Probing BERT in Hyperbolic Spaces

Boli Chen[1,3]*, Yao Fu[2]* (*Equal Contribution)
Guangwei Xu[1], Pengjun Xie[1], Chuanqi Tan[1], Mosha Chen[1], Liping Jing[3]

[1]Alibaba Group
[2]University of Edinburgh
[3]Beijing Jiaotong University

ICLR 2021

Motivation

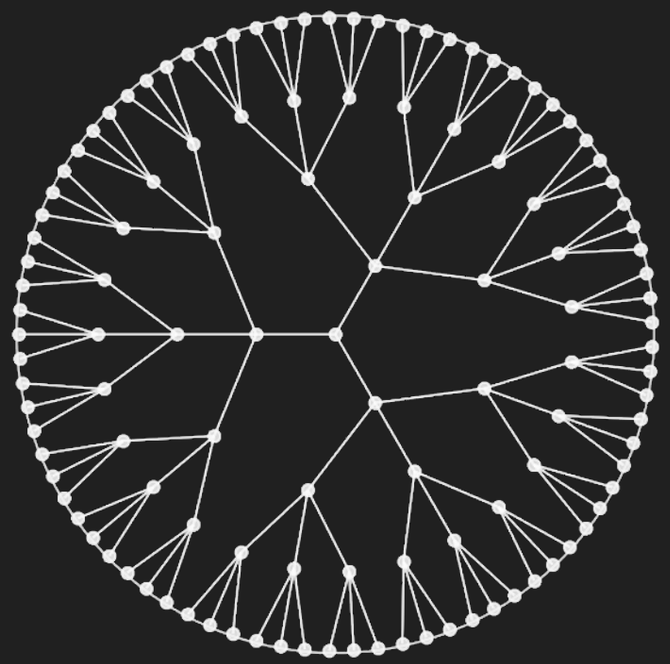
- To discover linguistic information encoded in contextualized representations

  - BERT embeddings ⟶
    - Syntax
    - Sentiment

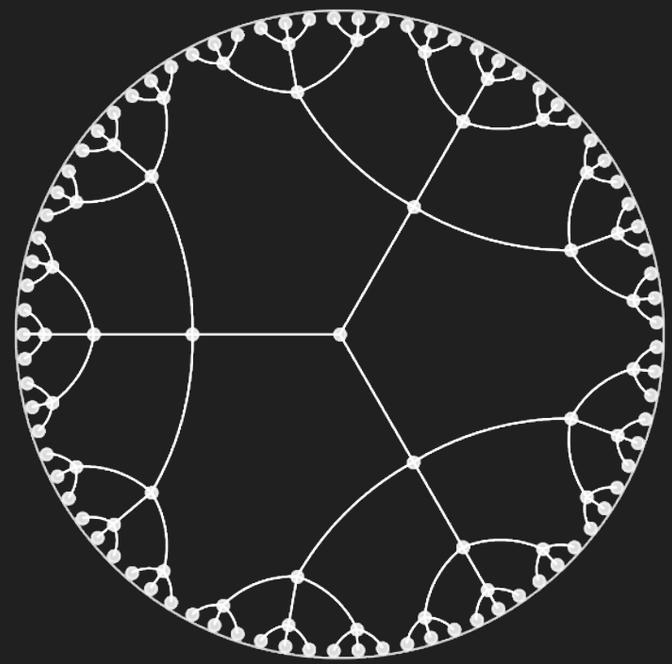
- Previously: structured probes to discover a Euclidean subspace where squared Euclidean dist. approx. tree dist.

- This work: Hyperbolic subspaces better encode/ recover tree/ hierarchical information from BERT

Why Hyperbolic Geometry?



Euclidean tree



Poincaré tree

- The volume of the Poincaré ball grows exponentially with its radius, similar to #children grows exponentially with tree depth (v.s. polynomially in Euclidean).
- The hyperbolic spaces have better inductive bias for capturing hierarchical information (see related work for more evidence).

