Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals

Yanai Elazar[1,2] **Shauli Ravfogel**[1,2] **Alon Jacovi**[1] **Yoav Goldberg**[1,2]
[1]Computer Science Department, Bar Ilan University
[2]Allen Institute for Artificial Intelligence
{yanaiela,shauli.ravfogel,alonjacovi,yoav.goldberg}@gmail.com

Abstract

A growing body of work makes use of *probing* in order to investigate the working of neural models, often considered black boxes. Recently, an ongoing debate emerged surrounding the limitations of the probing paradigm. In this work, we point out the inability to infer behavioral conclusions from probing results, and offer an alternative method that focuses on how the information is being used, rather than on what information is encoded. Our method, *Amnesic Probing*, follows the intuition that the utility of a property for a given task can be assessed by measuring the influence of a causal intervention that removes it from the representation. Equipped with this new analysis tool, we can ask questions that were not possible before, for example, is part-of-speech information important for word prediction? We perform a series of analyses on BERT to answer these types of questions. Our findings demonstrate that conventional probing performance is not correlated to task importance, and we call for increased scrutiny of claims that draw behavioral or causal conclusions from probing results.[1]

1 Introduction

What drives a model to perform a specific prediction? What information is being used for prediction, and what would have happen if that information went missing? Because neural representation is opaque and hard to interpret, answering these questions is challenging.

The recent advancements in Language Models (LMs) and their success in transfer learning of many NLP tasks (e.g., Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019b) spiked interest in understanding how these models work and what is being encoded in them. One prominent methodology that attempts to shed light on those questions is *probing* (Conneau et al., 2018) (also known as *auxilliary prediction* [Adi et al., 2016] and *diagnostic classification* [Hupkes et al., 2018]). Under this methodology, one trains a simple model —a *probe*—to predict some desired information from the latent representations of the pre-trained model. High prediction performance is interpreted as evidence for the information being encoded in the representation. A key drawback of such an approach is that while it may indicate that the information can be extracted from the representation, it provides no evidence for or against the actual use of this information by the model. Indeed, Hewitt and Liang (2019) have shown that under certain conditions, above-random probing accuracy can be achieved even when the information that one probes for is linguistically meaningless noise, which is unlikely to have any use by the actual model. More recently, Ravichander et al. (2020) showed that models encode linguistic properties, even when not required at all for solving the task, questioning the usefulness and common interpretation of probing. These results call for higher scrutiny of *causal* claims based on probing results.

In this paper, we propose a counterfactual approach that serves as a step towards causal attribution: *Amnesic Probing* (see Figure 1 for a schematic view). We build on the intuition that if a property Z (e.g., part-of-speech) is being used for a task T (e.g., language modeling), then the *removal* of Z should negatively influence the ability of the model to solve the task. Conversely, when the removal of Z has little or no influence on the ability to solve T, one can argue that knowing Z is not a significant contributing factor in the strategy the model employs in solving T.

As opposed to previous work that focused on intervention in the input space (Goyal et al., 2019;

[1]The code is available at: https://github.com/yanaiela/amnesic_probing.

Transactions of the Association for Computational Linguistics, vol. 9, pp. 160–175, 2021. https://doi.org/10.1162/tacl_a_00359
Action Editor: Radu Florian. Submission batch: 7/2020; Revision batch: 9/2020; Published 3/2021.

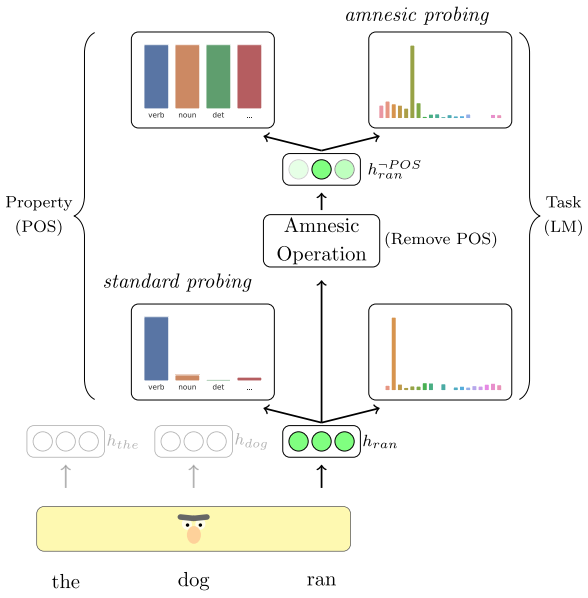


Figure 1: A schematic description of the proposed amnesic intervention: We transform the contextualized representation of the word ‘‘ran’’ so as to remove information (here, POS), resulting in a ‘‘cleaned’’ version $h_{ran}^{\neg POS}$. This representation is fed to the word-prediction layer and the behavioral influence of POS erasure is measured.

Kaushik et al., 2020; Vig et al., 2020) or in specific neurons (Vig et al., 2020), our intervention is done on the representation layers. This makes it easier than changing the input (which is non-trivial) and more efficient than querying hundred of neurons (which become combinatorial when considering the effect of multiple neurons simultaneously).

We demonstrate that amnesic probing can function as a debugging and analysis tool for neural models. Specifically, by using amnesic probing we show how to deduce whether a property is used by a given model in prediction.

In order to build the counterfactual representations, we need a function that operates on a pre-trained representation and returns a counterfactual version which no longer encodes the property we focus on. We use the recently proposed algorithm for neutralizing linear information: *Iterative Null-space Projection (INLP)* (Ravfogel et al., 2020). This approach allows us to ask the counterfactual question: ‘‘How will the prediction of a task differ without access to some property?’’ (Pearl and Mackenzie, 2018). This approach relies on the assumption that the usefulness of some information can be measured by neutralizing it from the representation, and witnessing the resulting *behavioral* change. It echoes the basic idea of ablation tests where one removes some component and measures the influence of that intervention.

We study several linguistic properties such as part-of-speech (POS) and dependency labels. Overall, we find that as opposed to the common belief, high probing performance does not mean that the probed information is used for predicting the main task (§4). This is consistent with the recent findings of Ravichander et al. (2020). Our analysis also reveals that the properties we examine are often being used differently in the *masked* setting (which is mostly used in LM training) and in the *non-masked* setting (which is commonly used for probing or fine-tuning) (§5). We then dive deeper into a more fine-grained analysis, and show that not all of the linguistic property labels equally influence prediction (§6). Finally, we re-evaluate previous claims about the way that BERT process the traditional NLP pipeline (Tenney et al., 2019a) with amnesic probing and provide a novel interpretation on the utility of different layers (§7).

