The Story about Probing

Huayang Li
28/04/2020

Before Talking about The Story

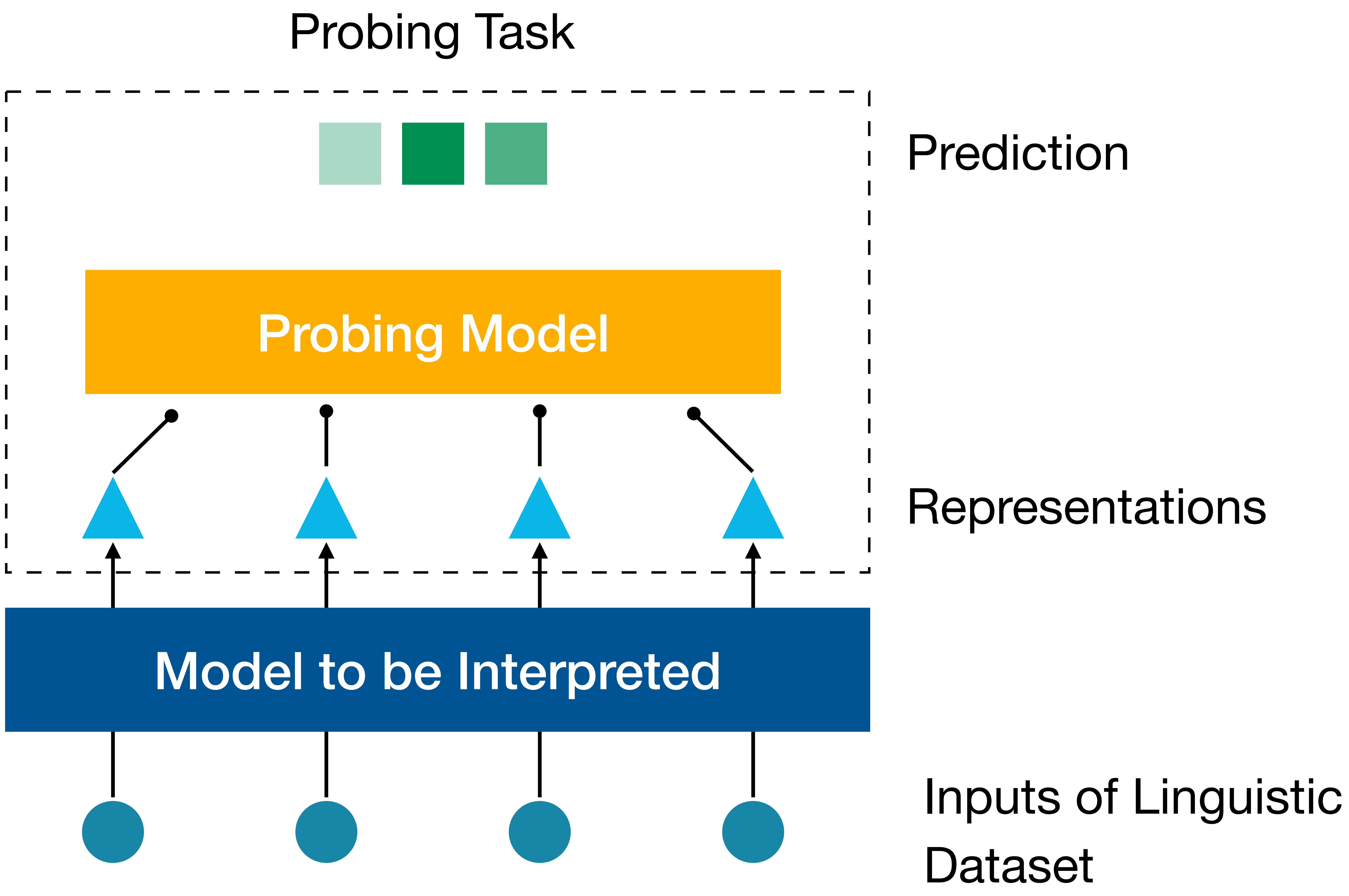
Some high-level backgrounds

- There are two ways to interpret NLP models
  - Understand the decision-making process of the model.
  - Understand the linguistic properties captured by the model.

