A Latent-Variable Model for Intrinsic Probing

Karolina Stańczak\textsuperscript{1}, Lucas Torroba Hennigen\textsuperscript{2}, Adina Williams\textsuperscript{3}, Ryan Cotterell\textsuperscript{4}, Isabelle Augenstein\textsuperscript{1}

\textsuperscript{1}University of Copenhagen
\textsuperscript{2}Massachusetts Institute of Technology
\textsuperscript{3}Meta AI
\textsuperscript{4}ETH Zürich

ks@di.ku.dk, lucastor@mit.edu, adinawilliams@meta.com, rcotterell@inf.ethz.ch, augenstein@di.ku.dk

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 suggest 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 yields tighter mutual information estimates than 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 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, Kovaleva, and Rumshisky 2020, \textit{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, Sennrich, and Nivre 2020; Ács, Kádár, and Kornai 2021), syntactic (Voita and Titov 2020; Hall Maudslay et al. 2020; Ács, Kádár, and Kornai 2021), and semantic (Vulić et al. 2020; Tang, Sennrich, and Nivre 2020) information.

In this paper, we focus on one type of probing known as intrinsic probing (Dalvi et al. 2019; Torroba Hennigen, Williams, and Cotterell 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 test 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 representations would require training $2^{768}$ 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 latent-variable probe for intrinsic probing. Our core idea, instead of training a different probe for each subset of neurons, is to 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). To do so, we train probes for 29 morphosyntactic properties across 6 languages\textsuperscript{1} 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

\textsuperscript{1}These authors contributed equally.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
that they are better at quantifying the informational content encoded in m-BERT representations (Devlin et al. 2019). We make two typological findings when applying our probe. We show that there is a difference in how information is structured depending on the language with certain language–attribute pairs requiring more dimensions to encode relevant information. 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.

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 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 \( \Pi \) to be the set of values that some property of interest can take, e.g., \( \Pi = \{\text{Singular, Plural}\} \) for the morphosyntactic number attribute. Let \( D = \{(\pi^{(n)}, h^{(n)})\}_{n=1}^{N} \) be a dataset of label–representation pairs: \( \pi^{(n)} \in \Pi \) is a linguistic property and \( h^{(n)} \in \mathbb{R}^d \) 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 \( h_C \) for the subvector of \( h \) which contains only the dimensions present in \( C \).

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

\[
C^* = \arg\max_{C \subseteq D, |C| = k} \sum_{n=1}^{N} \log p_\theta(\pi^{(n)} \mid h_C^{(n)})
\]

To exhaustively solve Eq. (1), we would have to train a probe \( p_\theta(\pi \mid h_C) \) for every one of the exponentially many subsets \( C \subseteq D \) of size \( k \). 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 \( p_\theta(\pi \mid h_C) \) for every subset \( C \subseteq D \) considered during the greedy selection procedure, i.e., we would end up training \( O(k |D|) \) 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 \( \theta \) 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 \( p_\theta(\pi \mid C, h) \) which predicts a linguistic attribute \( \pi \) given a subset \( C \) of the hidden dimensions; \( C \) is then used to subset \( h \) into \( h_C \). To avoid training a unique probe \( p_\theta(\pi \mid C, h) \) 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:

\[
p_\theta(\pi \mid h) = \sum_{C \subseteq D} p_\theta(\pi \mid C, h) p(C)
\]

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

Our goal is to estimate the parameters \( \theta \). We achieve this by maximizing the log-likelihood of the training data \( \sum_{n=1}^{N} \log \sum_{C \subseteq D} p_\theta(\pi^{(n)}, C \mid h^{(n)}) \) with respect to the parameters \( \theta \). 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 \( q_\phi(C) \) be a distribution over subsets, parameterized by parameters \( \phi \); we will use \( q_\phi(C) \) to approximate the true posterior distribution. Then, the log-likelihood is lower-bounded as follows:

\[
\sum_{n=1}^{N} \log \sum_{C \subseteq D} p_\theta(\pi^{(n)}, C \mid h^{(n)}) 
\geq \sum_{n=1}^{N} \left( \mathbb{E}_{q_\phi} \left[ \log p_\theta(\pi^{(n)}, C \mid h^{(n)}) \right] + H(q) \right)
\]

which follows from Jensen’s inequality, where \( H(q) \) is the entropy of \( q_\phi \). The derivation of the variational lower bound

