Amnesic Probing:
Behavioral Explanation with Amnesic Counterfactuals

Yanai Elazar, Shauli Ravfogel, Alon Jacovi and Yoav Goldberg

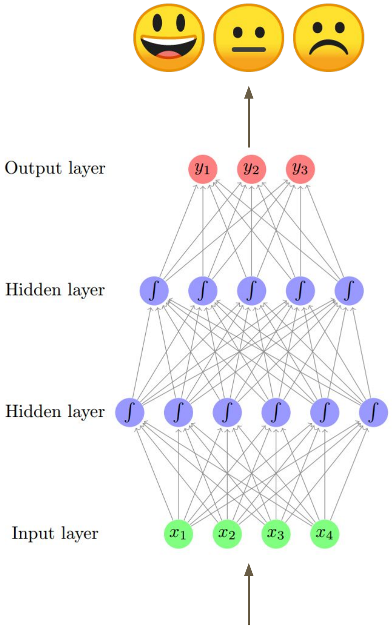
TACL 2021

Google Research, 3rd February, 2021

The State of NLP (ML?)

Output

Model



Input

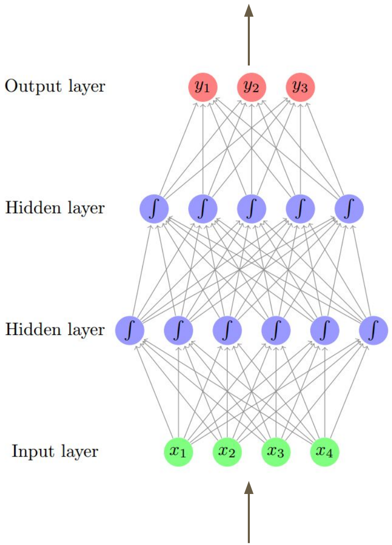
"Memories warm you up from the inside. But they also tear you apart."

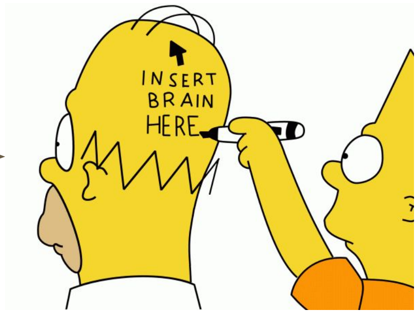
The State of NLP

We train these models for some task

Output

Model





Input

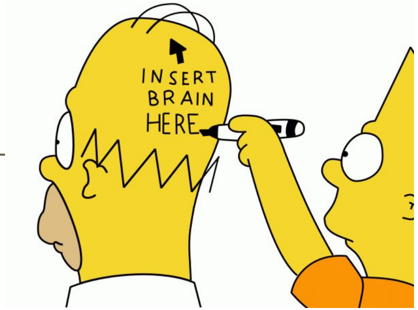
"Memories warm you up from the inside. But they also tear you apart."

The State of NLP

Output

And hope to get some "smart" model

Model



Input

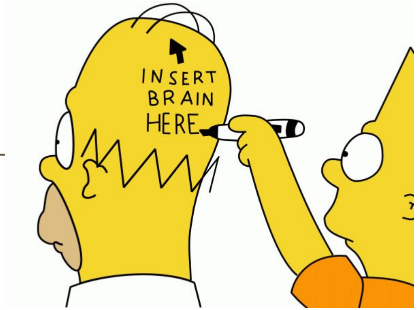
"Memories warm you up from the inside. But they also tear you apart."

The State of NLP

And hope to get some "smart" model

Output

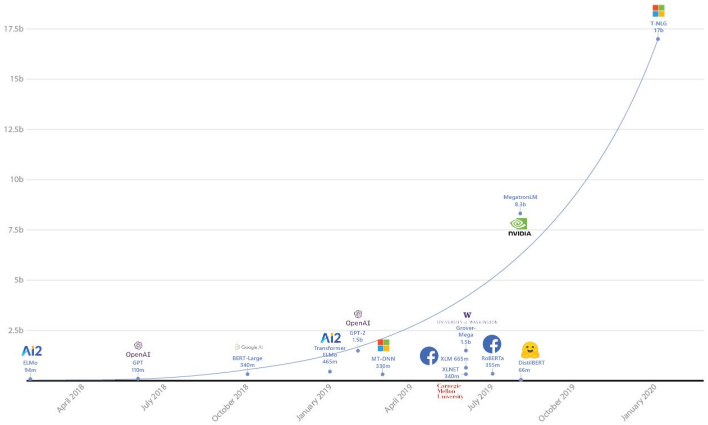
Model



Input

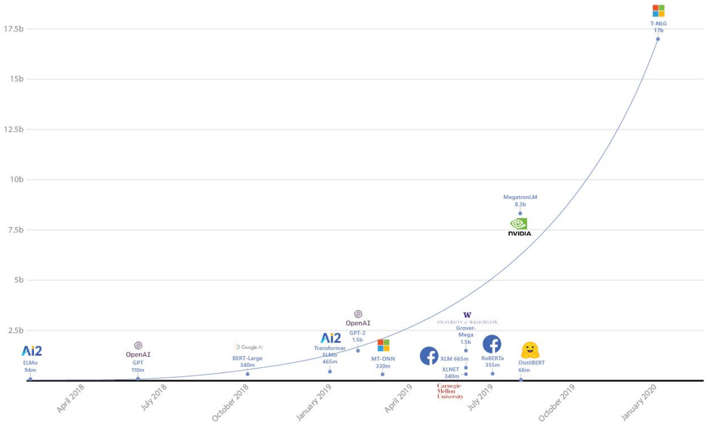
"Memories warm you up from the inside. But they also tear you apart."

The State of NLP: Sesame Street



Input *“Memories warm you up from the inside. But they also tear you apart.”*

The State of NLP: Inside Sesame Street



Input

"Memories warm you up from the inside. But they also tear you apart."

The State of NLP: Inside Sesame Street

Input *“Memories warm you up from the inside. But they also tear you apart.”*

Opening the BlackBox

you cannot cram the meaning of a whole %&!$# sentence into a single $&!# vector*

-- Ray Mooney

- So what can be crammed into that?

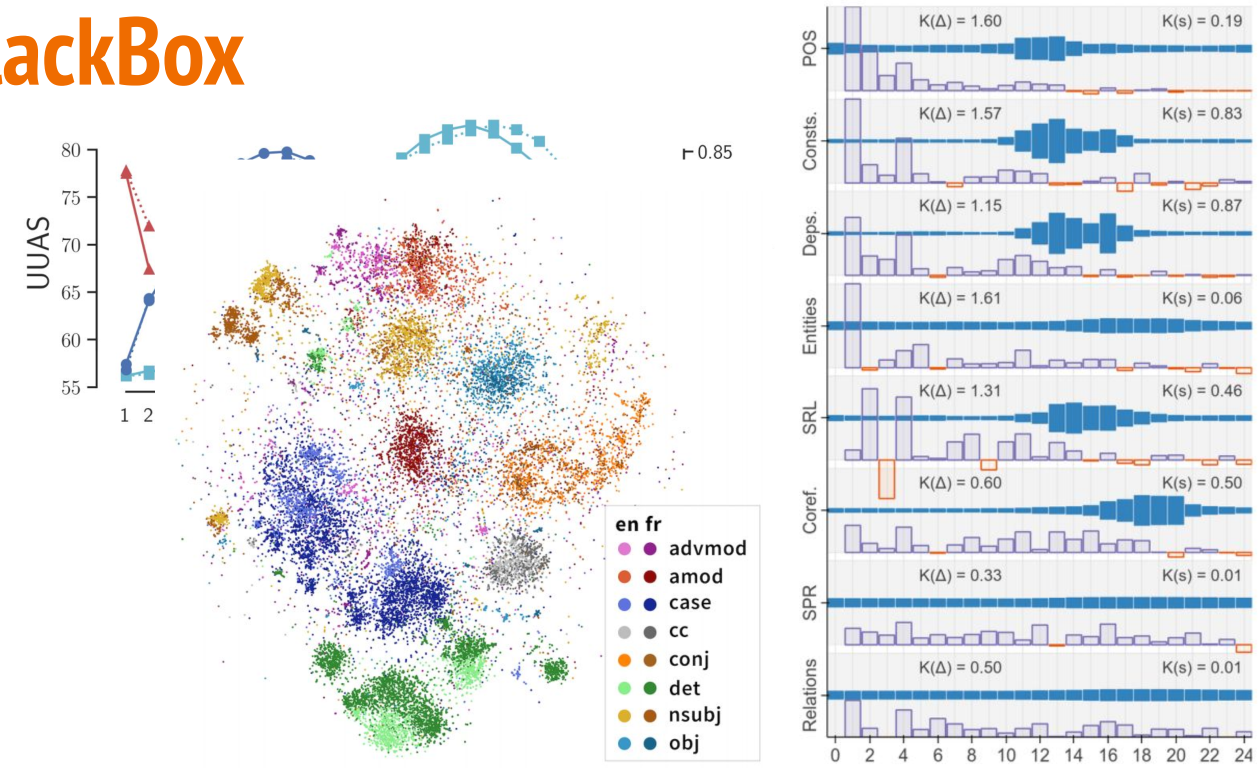
Opening the BlackBox

- Sentence Length
- Word Order
- Tense
- POS
- Tree depth
- Entities
- Coref.
- ...

Adi et al., 2016, Conneau et al., 2018, Hewitt and Manning, 2019, Tenney et al., 2019, Chi et al., 2020

Opening the BlackBox

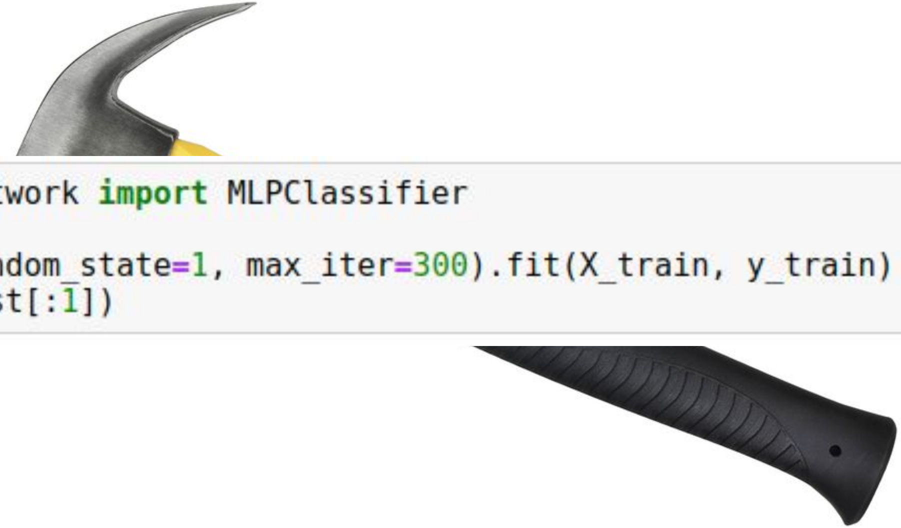
- Sentence Length
- Word Order
- Tense
- POS
- Tree depth
- Entities
- Coref.
- ...