2 *Amnesic Probing*

2.1 Setup and Formulation

Given a set of labeled data of data points $X = x_1, \ldots, x_n$[2] and task labels $Y = y_1, \ldots, y_n$ we analyze a model f that predicts the labels Y from X: $\hat{y}_i = f(x_i)$. We assume that this model is composed of two parts: an encoder h that transforms input x_i into a representation vector $\mathbf{h}_{x_i}$ and a classifier c that is used for predicting $\hat{y}_i$ based on $\mathbf{h}_{x_i}$: $\hat{y}_i = c(h(x_i))$. We refer by *model* to the component that follows the encoding function h and is used for the classification of the task of interest y. Each data point x_i is also associated with a *property* of interest z_i which represents additional information, which may or may not affect the decision of the classifier c.

In this work, we are interested in the change in prediction of the classifier c on the prediction $\hat{y}_i$ which is caused due to the removal of the property Z from the representation $h(x_i)$, that is $h(x_i)^{\neg Z}$.

2.2 *Amnesic Probing* with INLP

Under the counterfactual approach, we aim to evaluate the behavioral influence of a specific type of information Z (e.g., POS) on some task

[2]The data points can be words, documents, images, etc., based on the application.

(e.g., language modeling). To do so, we selectively remove this information from the representation and observe the change in the behavior of the model on the main task.

One commonly used method for information removal relies on adversarial training through the gradient reversal layer technique (Ganin et al., 2016). However, this techniques requires changing the original encoding by retraining the model, which is not desired in our case as we wish to study the original model's behavior. Additionally, Elazar and Goldberg (2018) found that this technique does not completely remove all the information from the learned representation.

Instead, we make use of a recently proposed algorithm called Iterative Nullspace Projection (INLP) (Ravfogel et al., 2020). Given a labeled dataset of representations H, and a property to remove, Z, INLP neutralizes the ability to linearly predict Z from H. It does so by training a sequence of linear classifiers (probes) $c_1, \ldots, c_k$ that predict Z, interpreting each one as conveying information on a unique direction in the latent space that corresponds to Z, and iteratively removing each of these directions. Concretely, we assume that the ith probe c_i is parameterized by a matrix W_i. In the ith iteration, c_i is trained to predict Z from H[3], and the data is projected onto its nullspace using a projection matrix $P_{N(W_i)}$. This operation guarantees $W_i P_{N(W_i)} H = 0$, i.e., it neutralizes the features in the latent space which were found by W_i to be indicative to Z. By repeating this process until no classifier achieves above-majority accuracy, INLP removes *all* such features.[4]

Amnesic Probing vs. Probing Note that amnesic probing *extends* conventional probing, as it is only relevant in cases where the property of interest can be predicted from the representation. If a probe gets random accuracy, the information cannot be used by the model to begin with. As such, amnesic probing can be seen as a complementary method, which inspects probe accuracy as a first step, but then proceeds to derive behavioral outcomes from the directions associated with the probe, with respect to a specific task.

[3]Concretely, we use linear SVM (Pedregosa et al., 2011).

[4]All relevant directions are removed to the extent they are identified by the classifiers we train. Therefore, we run INLP until the last linear classifier achieves a score within one point above majority accuracy on the development set.

2.3 Controls

The usage of INLP in this setup involves some subtleties we aim to account for: (1) Any modification to the representation, regardless of whether it removes information necessary to the task, may cause a decrease in performance. Can the drop in performance be attributed solely to the modification of the representation? (2) The removal of any property using INLP may also cause removal of correlating properties. Does the removed information only pertain to the property in question?

Control Over Information In order to control for the information loss of the representations, we make use of a baseline that removes the same number of directions as INLP does, but randomly.

For every INLP iteration the data matrix's rank decreases by the number of labels of the inspected property. This operation removes information from the representation which might be used for prediction. Using this control, *Rand*, instead of finding the directions using a classifier that learned some task, we generate random vectors from a uniform distribution, that accounts for random directions. Then, we construct the projection matrix as in INLP, by finding the intersection of nullspaces.

If the *Rand* impact on performance is lower than the impact of amnesic probing for some property, we conclude that we removed important directions for the main task. Otherwise, when the *Rand* control has a similar or higher impact, we conclude that there is no evidence for property usage for the main task.

Control over Selectivity[5] The result of the amnesic probing is taken as an indication to whether or not the model we query makes use of the inspected property for prediction. However, the removed features might solely correlate with the property (e.g., word position in the sentence has a nonzero correlation to syntactic function). To what extent is the information removal process we employ selective to the property in focus?

We test that by explicitly providing the gold information that has been removed from the

[5]Not to be confused with Hewitt and Liang (2019) Selectivity. Although recommended to use when performing standard probing, we argue it does not fit as a control for *amnesic probing* and provide a detailed explanation in Appendix B.

representation, and finetuning the subsequent layers (while the rest of the network is frozen). Restoring the original performance is taken as evidence that the property we aimed to remove is enough to account for the damage sustained by the amnesic intervention (it may still be the case that the intervention removes unrelated properties; but given the explicitly-provided property information, the model can make up for the damage). However, if the original performance is not restored, this indicates that the intervention removed more information than intended, and this cannot be accounted for by merely explicitly providing the value of the single property we focused on.

Concretely, we concatenate feature vectors of the studied property to the amnesic representations. Those vectors are 32-dimensional, and are initialized randomly, with a unique vector for each value of the property of interest. Those are fine-tuned until convergence. We note that as the new representation vectors are of a higher dimension than the original ones, we cannot use the original matrix. For an easier learning process, we use the original embedding matrix and concatenate it with a new embedding matrix, randomly initialized, and treat it as the new decision function.

3 Studying BERT: Experimental Setup

3.1 Model

We use our proposed method to investigate BERT (Devlin et al., 2019),[6] a popular and competitive masked language model (MLM) that has recently been the subject of many analysis works (e.g., Hewitt and Manning, 2019; Liu et al., 2019a; Tenney et al., 2019a). While most probing works focus on the ability to *decode* a certain linguistic property of the input text from the representation, we aim to understand which information is being *used* by it when predicting words from context. For example, we seek to answer questions such as the following: ''Is POS information used by the model in word prediction?'' The following experiments focus on language modeling, as a basic and popular task, but our method is more widely applicable.