The Beginning of This Story

What is probing

- Probing task is a classification task based on a dataset that can reveal some linguistic properties and a probing model
- The accuracy of the probing task can be regarded as a reflection of the linguistic properties captured by the representations



One Concern about Probing

But when a probe achieves high accuracy on a linguistic task using a representation, can we conclude that the representation encodes linguistic structure, or has the probe just learned the task?

—Percy Liang

Fix The Concern!

Introduce the control task

- The core idea of the **control task**: use a pseudo dataset with nonsense mappings between inputs and labels to evaluate the probing model's competence
- The gap between the performance of the control task and the performance of the real probing task is called **selectivity**
- higher selectivity means the probing model is better

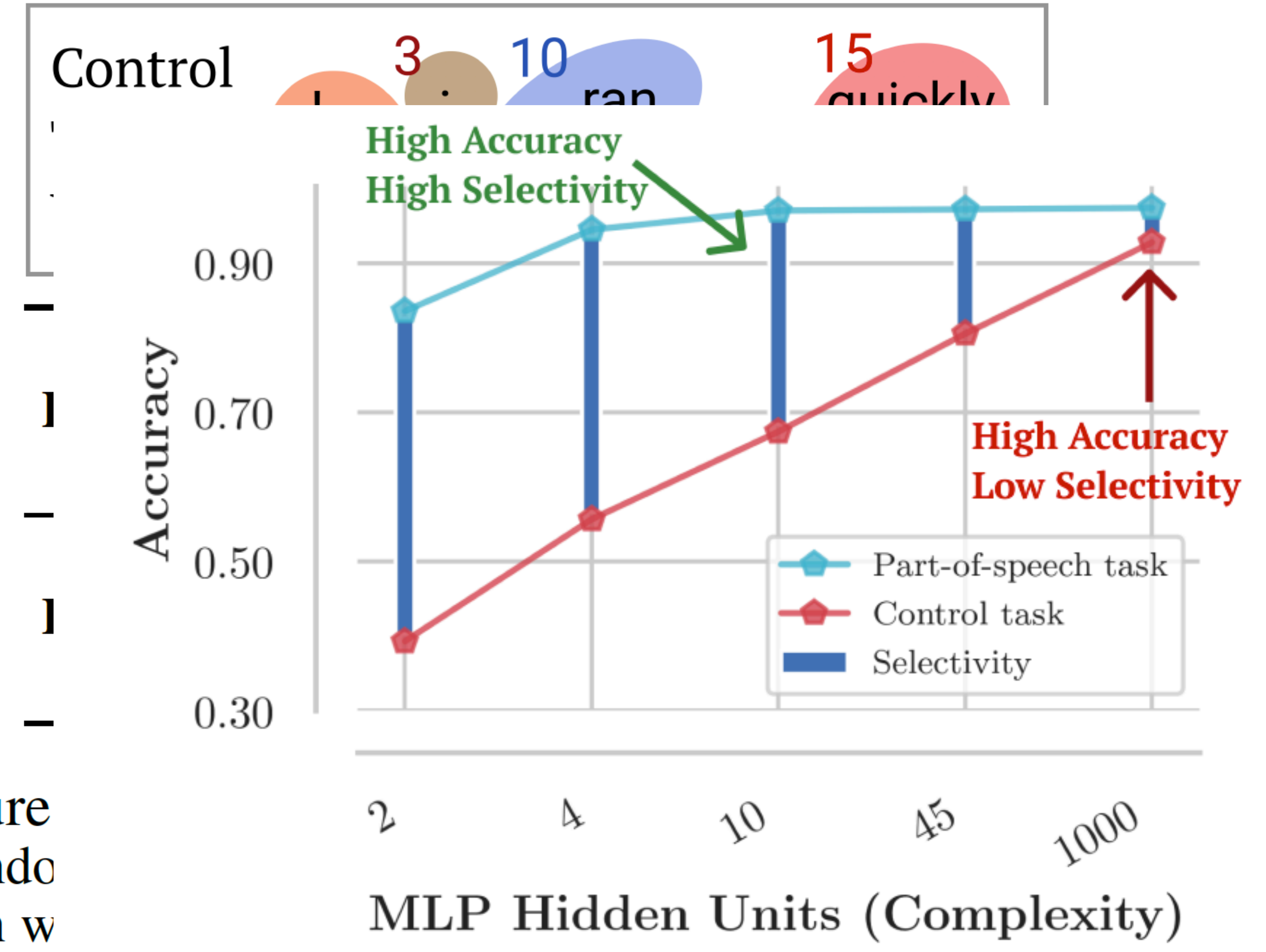


Figure
a rando
Each w
context
and out
can onl

Figure 2: Selectivity is defined as the difference between linguistic task accuracy and control task accuracy, and can vary widely, as shown, across probes which achieve similar linguistic task accuracies. These results taken from § 3.5.

