A Latent-Variable Model for Intrinsic Probing

Karolina Stańczak[*1], **Lucas Torroba Hennigen**[*2], **Adina Williams**[3],
Ryan Cotterell[4], **Isabelle Augenstein**[1]

[1]University of Copenhagen
[2]Massachusetts Institute of Technology
[3]Meta AI
[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 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 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, *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[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

---

---

[1]Arabic, English, Finnish, Polish, Portuguese, and Russian

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]

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 Π to be the set of values that some property of interest can take, e.g., $\Pi = \{\text{SINGULAR}, \text{PLURAL}\}$ 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 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):

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

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 $C \subseteq D$ 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* $p_{\boldsymbol{\theta}}(\pi \mid \boldsymbol{h}_C)$ 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:

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

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 $\boldsymbol{\theta}$. We achieve this 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:

$$\sum_{n=1}^{N} \log \sum_{C \subseteq D} p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)}) \qquad (3)$$
$$\geq \sum_{n=1}^{N} \left(\mathbb{E}_{q_{\boldsymbol{\phi}}}\left[\log p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)})\right] + \mathrm{H}(q) \right)$$

which follows from Jensen's inequality, where $\mathrm{H}(q_{\boldsymbol{\phi}})$ is the entropy of $q_{\boldsymbol{\phi}}$. The derivation of the variational lower bound

is shown below:

$$\sum_{n=1}^{N} \log \sum_{C \subseteq D} p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)}) \qquad (4)$$

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

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

$$\geq \sum_{n=1}^{N} \mathbb{E}_{q_{\boldsymbol{\phi}}} \left[\log \frac{p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)})}{q_{\boldsymbol{\phi}}(C)} \right] \qquad (5)$$

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

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 $q_{\boldsymbol{\phi}}$. 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 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 $q_{\boldsymbol{\phi}}(C)$, 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 $\boldsymbol{\theta}$ and the variational parameters $\boldsymbol{\phi}$, 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:

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{q_{\boldsymbol{\phi}}} \left[\log p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)}) \right] \qquad (6)$$

$$= \mathbb{E}_{q_{\boldsymbol{\phi}}} \left[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)}) \right]$$

$$\approx \sum_{m=1}^{M} \left[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\pi^{(n)}, C^{(m)} \mid \boldsymbol{h}^{(n)}) \right]$$

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 trick (Williams 1992):

$$\nabla_{\boldsymbol{\phi}} \mathbb{E}_{q_{\boldsymbol{\phi}}} \left[\log p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)}) \right] \qquad (7)$$

$$= \mathbb{E}_{q_{\boldsymbol{\phi}}} \left[\log p_{\boldsymbol{\theta}}(\pi^{(n)}, C \mid \boldsymbol{h}^{(n)}) \nabla_{\boldsymbol{\phi}} \log q_{\boldsymbol{\phi}}(C) \right]$$

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

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:

$$q_{\boldsymbol{\phi}}(C) = \prod_{d \in C} \frac{w_d}{1 + w_d} \prod_{d \notin C} \frac{1}{1 + w_d} \qquad (8)$$

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 $\mathcal{O}(|D|)$ time:

$$\mathrm{H}(q_{\boldsymbol{\phi}}) = \log Z - \sum_{d=1}^{|D|} \frac{w_d}{1 + w_d} \log w_d \qquad (9)$$

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:

$$q_{\boldsymbol{\phi}}(C) = \underbrace{q_{\boldsymbol{\phi}}^{\mathrm{CP}}(C \mid |C| = k)}_{\text{Conditional Poisson}} q_{\boldsymbol{\phi}}^{\mathrm{size}}(k) \qquad (10)$$

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:

$$q_{\boldsymbol{\phi}}^{\mathrm{CP}}(C) = \mathbb{1}\{|C| = k\} \frac{\prod_{d \in C} w_d}{Z^{\mathrm{CP}}} \qquad (11)$$

