high accuracy on a linguistic task ... can we conclude that the representation encodes linguistic structure, or has the probe just learned the task?"

Perturbed Masking

Uses a parameter-free approach:

- Mask up pairs of tokens.
- Calculate pairwise impact between tokens.
- Induce dependency trees with a matrix-based top-down parsing algorithm (MART).

Reappraising Perturbed Masking

MART's performance compared to different naïve baselines:

	MART	RB Tree	RH	Random
WSJ10	58.0	56.7	67.04	51.6
WSJ23	42.1	39.8	50.08	29.69

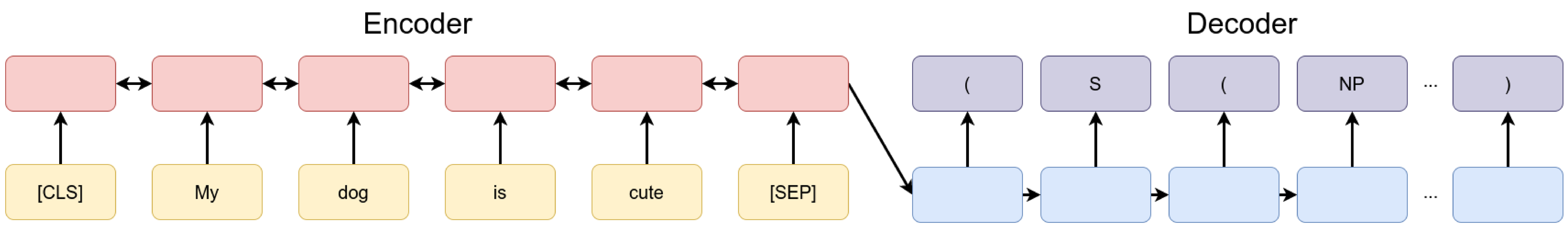
MART "performs" better when evaluated using RB trees as gold standard. It generates trees more closely resembling RB trees than constituency trees.

F1 MART vs.	Const. Tree	RB Tree
WSJ10	58.0	78.6
WSJ23	42.1	56.1

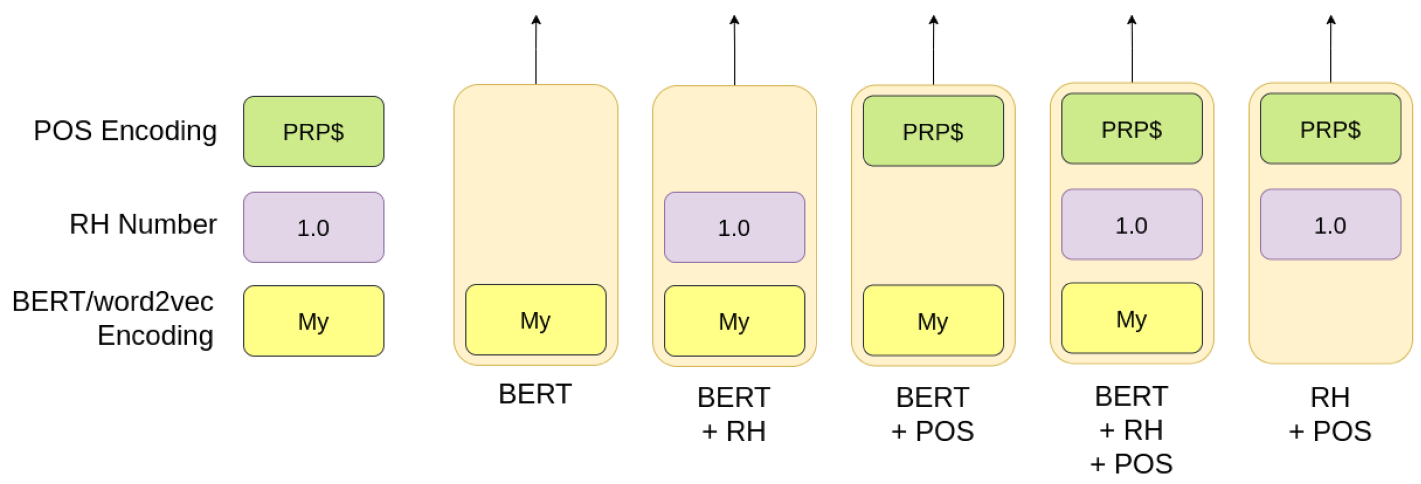
Wu et. al (2020):

"There is actually no guarantee that our probe will find a strong correlation with **human-designed syntax** ... What we found is the **'natural' syntax** inherent in BERT, which is acquired from self-supervised learning on plain text."