---

2 The code necessary to replicate our experiments is available at https://github.com/copenlu/flexible-probing.
is shown below:

\[
\sum_{n=1}^{N} \log \sum_{C \subseteq D} p_{\theta}(\pi^{(n)}, C \mid h^{(n)}) = \sum_{n=1}^{N} \log \sum_{C \subseteq D} q_{\phi}(C) \frac{p_{\theta}(\pi^{(n)}, C \mid h^{(n)})}{q_{\phi}(C)} = \sum_{n=1}^{N} \log \mathbb{E}_{q_{\phi}} \left[ \frac{p_{\theta}(\pi^{(n)}, C \mid h^{(n)})}{q_{\phi}(C)} \right] \geq \sum_{n=1}^{N} \mathbb{E}_{q_{\phi}} \left[ \log p_{\theta}(\pi^{(n)}, C \mid h^{(n)}) \right] + H(q)
\]  

(4)

Our likelihood is general and can take the form of any objective function. Thus, 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. We note that Eq. (3) is valid for any choice of \( q_{\phi} \). We explore two variational families for \( q_{\phi} \), 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 1999) corresponds to conditional Poisson sampling, which can be defined as conditioning a Poisson sample by a fixed sample size.

In the following, we introduce in detail the model training procedure in §3.1 and discuss our selection of the variational family \( q_{\phi}(C) \), in §3.2.

### 3.1 Parameter Estimation

As mentioned above, the 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. (3). We optimize the bound through stochastic gradient descent with respect to the model parameters \( \theta \) and the variational parameters \( \phi \), a technique known as stochastic variational inference (Hoffman et al. 2013). However, one final trick is necessary, since the variational bound still includes a sum over all subsets in the first term:

\[
\nabla_{\theta} \mathbb{E}_{q_{\phi}} \left[ \log p_{\theta}(\pi^{(n)}, C \mid h^{(n)}) \right] = \mathbb{E}_{q_{\phi}} \left[ \nabla_{\theta} \log p_{\theta}(\pi^{(n)}, C \mid h^{(n)}) \right] \approx \sum_{m=1}^{M} \left[ \nabla_{\theta} \log p_{\theta}(\pi^{(n)}, C^{(m)} \mid h^{(n)}) \right]
\]

(6)

where we take \( M \) Monte Carlo samples to approximate the sum. In the case of the gradient with respect to \( \phi \), we also have to apply the REINFORCE trick (Williams 1992):

\[
\nabla_{\phi} \mathbb{E}_{q_{\phi}} \left[ \log p_{\theta}(\pi^{(n)}, C \mid h^{(n)}) \right] = \mathbb{E}_{q_{\phi}} \left[ \log p_{\theta}(\pi^{(n)}, C \mid h^{(n)}) \nabla_{\phi} \log q_{\phi}(C) \right] \approx \sum_{m=1}^{M} \left[ \log p_{\theta}(\pi^{(n)}, C^{(m)} \mid h^{(n)}) \nabla_{\phi} \log q_{\phi}(C) \right]
\]

(7)

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 \( q_{\phi}(C) \)

We consider two choices of variational family \( q_{\phi}(C) \), 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:

\[
q_{\phi}(C) = \prod_{d \in C} \frac{w_{d}}{1 + w_{d}} \prod_{d \notin C} \frac{1}{1 + w_{d}}
\]

(8)

The formulation in Eq. (8) 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 \( O(|D|) \) time:

\[
H(q_{\phi}) = \log Z - \sum_{d=1}^{|D|} \frac{w_{d}}{1 + w_{d}} \log w_{d}
\]

(9)

where \( \log Z = \sum_{d=1}^{|D|} \log(1 + w_{d}) \). The gradient of Eq. (9) may be computed automatically through backpropagation.