Adi et al., 2016, Conneau et al., 2018, Hewitt and Manning, 2019, Tenney et al., 2019, Chi et al., 2020

Opening the BlackBox

But with what tool?



```
from sklearn.neural_network import MLPClassifier

clf = MLPClassifier(random_state=1, max_iter=300).fit(X_train, y_train)
clf.predict_proba(X_test[:1])
```

Opening the BlackBox

We (the community) found out that:

- Random Models perform surprisingly well (architecture inductive bias?)
- The hammer may have been too big

```
from sklearn.linear_model import LogisticRegression
clf = LogisticRegression().fit(X_train, y_train)
clf.predict_proba(X_test[:1])
```

```
from sklearn.neural_network import MLPClassifier

clf = MLPClassifier(random_state=1, max_iter=300).fit(X_train, y_train)
clf.predict_proba(X_test[:1])
```

Conneau et al., 2018, Hewitt and Liang, 2019

Opening the BlackBox

- Looking at the representations we find that different properties are encoded
- The first step of a long journey

Opening the BlackBox - This Work

- We ask a behavioral question:

  What information does a model use in order to make a prediction?

Also,

Is there a connection between the structural analysis (standard probes) to behavioral analysis (e.g. this work)

Amnesic Probing: A Behavioral Probe

Amnesic Probing: A Behavioral Analysis

- Understanding how our models work
- Interpretability and analysis tool, which allows to answer scientific questions (e.g. Does a LM use POS information?)
- Answer sensitive questions (e.g. Does the model use gender for making a decision)

Amnesic Probing: The Intuition

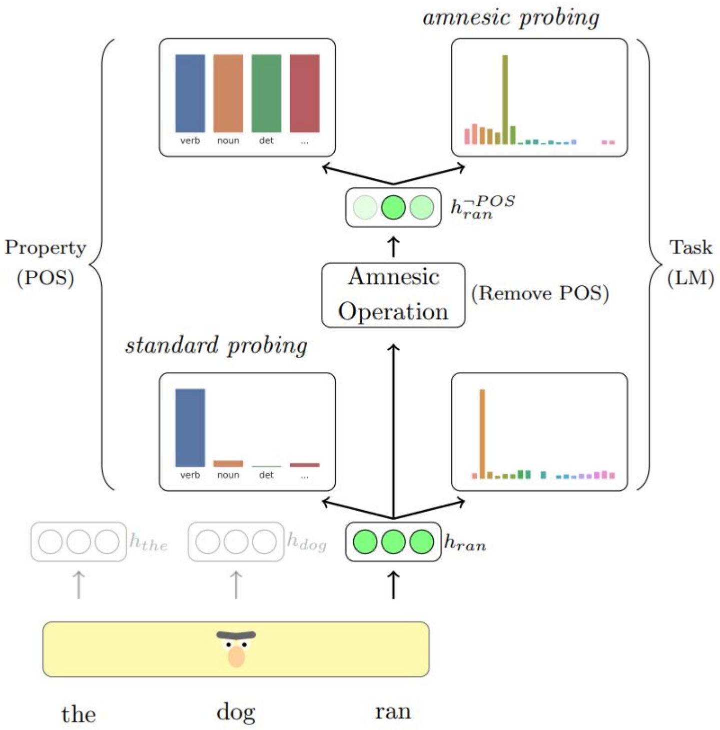
- Ablation Test (or *Counterfactuals*):
  - Removes a certain component
  - Test how it affects the results

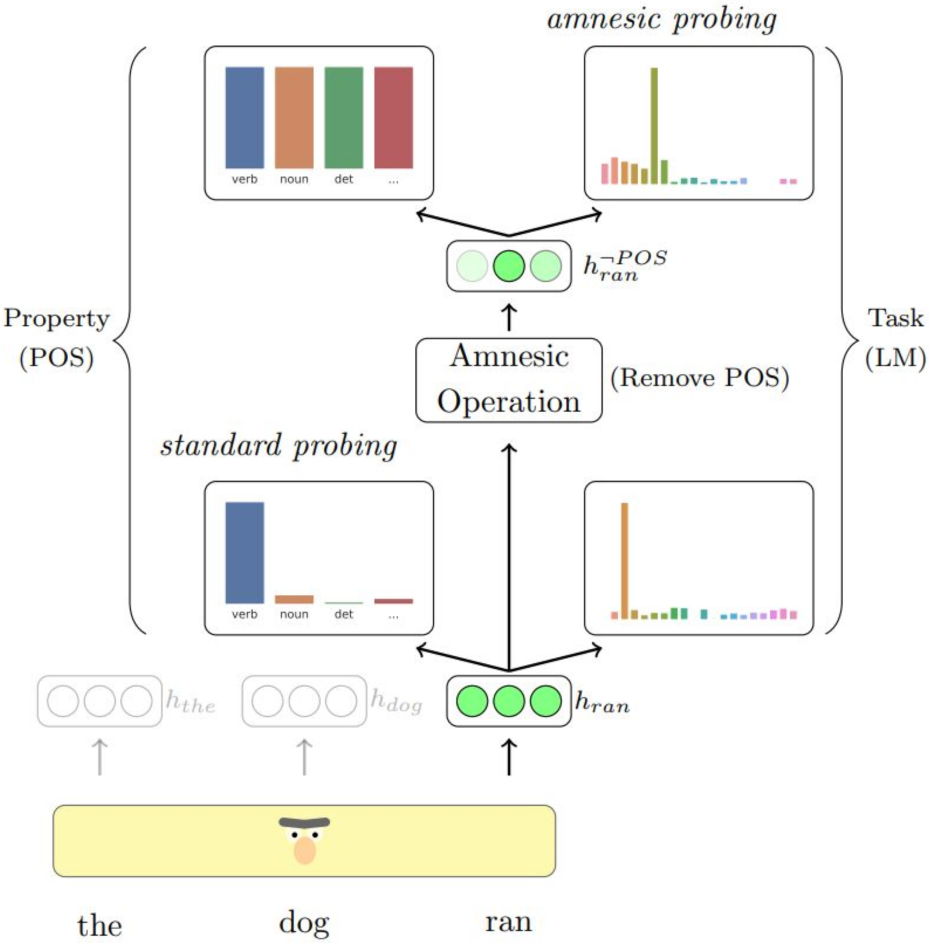
Amnesic Probing: The Intuition

- We remove a feature from the representation
- Does the model change its behavior?

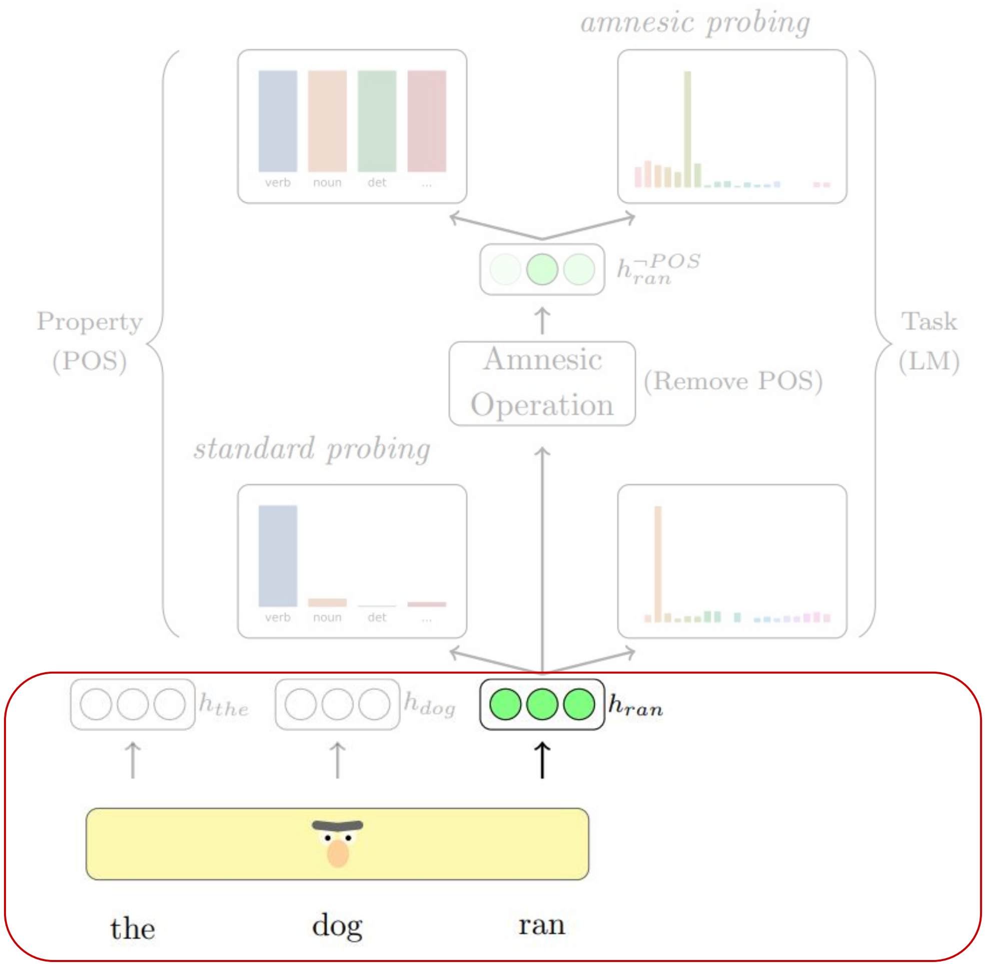
- Yes:
  - The model uses this information for its predictions
- No:
  - The model does **not** use this information for its predictions