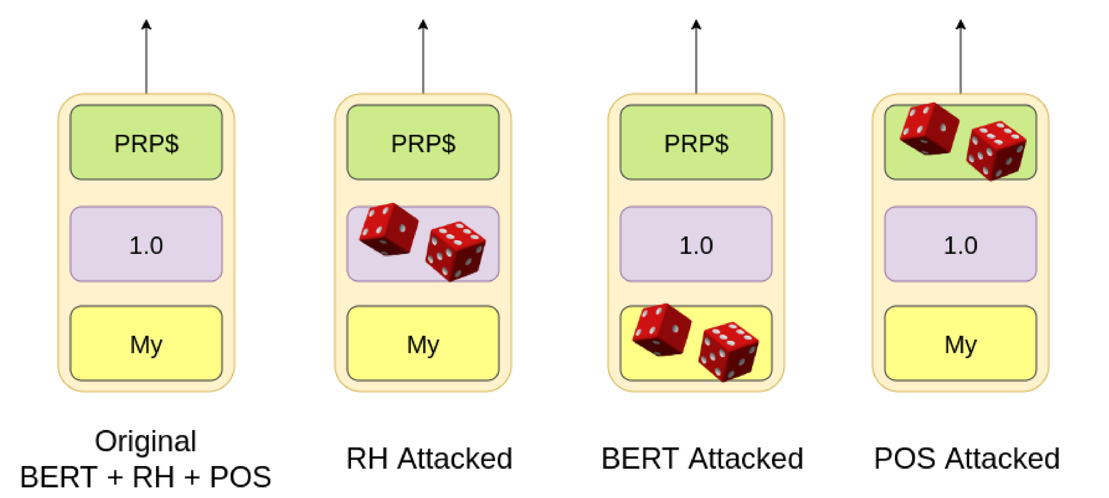
Experimental Design



RH Probe Encoder-Decoder Architecture. We use this simple probing architecture to avoid providing structural hints to the probe itself.

