A Latent-Variable Model for Intrinsic Probing

Anonymous ACL-IJCNLP submission

Abstract

The success of pre-trained contextualized representations has prompted researchers to analyze them for the presence of linguistic information. Indeed, it is natural to assume that these pre-trained representations do encode some level of linguistic knowledge as they have brought about large empirical improvements on a wide variety of NLP tasks, which suggests they are learning true linguistic generalization. In this work, we focus on intrinsic probing, an analysis technique where the goal is not only to identify whether a representation encodes a linguistic attribute, but also to pinpoint *where* this attribute is encoded. We propose a novel latent-variable formulation for constructing intrinsic probes and derive a tractable variational approximation to the log-likelihood. Our results show that our model is versatile and outperforms two intrinsic probes previously proposed in the literature. Finally, we find empirical evidence that pre-trained representations develop a cross-lingually entangled notion of morphosyntax.[1]

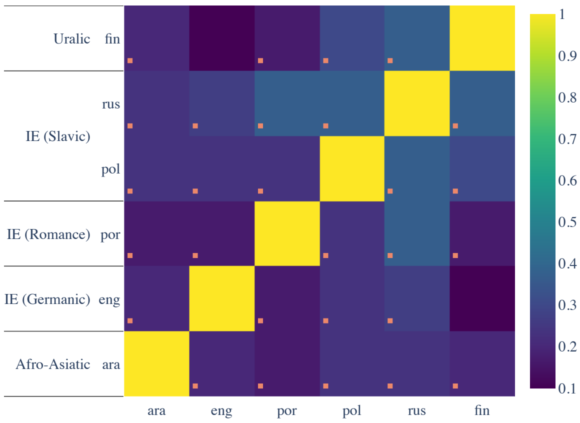


Figure 1: The percentage overlap between the top-30 most informative number dimensions in BERT for the probed languages. Statistically significant overlap, after Holm–Bonferroni family-wise error correction (Holm, 1979), with $\alpha = 0.05$, is marked with an orange square.

1 Introduction

There have been considerable improvements to the quality of pre-trained contextualized representations in recent years (e.g., Peters et al., 2018; Devlin et al., 2019; Raffel et al., 2020). These advances have sparked an interest in understanding what linguistic information may be lurking within the representations themselves (Poliak et al., 2018; Zhang and Bowman, 2018; Rogers et al., 2020, *inter alia*). One philosophy that has been proposed to extract this information is called **probing**, the task of training an external classifier to predict the linguistic property of interest directly from the representations. The hope of probing is that it sheds light onto how much linguistic knowledge is present in representations and, perhaps, how that information is structured. Probing has grown to be a fruitful area of research, with researchers probing for morphological (Tang et al., 2020; Ács et al., 2021), syntactic (Voita and Titov, 2020; Hall Maudslay et al., 2020; Ács et al., 2021), and semantic (Vulić et al., 2020; Tang et al., 2020) information.

In this paper, we focus on one type of probing known as **intrinsic probing** (Dalvi et al., 2019; Torroba Hennigen et al., 2020), a subset of which specifically aims to ascertain how information is structured within a representation. This means that we are not solely interested in determining whether a network encodes the tense of a verb, but also in pinpointing exactly *which* neurons in the network are responsible for encoding the property. Unfortunately, the naïve formulation of intrinsic probing requires one to analyze all possible combinations of neurons, which is intractable even for the smallest representations used in modern-day NLP. For example, analyzing all combinations of 768-dimensional BERT word representations would require us to train 2^{768} different probes, one for each combination of neurons, which far exceeds the estimated number of atoms in the observable universe.

To obviate this difficulty, we introduce a novel

[1] Code is available at: `http://anonymized`.

latent-variable probe for discriminative intrinsic probing. The core idea of this approach is that instead of training a different probe for each combination of neurons, we introduce a subset-valued latent variable. We approximately marginalize over the latent subsets using variational inference. Training the probe in this manner results in a set of parameters which work well across all possible subsets. We propose two variational families to model the posterior over the latent subset-valued random variables, both based on common sampling designs: Poisson sampling, which selects each neuron based on independent Bernoulli trials, and conditional Poisson sampling, which first samples a fixed number of neurons from a uniform distribution and then a subset of neurons of that size (Lohr, 2019). Conditional Poisson sampling offers the modeler more control over the distribution over subset sizes; they may pick the parametric distribution themselves.

We compare both variants to the two main intrinsic probing approaches we are aware of in the literature (§5.1). To do so, we train probes for 29 morphosyntactic properties across 6 languages (English, Portuguese, Polish, Russian, Arabic, and Finnish) from the Universal Dependencies (UD; Nivre et al. 2017) treebanks. We show that, in general, both variants of our method yield tighter estimates of the mutual information, though the model based on conditional Poisson sampling yields slightly better performance. This suggests that they are better at quantifying the informational content encoded in m-BERT contextual representations (Devlin et al., 2019). Further, we conduct a qualitative analysis of the most informative neurons (§5.2). We also analyze whether neural representations are able to learn cross-lingual abstractions from multilingual corpora. We confirm this statement and observe a strong overlap in the most informative dimensions, especially for number (Fig. 1). Additionally, we show that our method supports training deeper probes (App. B.1), though the advantages of non-linear probes over their linear counterparts are modest.

2 Intrinsic Probing

The success behind pre-trained contextual representations such as BERT (Devlin et al., 2019) suggests that they may offer a continuous analogue of the discrete structures in language, such as morphosyntactic attributes number, case, or tense. Intrinsic probing aims to recognize the parts of a network (assuming they exist) which encode such structures. In this paper, we will operate exclusively at the level of the neuron—in the case of BERT, this is one component of the 768-dimensional vector the model outputs. However, our approach can easily generalize to other settings, e.g., the layers in a transformer or filters of a convolutional neural network. Identifying individual neurons responsible for encoding linguistic features of interest has previously been shown to increase model transparency (Bau et al., 2019). In fact, knowledge about which neurons encode certain properties has also been employed to mitigate potential biases (Vig et al., 2020), for controllable text generation (Bau et al., 2019), and to analyze the linguistic capabilities of language models (Lakretz et al., 2019).

To formally describe our intrinsic probing framework, we first introduce some notation. We define Π to be the set of values that some property of interest can take, e.g., $\Pi = \{\textsc{Singular}, \textsc{Plural}\}$ for the morphosyntactic number attribute. Let $\mathcal{D} = \{(\pi^{(n)}, \boldsymbol{h}^{(n)})\}_{n=1}^{N}$ be a dataset of label–representation pairs: $\pi^{(n)} \in \Pi$ is a linguistic property and $\boldsymbol{h}^{(n)} \in \mathbb{R}^d$ is a representation. Additionally, let D be the set of all neurons in a representation; in our setup, it is an integer range. In the case of BERT, we have $D = \{1, \ldots, 768\}$. Given a subset of dimensions $C \subseteq D$, we write $\boldsymbol{h}_C$ for the subvector of $\boldsymbol{h}$ which contains only the dimensions present in C.