Poincaré Probe

Project and exponential map

$$p = \exp_0(Ph)$$

Hyperbolic space — Bert embedding space

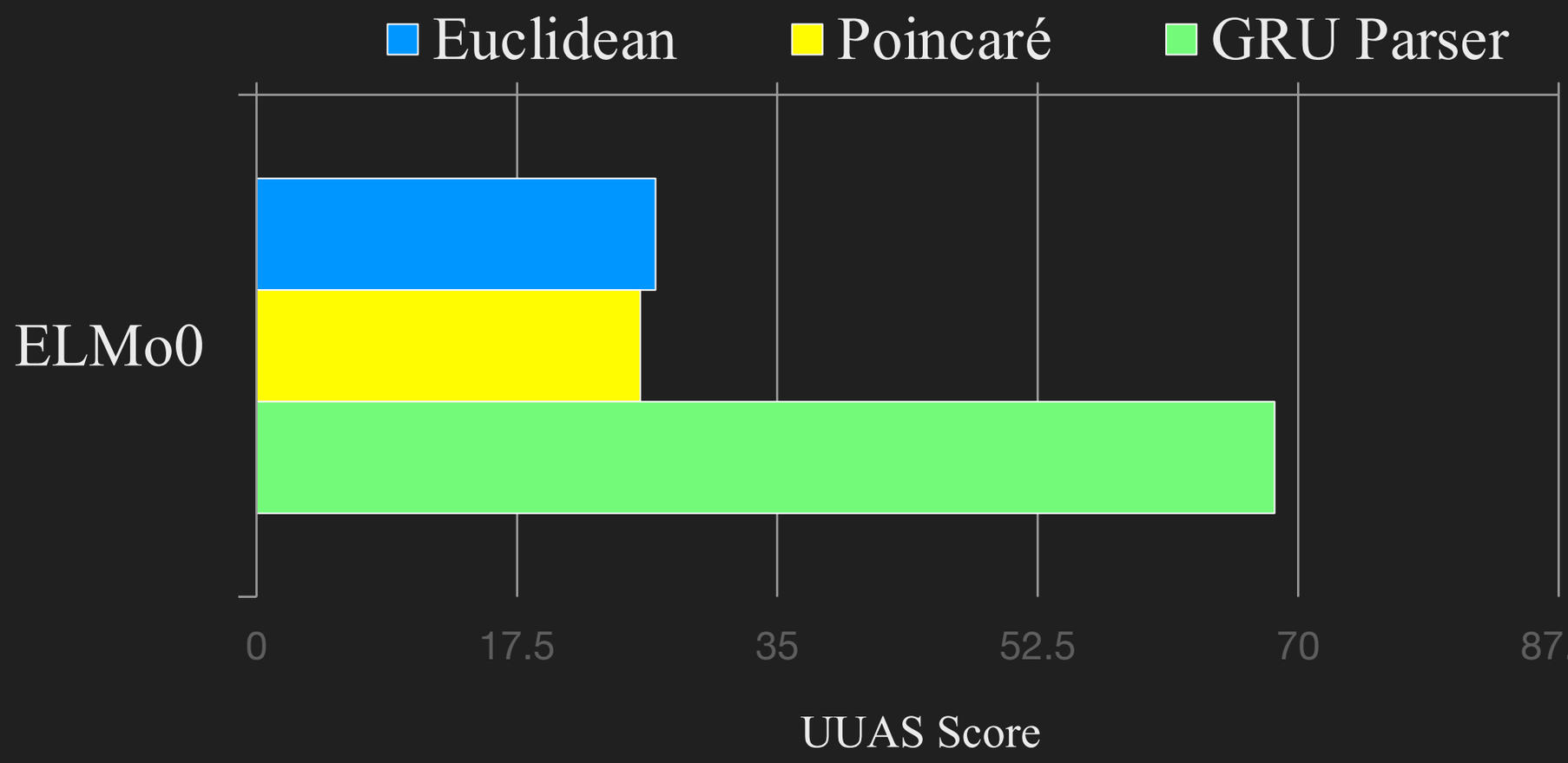
Proj matrix — Hyperbolic linear projection