Amnesic Probing: Overview



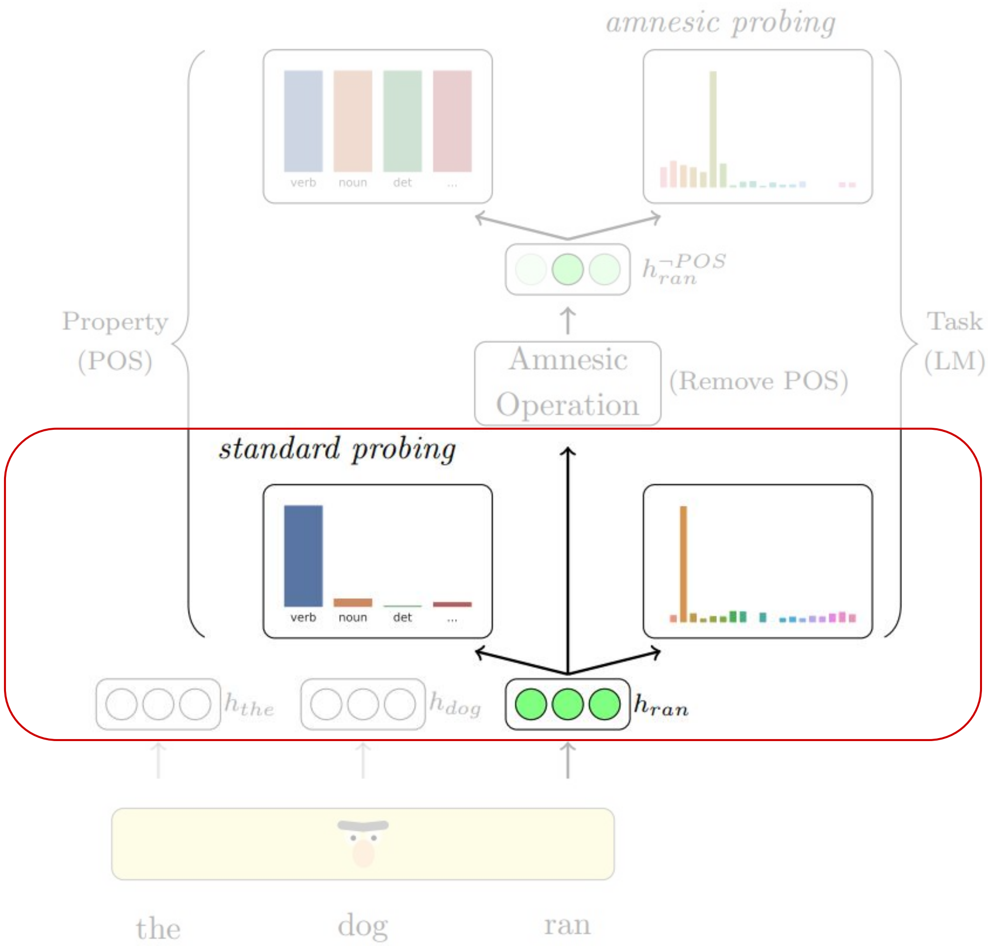

amnesic probing
verb
noun
det
...
$h_{ran}^{\neg POS}$
Property
(POS)
Amnesic
Operation
(Remove POS)
Task
(LM)
standard probing
verb
noun
det
...
h_{the}
h_{dog}
h_{ran}
the
dog
ran

1. Encode

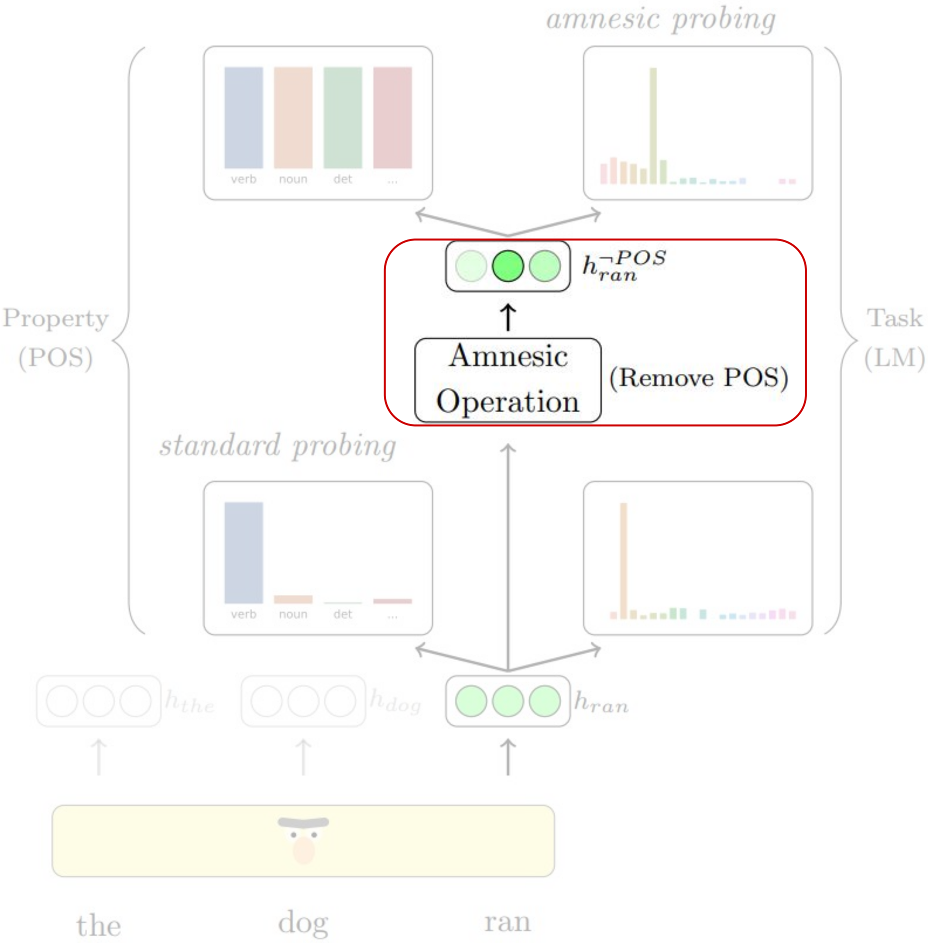


amnesic probing
verb
noun
det
...
$h_{ran}^{\neg POS}$
Property
(POS)
Task
(LM)
Amnesic
Operation
(Remove POS)
standard probing
verb
noun
det
...
h_{the}
h_{dog}
h_{ran}
the
dog
ran

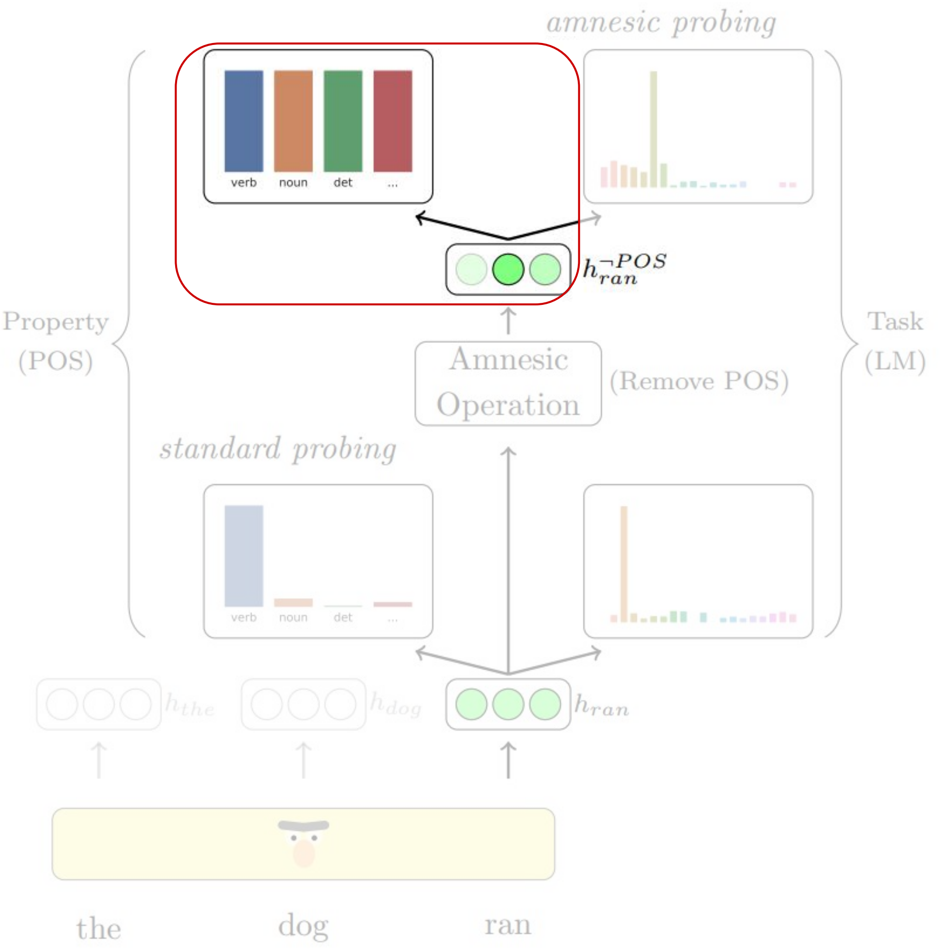
1. Encode
2. Probe



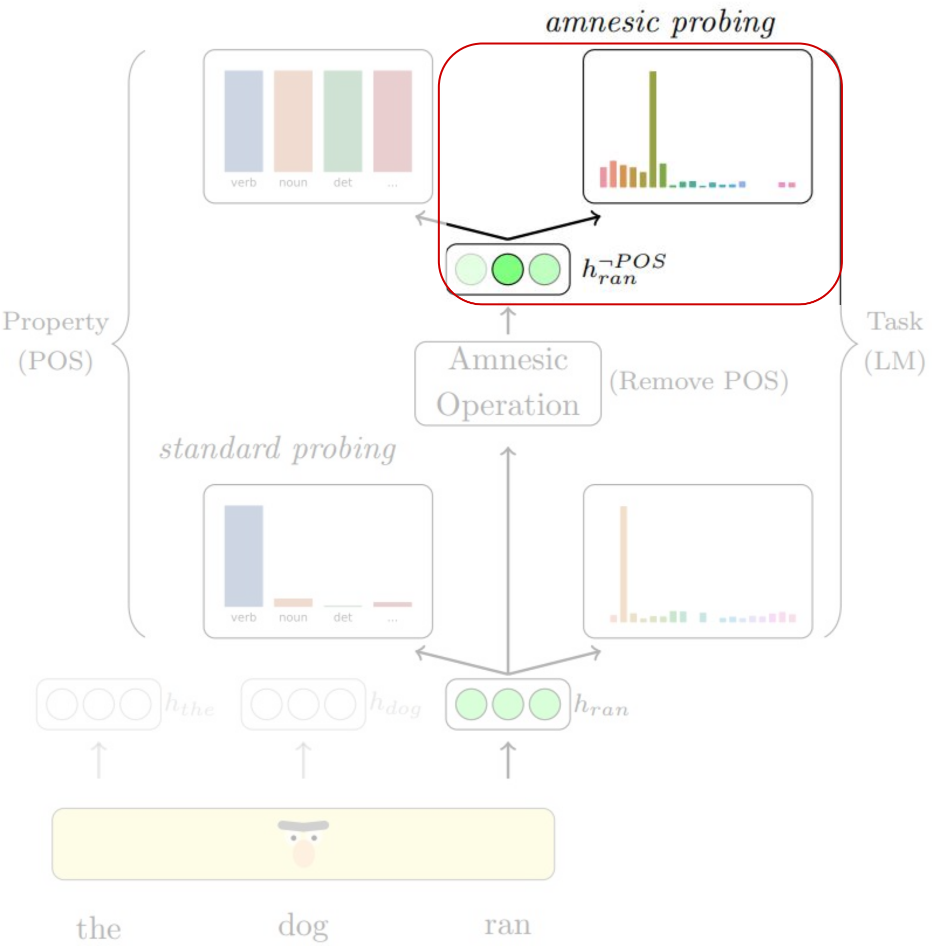
1. Encode
2. Probe
3. Amnesia



1. Encode
2. Probe
3. Amnesia
    3.1. Verify



1. Encode
2. Probe
3. Amnesia
   3.1. Verify
4. *Amnesic Probing*



The Amnesia

Amnesic Probing: The Amnesia

One option: Adversarial Training

Adversarial Removal of Demographic Attributes from Text Data

Yanai Elazar[†] and **Yoav Goldberg**[†*]