Let $p_{\boldsymbol{\theta}}(\pi^{(n)} \mid \boldsymbol{h}_C^{(n)})$ be a probe—a classifier trained to predict $\pi^{(n)}$ from a subvector $\boldsymbol{h}_C^{(n)}$. In intrinsic probing, our goal is to find the size k subset of neurons $C \subseteq D$ which are most informative about the property of interest. This may be written as the following combinatorial optimization problem (Torroba Hennigen et al., 2020):

$$C^\star = \underset{\substack{C \subseteq D,\\ |C| = k}}{\operatorname{argmax}} \sum_{n=1}^{N} \log p_{\boldsymbol{\theta}}\left(\pi^{(n)} \mid \boldsymbol{h}_C^{(n)}\right) \quad (1)$$

To exhaustively solve eq. (1), we would have to train a probe $p_{\boldsymbol{\theta}}(\pi \mid \boldsymbol{h}_C)$ for every one of the exponentially many subsets $C \subseteq D$. Thus, exactly solving eq. (1) is infeasible, and we are forced to rely on an approximate solution, e.g., greedily selecting the dimension that maximizes the objective. However, greedy selection alone is not enough to make solving eq. (1) manageable; because we must *retrain* $p_{\boldsymbol{\theta}}(\pi \mid \boldsymbol{h}_C)$ for *every* subset $C \subseteq D$

considered during the greedy selection procedure, i.e., we would end up training $\mathcal{O}(k\,|D|)$ classifiers. As an example, consider what would happen if one used a greedy selection scheme to find the 50 most informative dimensions for a property on 768-dimensional BERT representations. To select the first dimension, one would need to train 768 probes. To select the second dimension, one would train an additional 767, and so forth. After 50 dimensions, one would have trained 37893 probes. To address this problem, our paper introduces a latent-variable probe, which identifies a $\boldsymbol{\theta}$ that can be used for any combination of neurons under consideration allowing a greedy selection procedure to work in practice.

3 A Latent-Variable Probe

The technical contribution of this work is a novel latent-variable model for intrinsic probing. Our method starts with a generic probabilistic probe $p_{\boldsymbol{\theta}}(\pi \mid C, \boldsymbol{h})$ which predicts a linguistic attribute π given a subset C of the hidden dimensions; C is then used to subset $\boldsymbol{h}$ into $\boldsymbol{h}_C$. To avoid training a unique probe $p_{\boldsymbol{\theta}}(\pi \mid C, \boldsymbol{h})$ for every possible subset $C \subseteq D$, we propose to integrate a prior over subsets $p(C)$ into the model and then to marginalize out all possible subsets of neurons:

$$p_{\boldsymbol{\theta}}(\pi \mid \boldsymbol{h}) = \sum_{C \subseteq D} p_{\boldsymbol{\theta}}(\pi \mid C, \boldsymbol{h})\, p(C) \tag{2}$$

Due to this marginalization, our likelihood is *not* dependent on any specific subset of neurons C. Throughout this paper we will take $p(C)$ to be uniform, but other distributions are also possible; in this work, we opted for a non-informative prior.

Our goal is to estimate the parameters $\boldsymbol{\theta}$. We achieve it by maximizing the log-likelihood of the training data $\sum_{n=1}^{N} \log \sum_{C \subseteq D} p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)})$ with respect to the parameters $\boldsymbol{\theta}$. Unfortunately, directly computing this involves a sum over all possible subsets of D—a sum with an exponential number of summands. Thus, we resort to a variational approximation. Let $q_{\boldsymbol{\phi}}(C)$ be a distribution over subsets, parameterized by parameters $\boldsymbol{\phi}$; we will use $q_{\boldsymbol{\phi}}(C)$ to approximate the true posterior distribution. Then, the log-likelihood is lower-bounded as follows using Jensen's inequality:

$$\sum_{n=1}^{N} \log \sum_{C \subseteq D} p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)}) \tag{3}$$

$$\geq \sum_{n=1}^{N} \Big(\mathbb{E}_q \Big[\log p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)}) \Big] + \mathrm{H}(q) \Big) \tag{4}$$

where $\mathrm{H}(q_{\boldsymbol{\phi}})$ is the entropy of $q_{\boldsymbol{\phi}}$.[2]

Our likelihood is general, and can take the form of any objective function. This means that we can use this approach to train intrinsic probes with any type of architecture amenable to gradient-based optimization, e.g., neural networks. However, in this paper, we use a linear classifier, unless stated otherwise. Further, note that eq. (13) is valid for any choice of $q_{\boldsymbol{\phi}}$. We explore two variational families for $q_{\boldsymbol{\phi}}$, each based on a common sampling technique. The first (herein POISSON) applies Poisson sampling (Hájek, 1964), which assumes each neuron to be subjected to an independent Bernoulli trial. The second one (CONDITIONAL POISSON; Aires, 2000) corresponds to conditional Poisson sampling, which can be defined as conditioning a Poisson sample by a fixed sample size.

3.1 Parameter Estimation

As mentioned above, exact computation of the log-likelihood is intractable due to the sum over all possible subsets of D. Thus, we optimize the variational bound presented in eq. (13). We optimize the bound through stochastic gradient descent with respect to the model parameters $\boldsymbol{\theta}$ and the variational parameters $\boldsymbol{\phi}$, a technique known as stochastic variational inference (Hoffman et al., 2013). One final trick is necessary, however: The variational bound itself still includes a sum over all subsets in the first term. Thus, we have

$$\begin{aligned} &\nabla_{\boldsymbol{\theta}} \mathbb{E}_q \Big[\log p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)}) \Big] \qquad (5) \\ &\quad = \mathbb{E}_q \Big[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)}) \Big] \\ &\quad \approx \sum_{m=1}^{M} \Big[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\pi^{(n)}, C^{(m)} \mid \boldsymbol{h}^{(n)}) \Big] \end{aligned}$$

where we take M Monte Carlo samples to approximate the sum. In the case of the gradient with respect to $\boldsymbol{\phi}$, we also have to apply the REINFORCE

[2] See App. A for the full derivation.

trick (Williams, 1992):

$$\begin{aligned}\nabla_{\boldsymbol{\phi}} \mathbb{E}_q &\left[\log p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)})\right] \quad (6)\\ &= \mathbb{E}_q \left[\log p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)}) \nabla_{\boldsymbol{\phi}} \log q_{\boldsymbol{\phi}}(C)\right]\\ &\approx \sum_{m=1}^{M} \left[\log p_{\boldsymbol{\theta}}(\pi^{(n)}, C^{(m)} \mid \boldsymbol{h}^{(n)}) \nabla_{\boldsymbol{\phi}} \log q_{\boldsymbol{\phi}}(C)\right]\end{aligned}$$

where we again take M Monte Carlo samples. This procedure leads to an unbiased estimate of the gradient of the variational approximation.

3.2 Choice of Variational Family $q_{\boldsymbol{\phi}}(C)$.

We consider two choices of variational family $q_{\boldsymbol{\phi}}(C)$, both based on sampling designs (Lohr, 2019). Each defines a parameterized distribution over all subsets of D.