Again, What is probing

From the viewpoint of information theory

$$\mathrm{I}(T; R) = \mathrm{H}(T) - \mathrm{H}(T \mid R)$$

How to Understand Probing

From the viewpoint of information theory

$$\mathrm{H}(T \mid R) := - \underset{(t,\mathbf{r})\sim p(\cdot,\cdot)}{\mathbb{E}} \left[\log p(t \mid \mathbf{r})\right] \tag{9}$$

$$= - \underset{(t,\mathbf{r})\sim p(\cdot,\cdot)}{\mathbb{E}} \left[\log \frac{p(t \mid \mathbf{r}) q_{\boldsymbol{\theta}}(t \mid \mathbf{r})}{q_{\boldsymbol{\theta}}(t \mid \mathbf{r})}\right]$$

$$= - \underset{(t,\mathbf{r})\sim p(\cdot,\cdot)}{\mathbb{E}} \left[\log q_{\boldsymbol{\theta}}(t \mid \mathbf{r}) + \log \frac{p(t \mid \mathbf{r})}{q_{\boldsymbol{\theta}}(t \mid \mathbf{r})}\right]$$

$$= \underbrace{\mathrm{H}_{q_{\boldsymbol{\theta}}}(T \mid R)}_{\textit{estimate}} - \underbrace{\underset{\mathbf{r}\sim p(\cdot)}{\mathbb{E}} \mathrm{KL}(p(\cdot \mid \mathbf{r}) \mid\mid q_{\boldsymbol{\theta}}(\cdot \mid \mathbf{r}))}_{\textit{expected estimation error}}$$

How to Understand Probing

Bigger probes are better

$$
\begin{aligned}
\mathrm{I}(T; R) &:= \mathrm{H}(T) - \mathrm{H}(T \mid R) \\
&\geq \mathrm{H}(T) - \mathrm{H}_{q_{\boldsymbol{\theta}}}(T \mid R)
\end{aligned}
$$

How to Understand Probing

Results from the original paper

	# Tokens				bert	fastText		onehot	
Language	Train	Test	# POS	$\mathrm{H}(T)$	$\mathrm{H}(T \mid R)$	$\mathrm{H}(T \mid \mathbf{c}(R))$	$\mathcal{G}(T, R, \mathbf{c})$	$\mathrm{H}(T \mid \mathbf{c}(R))$	$\mathcal{G}(T, R, \mathbf{c})$
Basque	71,483	23,959	16	3.18	0.31	0.30	-0.01 (0.3%)	0.82	0.51 (16.0%)
Czech	1,164,956	172,420	18	3.33	0.08	0.14	0.06 (1.8%)	0.36	0.28 (08.4%)
English	177,583	22,044	17	3.62	0.21	0.39	0.18 (5.0%)	0.64	0.43 (11.9%)
Finnish	138,695	18,263	16	3.16	0.24	0.20	-0.04 (1.3%)	0.86	0.62 (19.6%)
Tamil	5,460	1,656	14	3.21	0.58	0.69	0.11 (3.4%)	1.65	1.05 (32.7%)
Turkish	36,562	9,567	15	3.02	0.33	0.27	-0.09 (3.0%)	0.86	0.50 (16.6%)

Table 1: Amount of information shared by BERT, fastText or onehot embeddings and a POS tagging task. When put into context, multilingual BERT does not tell us much more about syntax than trivial baselines. $\mathrm{H}(T)$ is estimated with a plug-in estimator from same treebanks we use to train the POS labelers.