[†]Computer Science Department, Bar-Ilan University, Israel

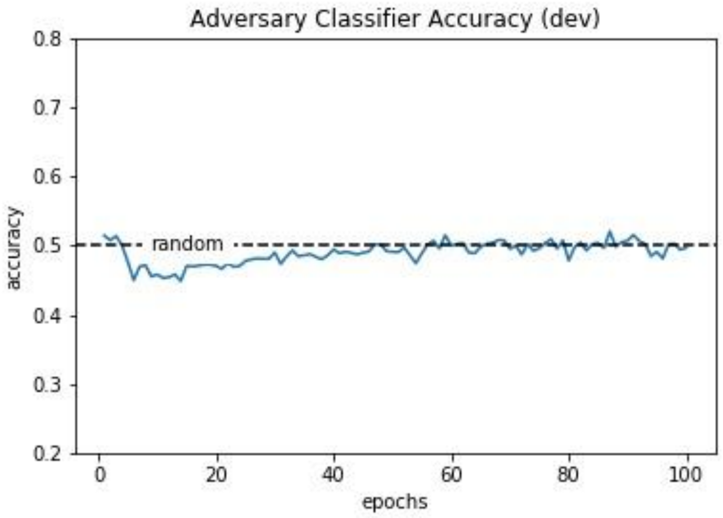
[*]Allen Institute for Artificial Intelligence

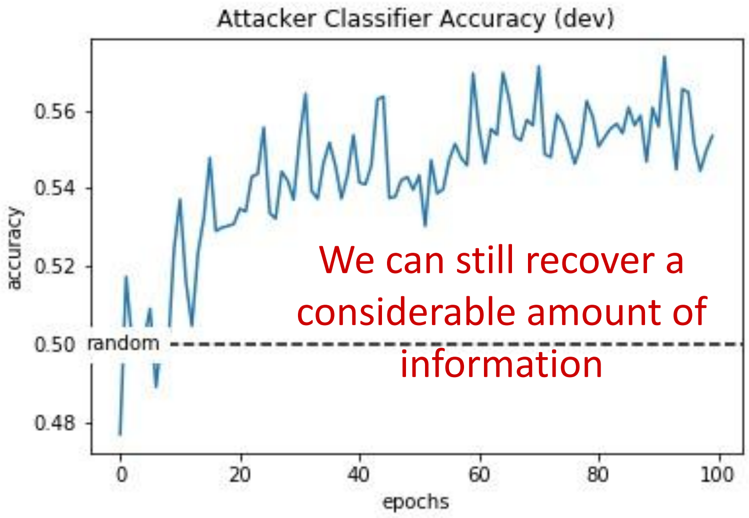
{yanaiela,yoav.goldberg}@gmail.com

EMNLP 2018

Amnesic Probing: The Amnesia

One option: Adversarial Training





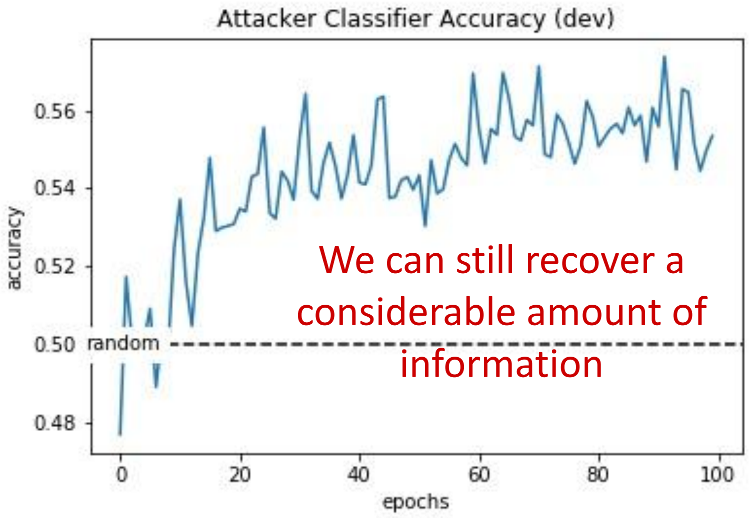
EMNLP 2018

Amnesic Probing: The Amnesia

One option: Adversarial Training

But also:

- Slow & unstable training
- Is it the same model afterwards?



EMNLP 2018

Amnesic Probing: The Amnesia

Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection

Shauli Ravfogel[1,2] **Yanai Elazar**[1,2] **Hila Gonen**[1] **Michael Twiton**[3] **Yoav Goldberg**[1,2]
[1]Computer Science Department, Bar Ilan University
[2]Allen Institute for Artificial Intelligence
[3]Independent researcher

ACL 2020

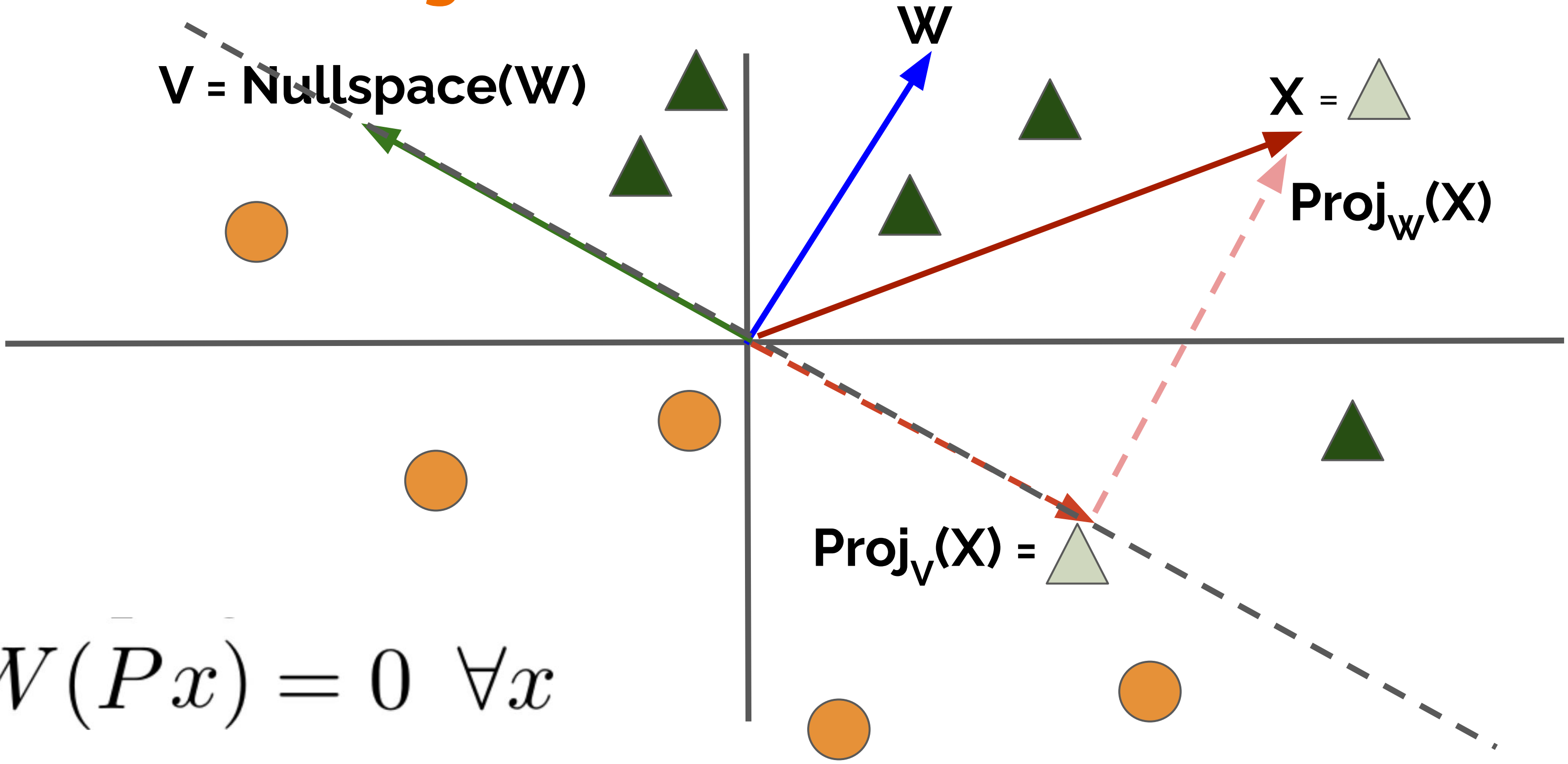
Amnesic Probing: The Amnesia

INLP: Iterative Nullspace Projection

- Find a projection matrix P, which projects into the nullspace

$$N(W) = \{x | Wx = 0\}$$

Amnesic Probing: The Amnesia



$$W(Px) = 0 \ \forall x$$

Amnesic Probing: The Amnesia

- Each projection only removes a single direction
- Therefore the “iterative” part:
- We repeat this process until convergence

Amnesic Probing: The Amnesia

- Debiasing applications (Ravfogel et al., 2020)

		BoW	FastText	BERT
Accuracy (profession)	Original	78.2	78.1	80.9
	+INLP	80.1	73.0	75.2
$GAP^{TPR,RMS}_{male}$	Original	0.203	0.184	0.184
	+INLP	0.124	0.089	0.095

Table 2: Fair classification on the Biographies corpus.

Check it out!

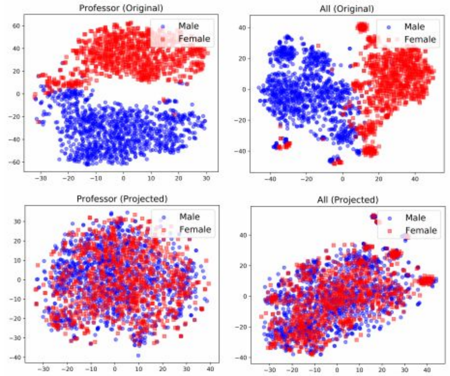


Figure 3: t-SNE projection of BERT representations for the profession "professor" (left) and for a random sample of all professions (right), before and after the projection.

Amnesic Probing: Properties

- The removed information is linear
- Guaranteed to remove linear all the linear information, given the usage of a good classifier
- No need to retrain a model (e.g. adversarial training)
- Post-hoc operation on a learned representation

Amnesic Probing: Using INLP

- We use INLP in this work, but this is a component that can be replaced with a future (non-linear) alternative

Feder et al., 2020

Amnesic Probing: Setup

- We take a trained model
- Choose properties/features of interest
- Measure the difference (Behavioral!)
  - Accuracy (of predicting the “right” label)
  - KL Divergence (DKL): softer metric, but on the entire distribution

Amnesic Probing: Controls

- Did the amnesic operation really had an effect?
- Did the amnesic operation remove too much information?

Amnesic Probing: Controls

- Did the amnesic operation really had an effect?
- Did the amnesic operation remove too much information?