Poisson Sampling. Poisson sampling is one of the simplest sampling designs. In our setting, each neuron d is given a unique non-negative weight $w_d = \exp(\phi_d)$. This gives us the following parameterized distribution over subsets:

$$q_{\boldsymbol{\phi}}(C) = \prod_{d \in C} \frac{w_d}{1 + w_d} \prod_{d \notin C} \frac{1}{1 + w_d} \quad (7)$$

The formulation in eq. (7) shows that taking a sample corresponds to $|D|$ independent coin flips—one for each neuron—where the probability of heads is $\frac{w_d}{1+w_d}$. The entropy of a Poisson sampling may be computed in $\mathcal{O}(|D|)$ time:

$$\mathrm{H}(q_{\boldsymbol{\phi}}) = \log Z - \sum_{d=1}^{|D|} \frac{w_d}{1 + w_d} \log w_d \quad (8)$$

where $\log Z = \sum_{d=1}^{|D|} \log(1 + w_d)$. The gradient of eq. (8) may be computed automatically through backpropagation. Poisson sampling automatically modules the size of the sampled set $C \sim q_{\boldsymbol{\phi}}(\cdot)$ and we have the expected size $\mathbb{E}[|C|] = \sum_{d=1}^{|D|} \frac{w_d}{1+w_d}$.

Conditional Poisson Sampling. We also consider a variational family that factors as follows:

$$q_{\boldsymbol{\phi}}(C) = \underbrace{q_{\boldsymbol{\phi}}^{\mathrm{CP}}(C \mid |C| = k)}_{\text{Conditional Poisson}} \; q_{\boldsymbol{\phi}}^{\mathrm{size}}(k) \quad (9)$$

In this paper, we take $q_{\boldsymbol{\phi}}^{\mathrm{size}}(k) = \mathrm{Uniform}(D)$, but a more complex distribution, e.g., a Categorical, could be learned. We define $q_{\boldsymbol{\phi}}^{\mathrm{CP}}(C \mid |C| = k)$ as a conditional Poisson sampling design. Similarly to Poisson sampling, conditional Poisson sampling starts with a unique positive weight associated with every neuron $w_d = \exp(\phi_d)$. However, an additional cardinality constraint is introduced. This leads to the following distribution

$$q_{\boldsymbol{\phi}}^{\mathrm{CP}}(C) = \mathbb{1}\{|C| = k\} \frac{\prod_{d \in C} w_d}{Z^{\mathrm{CP}}} \quad (10)$$

A more elaborate dynamic program which runs in $\mathcal{O}(k\,|D|)$ may be used to compute Z^{CP} efficiently (Aires, 2000). We may further compute the entropy $\mathrm{H}(q_{\boldsymbol{\phi}})$ and its the gradient in $\mathcal{O}(k\,|D|)$ time using the expectation semiring (Eisner, 2002; Li and Eisner, 2009). Sampling from $q_{\boldsymbol{\phi}}^{\mathrm{CP}}$ can be done efficiently using quantities computed when running the dynamic program used to compute Z^{CP} (Kulesza, 2012). In practice, we use the semiring implementations by Rush (2020).

4 Experimental Setup

Our setup is virtually identical to the morphosyntactic probing setup of Torroba Hennigen et al. 2020. This consists of first automatically mapping treebanks from UD v2.1 (Nivre et al., 2017) to the UniMorph (McCarthy et al., 2018) schema.[3] Then, we compute multilingual BERT (m-BERT) representations[4] for every sentence in the UD treebanks. After computing the m-BERT representations for the entire sentence, we extract representations for individual words in the sentence and pair them with the UniMorph morphosyntactic annotations. We estimate our probes' parameters using the UD training set and conduct greedy selection to approximate the objective in eq. (1) on the validation set; finally, we report the results on the test set, i.e., we test whether the set of neurons we found on the development set generalizes to held-out data. Additionally, we discard values that occur fewer than 20 times across splits. Finally, when feeding $\boldsymbol{h}_C$ as input to our probes, we set any dimensions that are not present in C to zero.

4.1 Baselines

We compare our latent-variable probe against two other recently proposed intrinsic probing methods as baselines.

- **Torroba Hennigen et al. (2020):** Our first baseline is generative probe, which models the

[3]We use the code available at: `https://github.com/unimorph/ud-compatibility`.

[4]We use the implementation by Wolf et al. (2020).

joint distribution of representations and their properties $p(\boldsymbol{h}, \pi) = p(\boldsymbol{h} \mid \pi)\, p(\pi)$, where the representation distribution $p(\boldsymbol{h} \mid \pi)$ is assumed to be Gaussian. Torroba Hennigen et al. (2020) report that a major limitation of this probe is that if certain dimensions of the representations are not distributed according to a Gaussian distribution, then probe performance will suffer.

- **Dalvi et al. (2019):** Our second baseline is a linear classifier, where dimensions not under consideration are zeroed out during evaluation (Dalvi et al., 2019; Durrani et al., 2020).[5] Their approach is a special case of our proposed latent-variable model, where q_{ϕ} is fixed, so that on every training iteration the entire set of dimensions is sampled.

4.2 Metrics

We compare our proposed method to the baselines above under two metrics: accuracy and mutual information (MI). Accuracy is a standard measure for evaluating probes as it is for evaluating classifiers in general. Next, we also report mutual information, which has recently been proposed as an evaluation metric for evaluating probes (Pimentel et al., 2020). More formally, mutual information (MI) is a function between a random variable over a Π-valued random variable P and a $\mathbb{R}^{|C|}$-valued random variable H_C over masked representations:

$$\mathrm{MI}(P; H_C) = \mathrm{H}(P) - \mathrm{H}(P \mid H_C) \tag{11}$$

where $\mathrm{H}(P)$ is the inherent entropy of the property being probed and is constant with respect to H_C; $\mathrm{H}(P \mid H_C)$ is the entropy over the property given the representations H_C. Exact computation of the mutual information is intractable, however; luckily, we can lower-bound the MI by approximating $\mathrm{H}(P \mid H_C)$ using our probe's average negative log-likelihood: $-\frac{1}{N}\sum_{n=1}^{N} \log p_{\boldsymbol{\theta}}(\pi^{(n)} \mid C, \boldsymbol{h}^{(n)})$ on held-out data. See Brown et al. (1992) for a derivation; $\mathrm{H}(P)$ is constant.

We also normalize the mutual information (NMI) by dividing the MI by the entropy which turns it into a percentage and is, arguably, more interpretable. We refer the reader to Gates et al. (2019) for a discussion of the normalization of MI.

4.3 What Makes a Good Probe?

Since we report a lower bound on the mutual information (§4), we deem the best probe to be the one that yields the tightest mutual information estimate, or, in other words, the one that achieves the highest mutual information estimate; this is a equivalent to having the best cross-entropy on held-out data, which is the standard evaluation metric for language modeling.