[6]Specifically, BERT-BASE-UNCASED (Wolf et al., 2019).

3.2 Studied Properties

We focus on six tasks of sequence tagging: coarse and fine-grained part-of-speech tags (*c-pos* and *f-pos*, respectively); syntactic dependency labels (*dep*); named-entity labels (*ner*); and syntactic constituency boundaries[7] that mark the beginning and the end of a phrase (*phrase start*, and *phrase end*, respectively).

We use the training and dev data of the following datasets for each task: English UD Treebank (McDonald et al., 2013) for *c-pos*, *f-pos*, and *dep*; and English OntoNotes (Weischedel et al., 2013) for *ner*, *phrase start*, and *phrase end*. For training, we use 100,000 random tokens from those datasets.

3.3 Metrics

We report the following metrics:

LM accuracy: Word prediction accuracy.

Kullback-Leibler Divergence (D_{KL}): We calculate the D_{KL} between the distribution of the model over tokens, before and after the amnesic intervention. This measure focuses on the entire distribution, rather than the correct token only. Larger values implies a more significant change.

4 To Probe or Not to Probe?

By using the probing technique, different linguistic phenomenon such as POS, dependency information, and NER (Tenney et al., 2019a; Liu et al., 2019a; Alt et al., 2020) have been found to be ''easily extractable'' (typically using linear probes). A naive interpretation of these results may conclude that because information can be easily extracted by the probing model, this information is being used for the predictions. We show that this is not the case. Some properties such as syntactic structure and POS are very informative and are being used in practice to predict words. However, we also find some properties, such as phrase markers, which the model *does not* make use of when predicting tokens, in contrast to what one can naively deduce from probing results. This finding is in line with a recent work that observed the same behavior (Ravichander et al., 2020).

For each linguistic property, we report the probing accuracy using a linear model, as well as the

[7]Based on the Penn Treebank syntactic definitions.

		dep	*f-pos*	*c-pos*	*ner*	*phrase start*	*phrase end*
Properties	N. dir	738	585	264	133	36	22
	N. classes	41	45	12	19	2	2
	Majority	11.44	13.22	31.76	86.09	59.25	58.51
Probing	Vanilla	76.00	89.50	92.34	93.53	85.12	83.09
LM-Acc	Vanilla	94.12	94.12	94.12	94.00	94.00	94.00
	Rand	12.31	56.47	89.65	92.56	93.75	93.86
	Selectivity	73.78	92.68	97.26	96.06	96.96	96.93
	Amnesic	7.05	12.31	61.92	83.14	94.21	94.32
LM-D_{KL}	Rand	8.11	4.61	0.36	0.08	0.01	0.01
	Amnesic	8.53	7.63	3.21	1.24	0.01	0.01

Table 1: Property statistics, probing accuracies, and the influence of the amnesic intervention on the model's distribution over words. *dep*: dependency edge identity; *f-pos* and *c-pos*: fine-grained and coarse POS tags; *phrase start* and *phrase end*: beginning and end of phrases. *Rand* refers to replacing our INLP-based projection with removal of an equal number of random directions from the representation. The number of iterations per task can be inferred from: $N.dir/N.classes$.

word prediction accuracy after removing information about that property. The results are summarized in Table 1.[8] Probing achieves substantially higher performance over majority across all tasks. Moreover, after neutralizing the studied property from the representation, the performance on that task drops to majority (not presented in the table for brevity). Next, we compare the LM performance before and after the projection and observe a major drop for *dep* and *f-pos* information (decrease of 87.0 and 81.8 accuracy points, respectively), and a moderate drop for *c-pos* and *ner* information (decrease of 32.2 and 10.8 accuracy points, respectively). For these tasks, *Rand* performance on *LM-Acc* is lower than the original scores, but substantially higher than the Amnesic scores. Recall that the *Rand* experiment is done with respect to the amnesic probing, thus the number of removed dimension is the same, but each task may differ in the amount of dimensions removed. Furthermore, the D_{KL} metric shows the same trend (but in reverse, as a lower value indicates on a smaller change). We also report the selectivity results, where in most experiments the LM performance is restored, indicating amnesic probing works as expected. Note that the *dep* performance is not fully restored, thus some non-related features must have been coupled and removed with the dependency features. We believe that this happens in part due to the large number of removed directions.[9] These results suggests that to a large degree, the damage to LM performance is to be attributed to the specific information we remove, and not to rank-reduction alone. We conclude that dependency information, POS and NER are important for word prediction.

Interestingly, for *phrase start* and *phrase end* we observe a small *improvement* in accuracy of 0.21 and 0.32 points, respectively. The performance for the control on these properties is lower, therefore not only are these properties not important for the LM prediction at this part of the model, they slightly harm it. The last observation is rather surprising as phrase boundaries are coupled to the structure of sentences, and the words that form them. A potential explanation for this phenomenon is that this information is simply not being used at this part of the model, and is rather being processed in an earlier stage. We further inspect this hypothesis in Section 7. Finally, the probe accuracy does not correlate with task importance as measured by our method (Spearman correlation of 8.5, with a p-value of 0.871).

These results strengthen recent works that question the usefulness of probing as an analysis tool (Hewitt and Liang, 2019; Ravichander et al.,

[8]Note that because we use two different datasets, the UD Treebank for *dep*, *f-pos*, *c-pos*, and OntoNotes for *ner*, *phrase-start*, and *phrase-end*, the Vanilla LM-Acc performance differ between these setups.

[9]Since this experiment involves additional fine-tuning and is not entirely comparable to the vanilla setup (also due to the additional explicit information), we also experiment with concatenating the inspected features and finetuning. This results in an improvement of 3-4 points, above the vanilla experiment.

		dep	*f-pos*	*c-pos*	*ner*	*phrase start*	*phrase end*
Properties	N. dir	820	675	240	95	35	52
	N. classes	41	45	12	19	2	2
	Majority	11.44	13.22	31.76	86.09	59.25	58.51
Probing	Vanilla	71.19	78.32	84.40	90.68	85.53	83.21
LM-Acc	Vanilla	56.98	56.98	56.98	57.71	57.71	57.71
	Rand	4.67	24.69	54.55	56.88	57.46	57.27
	Selectivity	20.46	59.51	66.49	60.35	60.97	60.80
	Amnesic	4.67	6.01	33.28	48.39	56.89	56.19
LM-D$_{KL}$	Rand	7.77	6.10	0.45	0.10	0.02	0.04
	Amnesic	7.77	7.26	3.36	1.39	0.06	0.13

Table 2: Amnesic probing results for the *masked* representations. Properties statistics, word-prediction accuracy and D_{KL} results for the different properties inspected in this work. We report the vanilla word prediction accuracy and the Amnesic scores, as well as the Rand and 1-Hot controls which shows minimal information loss and high selectivity (except for the *dep* property which all information was removed). The D_{KL} is also reported for all properties in the last rows which show similar trends as the accuracy performance.