A more elaborate dynamic program which runs in $\mathcal{O}(k\,|D|)$ may be used to compute Z^{CP} efficiently (Aires 1999). We may further compute the entropy $\mathrm{H}(q_{\boldsymbol{\phi}})$ and its the gradient in $\mathcal{O}(|D|^2)$ 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).[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 Uni-Morph (McCarthy et al. 2018) schema.[4] Then, we compute multilingual BERT (m-BERT) representations[5] 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 $\boldsymbol{h}_C$ 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 $p(\boldsymbol{h}, \pi) = p(\boldsymbol{h} \mid \pi)\,p(\pi)$, where the representation distribution $p(\boldsymbol{h} \mid \pi)$ is assumed to be Gaussian. Torroba Hennigen, 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).[6] 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.

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).[7] Due to computational cost, we limit our comparisons with UPPER BOUND to 6 randomly chosen morphosyntactic attributes,[8] 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 Π-valued random variable P and a $\mathbb{R}^{|C|}$-valued random variable H_C over masked representations:

$$\mathrm{MI}(P; H_C) = \mathrm{H}(P) - \mathrm{H}(P \mid H_C) \qquad (12)$$

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, 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. 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

---

[4]We adopt the code available at: https://github.com/unimorph/ud-compatibility.

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

[6]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, Williams, and Cotterell 2020) and use the performance-based objective in Eq. (1).

[7]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 LINEAR experiments take a few minutes or even seconds to run, compared to UPPER BOUND which takes multiple hours.

[8]English–Number, Portuguese–Gender and Noun Class, Polish–Tense, Russian–Voice, Arabic–Case, Finnish–Tense

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 λ_1 and λ_2, respectively. We follow the experimental set-up proposed by Dalvi et al. (2019), where we set $\lambda_1, \lambda_2 = 10^{-5}$ 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 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).

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
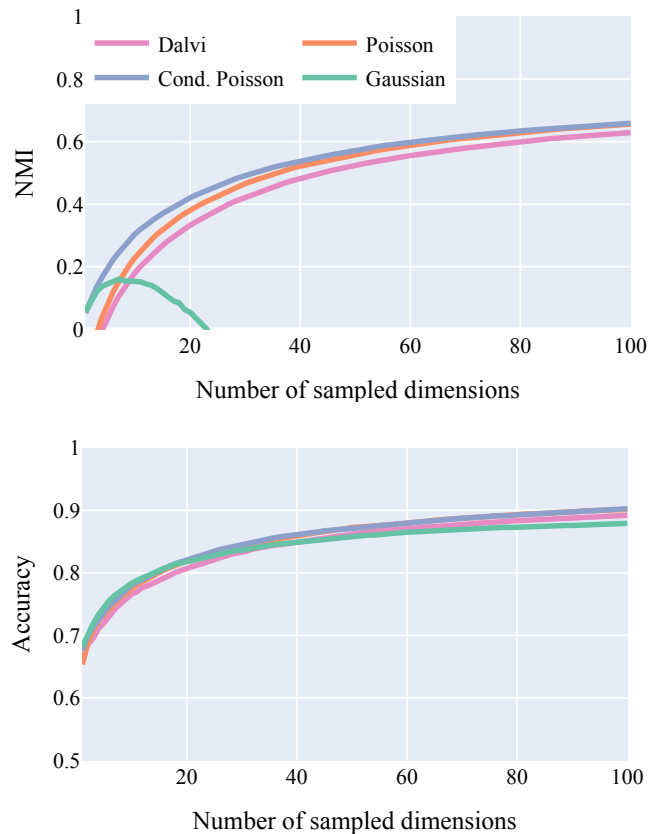

Figure 1: Comparison of the POISSON, CONDITIONAL POISSON, LINEAR (Dalvi et al. 2019) and GAUSSIAN (Torroba Hennigen, Williams, and Cotterell 2020) probes. We use the greedy selection approach in Eq. (1) to select the most informative dimensions, and average across all language–attribute pairs we probe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

probes to LINEAR (Dalvi et al. 2019) and GAUSSIAN (Torroba Hennigen, Williams, and Cotterell 2020). We refer to §4.3 for a discussion of the limitations of our method.