However, in the context of intrinsic probing, the topic of primary interest is what the probe reveals about the structure of the representations. For instance, does the probe reveal that the information encoded in the embeddings is focalized or dispersed across many neurons? Several prior works (e.g., Lakretz et al., 2019) focus on the single neuron setting, which is a special, very focal case. To engage with this prior work, we compare probes not only with respect to their performance (MI and accuracy), but also with respect to the size of the subset of dimensions being evaluated, i.e., the size of set C.

We acknowledge that there is a disparity between the quantitative evaluation we employ, in which probes are compared based on their MI estimates, and qualitative nature of intrinsic probing, which aims to identify the substructures of a model that encode a property of interest. However, it is non-trivial to evaluate fundamentally qualitative procedures in a large-scale, systematic, and unbiased manner. Therefore, we rely on the quantitative evaluation metrics presented in §4.2, while also including a qualitative analysis (§5.2).

4.4 Training and Hyperparameter Tuning

We train our probes for a maximum of 2000 epochs using the Adam optimizer (Kingma and Ba, 2015). We add early stopping with a patience of 50 as a regularization technique. Early stopping is conducted by holding out 10% of the training data; our development set is reserved for the greedy selection of subsets of neurons. Our implementation is built with PyTorch (Paszke et al., 2019). To execute a fair comparison with Dalvi et al. (2019), we train all probes other than the Gaussian probe using ElasticNet regularization (Zou and Hastie, 2005), which consists of combining both L_1 and L_2 regularization, where the regularizers are weighted by tunable regularization coefficients λ_1 and λ_2, respectively. We follow the experimental set-up proposed by Dalvi et al. (2019), where we set $\lambda_1, \lambda_2 = 10^{-5}$

[5]We note that they do not conduct intrinsic probing via dimension selection: Instead, they use the absolute magnitude of the weights as a proxy for dimension importance. In this paper, we adopt the approach of (Torroba Hennigen et al., 2020) and use the performance-based objective in eq. (1).

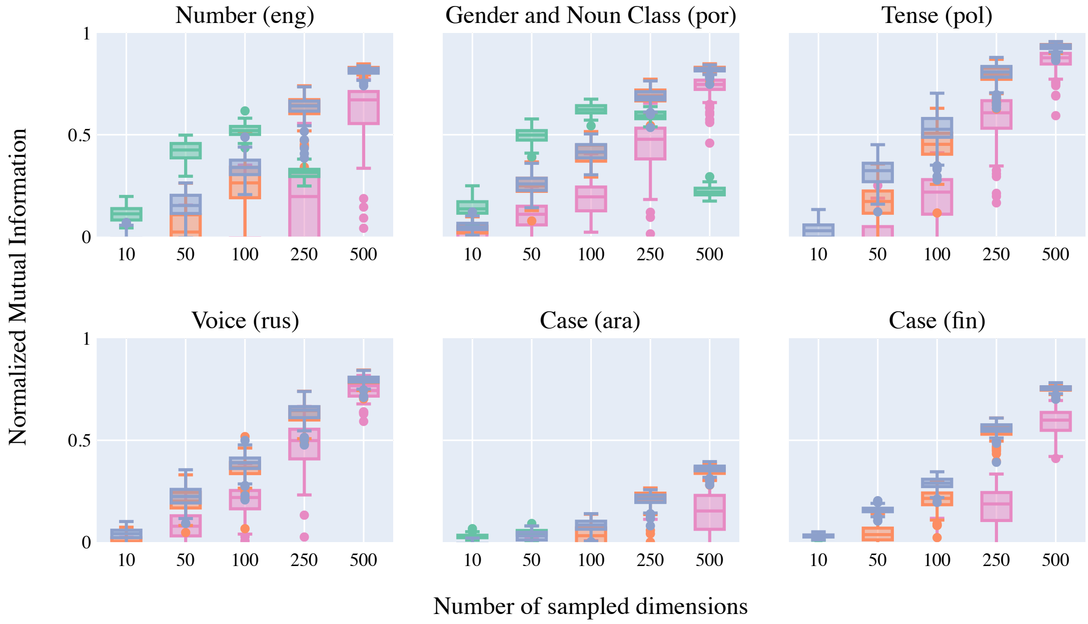


Figure 2: Comparison of the POISSON and CONDITIONAL POISSON methods to the DALVI (Dalvi et al., 2019) and GAUSSIAN, when probing selected multilingual BERT (Devlin et al., 2019) representations. For each of the subset sizes shown on the x-axis, we sampled 100 different subsets of BERT dimensions at random. Note that in some cases (e.g., Polish tense), GAUSSIAN does not obtain positive mutual information (§4) in any of dimensionalities, hence it does not appear on the graph.

for all probes. In a preliminary experiment, we performed a grid search over these hyperparameters to confirm that the probe is not very sensitive to the tuning of these values (unless they are extreme) as Dalvi et al. (2019) claims. For GAUSSIAN, we take the MAP estimate, with a weak data-dependent prior (Murphy, 2012, Chapter 4).

5 Results and Discussion

In this section, we present the results of our empirical investigation. First, we address our main research question: Does our latent-variable probe presented in §3 outperform previously proposed intrinsic probing methods (§5.1)? Second, we analyze the structure of the most informative m-BERT neurons for the different morphosyntactic attributes we probe for (§5.2). Finally, we investigate whether knowledge about morphosyntax encoded in neural representations is shared across languages (§5.3). In App. B.1, we show that our latent-variable probe is flexible enough to support deep neural probes.

5.1 How Do Our Methods Perform?

The main question we ask is how the performance of our models compares to existing intrinsic probing approaches. To investigate this research question, we compare the performance of the POISSON and CONDITIONAL POISSON probes to DALVI (Dalvi et al., 2019) and GAUSSIAN (Torroba Hennigen et al., 2020). Refer to §4.3 for a discussion of the limitations of our method.

Experimental Setup. Since the performance of a probe on a specific subset of dimensions is related to both the subset itself (e.g., whether it is informative or not) and the number of dimensions being evaluated (e.g., if a probe is trained to expect 768 dimensions as input, it might work best when few or no dimensions are filled with zeros), we sample 100 subsets of dimensions with 5 different possible sizes (we considered 10, 50, 100, 250, 500 dim.) and compare every model's performance on each of those subset sizes. As the UPPER BOUND baseline needs to be retrained for every set of dimensions under consideration,[6] we limit our comparisons with UPPER BOUND to 6 randomly chosen morphosyntactic attributes, each in a different language.

[6]The UPPER BOUND yields the tightest estimate on the mutual information, however as mentioned in §2, this is unfeasible since it requires retraining for every different combination of neurons. For comparison, in English number, on an Nvidia RTX 2070 GPU, our POISSON, GAUSSIAN and DALVI experiments take a few minutes or even seconds to run, compared to UPPER BOUND which takes multiple hours.