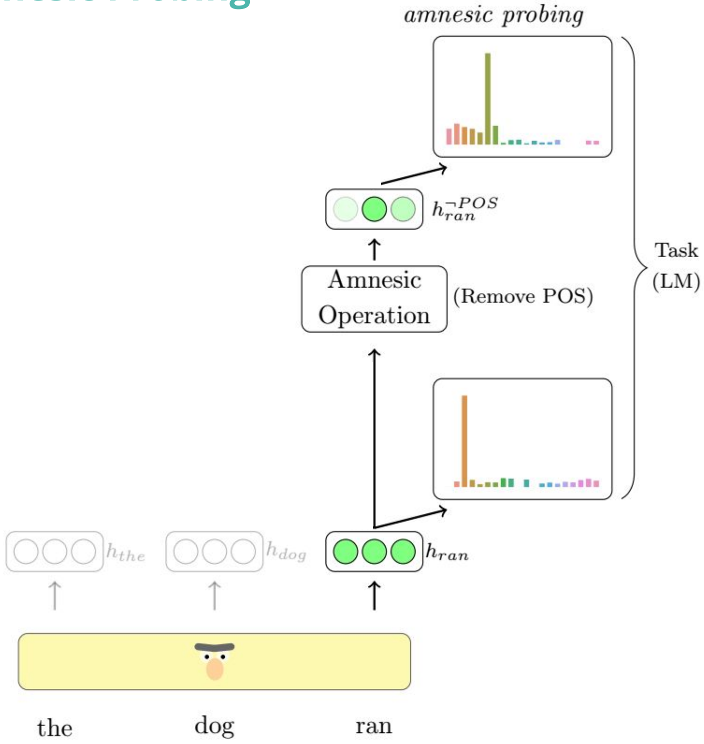
- Control over Information
    - Removing random directions

Amnesic Probing: Controls

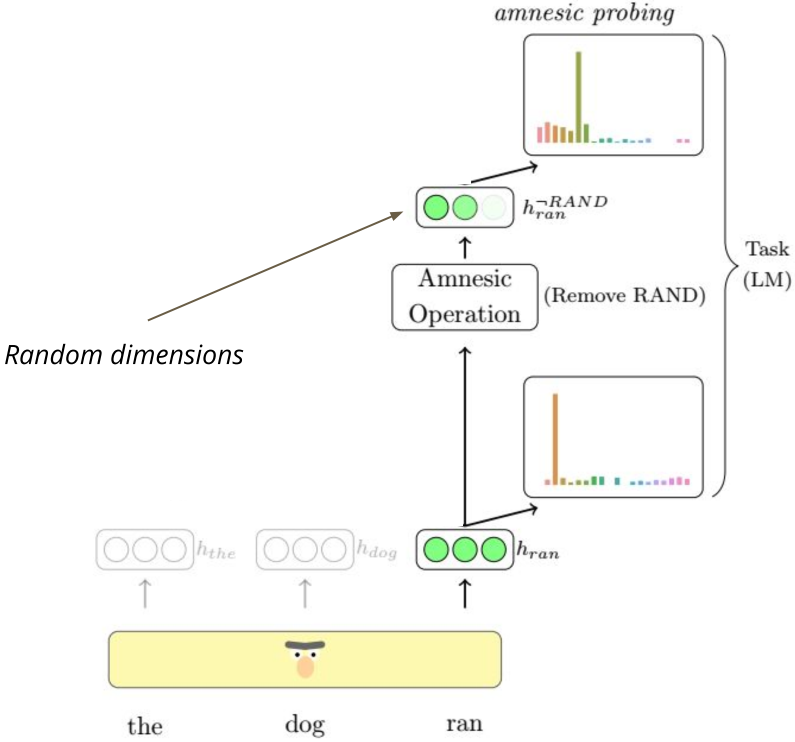
- Did the amnesic operation really had an effect?
- Did the amnesic operation remove too much information?

- Control over Information
  - Removing random directions
- Control over Selectivity
  - Add back the “real” features, and retrain