2020), but measure it from the usefulness of properties on the main task. We conclude that high probing performance does not entail this information is being used at a later part of the network.

5 What Properties are Important for the Pre-Training Objective?

Probing studies tend to focus on representations that are used for an end-task (usually the last hidden layer before the classification layer). In the case of MLM models, the words are not masked when encoding them for downstream tasks.

However, these representations are different from those used during the pre-training LM phase (of interest to us), where the input words are masked. It is therefore unclear if the conclusions drawn from conventional probing also apply to the way that the pre-trained model operates.

From this section on, unless mentioned otherwise, we report our experiments on the masked words. That is, given a sequence of tokens $x_1, \ldots, x_i, \ldots, x_n$ we encode the representation of each token x_i using its context, as follows: $x_1, \ldots, x_{i-1}, [MASK], x_{i+1}, \ldots, x_n$. The rest of the tokens remain intact. We feed these input tokens to BERT, and only use the masked representation of each word in its context $h(x_1, \ldots, x_{i-1}, [MASK], x_{i+1}, \ldots, x_n)_i$.

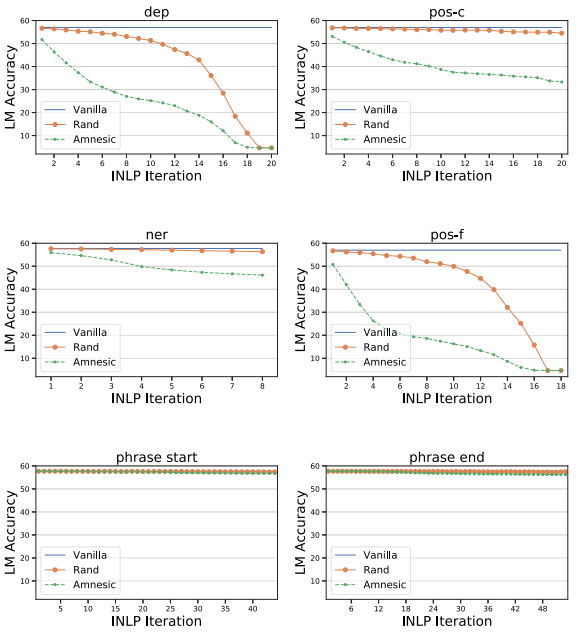


Figure 2: LM accuracy over INLP predictions, for the *masked* tokens version. We present both the vanilla word-prediction score (straight, blue line), as well as the control (orange, large circles) and INLP (green, small circles). Note that the number of removed dimensions per iteration differs, based on the number of classes of that property.

We repeat the experiments from Section 4 and report the results in Table 2. As expected, the LM accuracy drops significantly, as the model does not have access to the original word, and it has to infer it only based on context. Overall, the

c-pos	Vanilla	Rand	Amnesic	Δ
verb	46.72	44.85	34.99	11.73
noun	42.91	38.94	34.26	8.65
adposition	73.80	72.21	37.86	35.93
determiner	82.29	83.53	16.64	65.66
numeral	40.32	40.19	33.41	6.91
punctuation	80.71	81.02	47.03	33.68
particle	96.40	95.71	18.74	77.66
conjunction	78.01	72.94	4.28	73.73
adverb	39.84	34.11	23.71	16.14
pronoun	70.29	61.93	33.23	37.06
adjective	46.41	42.63	34.56	11.85
other	70.59	76.47	52.94	17.65

Table 3: Masked, c-pos removal, fine-grained LM analysis. Removing *c-pos* information and testing the accuracy performance of words, accumulating by their label. Δ is the difference in performance between the vanilla and Amnesic scores.

trends in the *masked* setting are similar to the *non-masked* setting. However, this is not always the case, as we show in Section 7. We also report the selectivity control. Notice that the performance for this experiment was improved across all tasks. In the case of *dep* and *f-pos*, where we had to neutralize most of the dimensions the performance does not fully recover. Note that the number of classes in those experiments might be a factor in the large performance gaps (expressed by the number of removed dimensions, N. dir, in the table). While not part of this study, it would be interesting to control for this factor in future work. However for the rest of the properties (*c-pos*, *ner*, and the *phrase-markers*) the performance is fully recovered, showing our methods' selectivity.

To further study the effect of INLP and inspect how the different dimensions removal affect performance, we display in Figure 2 the LM performance after each iteration, both with the amnesic probing and the control, and observe a consistent gap between them. Moreover, we highlight the difference in the slope for our method and the random direction removal. The amnesic probing exemplifies a much steeper slope than the random direction, indicating that the studied properties are indeed correlated with words prediction. We also provide the main task performance after each iteration in Figure 5 in the Appendix, which steadily decreases with each iteration.

c-pos	Vanilla	Amnesic	Δ
verb	56.98	55.60	1.38
noun	56.98	55.79	1.19
adposition	56.98	53.40	3.58
determiner	56.98	51.04	5.94
numeral	56.98	55.88	1.10
punctuation	56.98	53.12	3.86
particle	56.98	55.26	1.72
conjunction	56.98	54.29	2.69
adverb	56.98	55.64	1.34
pronoun	56.98	54.97	2.02
adjective	56.98	55.95	1.03

Table 4: Word prediction accuracy after fine-grained tag distinction removal, *masked* version. Rand control performance are all between 56.05 and 56.49 accuracy (with a maximum difference from vanilla of 0.92 points).

6 Specific Labels and Word Prediction

In the previous sections we observed the impact (or lack thereof) of different properties on word prediction. But when a property affects words prediction, are all words affected similarly? In this section, we inspect a more fine-grained version of the properties of interest, and study the impact of those on word predictions.