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 regime 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

	Number of dimensions				
	10	50	100	250	500
GAUSSIAN					
C. POISSON	0.50	0.58	0.70	0.99	1.00
POISSON	0.21	0.49	0.66	0.98	1.00
LINEAR					
C. POISSON	0.99	1.00	1.00	1.00	0.98
POISSON	0.95	0.99	1.00	1.00	0.97

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.

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,
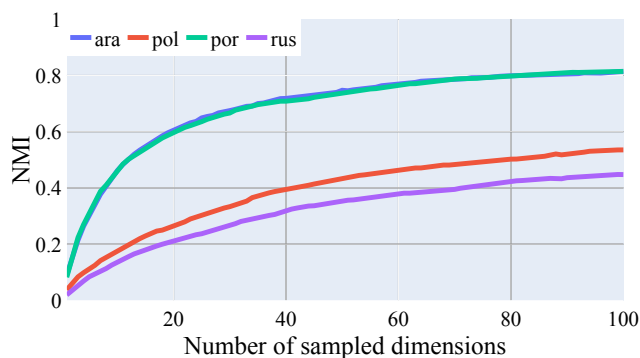

Figure 2: Comparison of the average NMI for gender dimensions in BERT for each of the available languages. We use the greedy selection approach in Eq. (1) to select the most informative dimensions, and average across all language–attribute pairs we probe for.

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]

---

[9]While 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).

[10]Recently, both Stańczak et al. (2022), who utilize the POISSON probe, and Antverg and Belinkov (2021) find evidence supporting a similar phenomenon.

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

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.
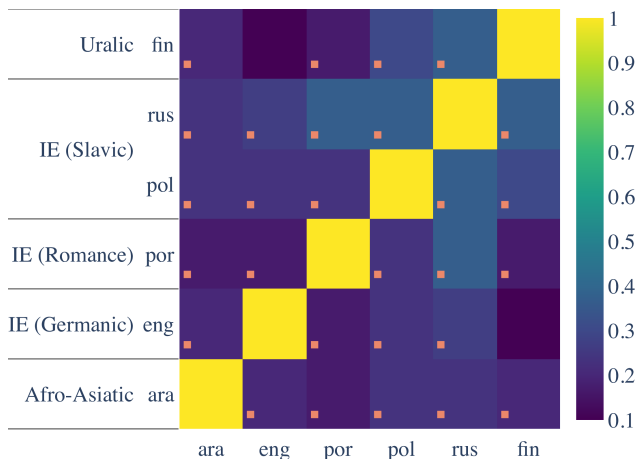

Figure 3: The percentage overlap between the top 30 most informative number dimensions in BERT for the probed languages. Statistically significant overlap, after Holm–Bonferroni family-wise error correction (Holm 1979), with $\alpha = 0.05$, is marked with an orange square.

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

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.

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.

References

Ács, J.; Kádár, Á.; and Kornai, A. 2021. Subword Pooling Makes a Difference. In *ACL*, 2284–2295.

Aires, N. 1999. Algorithms to find exact inclusion probabilities for conditional Poisson sampling and Pareto πs sampling designs. *Methodol. Comput. Appl. Probab.*, 1(4): 457–469.

Antverg, O.; and Belinkov, Y. 2021. On the Pitfalls of Analyzing Individual Neurons in Language Models. *arXiv:2110.07483*.

Bau, A.; Belinkov, Y.; Sajjad, H.; Durrani, N.; Dalvi, F.; and Glass, J. 2019. Identifying and Controlling Important Neurons in Neural Machine Translation. In *ICLR*.