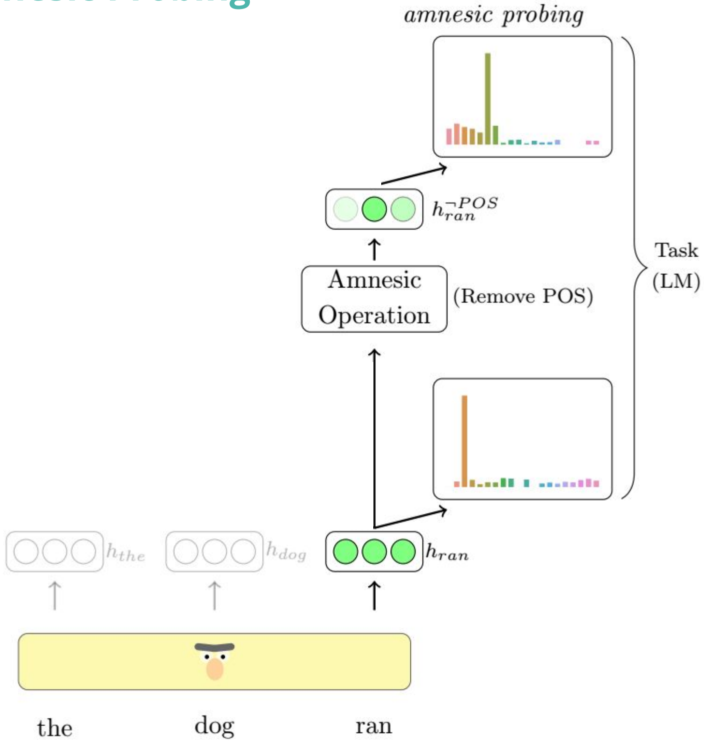
Amnesic Probing



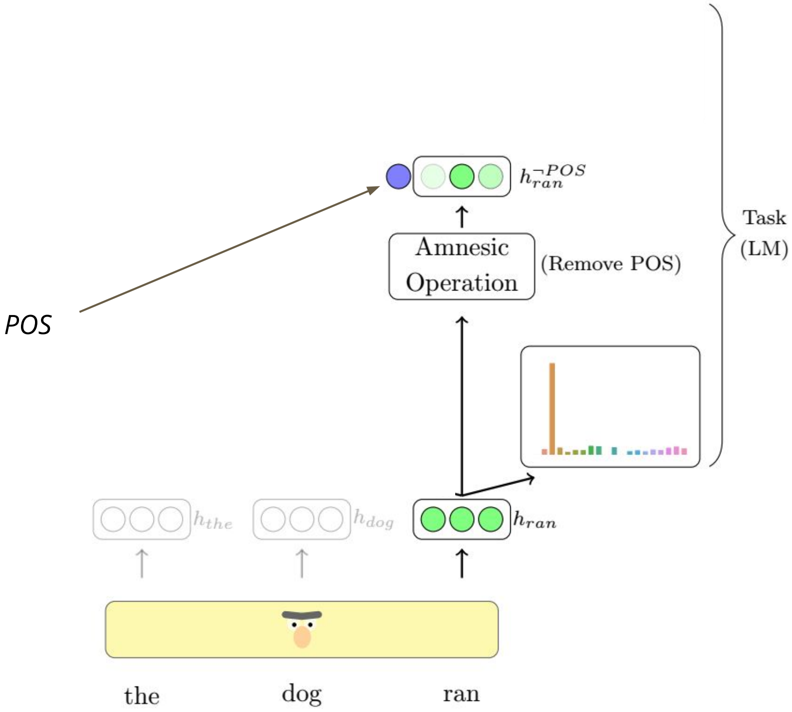
Control over Information



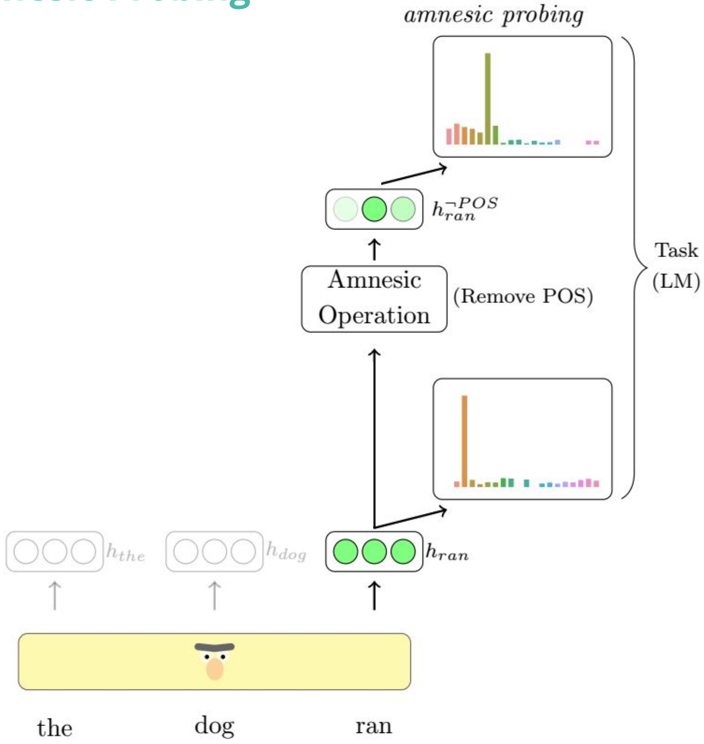
Amnesic Probing



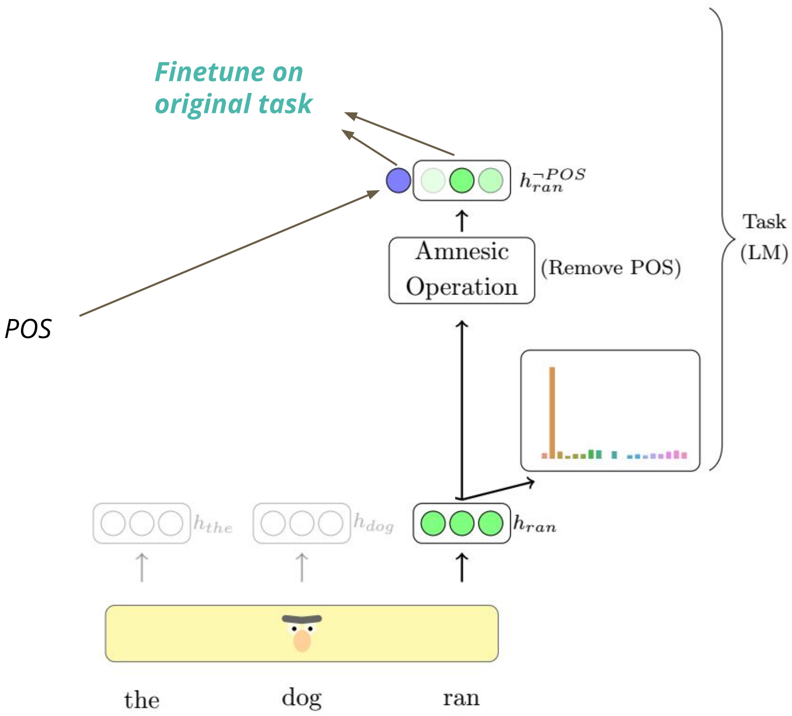
Control over Selectivity



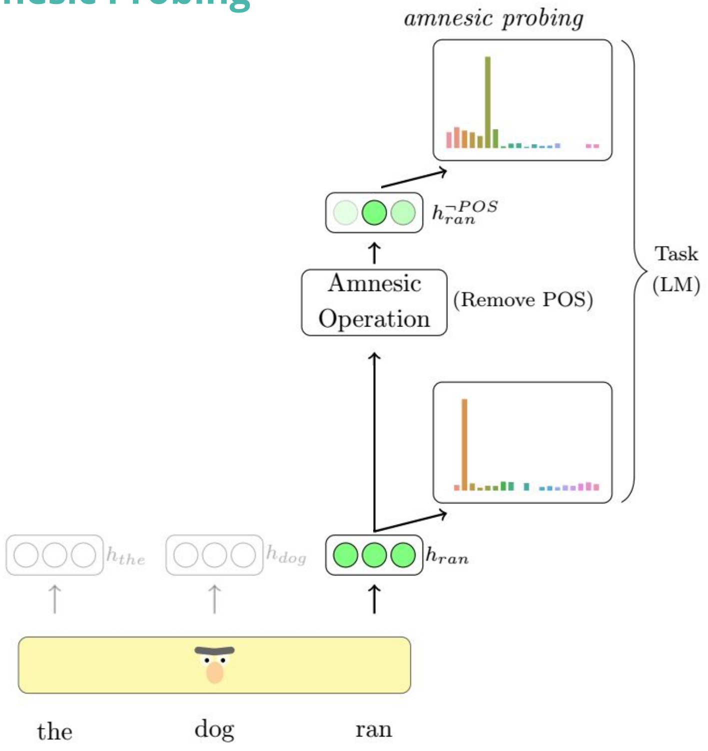
Amnesic Probing



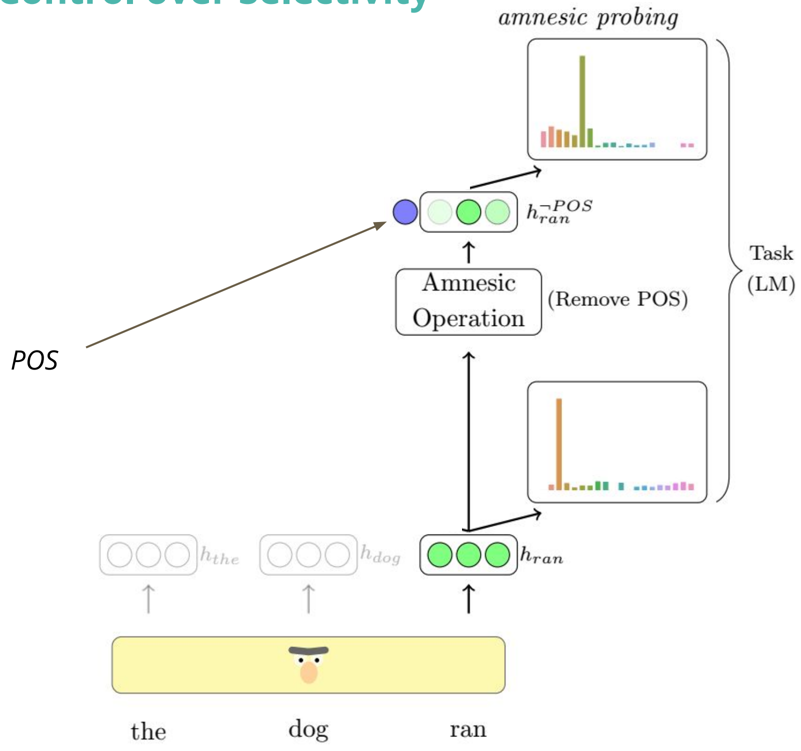
Control over Selectivity



Amnesic Probing



Control over Selectivity



Amnesic Operation - Randomization

First Align, then Predict:
Understanding the Cross-Lingual Ability of Multilingual BERT

Benjamin Muller[1,2] **Yanai Elazar**[3,4] **Benoît Sagot**[1] **Djamé Seddah**[1]
[1]Inria, Paris, France [2]Sorbonne Université, Paris, France
[3]Computer Science Department, Bar Ilan University
[4]Allen Institute for Artificial Intelligence
{benjamin.muller, benoit.sagot,djame.seddah}@inria.fr
yanaiela@gmail.com

EACL 2021

Amnesic Operation - Randomization

- Study the role of a model component (e.g. MLP) in pretraining
- Randomize the layer and then fine-tune
- Measure difference in performance w/ and w/o the randomization.

Muller et al., EACL 2021

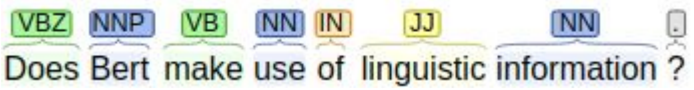
Case Study: Pre-trained BERT

Amnesic Probing: Results

- The Model: BERT-base

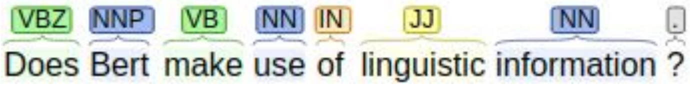
Amnesic Probing: Results

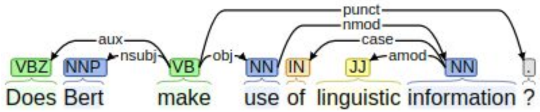
- The Model: BERT-base
- Properties:
    - POS



Amnesic Probing: Results

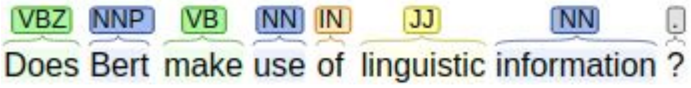
- The Model: BERT-base
- Properties:
  - POS
  - Dependency edges

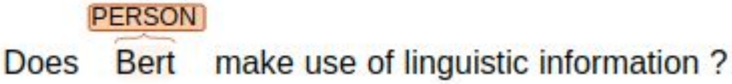


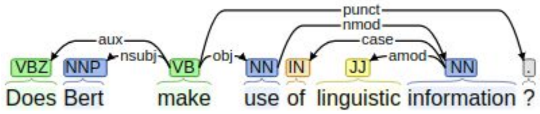


Amnesic Probing: Results

- The Model: BERT-base
- Properties:
    - POS
    - Dependency edges
    - NER

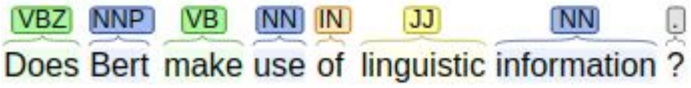


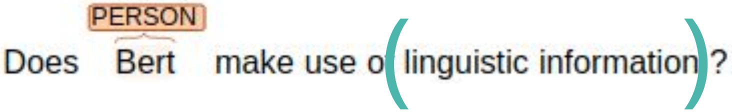


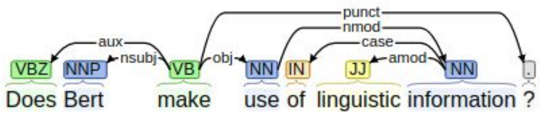


Amnesic Probing: Results

- The Model: BERT-base
- Properties:
    - POS
    - Dependency edges
    - NER
    - Constituency boundaries







Amnesic Probing: Results

		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

Amnesic Probing: Results

Linguistic Properties

		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

Amnesic Probing: Results

Standard Probing

		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

Amnesic Probing: Results

LM Accuracy Results

		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

Amnesic Probing: Results

Amnesic Comparison

		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

Amnesic Probing: Results

Comparison to Control:
Information

		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

Amnesic Probing: Results

Comparison to Control:
Selectivity

		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

Amnesic Probing: Results

		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

DKL Results

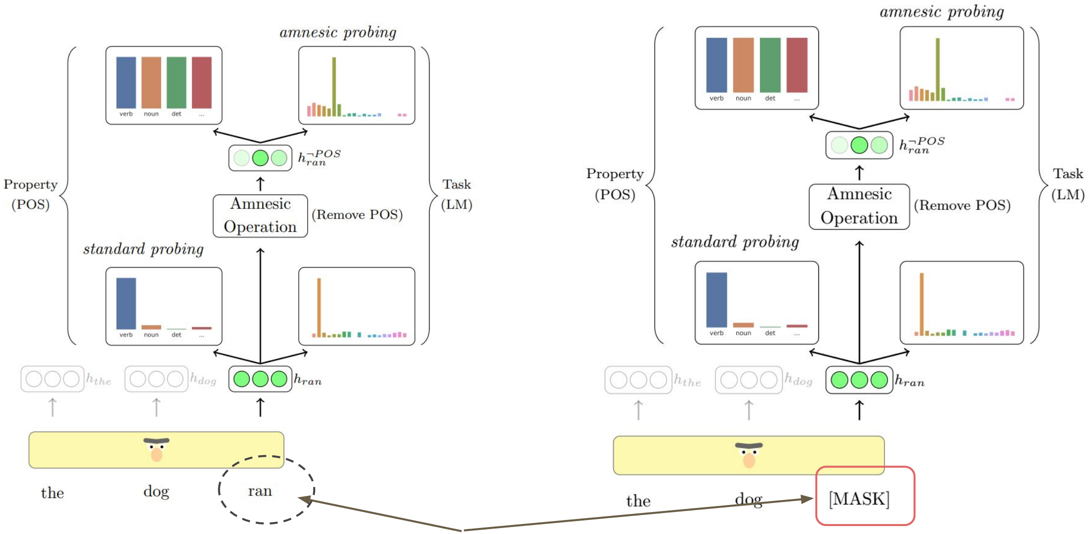
Amnesic Probing: Results

		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

DKL Results

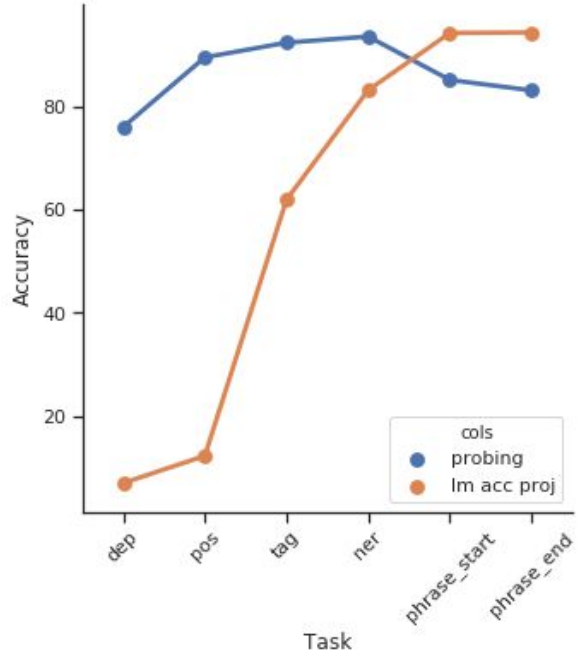
Amnesic Probing: Results

- We perform the same experiments on another setup, where the words are masked
  - (Similar results, but we'll come back to it)