Fine-Grained Analysis When we remove the POS information from the representation, are nouns affected to the same degree as conjunctions? We repeat the *masked* experimental setting from Section 5, but this time we inspect the word prediction performance for the different labels. We report the results for the *c-pos* tagging in Table 3. We observe large differences in the word prediction performance before and after the POS removal between the labels. Nouns, numbers, and verbs show a relatively small impact in performance (8.64, 6.91, and 11.73 respectively), while conjunctions, particles and determiners demonstrate large performance drops (73.73, 77.66, and 65.65, respectively). We see that the information about POS labels at the word-level prediction is much more important in closed-set vocabularies (such as conjunctions and determiners) than with open vocabularies (such as nouns and verbs).

A manual inspection of predicted words after removing the POS information reveals that many of the changes are due to the transformation of function words to content words. For example, the words 'and', 'of', and 'a' become 'rotate', 'say', and 'final', respectively, in the inspected sentences. For a more quantitative analysis, we use a POS tagger in order to measure the POS label confusion before and after the intervention. Out of the 12,700 determiners conjuncions and punctuations, 200 of the predicted words by BERT were tagged as nouns and verbs before the intervention, compared to 3,982 after.

Removal of Specific Labels Following the observation that classes are affected differently when predicting words, we further investigate the differences of specific label removal. To this end, we repeat the amnesic probing experiments, but instead of removing the fine-grained information of a linguistic property, we make a cruder removal: The distinction between a specific label and the rest. For example, with POS as the general property, we now investigate whether the information of noun vs. the rest is important for predicting a word. We perform this experiment for all of the *pos-c* labels, and report the results in Table 4.[10]

We observe big performance gaps when removing different labels. For example, removing the distinctions between nouns and the rest, or verbs and the rest has minimal impact on performance. On the other hand, determiners and punctuations are highly affected. This is consistent with the previous observation on removing specific information. These results call for more detailed observations and experiments when studying a phenomenon as the fine-grained property distinction does not behave the same across labels.[11]

7 Behavior Across Layers

The results up to this section treat all of BERT's 'Transformer blocks' (Vaswani et al., 2017) as the encoding function and the embedding matrix as the model. But what happens when we remove the information of some linguistic property from earlier layers?

By using INLP to remove a property from an intermediate layer, we prevent the subsequent layer from using linearly present information originally stored in that layer. Though this operation does not erase all the information correlative with the studied property (as INLP only removes linearly present information), it makes it harder for the model to use this information. Concretely, we begin by extracting the representation of some text from the first k layers of BERT and then run INLP on these representations to remove the property of interest. Given that we wish to study the effect of a property on layer i, we project the representation using the corresponding projection matrix P_i that was learned on those representation, and then continue the encoding of the following layers.[12]

7.1 Property Recovery After an Amnesic Operation

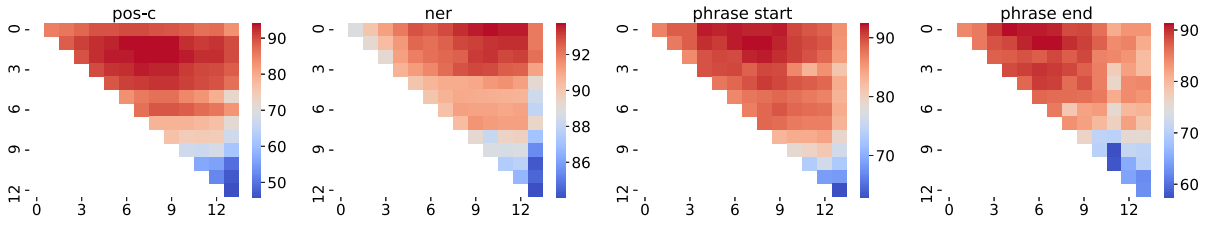
Is the property we linearly remove from a given layer recoverable by subsequent layers? We remove the information about some linguistic property from layer i, and learn a probe classifier on all subsequent layers $i+1, \ldots, n$. This tests how much information about this property the following layers have recovered. We experiment with the properties that could be removed without reducing too many dimensions: *pos-c*, *ner*, *phrase start*, and *phrase end*. These results are summarized in Figure 3, both for the non-masked version (upper row) and the masked version (lower row).

Notably, for the *pos-c*, non-masked version, the information is highly recoverable in subsequent layers when applying the amnesic operation on the first seven layers: the performance drops from the regular probing of that layer between 5.72 and 12.69 accuracy points. However, in the second part of the network, the drop is substantially larger: between 16.57 and 46.39 accuracy points. For the masked version, we witness an opposite trend: The *pos-c* information is much less recoverable in the lower parts of the network than the upper parts. In particular, the removal of *pos-c* from the second

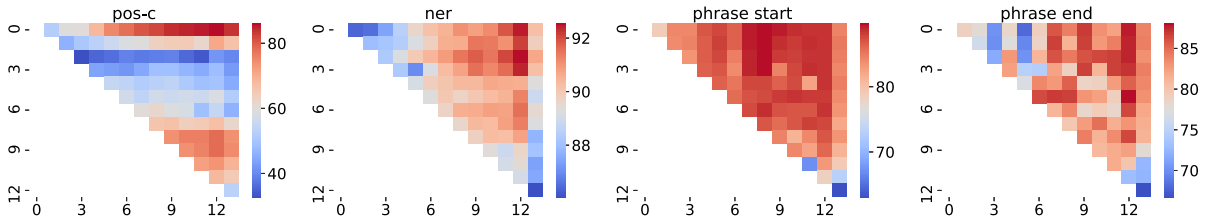
[10] In order to properly compare the different properties, we run INLP for solely 60 iterations, for each property. Since the 'other' tag is not common, we omit it from this experiment.

[11] We repeat these experiments with the other studied properties and observe similar trends.

[12] As the representations used to train INLP do not include BERTs' special tokens (e.g., 'CLS', 'SEP'), we also don't use the projection matrix on those tokens.



(a) Non-Masked version



(b) Masked version

Figure 3: Layer-wise removal. Removing from layer i (the rows) and testing probing performance on layer j (the columns). Top row (3a) is non-masked version, bottom row (3b) is masked.

layer appears to affect the rest of the layers, which do not manage to recover a high score on this task, ranging from 32.7 to 42.1 accuracy.

For all of the non-masked experiments the upper layers seem to make it harder for the subsequent layers to extract the property. In the masked version however, there is no consistent trend. It is harder to extract properties after the lower parts for *pos-c* and *ner*. For *phrase start* the upper part makes it harder for further extraction and for *phrase end* both the lower and upper parts make it harder, as opposed to the middle layers. Further research is needed in order to understand the significance of those findings, and whether or not they are related to information usage across layers.