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

\[
q_{\phi}(C) = \prod_{d \in C} \left[ \frac{\text{Uniform}(D)}{w_{d}} \right] \prod_{d \notin C} \left[ \frac{\text{Uniform}(D)}{1 + w_{d}} \right] q_{\phi}^{\text{size}}(k)
\]

where \( \text{Uniform}(D) \) is a uniform distribution over the set \( D \), \( q_{\phi}^{\text{size}}(k) \) is a uniform distribution over the size of the subset, and \( |C| = k \) is the cardinality of the subset.

In this paper, we take \( q_{\phi}^{\text{size}}(k) = \text{Uniform}(D) \), but a more complex distribution, e.g., a Categorical, could be learned. We define \( q_{\phi}^{\text{CP}}(C \mid |C| = k) \) 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:

\[
q_{\phi}^{\text{CP}}(C) = \prod_{d \in C} \frac{w_{d}}{1 + w_{d}} \prod_{d \notin C} \frac{1}{1 + w_{d}}
\]

(10)

A more elaborate dynamic program which runs in \( O(k |D|) \) may be used to compute \( Z^{\text{CP}} \) efficiently (Aires 1999). We may further compute the entropy \( H(q_{\phi}) \) and its the gradient in \( O(|D|^{2}) \) time using the expectation semiring (Eisner 2002; Li and Eisner 2009). Sampling from \( q_{\phi}^{\text{CP}} \) can be done efficiently using quantities computed when running the dynamic program used to compute \( Z^{\text{CP}} \) (Kulesza 2012).\(^3\)

\(^3\)We use the semiring implementation by Rush (2020).
4 Experimental Setup

Our setup is virtually identical to the morphosyntactic probing setup of Torroba Hennigen, Williams, and Cotterell (2020). This consists of first automatically mapping treebanks from UD v2.1 (Nivre et al. 2017) to the UniMorph (McCarthy et al. 2018) schema. Then, we compute multilingual BERT (m-BERT) representations 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. When feeding $h_C$ as input to our probes, we set any dimensions that are not present in $C$ to zero. We select $M = 5$ as the number of Monte Carlo samples since we found this to work adequately in small-scale experiments. We compare the performance of the probes on 29 language–attribute pairs.

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.

4.1 Baselines

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

- Torroba Hennigen, Williams, and Cotterell (2020): Our first baseline is a generative probe that models the joint distribution of representations and their properties $p(h, \pi) = p(h | \pi)p(\pi)$, where the representation distribution $p(h | \pi)$ is assumed to be Gaussian. Torroba Hennigen, Williams, and Cotterell (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).

Their approach is a special case of our proposed latent-variable model, where $g_{\theta}$ is fixed so that on every training iteration the entire set of dimensions is sampled.

Additionally, we compare our methods to a naïve approach, a probe that is re-trained for every set of dimensions under consideration selecting the dimension that maximizes the objective (herein Upper Bound). Due to computational cost, we limit our comparisons with Upper Bound to 6 randomly chosen morphosyntactic attributes, each in a different language.

4.2 Metrics

We compare our proposed method to the baselines above under two metrics: accuracy and mutual information (MI). We report mutual information, which has recently been proposed as an evaluation metric for probes (Pimentel et al. 2020). Here, mutual information (MI) is a function between a II-valued random variable $P$ and a $\mathbb{R}^{|C|}$-valued random variable $H_C$ over masked representations:

$$\text{MI}(P; H_C) = H(P) - \mathbb{H}(P | H_C)$$

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

We also report accuracy which is a standard measure for evaluating probes as it is for evaluating classifiers in general. However, accuracy can be a misleading measure, especially on imbalanced datasets since it considers solely correct predictions.