$$q = Q \otimes p$$

Hyperbolic syntax/ sentiment subspace

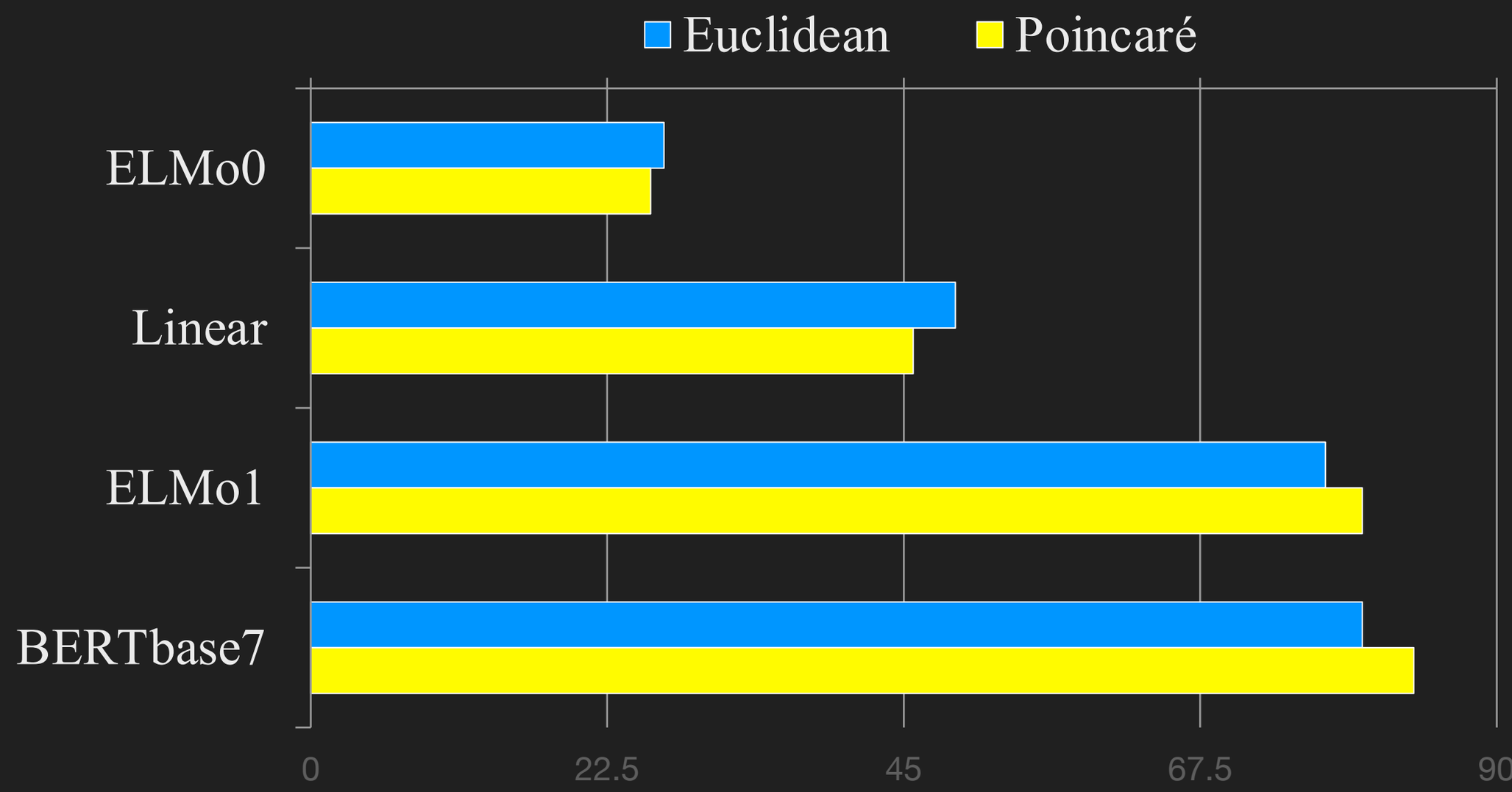
- Syntax subspace: hyperbolic distance approx. tree distance
- Sentiment subspace: hyperbolic distance approx. word polarity