This lead us to the final experiment where we test for the main task performance after an *amnesic* operation at the intermediate layers.

7.2 Re-rediscovering the NLP Pipeline

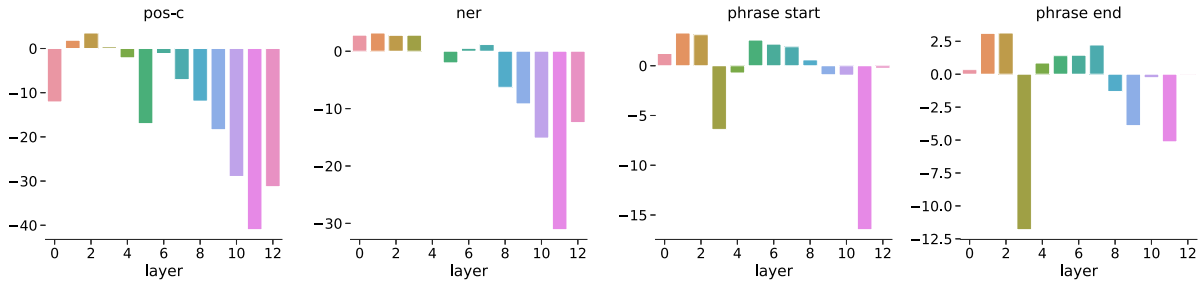
In the previous set of experiments, we measured how much of the signal removed in layer i is recovered in subsequent layers. We now study how the removal of information in layer i affects the word prediction accuracy at the final layer, in order to get a complementary measure for layer importance with respect to a property. The results for the different properties are presented in Figure 4, where we plot the difference in word prediction performance between the control and the amnesic probing when removing a linguistic property from a certain layer.

These results provide a clear interpretation on the internal function of BERT's layers. For the masked version (Figure 4), we observe that the *pos-c* properties are mostly important in layer 3 and its surrounding layers, as well as layer 12. However, this information is accurately extractable only towards the last layers. For *ner*, we observe that the main performance loss occurs at layer 4. For *phrase-markers* the middle layers are important: layers 5 and 7 for *phrase start* (although the absolute performance loss is not big) and layer 6 for *phrase end* contributes the most for the word prediction performance.

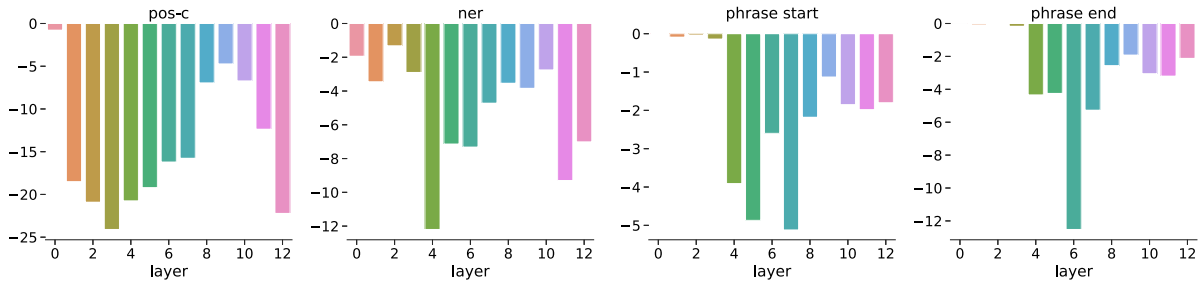
The story with the *non-masked* version is quite different (Figure 4). First, notice that the amnesic operation *improves* the LM performance for all properties, in some layers.[13] Second, the drop in performance peak across all properties is different than the *masked* version experiments. Particularly, it seems that for *pos-c*, when the words are non-masked in the input, the most important layer for *pos-c* is 11 (and not layer 3, as in the masked version), while this information is easily extractable (by standard probing) across all layers (above 80% accuracy).

Interestingly, the conclusions we draw on layer-importance from amnesic probing partly differ

[13] Giulianelli et al. (2018) observed a similar behavior by performing an intervention on LSTM activations.



(a) Non-Masked version



(b) Masked version

Figure 4: The influence of the different properties, from each layer on LM predictions. Top figure (4a) shows the results on the regular, *non-masked* version, bottom figure (4b) for the *masked* version. Colors allow ease of layer comparison across graphs.

from the ones in the ‘‘Pipeline processing’’ hypothesis (Tenney et al., 2019a), which aims to localize and attribute information processing of linguistic properties to parts of BERT (for the non-masked version).[14] On one hand the *ner* experiment trends are similar: the last layers are much more important than earlier ones (in particular, layer 11 is the most affected in our case, with a decrease of 31.09 accuracy points. On the other hand, in contrast to their hypotheses, we find that POS information, *pos-c* (which was considered to be more important in the earlier layers) affects the word prediction performance much more in the upper layers (40.99 accuracy loss in the 11th layer). Finally, we note that our approach performs an ablation of these properties in the representation space, which reveals which layers are actually responsible for processing properties, as opposed to Tenney et al. (2019a), who focused on where this information is easily extractable.

We note the big differences in behavior when analyzing the masked vs. the non-masked version of BERT, and call for future work to make a clearer distinctions between the two. Finally, we stress that the different experiments should not be compared between one setting to the other, and thus the different y-scales in the figures. This is due to confounding variables (e.g., the number of removed dimensions from the representations), which we do not control for in this work.

8 Related Work

With the established impressive performance of large pre-trained language models (Devlin et al., 2019; Liu et al., 2019b), based on the Transformer architecture (Vaswani et al., 2017), a large body of work started studying and gaining insight into how these models work and what do they encode.[15] For a thorough summary of these advancements we refer the reader to a recent primer on the subject (Rogers et al., 2020).

[14]We note that this work analyzes BERT-base, in contrast to Tenney et al. (2019a) who analyzed BERT-Large.

[15]These works cover a wide variety of topics, including grammatical generalization (Goldberg, 2019; Warstadt et al., 2019), syntax (Tenney et al., 2019b; Lin et al., 2019; Reif et al., 2019; Hewitt and Manning, 2019; Liu et al., 2019a), world knowledge (Petroni et al., 2019; Jiang et al., 2020), reasoning (Talmor et al., 2019), and common sense (Forbes et al., 2019; Zhou et al., 2019; Weir et al., 2020).