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 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 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 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 the 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 qualitatively inspecting the implications of our probes.

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 $\lambda_1$ and $\lambda_2$, respectively. We follow the experimental set-up proposed by Dalvi et al. (2019), where we set $\lambda_1, \lambda_2 = 10^{-5}$ 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) which aligns with the claim presented in Dalvi et al. (2019).

For Gaussian, we take the MAP estimate, with a weak data-dependent prior (Murphy 2012, Chapter 4). In addition, we found that a slight improvement in the performance of Poisson and Conditional Poisson was obtained by scaling the entropy term in Eq. (3) by a factor of 0.01.

5 Results

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 mBERT 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).

5.1 How Do Our Methods Perform?

To investigate how the performance of our models compares to existing intrinsic probing approaches, we compare the performance of the Poisson and Conditional Poisson probes to LINEAR (Dalvi et al. 2019) and Gaussian (Torroba Hennigen, Williams, and Cotterell 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.

In general, Conditional Poisson tends to outperform Poisson at lower dimensions, however, Poisson tends to catch up as more dimensions are added. Our results suggest that both variants of our latent-variable model from §3 are effective and generally outperform the LINEAR baseline as shown in Tab. 1. The Gaussian baseline tends to perform similarly to Conditional Poisson when we consider subsets of 10 dimensions, and it outperforms Poisson substantially. However, for subsets of size 50 or more, both Conditional Poisson and Poisson are preferable. We believe that the robust performance of Gaussian in the low-dimensional regimen can be attributed to its ability to model non-linear decision boundaries (Murphy 2012, Chapter 4).

The trends above are corroborated by a comparison of the mean NMI (Tab. 2, top) achieved by each of these probes for different subset sizes. Notwithstanding, Gaussian’s performance (in terms of NMI) is not stable and
can yield low or even negative mutual information estimates across all subsets of dimensions. 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 validates Torroba Hennigen, Williams, and Cotterell’s claim that some dimensions are not adequately modeled by the Gaussian assumption. While these results suggest that GAUSSIAN may be preferable if performing a comparison based on accuracy, the instability of GAUSSIAN when considering NMI suggests that this edge in terms of accuracy comes at a hefty cost in terms of calibration (Guo et al. 2017).9

Further, 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 and indeed, this assumption is confirmed by the results in Tab. 2 (bottom). 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.

We also conduct a direct comparison of LINEAR, GAUSSIAN, POISSON and CONDITIONAL POISSON when used to identify the most informative subsets of dimensions. The average MI and accuracy reported by each model across all 29 morphosyntactic language–attribute pairs is presented in Fig. 1. 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 MI. However, as more dimensions are taken into consideration, our models vastly outperform GAUSSIAN. Our models perform comparably at high dimensions, but CONDITIONAL POISSON performs slightly better for 1–20 dimensions. POISSON outperforms LINEAR at high dimensions, and CONDITIONAL POISSON outperforms LINEAR for all dimensions considered. These effects are less pronounced for accuracy, which we believe to be due to accuracy’s insensitivity to a probe’s confidence in its prediction. Finally, while CONDITIONAL POISSON achieves a tighter bound on NMI than POISSON, we recommend the POISSON probe for larger experimental setups due to its computational efficiency.

### 5.2 Information Distribution

We compare the performance of the CONDITIONAL POISSON probe for each attribute for all available languages in order to better understand the relatively high NMI variance across results (see Tab. 2). In Fig. 2, we plot the average NMI for gender and observe that languages with two genders present (Arabic and Portuguese) achieve higher performance than languages with three genders (Russian and Polish) which is an intuitive result due to increased task complexity. Further, we see that the slopes for both Russian and Polish are flatter, especially at lower dimensions. This implies that the information for Russian and Polish is more dispersed and more dimensions are needed to capture the typological information.