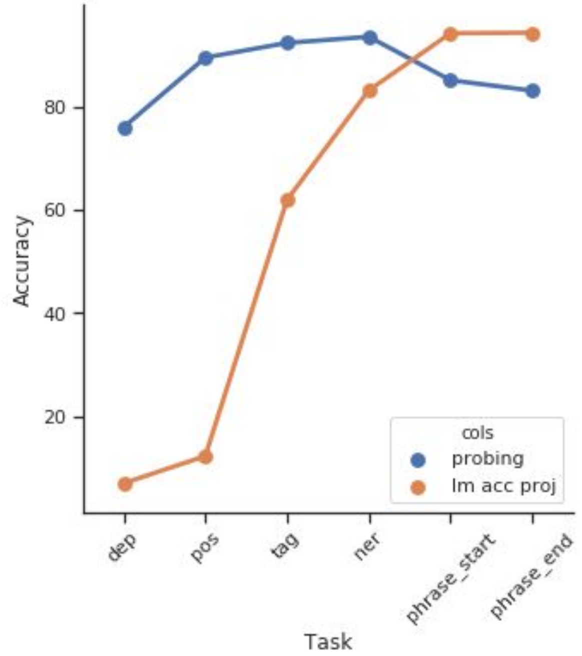
Amnesic Probing vs. Standard Probing

- We plot the performance of regular probing vs. *amnesic probing*
- We observe no correlation between the two metrics



Amnesic Probing vs. Standard Probing

- We plot the performance of regular probing vs. *amnesic probing*
- We observe no correlation between the two metrics
- Behavioural conclusions cannot be made from standard probing results



Ravichander et al., 2020, Tamkin et al., 2020

Amnesic Probing: Diving In

Amnesic Probing Fine Grained

- How individuals POS are affected by the removal of POS information?
- Open vs. Closed vocabulary

Small changes

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

Large changes

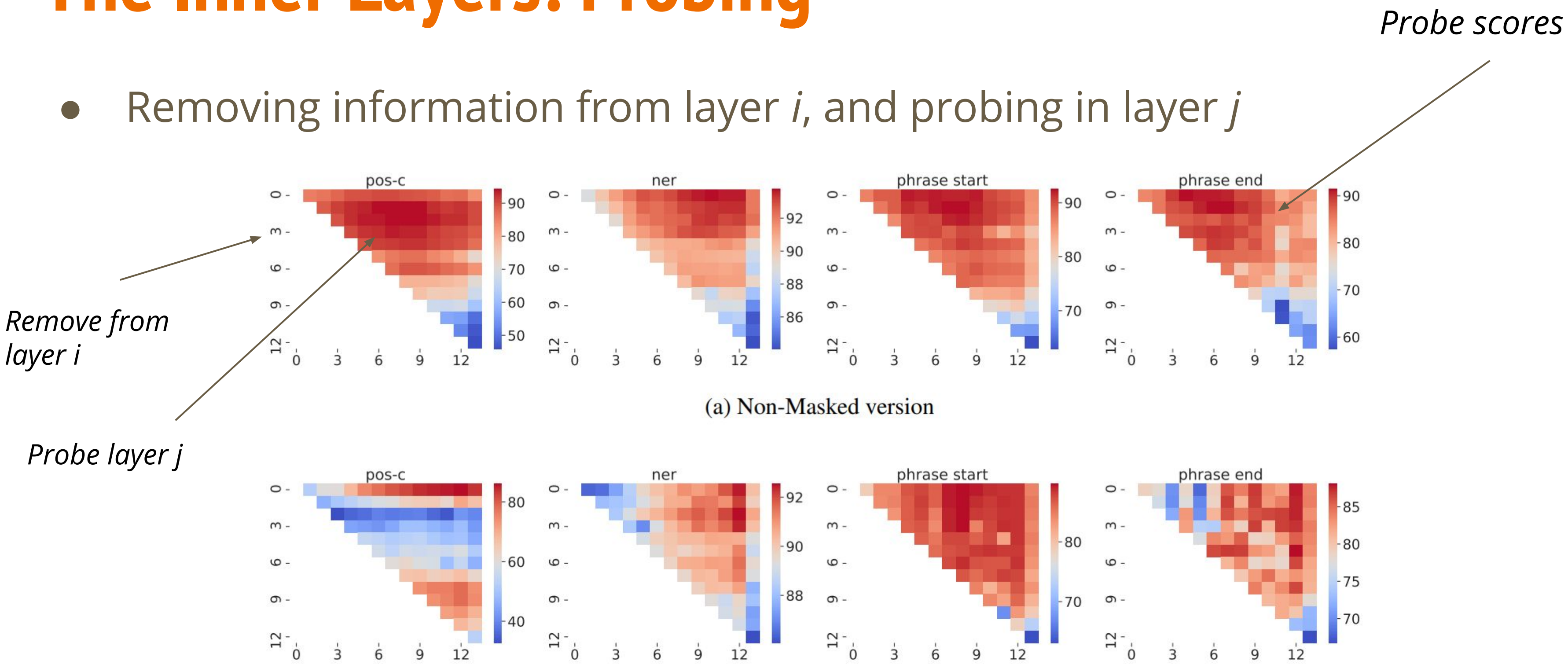
Amnesic Probing: Inside The Model

The Inner Layers

- Until now, querying the last layer
  - INLP removes linear information, last layer is only multiplied by a matrix
- We perform the same analysis on the Inner layers
- Standard Probe (after the amnesic operation)
- Behavioral Probe