A particularly popular and easy-to-use interpretation method is probing (Conneau et al., 2018). Despite its popularity, recent works have questioned the use of probing as an interpretation tool. Hewitt and Liang (2019) have emphasized the need to distinguish between decoding and learning the probing tasks. They introduced *control tasks*, a consistent but linguistically meaningless attribution of labels to tokens, and have shown that probes trained on the control tasks often perform well, due to the strong lexical information held in the representations and learned by the probe. This leads them to propose a selectivity measure that aims to choose probes which achieve high accuracy only on linguistically-meaningful tasks. Tamkin et al. (2020) claim that probing cannot serve as an explanation of downstream task success. They observe that the probing scores do not correlate with the transfer scores achieved by fine-tuning.

Finally, Ravichander et al. (2020) show that probing can achieve non-trivial results for linguistic properties that were not needed for the task the model was trained on. In this work, we observe a similar phenomenon, but from a different angle. We actively remove some property of interest from the queried representation, and measure the impact of the *amnesic* representation of the property on the main task.

Two recent works study the probing paradigm from an information-theory perspective. Pimentel et al. (2020) emphasize that under a mutual-information maximization objective, ''better'' probes are increasingly more accurate, regardless of complexity. They use the data-processing inequality to question the rationale behind methods that focus on encoding, and propose *ease of extractability* as an alternative criterion. Voita and Titov (2020) follow this direction, using the concept of minimum description length (MDL, Rissanen, 1978) to quantify the total information needed to transmit both the probing model and the labels it predicts. Our discussion here is somewhat orthogonal to those on the meaning of encoding and probe complexity, as we focus on the information influence on the model's behavior, rather than on the ability to extract it from the representation.

Finally and concurrent to this work, Feder et al. (2020) have studied a similar question of a *causal* attribution of concepts to representations, using adversarial training guided by causal graphs.

9 Discussion

Intuitively, we would like to completely neutralize the abstract property we are interested in—e.g., POS information (*completeness*), as represented by the model—while keeping the rest of the representation intact (*selectivity*). This is a nontrivial goal, as it is not clear whether neural models actually have abstract and disentangled representations of properties such as POS, which are independent of other properties of the text. It may be the case that the representation of many properties is intertwined. Indeed, there is an ongoing debate on the assertion that certain information is ''encoded'' in the representation (Voita and Titov, 2020; Pimentel et al., 2020). However, even if a disentangled representation of the information we focus on exists, it is not clear how to detect it.

We implement the information removal operation with INLP, which gives a first order approximation using linear classifiers; we note, however, that one can in principle use other approaches to achieve the same goal. While we show that we do remove the linear ability to predict the properties and provide some evidence to the selectivity of this method (§2), one has to bear in mind that we remove only linearly-present information, and that the classifiers can rely on arbitrary features that happen to correlate with the gold label, be it a result of spurious correlations or inherent encoding of the direct property. Indeed, we observe this behavior in Section 7.1 (Figure 3), where we neutralize the information from certain layers, but occasionally observe higher probing accuracy in following layers. We thus stress that the information we remove in practice should be seen only as an approximation for the abstract information we are interested in, and that one has to be cautious of *causal* interpretations of the results. Although in this paper we use the INLP algorithm in order to remove linear information, *amnesic probing* is not restricted to removing linear information. When non-linear removal methods become available, they can be swapped instead of INLP. This stresses the importance of creating algorithms for non-linear information removal.

Another unanswered question is how to quantify the *relative* importance of different properties encoded in the representation for the word prediction task. The different erasure portion for different properties makes it hard to draw

conclusions on which property is more important for the task of interest. Although we do not make claims such as ‘‘dependency information is more important than POS’’, these are interesting questions that should be further discussed and researched.

10 Conclusions

In this work, we propose a new method, *Amnesic Probing*, which aims to quantify the influence of specific properties on a model that is trained on a task of interest. We demonstrate that conventional probing falls short in answering such behavioral questions, and perform a series of experiments on different linguistic phenomenon, quantifying their influence on the masked language modeling task. Furthermore, we inspect both unmasked and masked BERT's representation and detail the differences between them, which we find to be substantial. We also highlight the different influence of specific fine-grained properties (e.g., nouns and determiners) on the final task. Finally, we use our proposed method on the different layers of BERT, and study which parts of the model make use of the different properties. Taken together, we argue that compared with probing, counterfactual intervention—such as the one we present here—can provide a richer and more refined view of the way symbolic linguistic information is encoded and used by neural models with distributed representations.[16]

Acknowledgments

We would like to thank Hila Gonen, Amit Moryossef, Divyansh Kaushik, Abhilasha Ravichander, Uri Shalit, Felix Kreuk, Jurica Ševa, and Yonatan Belinkov for their helpful comments and discussions. We also thank the anonymous reviewers and the action editor, Radu Florian, for their valuable suggestions.

This project has received funding from the Europoean Research Council (ERC) under the Europoean Union's Horizon 2020 research and innovation programme, grant agreement no. 802774 (iEXTRACT). Yanai Elazar is grateful to be partially supported by the PBC fellowship for outstanding PhD candidates in Data Science.

[16]All of the experiments were logged and tracked using Weights and Biases (Biewald, 2020).

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *CoRR*, abs/1608.04207.

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. Probing linguistic features of sentence-level representations in relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1534–1545, Online. Association for Computational Linguistics.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from `wandb.com`.

Alexis Conneau, Germán 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. **DOI:** `https://doi.org/10.18653/v1/P18-1198`

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/D18-1002`

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2020. Causalm: Causal model explanation through counterfactual language models.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? *Proceedings of the 41st Annual Conference of the Cognitive Science Society*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1):2096–2030.

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 Blackbox NLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248. **DOI:** https://doi.org/10.18653/v1/W18-5426

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Yash Goyal, Uri Shalit, and Been Kim. 2019. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Empirical Methods in Natural Language Processing (EMNLP)*. **DOI:** https://doi.org/10.18653/v1/D19-1275

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 4129–4138.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926. **DOI:** https://doi.org/10.1613/jair.1.11196

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438. **DOI:** https://doi.org/10.1162/tacl_a_00324

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside berts linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. 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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic Books.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. 2011. Scikit-learn: Machine learning in python.