### 5.3 Cross-Lingual Overlap

We compare the most informative m-BERT dimensions recovered by our probe across languages and find that, in many cases, the same set of neurons express the same morphosyntactic phenomena across languages. For example, we find that Russian, Polish, Portuguese, English and Arabic have statistically significant overlap in the top 30 most informative neurons for number (Fig. 3). These results indicate that BERT may be leveraging data from other languages to develop a cross-lingually entangled notion of morphosyntax (Torroba Hennigen, Williams, and Cotterell 2020) and that this effect may be particularly strong between typologically similar languages.10

<table>
<thead>
<tr>
<th>Number of dimensions</th>
<th>10</th>
<th>50</th>
<th>100</th>
<th>250</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GAUSSIAN</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. POISSON</td>
<td>0.50</td>
<td>0.58</td>
<td>0.70</td>
<td>0.99</td>
<td>1.00</td>
</tr>
<tr>
<td>POISSON</td>
<td>0.21</td>
<td>0.49</td>
<td>0.66</td>
<td>0.98</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>LINEAR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. POISSON</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.98</td>
</tr>
<tr>
<td>POISSON</td>
<td>0.95</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Table 1: Proportion of experiments where CONDITIONAL POISSON (C. POISSON) and POISSON beat the benchmark models LINEAR and GAUSSIAN in terms of NMI. For each of the subset sizes, we sampled 100 different subsets of BERT dimensions at random.

---

9While accuracy only cares about whether predictions are correct, NMI penalizes miscalibrated predictions since it is proportional to the negative log likelihood (Guo et al. 2017).

10Recently, both Stafaćz et al. (2022), who utilize the POISSON probe, and Antverg and Belinkov (2021) find evidence supporting a similar phenomenon.
For each subset size considered, we take our averages over 100 randomly sampled subsets of BERT dimensions.

A growing interest in interpretability has led to a flurry of work in assessing what pre-trained representations know about language. To this end, diverse methods have been employed, such as the construction of challenge sets that evaluate how well representations model particular phenomena (Linzen, Dupoux, and Goldberg 2016; Gulordava et al. 2018; Goldberg 2019; Goodwin, Sinha, and O’Donnell 2020), and visualization methods (Kádár, Chrupała, and Alishahi 2017; Rethmeier, Saxena, and Augenstein 2020). Work on probing comprises a major share of this endeavor (Belinkov and Glass 2019; Belinkov 2021). This has taken the form of 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; Ravichander, Belinkov, and Hovy 2021; Geva et al. 2022, *inter alia*).

Efforts have ranged widely, but most of these focus on extrinsically rather than intrinsic probing. Most work on the latter has focused primarily on ascribing roles to individual neurons through methods such as visualization (Karpathy, Johnson, and Fei-Fei 2015; Li et al. 2016) and ablation (Li, Monroe, and Jurafsky 2016). For example, recently Lakretz et al. (2019) conduct an in-depth study of how 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 information related to gender and societal biases is encoded in individual neurons and how it is being propagated through different model components.

### 6 Related Work

A growing interest in interpretability has led to a flurry of work in assessing what pre-trained representations know about language. To this end, diverse methods have been employed, such as the construction of challenge sets that evaluate how well representations model particular phenomena (Linzen, Dupoux, and Goldberg 2016; Gulordava et al. 2018; Goldberg 2019; Goodwin, Sinha, and O’Donnell 2020), and visualization methods (Kádár, Chrupała, and Alishahi 2017; Rethmeier, Saxena, and Augenstein 2020). Work on probing comprises a major share of this endeavor (Belinkov and Glass 2019; Belinkov 2021). This has taken the form of 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; Ravichander, Belinkov, and Hovy 2021; Geva et al. 2022, *inter alia*).