Bau, D.; Zhu, J.-Y.; Strobelt, H.; Lapedriza, A.; Zhou, B.; and Torralba, A. 2020. Understanding the Role of Individual Units in a Deep Neural Network. *PNAS*.

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

Belinkov, Y.; and Glass, J. 2019. Analysis Methods in Neural Language Processing: A Survey. *TACL*, 7: 49–72.

Brown, P. F.; Della Pietra, S. A.; Della Pietra, V. J.; Lai, J. C.; and Mercer, R. L. 1992. An Estimate of an Upper Bound for the Entropy of English. *Comput. Linguist.*, 18(1): 31–40.

Conneau, A.; Kruszewski, G.; Lample, G.; Barrault, L.; and Baroni, M. 2018. What You Can Cram into a Single \$&!#* Vector: Probing Sentence Embeddings for Linguistic Properties. In *ACL*, 2126–2136.

Dalvi, F.; Durrani, N.; Sajjad, H.; Belinkov, Y.; Bau, A.; and Glass, J. 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.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, 4171–4186.

Durrani, N.; Sajjad, H.; Dalvi, F.; and Belinkov, Y. 2020. Analyzing Individual Neurons in Pre-trained Language Models. In *EMNLP*, 4865–4880.

Eisner, J. 2002. Parameter Estimation for Probabilistic Finite-State Transducers. In *ACL*, 1–8.

Gates, A. J.; Wood, I. B.; Hetrick, W. P.; and Ahn, Y.-Y. 2019. Element-Centric Clustering Comparison Unifies Overlaps and Hierarchy. *Sci. Rep.*, 9(1): 8574.

Geva, M.; Caciularu, A.; Wang, K. R.; and Goldberg, Y. 2022. Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space. *arXiv:2203.14680*.

Giulianelli, M.; Harding, J.; Mohnert, F.; Hupkes, D.; and Zuidema, W. 2018. Under the Hood: Using Diagnostic Classifiers to Investigate and Improve How Language Models Track Agreement Information. In *EMNLP*, 240–248.

Goldberg, Y. 2019. Assessing BERT's Syntactic Abilities. *arXiv:1901.05287*.

Goodwin, E.; Sinha, K.; and O'Donnell, T. J. 2020. Probing Linguistic Systematicity. In *ACL*, 1958–1969.

Gulordava, K.; Bojanowski, P.; Grave, E.; Linzen, T.; and Baroni, M. 2018. Colorless Green Recurrent Networks Dream Hierarchically. In *NAACL-HLT*, 1195–1205.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *ICML*, 1321–1330.

Hájek, J. 1964. Asymptotic Theory of Rejective Sampling with Varying Probabilities from a Finite Population. *Ann. Math. Stat.*, 35(4): 1491–1523.

Hall Maudslay, R.; Valvoda, J.; Pimentel, T.; Williams, A.; and Cotterell, R. 2020. A Tale of a Probe and a Parser. In *ACL*, 7389–7395.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Comput.*, 9(8): 1735–1780.

Hoffman, M. D.; Blei, D. M.; Wang, C.; and Paisley, J. 2013. Stochastic variational inference. *JMLR*, 14(4): 1303–1347.

Holm, S. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.*, 6(2): 65–70.

Kádár, Á.; Chrupała, G.; and Alishahi, A. 2017. Representation of Linguistic Form and Function in Recurrent Neural Networks. *Comput. Linguist.*, 43(4): 761–780.

Karpathy, A.; Johnson, J.; and Fei-Fei, L. 2015. Visualizing and Understanding Recurrent Networks. In *ICLR*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*. San Diego, CA.

Kulesza, A. 2012. Determinantal Point Processes for Machine Learning. *Found. Trends Mach. Learn.*, 5(2-3): 123–286.

Lakretz, Y.; Kruszewski, G.; Desbordes, T.; Hupkes, D.; Dehaene, S.; and Baroni, M. 2019. The Emergence of Number and Syntax Units in LSTM Language Models. In *NAACL-HLT*, 11–20.