Results. We compare the performance of the probes on 29 different language–attribute pairs (refer to App. C for a listing). Our results suggest that both variants of our latent-variable model from §3 are effective and generally outperform the two baselines we consider. In particular, CONDITIONAL POISSON tends to outperform POISSON at lower dimensions, however, POISSON tends to catch up as more dimensions are added. We plot these results for six randomly selected language–attribute pairs in Fig. 2 in terms of NMI. See Fig. 6 in the App. D an equivalent plot for accuracy.

When evaluating CONDITIONAL POISSON on few dimensions (e.g., 10), we find that it generally provides a low but positive mutual information estimate, whereas DALVI and POISSON can yield negative mutual information estimates. Notably, negative mutual information only arises because the model underperforms a random-guessing baseline. In contrast, the GAUSSIAN method tends to perform well at low dimensions, and it even outperforms CONDITIONAL POISSON for language–attribute pairs such as English number and Portuguese gender. We assume this can be attributed to GAUSSIAN's ability to model non-linear decision boundaries (Murphy, 2012, Chapter 4). However, GAUSSIAN's performance is not stable and can yield low or even negative mutual information estimates across all subsets of dimensions, e.g., for Polish tense and Russian voice representations. Adding a new dimension can never decrease the mutual information, so the observable decreases occur because the generative model deteriorates upon adding another dimension, which corroborates Torroba Hennigen et al.'s claim that some dimensions are not adequately modeled by the Gaussian assumption. We include some additional comparisons in App. B.

Finally, we compare the POISSON and CONDITIONAL POISSON probes to the UPPER BOUND baseline. This is expected to be the highest performing since it is re-trained for *every* subset under consideration. This is feasible because we only evaluate subsets discovered by the greedy selection procedure. The difference between our probes' performance and the UPPER BOUND baseline's performance can be seen as the cost of sharing parameters across all subsets of dimensions, and an effective intrinsic probe should minimize this. This is illustrated in Fig. 7 in the Appendix. As expected, our results suggest that both methods achieve performance that is close to the UPPER BOUND method. This tells us that the latent-variable approach is nearly as good as if we retrained our probe from scratch knowing the subset of neurons of interest *a priori*.

5.2 A Taste of Analysis

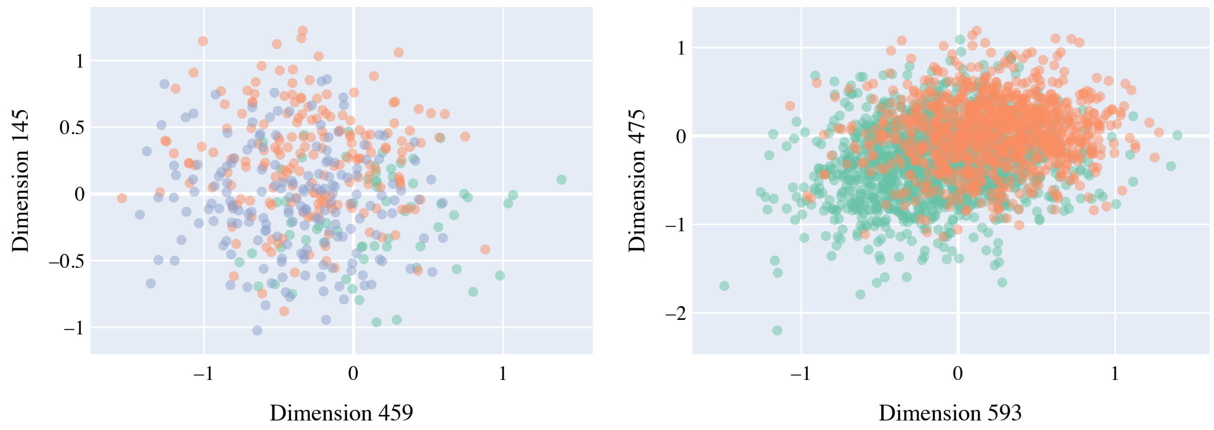
To better understand the behavior of our probes, we follow Torroba Hennigen et al. (2020) in investigating the structure of the top two most informative neurons in the final layer selected by our CONDITIONAL POISSON probe for particular language–attribute pairs. We observe that the activations of the two neurons for the majority of language–attribute pairs are largely overlapping, regardless of how many values are in the set Π. While tense in Finnish shows strong separation of values for all sets Π, we observe that Russian voice is the most dispersed of all language-attribute pairs. We present selected results in Fig. 3.

5.3 Cross-lingual Overlap

We use our probe to analyze whether the most informative dimensions are shared in m-BERT embeddings across languages in order to validate the hypothesis by Torroba Hennigen et al. (2020) of BERT leveraging data from other languages to develop a cross-lingually entangled notion of morphosyntax. Indeed, an inspection of the overlap in informative dimensions in BERT across languages reveals evidence of cross-lingual neuron reuse when encoding morphosyntactic attributes. We observe a strong overlap in the most informative dimensions, especially for number (Fig. 1) and to a lesser extent in other attributes such as gender and case (Fig. 8). This suggests that BERT may be leveraging data from other languages to develop a cross-lingually entangled notion of morphosyntax. A significant overlap in the salient case neurons for Russian and Polish might indicate additionally that morphosyntactic representations are similar across languages within the same language genus.

6 Related Work

A growing interest in interpretability has led to a flurry of work in trying to assess exactly what pre-trained representations know about language. To this end, diverse methods have been employed, such as the construction of specific challenge sets that seek to evaluate how well representations model particular phenomena (Linzen et al., 2016;



(a) Russian–Voice (active, middle and passive)

(b) Finnish–Tense (past and present)

Figure 3: Scatter plots for the two most informative dimensions selected by the CONDITIONAL POISSON probe for m-BERT representations for a range of language–attribute pairs.

Gulordava et al., 2018; Goldberg, 2019; Goodwin et al., 2020), methods for determining whether certain capabilities help to achieve accurate models of particlar data (Perez et al., 2021), as well as visualization methods (Kádár et al., 2017; Rethmeier et al., 2020). Work on probing comprises a major share of this endeavor (Belinkov and Glass, 2019; Belinkov, 2021). This has taken the form of both focused studies on particular linguistic phenomena (e.g., subject-verb number agreement, Giulianelli et al., 2018) to broad assessments of contextual representations in a wide array of tasks (Şahin et al., 2020; Tenney et al., 2018; Conneau et al., 2018; Liu et al., 2019; Ravichander et al., 2021, *inter alia*).

Efforts have ranged widely, but most of these focus on extrinsic rather than intrinsic probing. Most work on the latter has focused primarily on ascribing roles to individual neurons through methods such as visualization (Karpathy et al., 2015; Li et al., 2016) and ablation (Li et al., 2017). For example, recently Lakretz et al. (2019) conduct an in-depth study of how long–short-term memory networks (LSTMs; Hochreiter and Schmidhuber, 1997) capture subject–verb number agreement, and identify two units largely responsible for this phenomenon.