The Efforts Paid By Probing Models

Use Minimum Description Length as our tool

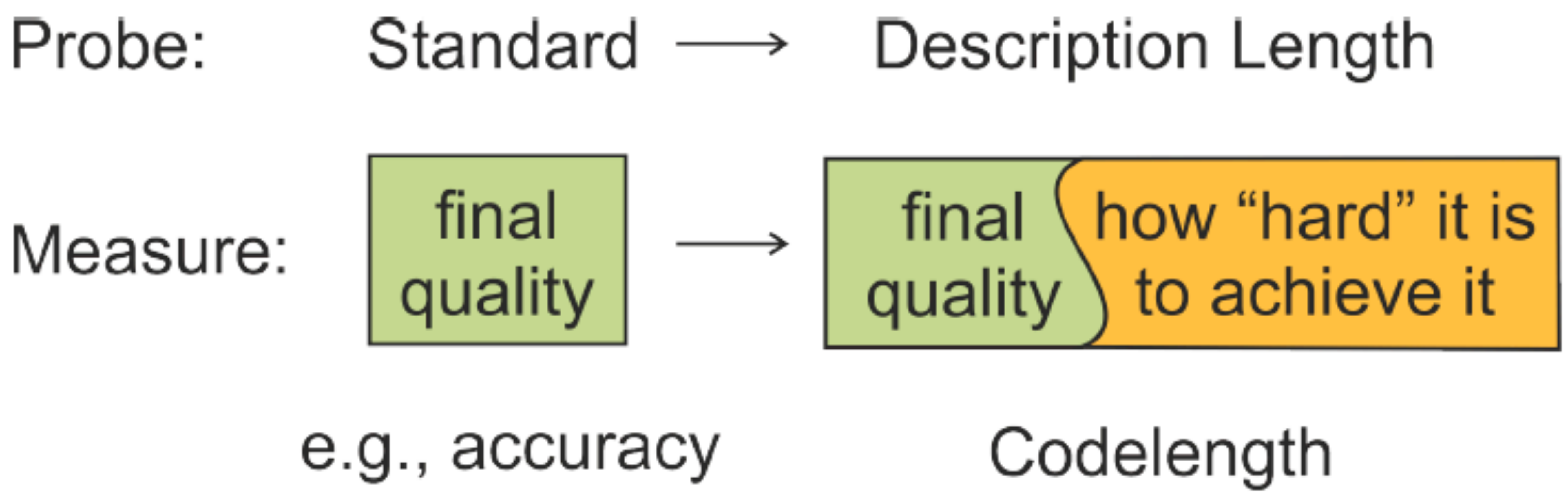


Figure 1: Illustration of the idea behind MDL probes.

The Efforts Paid By Probing Models

Use Minimum Description Length as our tool

- A communicate game between Alice and Bob
    - Alice knows all (x, y) pairs from dataset D
    - Bob just knows x from D
    - Alice want to communicate y to Bob
- Transmission: Data and Model
- The bits they need is the efforts of the probing model need to be paid

The Efforts Paid By Probing Models

Use Minimum Description Length as our tool

$$
\begin{aligned}
&L^{2\text{-}part}_{\theta^*}(y_{1:n}|x_{1:n}) = \\
&= L_{param}(\theta^*) + L_{p_{\theta^*}}(y_{1:n}|x_{1:n}) \\
&= L_{param}(\theta^*) - \sum_{i=1}^{n} \log_2 p_{\theta^*}(y_i|x_i).
\end{aligned}
$$

The Efforts Paid By Probing Models

Use Minimum Description Length as our tool

$$
\begin{aligned}
&L_{\beta}^{var}(y_{1:n}|x_{1:n}) = \\
&= -\mathbb{E}_{\theta\sim\beta}\left[\log_2\alpha(\theta) - \log_2\beta(\theta) + \sum_{i=1}^{n}\log_2 p_\theta(y_i|x_i)\right] \\
&= KL(\beta \parallel \alpha) - \mathbb{E}_{\theta\sim\beta}\sum_{i=1}^{n}\log_2 p_\theta(y_i|x_i), \qquad (3)
\end{aligned}
$$