Li, J.; Chen, X.; Hovy, E.; and Jurafsky, D. 2016. Visualizing and Understanding Neural Models in NLP. In *NAACL-HLT*, 681–691.

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

Li, Z.; and Eisner, J. 2009. First- and Second-Order Expectation Semirings with Applications to Minimum-Risk Training on Translation Forests. In *EMNLP*, 40–51.

Linzen, T.; Dupoux, E.; and Goldberg, Y. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *TACL*, 4: 521–535.

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

McCarthy, A. D.; Silfverberg, M.; Cotterell, R.; Hulden, M.; and Yarowsky, D. 2018. Marrying Universal Dependencies and Universal Morphology. In *ACL*, 91–101.

Mu, J.; and Andreas, J. 2020. Compositional Explanations of Neurons. In *NeurIPS*, volume 33, 17153–17163.

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

Nivre, J.; Agić, Ž.; Ahrenberg, L.; Antonsen, L.; Aranzabe, M. J.; Asahara, M.; Ateyah, L.; Attia, M.; Atutxa, A.; Augustinus, L.; Badmaeva, E.; Ballesteros, M.; Banerjee, E.; Bank, S.; Barbu Mititelu, V.; Bauer, J.; Bengoetxea, K.; Bhat, R. A.; Bick, E.; Bobicev, V.; Börstell, C.; Bosco, C.; Bouma, G.; Bowman, S.; Burchardt, A.; Candito, M.; Caron, G.; Cebiroğlu Eryiğit, G.; Celano, G. G. A.; Cetin, S.; Chalub, F.; Choi, J.; Cinková, S.; Çöltekin, Ç.; Connor, M.; Davidson, E.; de Marneffe, M.-C.; de Paiva, V.; Diaz de Ilarraza, A.; Dirix, P.; Dobrovoljc, K.; Dozat, T.; Droganova, K.; Dwivedi, P.; Eli, M.; Elkahky, A.; Erjavec, T.; Farkas, R.; Fernandez Alcalde, H.; Foster, J.; Freitas, C.; Gajdošová, K.; Galbraith, D.; Garcia, M.; Gärdenfors, M.; Gerdes, K.; Ginter, F.; Goenaga, I.; Gojenola, K.; Gökırmak, M.; Goldberg, Y.; Gómez Guinovart, X.; Gonzáles Saavedra, B.; Grioni, M.; Grūzītis, N.; Guillaume, B.; Habash, N.; Hajič, J.; Hajič jr., J.; Hà Mỹ, L.; Harris, K.; Haug, D.; Hladká, B.; Hlaváčová, J.; Hociung, F.; Hohle, P.; Ion, R.; Irimia, E.; Jelínek, T.; Johannsen, A.; Jørgensen, F.; Kaşıkara, H.; Kanayama, H.; Kanerva, J.; Kayadelen, T.; Kettnerová, V.; Kirchner, J.; Kotsyba, N.; Krek, S.; Laippala, V.; Lambertino, L.; Lando, T.; Lee, J.; Lê Hồng, P.; Lenci, A.; Lertpradit, S.; Leung, H.; Li, C. Y.; Li, J.; Li, K.; Ljubešić, N.; Loginova, O.; Lyashevskaya, O.; Lynn, T.; Macketanz, V.; Makazhanov, A.; Mandl, M.; Manning, C.; Mărănduc, C.; Mareček, D.; Marheinecke, K.; Martínez Alonso, H.; Martins, A.; Mašek, J.; Matsumoto, Y.; McDonald, R.; Mendonça, G.; Miekka, N.; Missilä, A.; Mititelu, C.; Miyao, Y.; Montemagni, S.; More, A.; Moreno Romero, L.; Mori, S.; Moskalevskyi, B.; Muischnek, K.; Müürisep, K.; Nainwani, P.; Nedoluzhko, A.; Nešpore-Bērzkalne, G.; Nguyễn Thị, L.; Nguyễn Thị Minh, H.; Nikolaev, V.; Nurmi, H.; Ojala, S.; Osenova, P.; Östling, R.; Øvrelid, L.; Pascual, E.; Passarotti, M.; Perez, C.-A.; and Perrier, G. e. a. 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.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*, volume 32, 8024–8035.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *NAACL-HLT*, 2227–2237.