More recently, there has been a growing interest in extending intrinsic probing to collections of neurons. Bau et al. (2019) utilize unsupervised methods to identify important neurons, and then attempt to control a neural network's outputs by selectively modifying them. Bau et al. (2020) pursue a similar goal in a computer vision setting, but ascribe meaning to neurons based on how their activations correlate with particular classifications in images, and are able to control these manually with interpretable results. Aiming to answer questions on interpretability in computer vision and natural language inference, Mu and Andreas (2020) develop a method to create compositional explanations of individual neurons and investigate abstractions encoded in them. Vig et al. (2020) analyze how certain information is encoded in individual neurons and how it is being propagated through different model components such as neurons and attention heads and apply their method to study gender and other societal biases.

7 Conclusion

In this paper, we introduce a new method for training discriminative intrinsic probes that can perform well across any subset of dimensions. To do so, we train a probing classifier with a subset-valued latent variable and demonstrate how the latent subsets can be marginalized using variational inference. We propose two variational families, based on common sampling designs, to model the posterior over subsets: Poisson sampling and conditional Poisson sampling. We demonstrate that both variants outperform our baselines in terms of mutual information, and that using a conditional Poisson variational family gives optimal performance. Further, we demonstrate that our method has the flexibility to be used with linear and deeper probes. Finally, we find empirical evidence for overlap in the specific neurons used to encode morphosyntactic properties across languages.

References

Judit Ács, Ákos Kádár, and Andras Kornai. 2021. Subword pooling makes a difference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2284–2295, Online. Association for Computational Linguistics.

Nibia Aires. 2000. Comparisons between conditional Poisson sampling and Pareto πs sampling designs. *Journal of Statistical Planning and Inference*, 88(1):133–147.

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2019. Identifying and controlling important neurons in neural machine translation. In *International Conference on Learning Representations*.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. 2020. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and alternatives. *arXiv preprint arXiv:2102.12452*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Jennifer C. Lai, and Robert L. Mercer. 1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1):31–40.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. 2019. What is one grain of sand in the desert? Analyzing individual neurons in deep NLP models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:6309–6317.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. Analyzing individual neurons in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4865–4880, Online. Association for Computational Linguistics.

Jason Eisner. 2002. Parameter estimation for probabilistic finite-state transducers. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 1–8, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alexander J. Gates, Ian B. Wood, William P. Hetrick, and Yong-Yeol Ahn. 2019. Element-centric clustering comparison unifies overlaps and hierarchy. *Scientific Reports*, 9(1):8574.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *arXiv:1901.05287 [cs]*.

Emily Goodwin, Koustuv Sinha, and Timothy J. O'Donnell. 2020. Probing linguistic systematicity. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1958–1969, Online. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Jaroslav Hájek. 1964. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35(4):1491–1523.

Rowan Hall Maudslay, Josef Valvoda, Tiago Pimentel, Adina Williams, and Ryan Cotterell. 2020. A tale of a probe and a parser. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7389–7395, Online. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods* *in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. 2013. Stochastic variational inference. *Journal of Machine Learning Research*, 14(4):1303–1347.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Ákos Kádár, Grzegorz Chrupała, and Afra Alishahi. 2017. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4):761–780.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Workshop Proceedings*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, San Diego, CA.

Alex Kulesza. 2012. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2-3):123–286.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 11–20, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2017. Understanding neural networks through representation erasure. *arXiv:1612.08220 [cs]*.

Zhifei Li and Jason Eisner. 2009. First- and second-order expectation semirings with applications to minimum-risk training on translation forests. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 40–51, Singapore. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Sharon L. Lohr. 2019. *Sampling: Design and Analysis*, 2 edition. CRC Press.

Arya D. McCarthy, Miikka Silfverberg, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2018. Marrying universal dependencies and universal morphology. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 91–101, Brussels, Belgium. Association for Computational Linguistics.

Jesse Mu and Jacob Andreas. 2020. Compositional explanations of neurons. In *Advances in Neural Information Processing Systems*, volume 33, pages 17153–17163. Curran Associates, Inc.

Kevin P. Murphy. 2012. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning Series. MIT Press, Cambridge, MA.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 807–814, Madison, WI, USA.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Victoria Bobicev, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Aljoscha Burchardt, Marie Candito, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Silvie Cinková, Çağrı Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Ali Elkahky, Tomaž Erjavec, Richárd

Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Radu Ion, Elena Irimia, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, Václava Kettnerová, Jesse Kirchner, Natalia Kotsyba, Simon Krek, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shinsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Kaili Müürisep, Pinkey Nainwani, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Hanna Nurmi, Stina Ojala, Petya Osenova, Robert Östling, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Larissa Rinaldi, Laura Rituma, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Benoît Sagot, Shadi Saleh, Tanja Samardžić, Manuela Sanguinetti, Baiba Saulīte, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Jonathan North Washington, Mats Wirén, Tak-sum Wong, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 2017. Universal dependencies 2.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. Rissanen data analysis: Examining dataset characteristics via description length. *arXiv preprint arXiv:2103.03872*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

Nils Rethmeier, Vageesh Kumar Saxena, and Isabelle Augenstein. 2020. TX-Ray: Quantifying and explaining model-knowledge transfer in (un-)supervised NLP. In *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence*, page 197. AUAI Press.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Alexander Rush. 2020. Torch-Struct: Deep structured prediction library. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 335–342, Online. Association for Computational Linguistics.

Gözde Gül Şahin, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych. 2020. LINSPECTOR: Multilingual probing tasks for word representations. *Computational Linguistics*, 46(2):335–385.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2020. Understanding pure character-based neural machine translation: The case of translating Finnish into English. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4251–4262, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2018. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 197–216, Online. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2020. HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv:1910.03771 [cs]*.

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361, Brussels, Belgium. Association for Computational Linguistics.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320.

A Variational Lower Bound

The derivation of the variational lower bound is shown below:

$$\sum_{n=1}^{N} \log \sum_{C \subseteq D} p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)}) \tag{12}$$

$$= \sum_{n=1}^{N} \log \sum_{C \subseteq D} q_{\boldsymbol{\phi}}(C) \frac{p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)})}{q_{\boldsymbol{\phi}}(C)}$$

$$= \sum_{n=1}^{N} \log \mathbb{E}_q \left[\frac{p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)})}{q_{\boldsymbol{\phi}}(C)} \right]$$

$$\geq \sum_{n=1}^{N} \mathbb{E}_q \left[\log \frac{p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)})}{q_{\boldsymbol{\phi}}(C)} \right] \tag{13}$$

$$= \sum_{n=1}^{N} \left(\mathbb{E}_q \left[\log p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)}) \right] + \mathrm{H}(q) \right)$$

B Additional Intrinsic Probe Comparisons

We also conduct a direct comparison of DALVI, GAUSSIAN, POISSON and CONDITIONAL POISSON when used to identify the most informative subsets of dimensions. The average MI reported by each model across all 29 morphosyntactic language–attribute pairs is presented in Fig. 4. On average, CONDITIONAL POISSON offers comparable performance to GAUSSIAN at low dimensionalities for both NMI and accuracy, though the latter tends to yield a slightly higher (and thus a tighter) bound on the mutual information. However, as more dimensions are taken into consideration, our models vastly outperform GAUSSIAN. POISSON and CONDITIONAL POISSON perform comparably at high dimensions, but CONDITIONAL POISSON performs slightly better for 1–20 dimensions. POISSON outperforms DALVI at high dimensions, and CONDITIONAL POISSON outperforms DALVI for all dimensions considered.