Comparison between the Poincaré and Euclidean probes



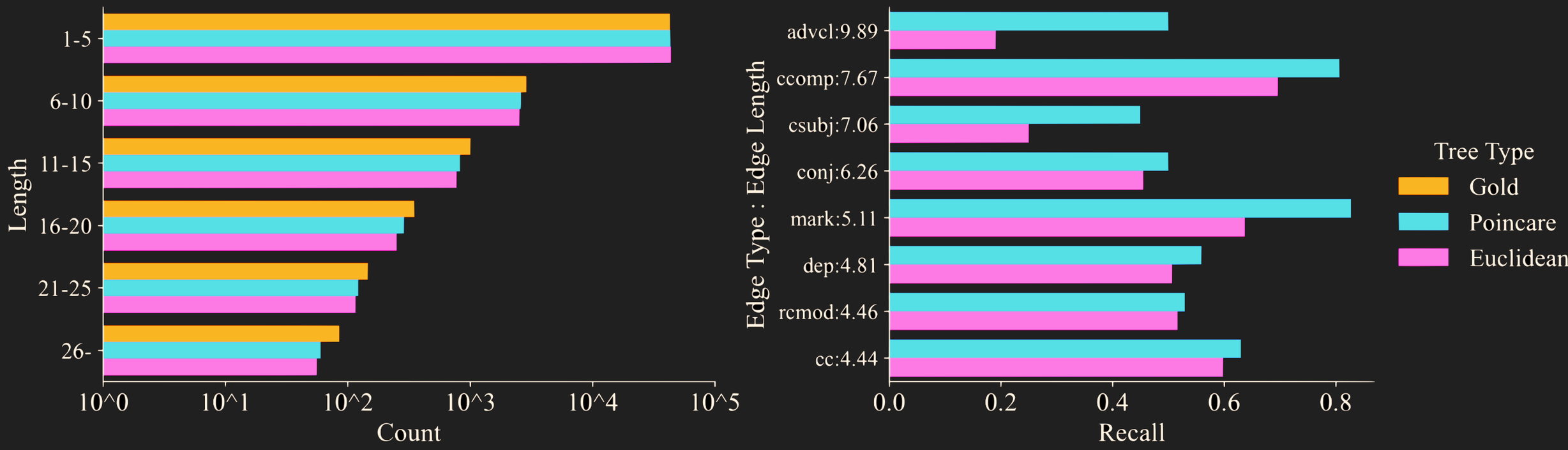
- How to evaluate probes and differentiate them with parsers?
- Evaluation: probe sensitivity
- For embeddings do not contain syntax information (like ELMo0), a probe cannot assign high scores, while a parser should

Comparison between the Poincaré and Euclidean probes



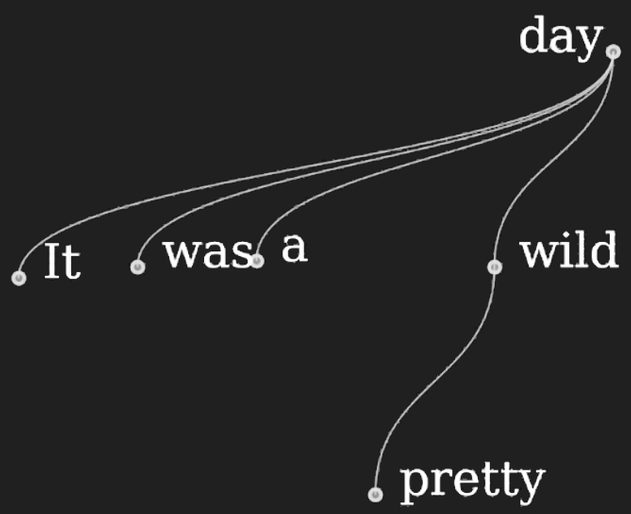
- Sensitivity cont': for embeddings contain syntax like BERTbase7, a probe should accurately recover the parsing scores, rather than underestimating
- Euclidean probes tend to underestimate deeper trees and longer edges

Comparison between the Poincaré and Euclidean probes

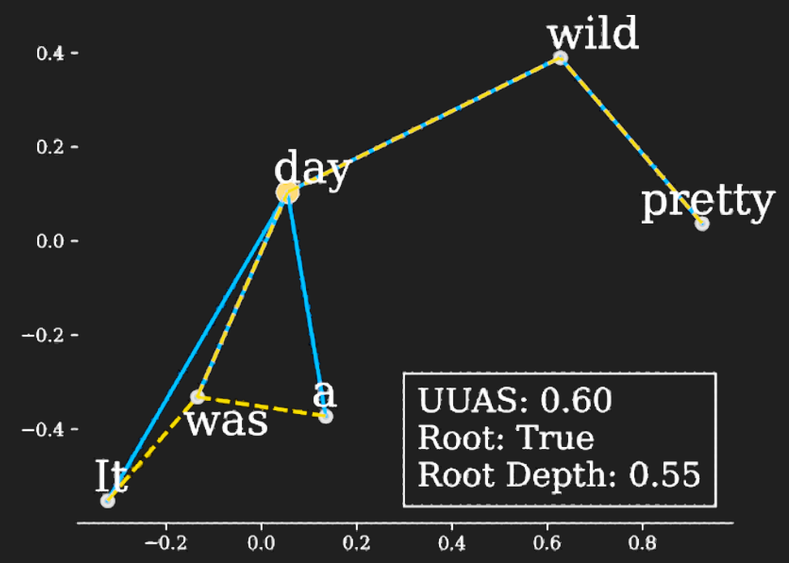


- Left: Poincaré recovers length dist. closer to gold
- Right: Poincaré better recovers longer edge types