Efforts have ranged widely, but most of these focus on extrinsic rather than intrinsic probing. Most work on the latter

### 7 Conclusion

In this paper, we introduce a new method for training intrinsic probes. We construct 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 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 generally gives optimal performance. Next, we investigate information distribution for each attribute for all available languages. Finally, we find empirical evidence for overlap in the specific neurons used to encode morphosyntactic properties across languages. Future work will focus on the application of this method to investigate the cross-lingual patterns in representations learned by pre-trained language models in a typologically diverse setting.

---

### Table 2: Mean and standard deviation of NMI for the Poisson, Conditional Poisson, Linear (Dalvi et al. 2019) and Gaussian (Torroba Hennigen, Williams, and Cotterell 2020) probes for all language–attribute pairs (top) and mean NMI and standard deviation for the Conditional Poisson, Poisson and Upper Bound for 6 selected language–attribute pairs (bottom). For each subset size considered, we take our averages over 100 randomly sampled subsets of BERT dimensions.

<table>
<thead>
<tr>
<th>Probe</th>
<th>10</th>
<th>50</th>
<th>100</th>
<th>250</th>
<th>500</th>
<th>768</th>
</tr>
</thead>
<tbody>
<tr>
<td>COND. POISSON</td>
<td>0.04 ± 0.03</td>
<td>0.18 ± 0.10</td>
<td>0.31 ± 0.14</td>
<td>0.54 ± 0.17</td>
<td>0.69 ± 0.15</td>
<td>0.71 ± 0.15</td>
</tr>
<tr>
<td>POISSON</td>
<td>−0.18 ± 0.28</td>
<td>0.03 ± 0.24</td>
<td>0.22 ± 0.21</td>
<td>0.53 ± 0.17</td>
<td>0.69 ± 0.16</td>
<td>0.71 ± 0.19</td>
</tr>
<tr>
<td>LINEAR</td>
<td>−0.28 ± 0.35</td>
<td>−0.18 ± 0.36</td>
<td>−0.06 ± 0.35</td>
<td>0.24 ± 0.33</td>
<td>0.59 ± 0.21</td>
<td>0.78 ± 0.14</td>
</tr>
<tr>
<td>GAUSSIAN</td>
<td>−0.15 ± 0.43</td>
<td>−1.20 ± 2.82</td>
<td>−3.97 ± 8.62</td>
<td>−61.70 ± 186.15</td>
<td>−413.80 ± 1175.31</td>
<td>−1067.08 ± 2420.08</td>
</tr>
<tr>
<td>COND. POISSON</td>
<td>0.04 ± 0.03</td>
<td>0.21 ± 0.11</td>
<td>0.35 ± 0.16</td>
<td>0.58 ± 0.2</td>
<td>0.77 ± 0.19</td>
<td>0.74 ± 0.16</td>
</tr>
<tr>
<td>POISSON</td>
<td>−0.10 ± 0.10</td>
<td>0.11 ± 0.13</td>
<td>0.28 ± 0.17</td>
<td>0.57 ± 0.20</td>
<td>0.73 ± 0.20</td>
<td>0.76 ± 0.18</td>
</tr>
<tr>
<td>UPPER BOUND</td>
<td>0.10 ± 0.06</td>
<td>0.36 ± 0.16</td>
<td>0.52 ± 0.19</td>
<td>0.70 ± 0.20</td>
<td>0.79 ± 0.17</td>
<td>0.81 ± 0.13</td>
</tr>
</tbody>
</table>

Figure 3: 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.

[Image of Table 2]

## Conclusion

In this paper, we introduce a new method for training intrinsic probes. We construct 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 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 generally gives optimal performance. Next, we investigate information distribution for each attribute for all available languages. Finally, we find empirical evidence for overlap in the specific neurons used to encode morphosyntactic properties across languages. Future work will focus on the application of this method to investigate the cross-lingual patterns in representations learned by pre-trained language models in a typologically diverse setting.
References


Li, J.; Monroe, W.; and Jurafsky, D. 2016. Understanding Neural Networks through Representation Erasure. CoRR, abs/1612.08220.