B.1 How Do Deeper Probes Perform?

Multiple papers have promoted the use of linear probes (Tenney et al., 2018; Liu et al., 2019), in part because they are ostensibly less likely to memorize patterns in the data (Zhang and Bowman, 2018; Hewitt and Liang, 2019), though this is subject to debate (Voita and Titov, 2020; Pimentel et al., 2020). Here we verify our claim from §3 that our probe can be applied to any kind of discriminative probe architecture as our objective function can be optimized using gradient descent.

Experimental Setup. We follow the setup of Hewitt and Liang (2019), and test MLP-1 and MLP-2 probes alongside a LINEAR probe. The MLP-1 and MLP-2 probes are multilayer perceptrons (MLP) with one and two hidden layer(s), respectively, and Rectified Linear Unit (ReLU; Nair and Hinton, 2010) activation function.

Results. In Fig. 5, we can see that our method not only works well for deeper probes, but also outperforms the linear probe in terms of NMI. However, at higher dimensionalities, the advantage of a deeper probe diminishes. We also find that the difference in performance between MLP-1 and MLP-2 is negligible.

C List of Probed Morphosyntactic Attributes

The 29 language–attribute pairs we probe for in this work are listed below:

- **Arabic**: Aspect, Case, Definiteness, Gender, Mood, Number, Voice
- **English**: Number, Tense
- **Finnish**: Case, Number, Person, Tense, Voice
- **Polish**: Animacy, Case, Gender, Number, Tense
- **Portuguese**: Gender, Number, Tense
- **Russian**: Animacy, Aspect, Case, Gender, Number, Tense, Voice

D Supplementary Results

Fig. 6 compares the accuracy of our two models, POISSON and CONDITIONAL POISSON, to the DALVI and GAUSSIAN baselines. The figure reflects the trends observed in §5.1: With the exception of the few dimension regimen of GAUSSIAN, POISSON and CONDITIONAL POISSON outperform the DALVI and GAUSSIAN baselines.

Fig. 7 compares the NMI of our two models, POISSON and CONDITIONAL POISSON, to the UPPER BOUND baseline. The figure reflects the trends observed in §5.1: POISSON and CONDITIONAL POISSON achieve performance that is close to the UPPER BOUND baseline.

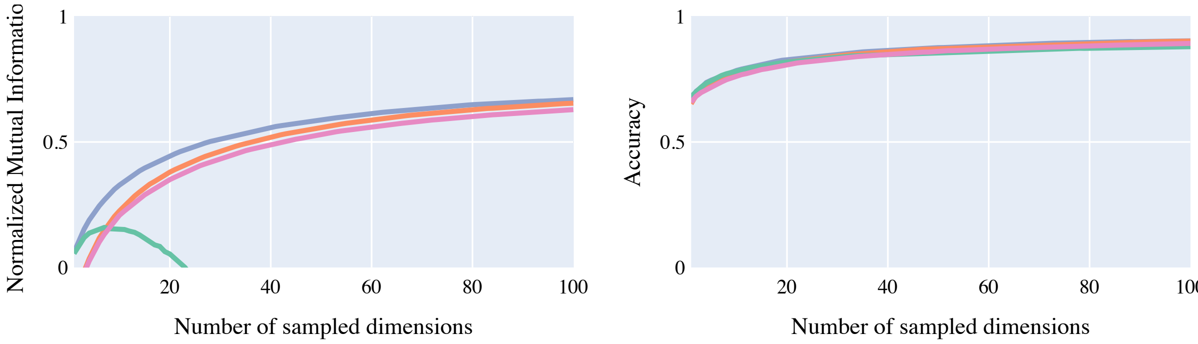


Figure 4: Comparison of the POISSON, CONDITIONAL POISSON, DALVI (Dalvi et al., 2019) and GAUSSIAN (Torroba Hennigen et al., 2020) probes. We use the greedy selection approach in eq. (1) to select the most informative dimensions, and average across all language–attribute pairs we probe for.

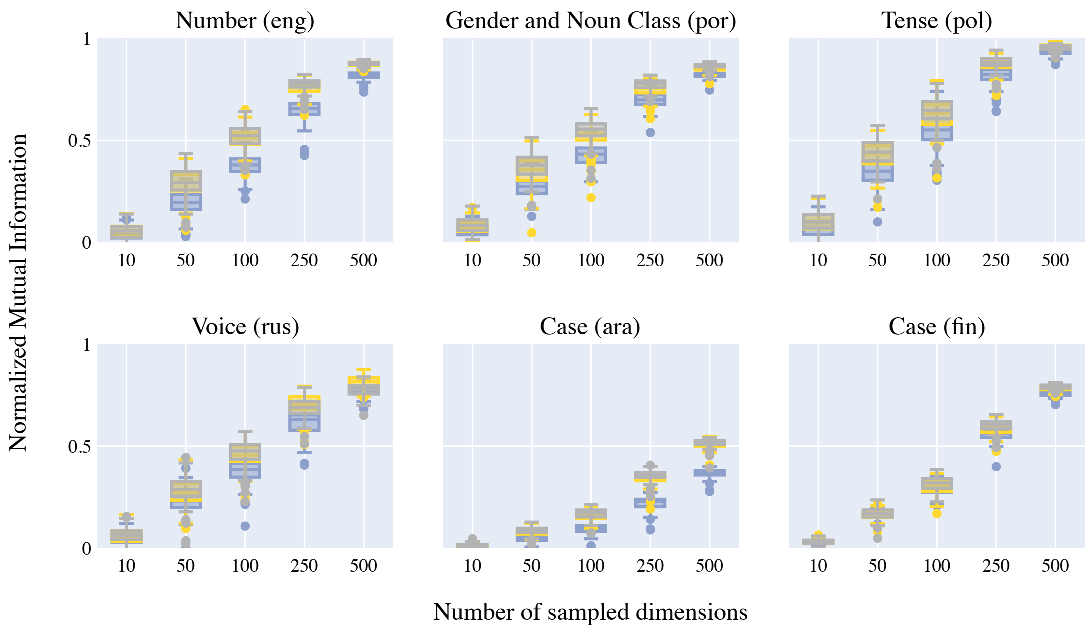


Figure 5: Comparison of a LINEAR probe to non-linear MLP-1 and MLP-2 probes for selected language-attribute pairs. For each of the subset sizes shown on the x-axis, we sampled 100 different subsets of BERT dimensions at random.