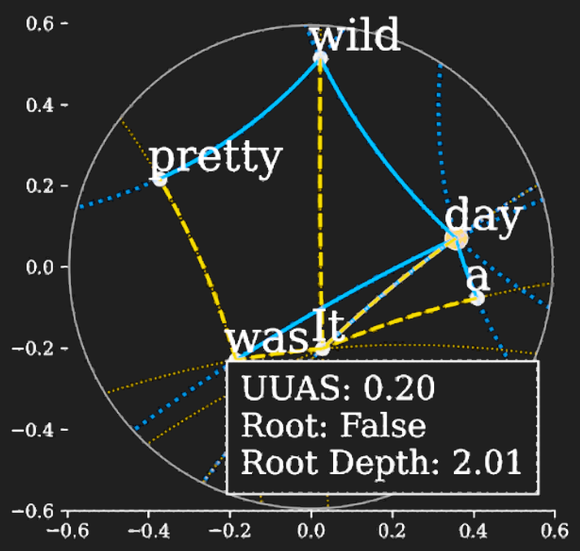
PCA projection of dependency trees



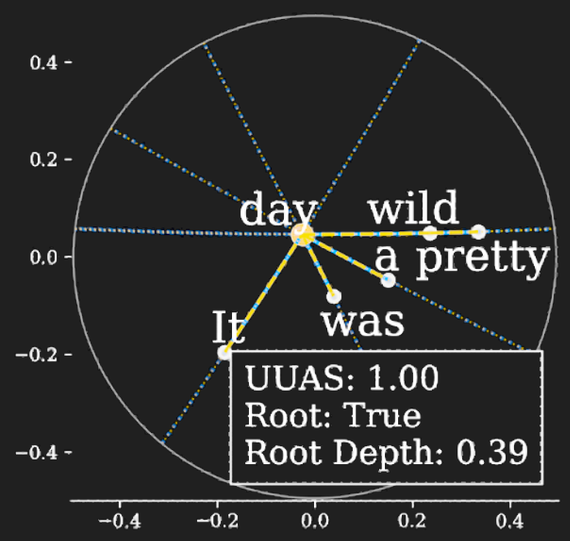
Syntax tree



Euclidean probe: BERTBASE7



Poincaré probe: ELMo0

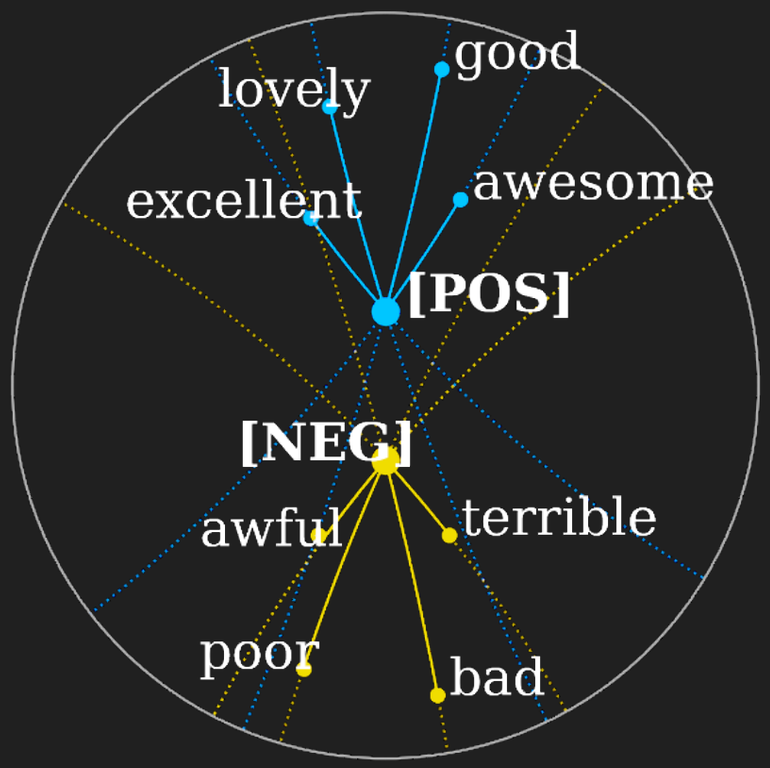


Poincaré probe: BERTBASE7

Probing Syntax

- See paper for more syntax results

Probing Sentiment



- Sentiment words embedded in a Poincaré ball.
- Hierarchy is defined as the sentiment polarity.
- We assume two meta [POS] and [NEG] embeddings at the highest level.
- Words with stronger sentiments are closer to their corresponding meta-embeddings.