Pimentel, T.; Valvoda, J.; Hall Maudslay, R.; Zmigrod, R.; Williams, A.; and Cotterell, R. 2020. Information-Theoretic Probing for Linguistic Structure. In *ACL*, 4609–4622.

Poliak, A.; Haldar, A.; Rudinger, R.; Hu, J. E.; Pavlick, E.; White, A. S.; and Van Durme, B. 2018. Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation. In *EMNLP*, 67–81.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR*, 21(140): 1–67.

Ravichander, A.; Belinkov, Y.; and Hovy, E. 2021. Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? In *EACL*, 3363–3377.

Rethmeier, N.; Saxena, V. K.; and Augenstein, I. 2020. TX-Ray: Quantifying and explaining model-knowledge transfer in (un-)supervised NLP. In Adams, R. P.; and Gogate, V., eds., *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence*, 197.

Rogers, A.; Kovaleva, O.; and Rumshisky, A. 2020. A Primer in BERTology: What We Know About How BERT Works. *TACL*, 8: 842–866.

Rush, A. 2020. Torch-Struct: Deep Structured Prediction Library. In *ACL*, 335–342.

Şahin, G. G.; Vania, C.; Kuznetsov, I.; and Gurevych, I. 2020. LINSPECTOR: Multilingual Probing Tasks for Word Representations. *Comput. Linguist.*, 46(2): 335–385.

Stańczak, K.; Ponti, E.; Torroba Hennigen, L.; Cotterell, R.; and Augenstein, I. 2022. Same Neurons, Different Languages: Probing Morphosyntax in Multilingual Pre-trained Models. In *NAACL-HLT*, 1589–1598.

Tang, G.; Sennrich, R.; and Nivre, J. 2020. Understanding Pure Character-Based Neural Machine Translation: The Case of Translating Finnish into English. In *CICLing*, 4251–4262.

Tenney, I.; Xia, P.; Chen, B.; Wang, A.; Poliak, A.; McCoy, R. T.; Kim, N.; Durme, B. V.; Bowman, S. R.; Das, D.; and Pavlick, E. 2018. What Do You Learn from Context? Probing for Sentence Structure in Contextualized Word Representations. In *ICLR*.

Torroba Hennigen, L.; Williams, A.; and Cotterell, R. 2020. Intrinsic Probing through Dimension Selection. In *EMNLP*, 197–216.

Vig, J.; Gehrmann, S.; Belinkov, Y.; Qian, S.; Nevo, D.; Singer, Y.; and Shieber, S. 2020. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *NeurIPS*, volume 33, 12388–12401.

Voita, E.; and Titov, I. 2020. Information-Theoretic Probing with Minimum Description Length. In *EMNLP*, 183–196.

Vulić, I.; Ponti, E. M.; Litschko, R.; Glavaš, G.; and Korhonen, A. 2020. Probing Pretrained Language Models for Lexical Semantics. In *EMNLP*, 7222–7240.

Williams, R. J. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Mach. Learn.*, 8: 229–256.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; and Brew, J. 2020. HuggingFace's Transformers: State-of-the-Art Natural Language Processing. *arXiv:1910.03771*.

Zhang, K.; and Bowman, S. 2018. Language Modeling Teaches You More than Translation Does: Lessons Learned Through Auxiliary Syntactic Task Analysis. In *EMNLP*, 359–361.

Zou, H.; and Hastie, T. 2005. Regularization and variable selection via the Elastic Net. *J R Stat Soc Series B Stat Methodol*, 67(2): 301–320.