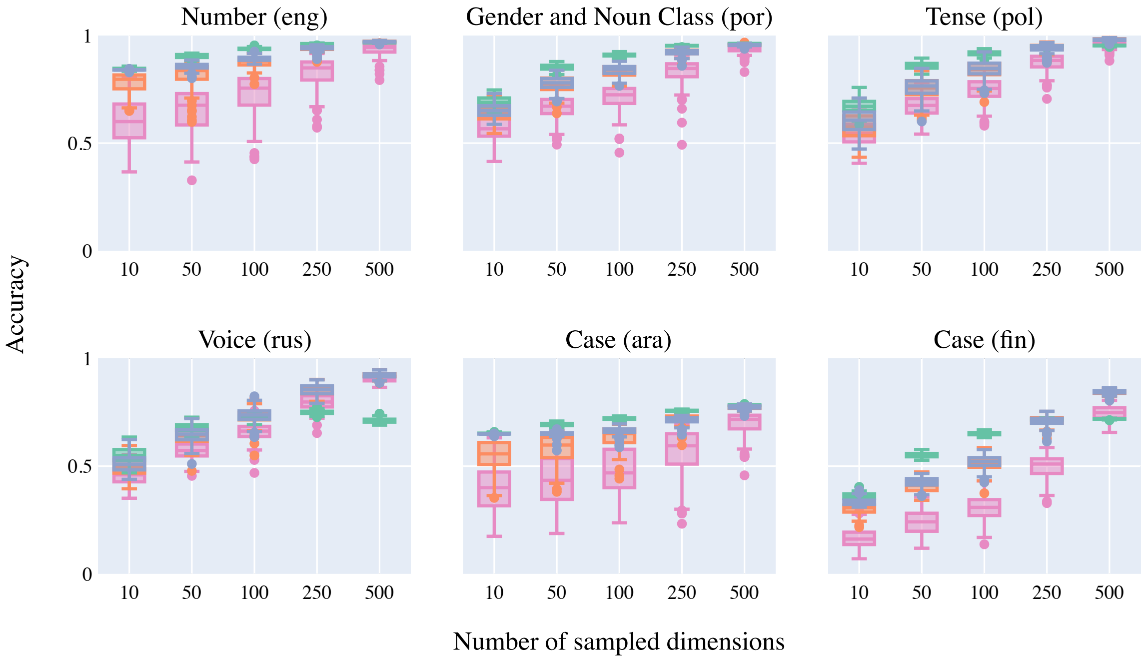


Figure 6: Accuracy comparison of the CONDITIONAL POISSON and POISSON methods to the DALVI (Dalvi et al., 2019) and GAUSSIAN baselines, when probing selected multilingual BERT (Devlin et al., 2019) representations. For each of the subset sizes shown on the x-axis, we sampled 100 different subsets of BERT dimensions at random.

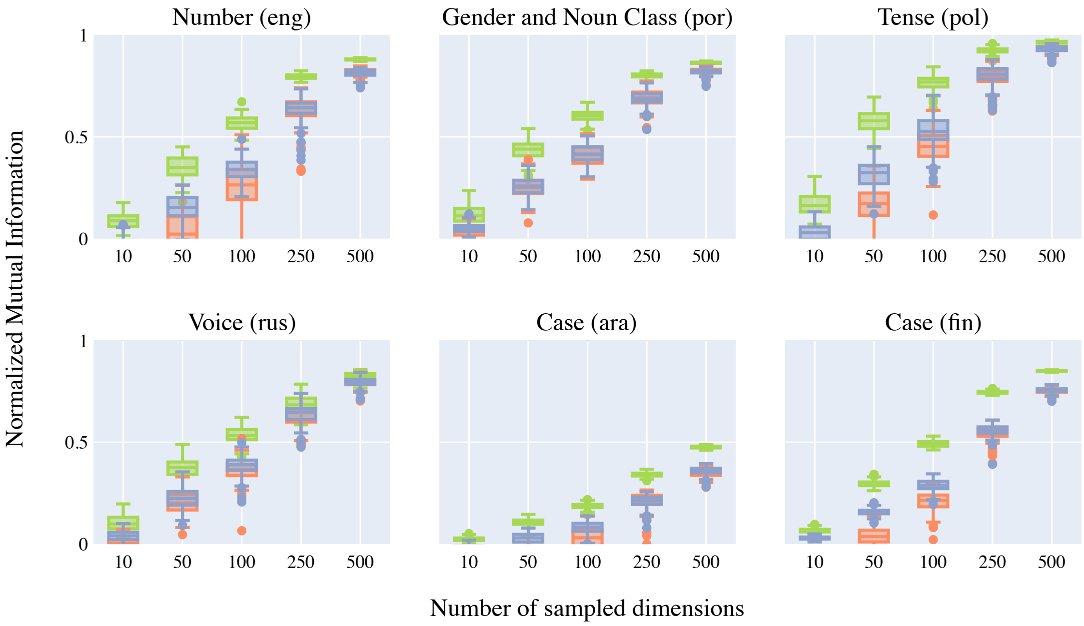


Figure 7: Comparison of the POISSON and CONDITIONAL POISSON methods to the UPPER BOUND baseline, when probing selected representations. For each of the subset sizes shown on the x-axis, we sampled 100 different subsets of m-BERT dimensions at random.

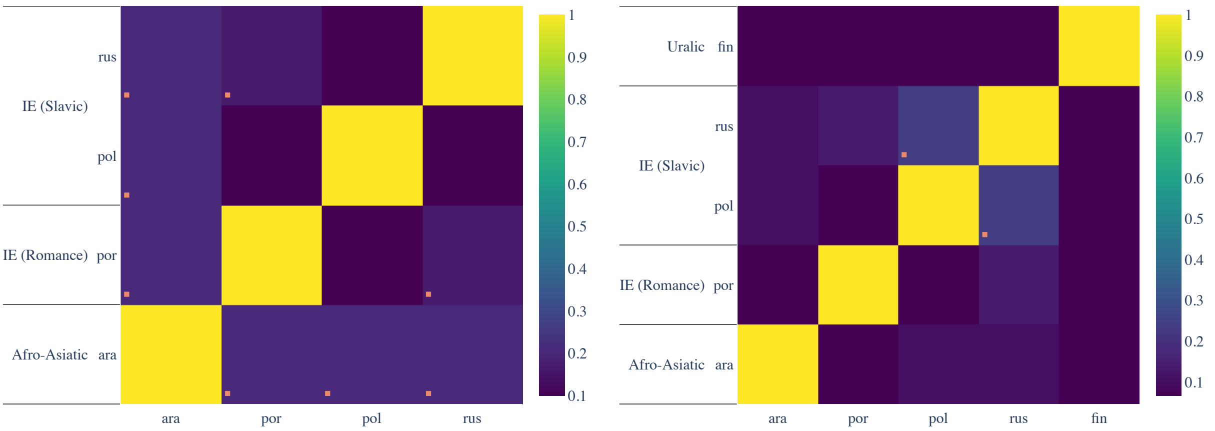


Figure 8: The percentage overlap between the top-30 most informative gender (left) and case (right) dimensions in BERT for the probed languages. Statistically significant overlap, after Holm–Bonferroni family-wise error correction (Holm, 1979), with $\alpha = 0.05$, is marked with an orange square.