Objective functions:

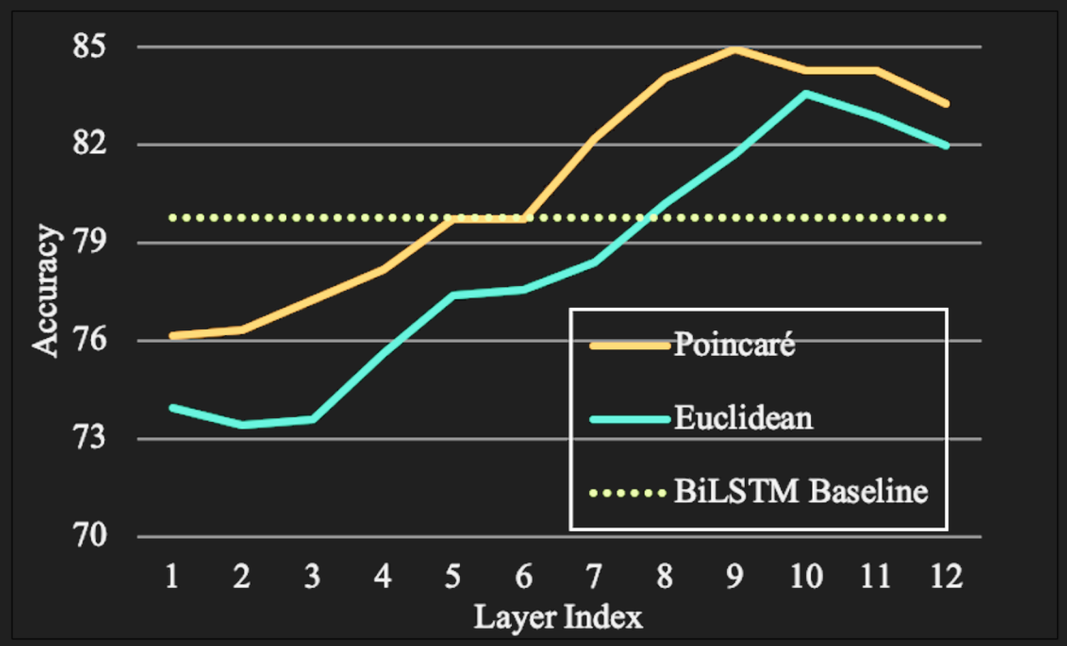
$$l_{pos} = \sum_{i=1}^{t} d_{\mathbb{D}^k}(\boldsymbol{q}_i, \boldsymbol{c}_{neg})$$

$$l_{neg} = \sum_{i=1}^{t} d_{\mathbb{D}^k}(\boldsymbol{q}_i, \boldsymbol{c}_{pos})$$

Comparison between the Poincaré and Euclidean probes

Classification accuracy on Movie Review dataset

	BiLSTM	LINEAR		BERTBASE9		BERTBASE10	
		Euclidean	Poincaré	Euclidean	Poincaré	Euclidean	Poincaré
Accuracy	79.7	48.4	48.4	81.7	**84.9**	83.5	84.2