Journal of Machine Learning Research, 12(Oct): 2825–2830.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237. **DOI:** https://doi.org/10.18653/v1/N18-1202

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473. **DOI:** https://doi.org/10.18653/v1/D19-1250

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.420

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.647

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2020. Probing the probing paradigm: Does probing accuracy entail task relevance? *arXiv preprint arXiv:2005.00719*.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B. Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. In *Advances in Neural Information Processing Systems*, pages 8592–8600.

Jorma Rissanen. 1978. Modeling by shortest data description. *Automatica*, 14(5):465–471. **DOI:** https://doi.org/10.1016/0005-1098(78)90005-5

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how Bert works. *arXiv preprint arXiv:2002.12327*. **DOI:** https://doi.org/10.1162/tacl_a_00349

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019. olmpics – on what language model pre-training captures. **DOI:** https://doi.org/10.1162/tacl_a_00342

Alex Tamkin, Trisha Singh, Davide Giovanardi, and Noah Goodman. 2020. Investigating transferability in pretrained language models. *arXiv preprint arXiv:2004.14975*. **DOI:** https://doi.org/10.18653/v1/2020.findings-emnlp.125

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the Conference of the Association for Computational Linguistics, ACL*, pages 4593–4601. **DOI:** https://doi.org/10.18653/v1/P19-1452

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural NLP: The case of gender bias. *arXiv preprint arXiv:2004.12265*.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. *arXiv preprint arXiv:2003.12298*. **DOI:** https://doi.org/10.18653/v1/2020.emnlp-main.14

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretič, and Samuel R. Bowman. 2019. Investigating BERTs knowledge of language: Five analysis methods with NPIs. 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 2870–2880. **DOI:** https://doi.org/10.18653/v1/D19-1286

Nathaniel Weir, Adam Poliak, and Benjamin Van Durme. 2020. Probing neural language models for human tacit assumptions. In *42nd Annual Virtual Meeting of theCognitive Science Society (CogSci)*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, and Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2019. Evaluating commonsense in pre-trained language models. *arXiv preprint arXiv:1911.11931*.

Appendix A

We provide additional experiments that depict the performance of the main task (e.g., POS) performance during the INLP iterations in Figure 5.

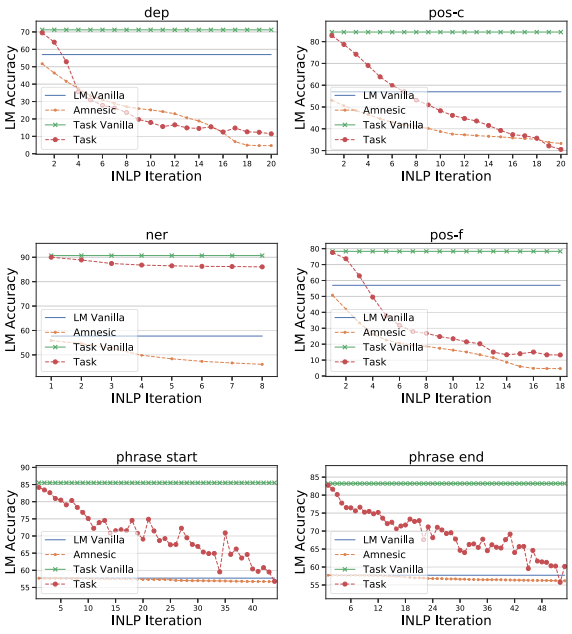


Figure 5: LM accuracy over INLP predictions, for the *masked* tokens version. We present both the Vanilla word-prediction score (straight, blue line), as well as Amnesic Probing (orange, small circles), and the main task performance (red, large circles). For reference we also provide the vanilla probing performance of each task (green, cross marks). Note that the number of removed dimensions per iteration differs, based on the number of classes of that property.

Appendix B Hewitt and Liang's Control Task

Control Task (Hewitt and Liang, 2019) has been suggested as a way to attribute the performance of the probe to extraction of *encoded* information, as opposed to lexical memorization. Our goal in this work, however, is not to extract information from the representation (as is done in conventional probing) but to measure a behavioural outcome. Since the control task is solved by lexical memorization, applying INLP on control task's classifiers erases lexical information (i.e., erases the ability to distinguish between arbitrary words), which is at the core of the LM objective and which is highly correlated with many of the other linguistic properties, such as POS. We argue that even if we do see a significant drop in performance with the control task, this says little on the validity of the results of removal of the linguistic property (e.g., POS). However, for completeness, we provide the results in Figure 6. As can be seen from this figure, this control's slope is smaller than

the one of the amnesic probing, suggesting that those directions have less behavioral influence. However, the slopes are steeper than the 'Rand' experiment. This is due to the identity removal of groups of words, due to the label shuffle, as suggested in their setup. This is the reason we believe this test is not adequate in our case, and why we provide other tests to control for our method: Rand and Selectivity (§2.3).

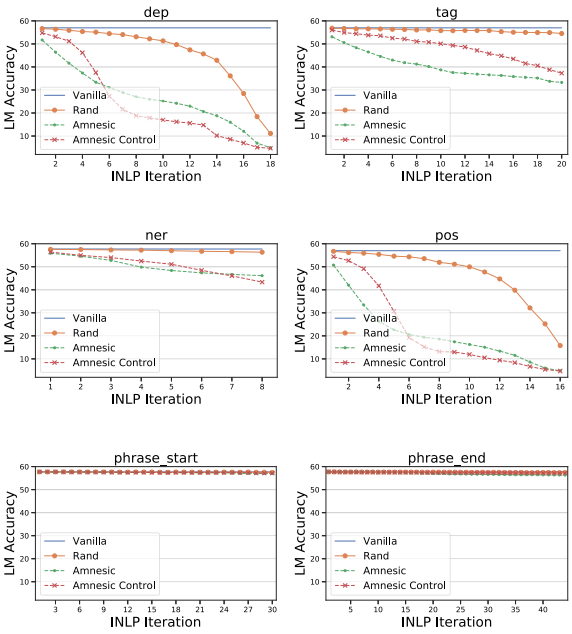


Figure 6: LM accuracy over INLP predictions, for the *masked* tokens version. We present both the Vanilla word-prediction score (straight, blue line), as well as Amnesic Probing (orange, small circles), and the control performance (orange, large circles). We also provide the Control results for selectivity test, proposed by Hewitt and Liang (2019) (red, crosses). Note that the number of removed dimensions per iteration differs, based on the number of classes of that property.