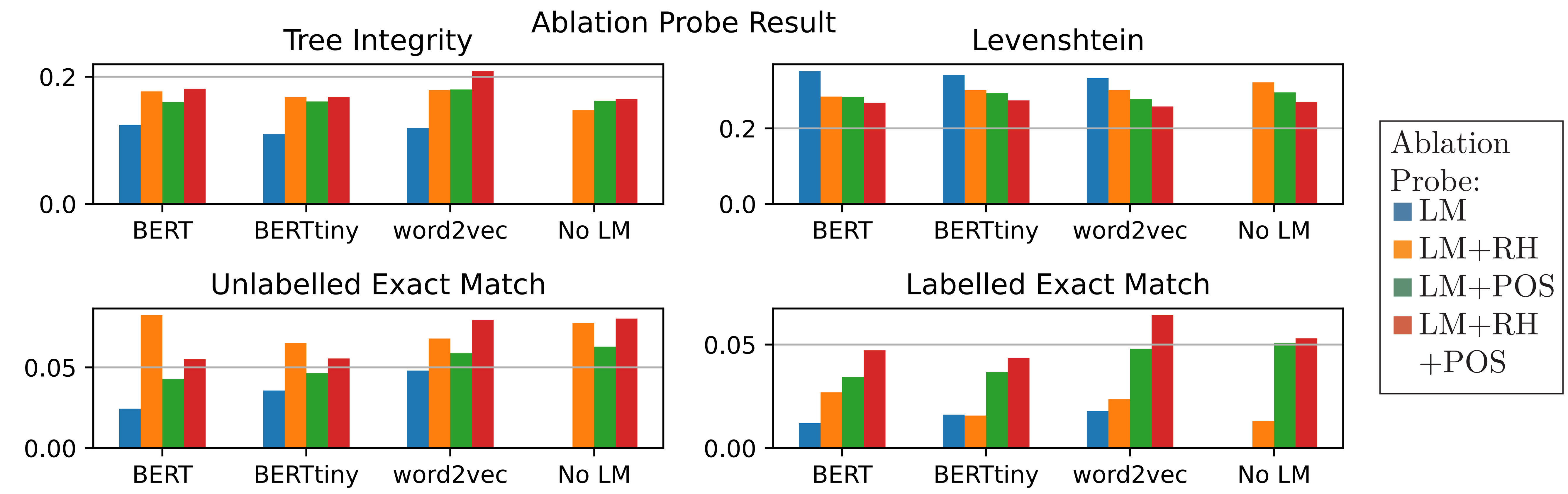


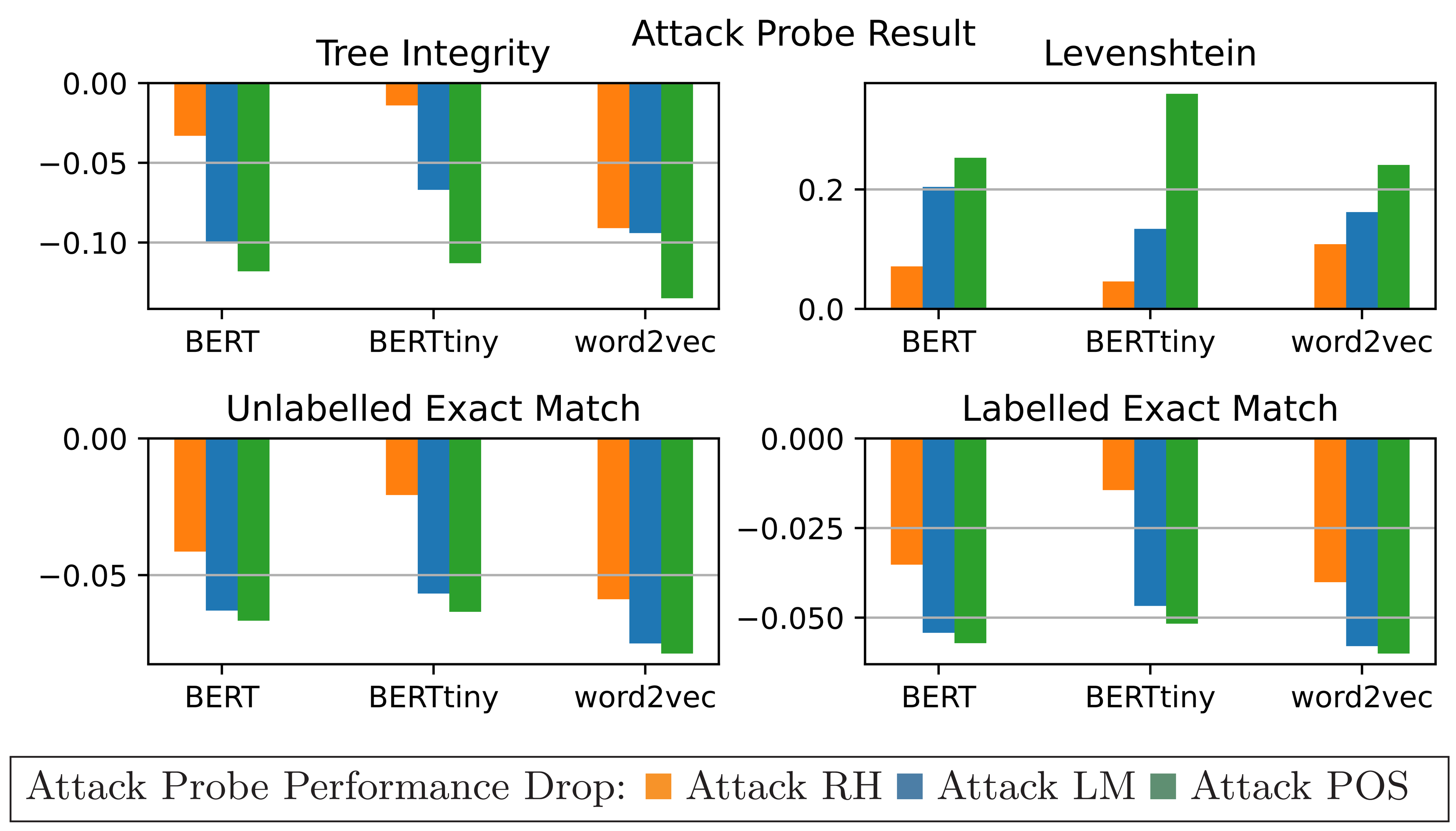
Ablation Probe: whether the addition of a feature type during training can increase the probe's performance.



Attack Probe: whether randomizing (attacking) certain features during testing can cause the performance to drop.

Experimental Results & Analysis





Ablation Probe

- Language models provide useful information for parsing.
- RH distance increases performance across the board – even on top of what POS provides.
- Better language model ≠ More syntactic knowledge.

Attack Probe

- Higher dimensionality = Easier to extract.
- Better language model = Easier to extract.