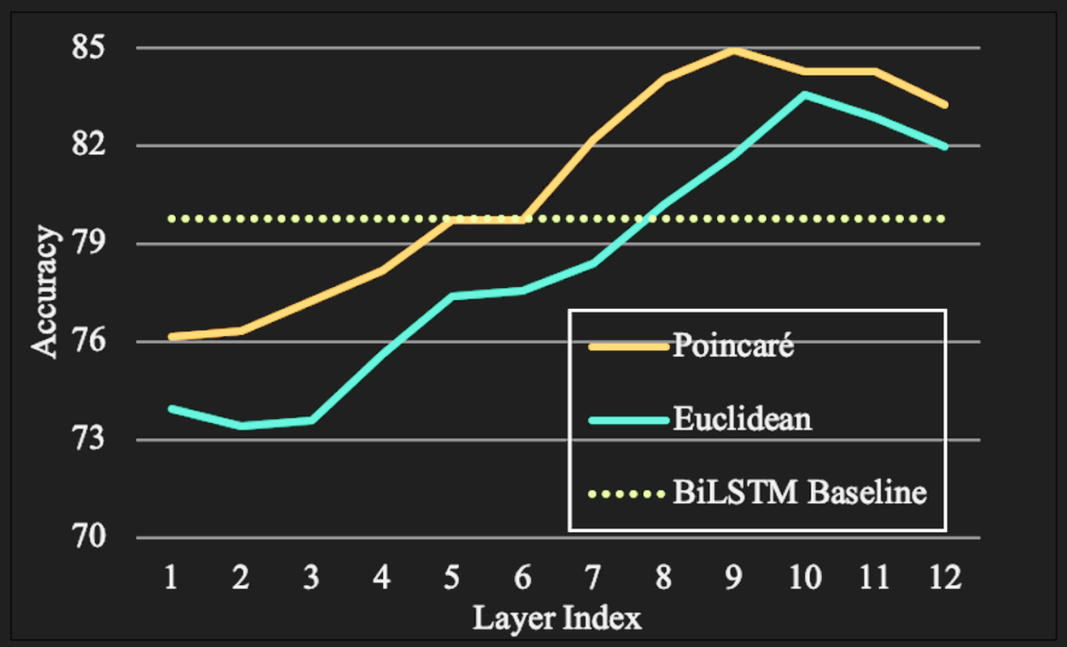


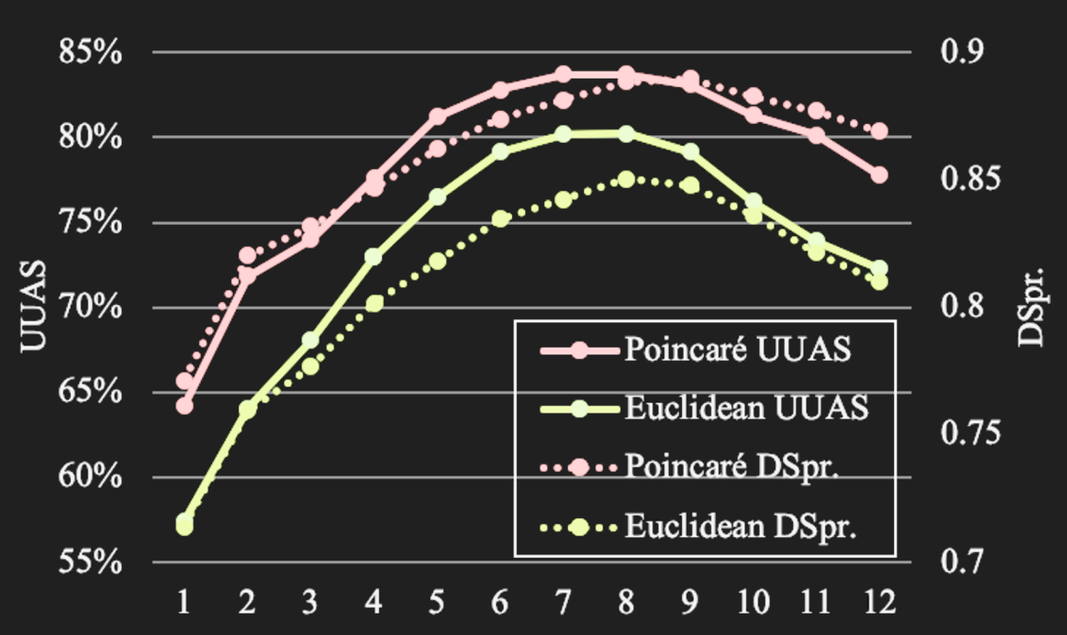
Accuracy across BERTBASE layers

Comparison between the Poincaré and Euclidean probes

Classification accuracy on Movie Review dataset

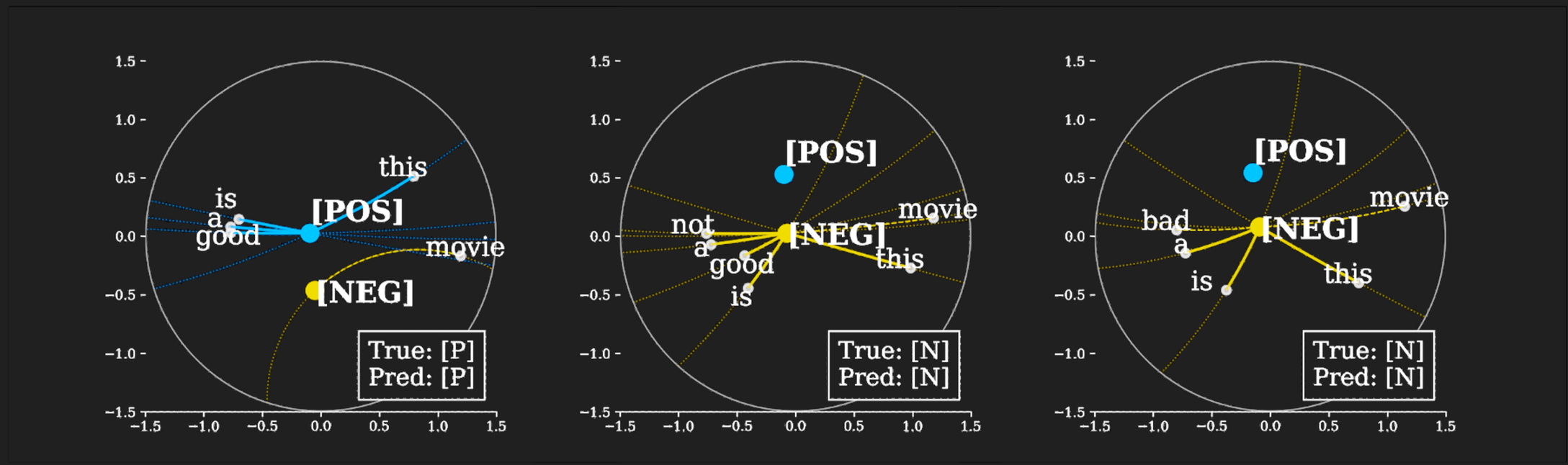
	BiLSTM	LINEAR		BERTBASE9		BERTBASE10	
		Euclidean	Poincaré	Euclidean	Poincaré	Euclidean	Poincaré
Accuracy	79.7	48.4	48.4	81.7	**84.9**	83.5	84.2





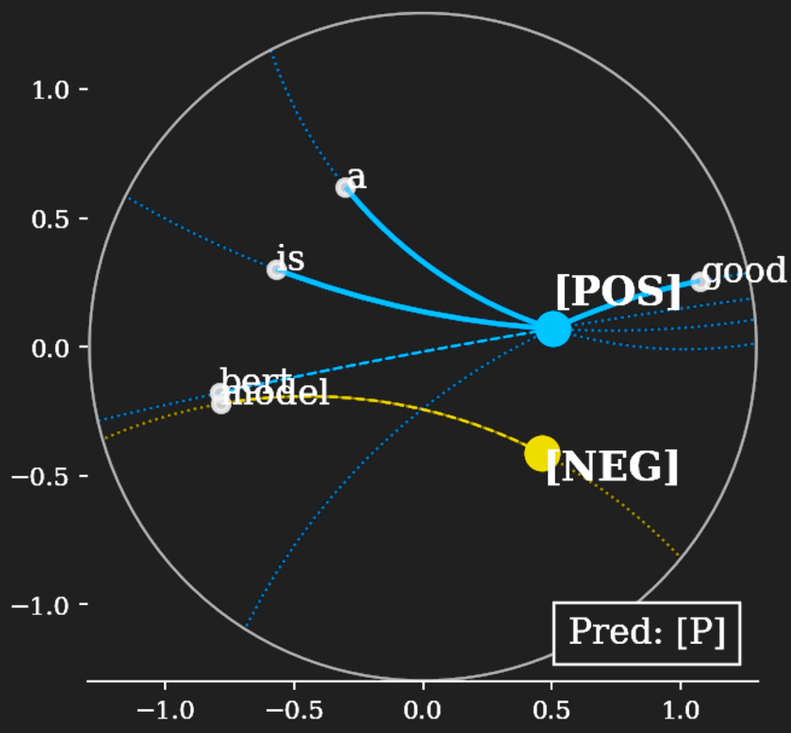
Comparison between the sentiment (left) and syntax (right) probing task

Lexically-controlled contextualization



Conclusion

- Poincaré probe can recover hyperbolic subspaces for hierarchical information encoded in BERT.
- The syntactic probe shows that BERT may encode syntax geometrically different from the Euclidean space.
- The sentiment probe further reveals the geometry of BERT embeddings by studying their localization with different contextualization.
- Our exploration brings up new possibilities about the geometry of BERT embeddings with detailed discussions and extensive visualizations.



BERT is a good model

Thanks