The Efforts Paid By Probing Models

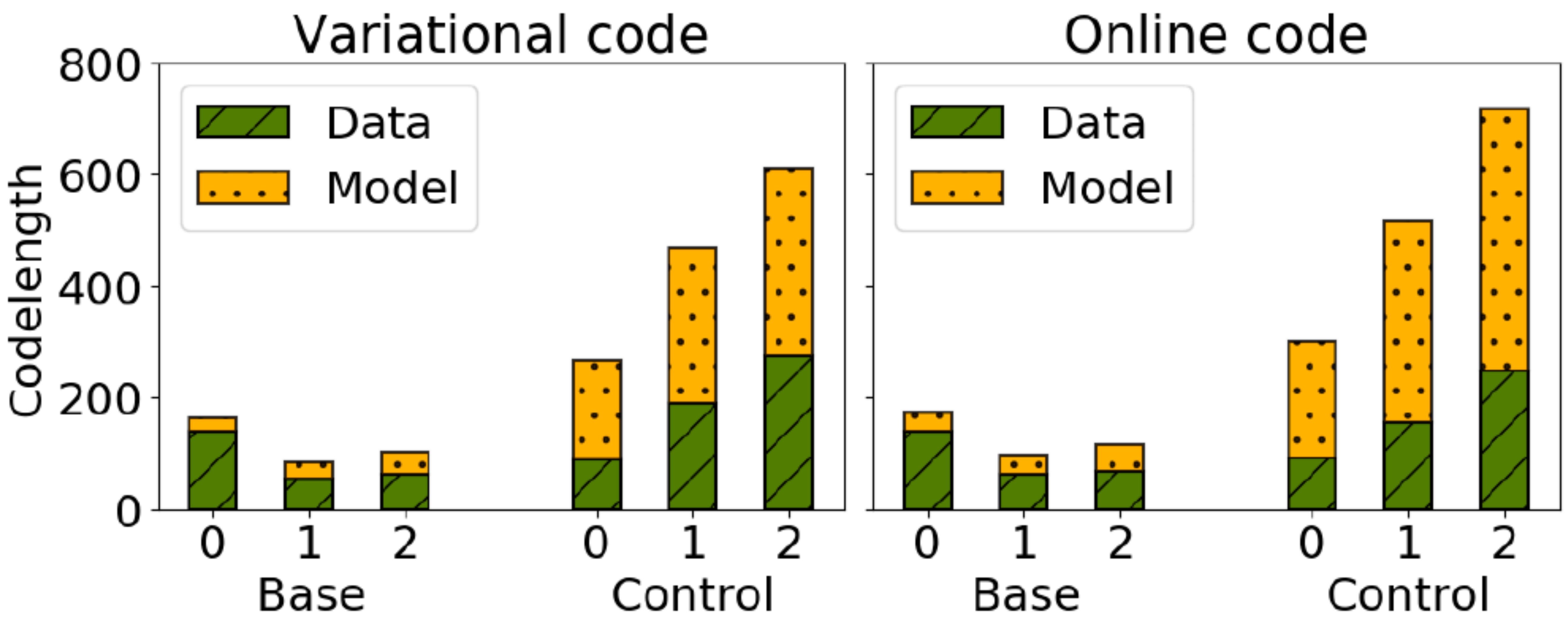
Results

	Accuracy	Description Length			
		variational code		online code	
		codelength	compression	codelength	compression
MLP-2, h=1000					
LAYER 0	93.7 / 96.3	163 / 267	31.32 / 19.09	173 / 302	29.5 / 16.87
LAYER 1	97.5 / 91.9	85 / 470	59.76 / 10.85	96 / 515	53.06 / 9.89
LAYER 2	97.3 / 89.4	103 / 612	49.67 / 8.33	115 / 717	44.3 / 7.11

Table 2: Experimental results; shown in pairs: linguistic task / control task. Codelength is measured in kbits (variational codelength is given in equation (3), online – in equation (4)). Accuracy is shown for the standard probe as in Hewitt and Liang (2019); for the variational probe, scores are similar (see Table 3).

The Efforts Paid By Probing Models

Results of model codelength and data codelength



(a) (b)

Reference

- Information-Theoretic Probing with Minimum Description Length, arXiv
- Information-Theoretic Probing for Linguistic Structure, ACL2020
- Designing and Interpreting Probes with Control Tasks, EMNLP2019
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL2019
- Deep contextualized word representations, NAACL2018

THANKS FOR YOUR TIME