The Inner Layers: Probing

- Removing information from layer *i*, and probing in layer *j*



(a) Non-Masked version

(b) Masked version

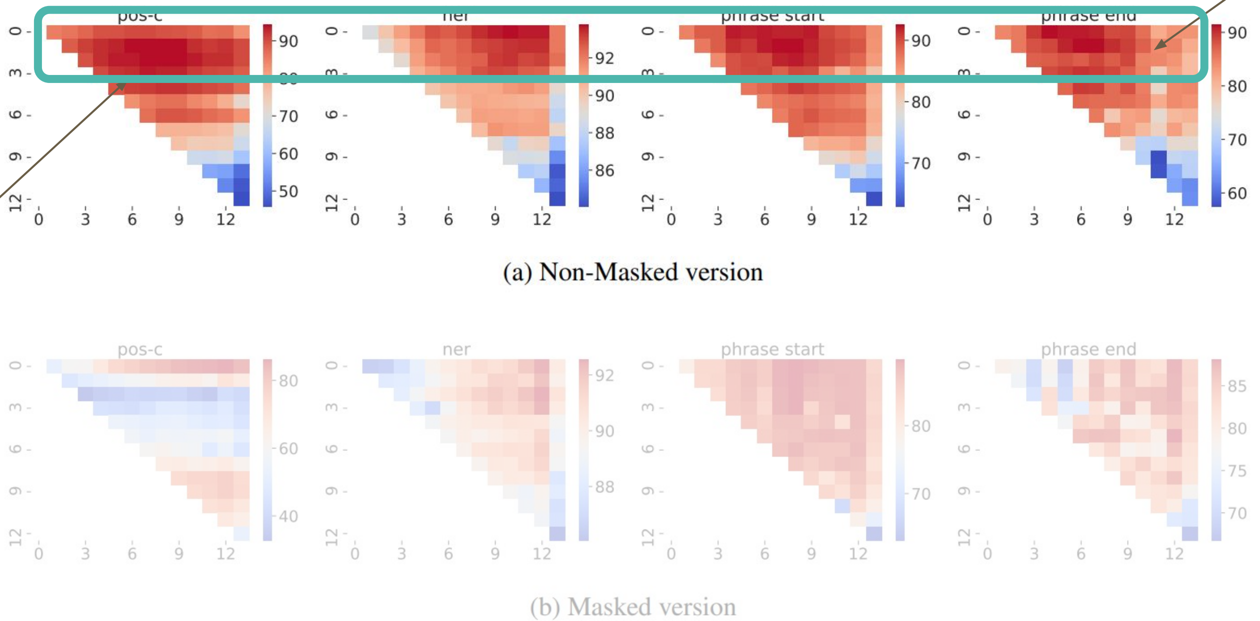
The Inner Layers: Probing

- Removing information from layer *i*, and probing in layer *j*

Probe scores

Remove from layer i

Probe layer j

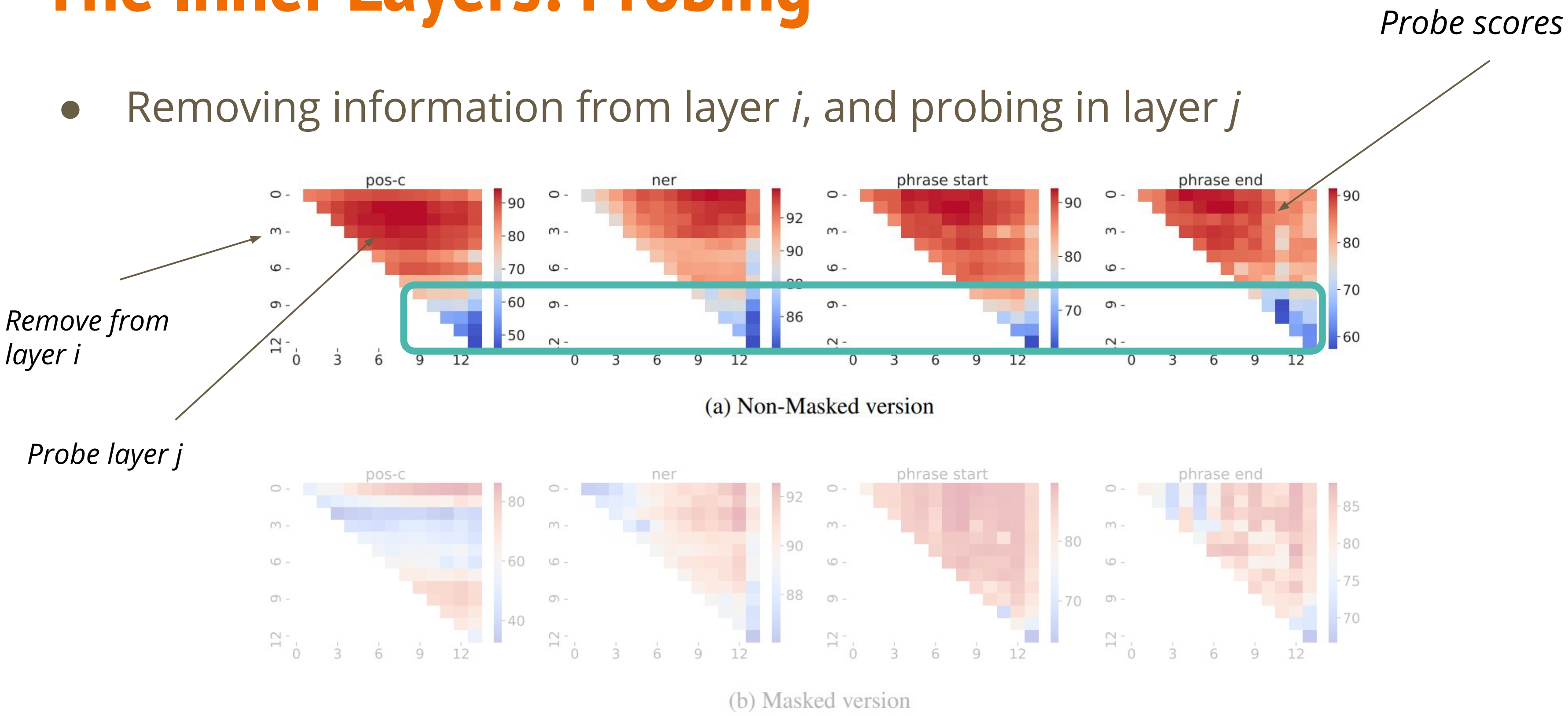


(a) Non-Masked version

(b) Masked version

The Inner Layers: Probing

- Removing information from layer *i*, and probing in layer *j*



(a) Non-Masked version

(b) Masked version

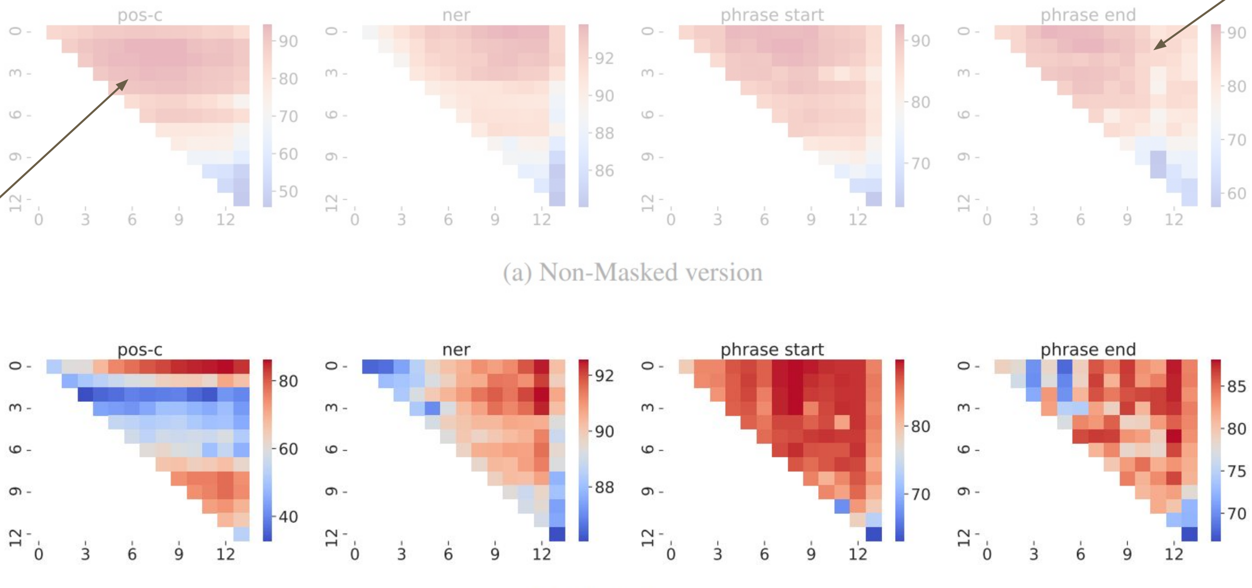
The Inner Layers: Probing

- Removing information from layer *i*, and probing in layer *j*

Probe scores

Remove from layer i

Probe layer j

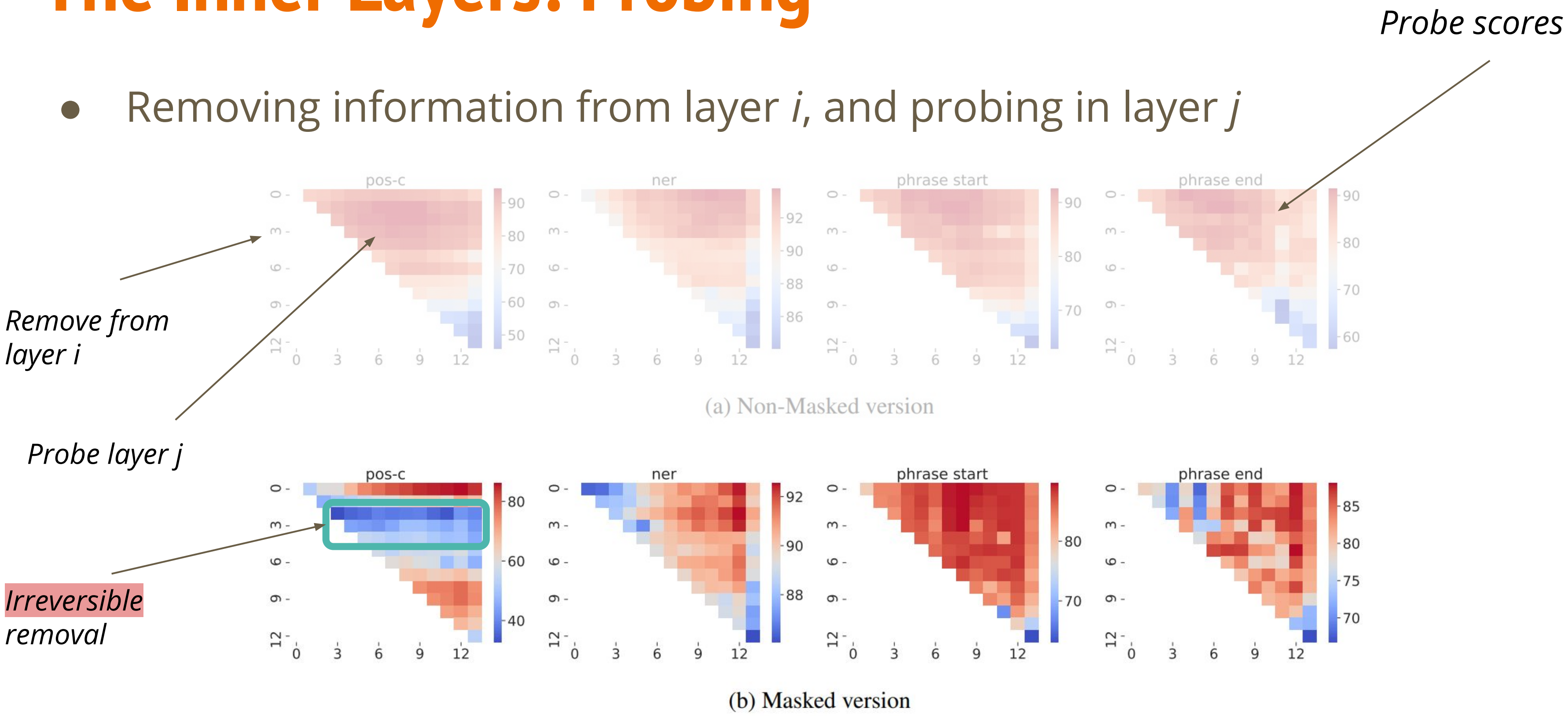


(a) Non-Masked version

(b) Masked version

The Inner Layers: Probing

- Removing information from layer *i*, and probing in layer *j*



The Inner Layers: Probing

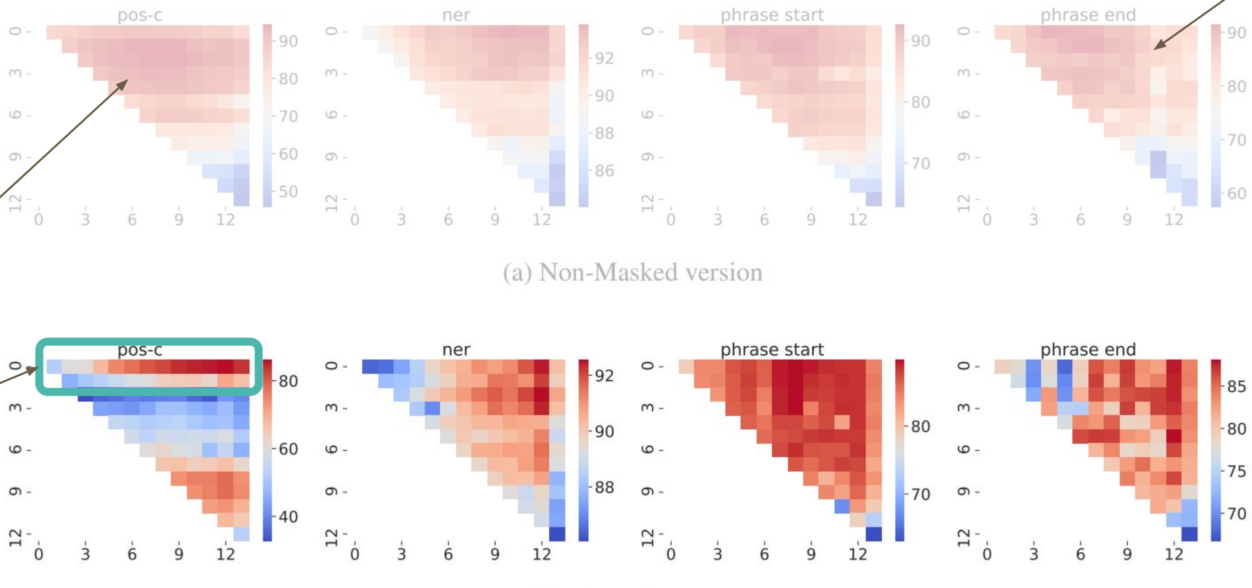
- Removing information from layer *i*, and probing in layer *j*

Probe scores

Remove from layer i

Probe layer j

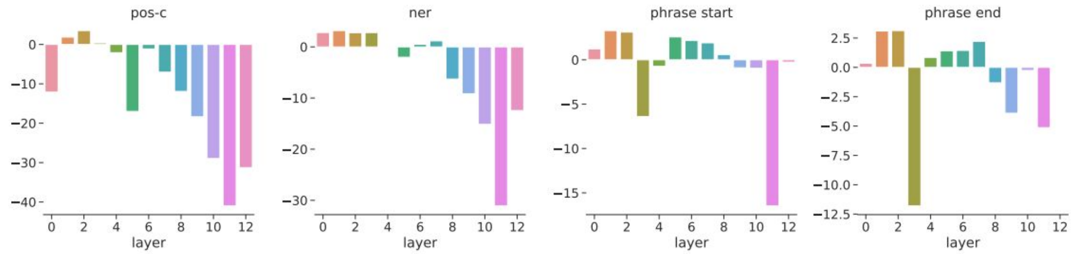
Reversible removal



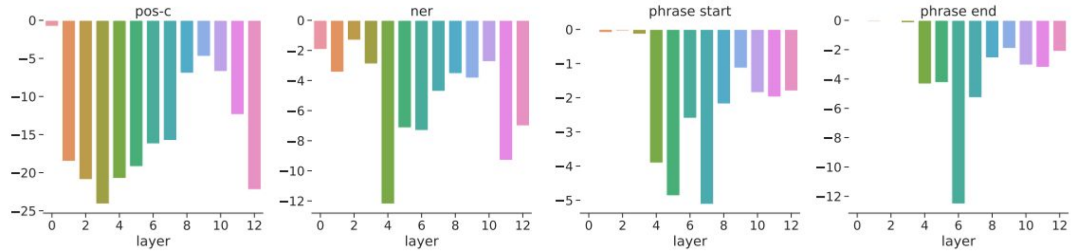
(b) Masked version

The Inner Layers: Amnesic Probing

- Removing information from layer *i*, and inspecting the model's predictions



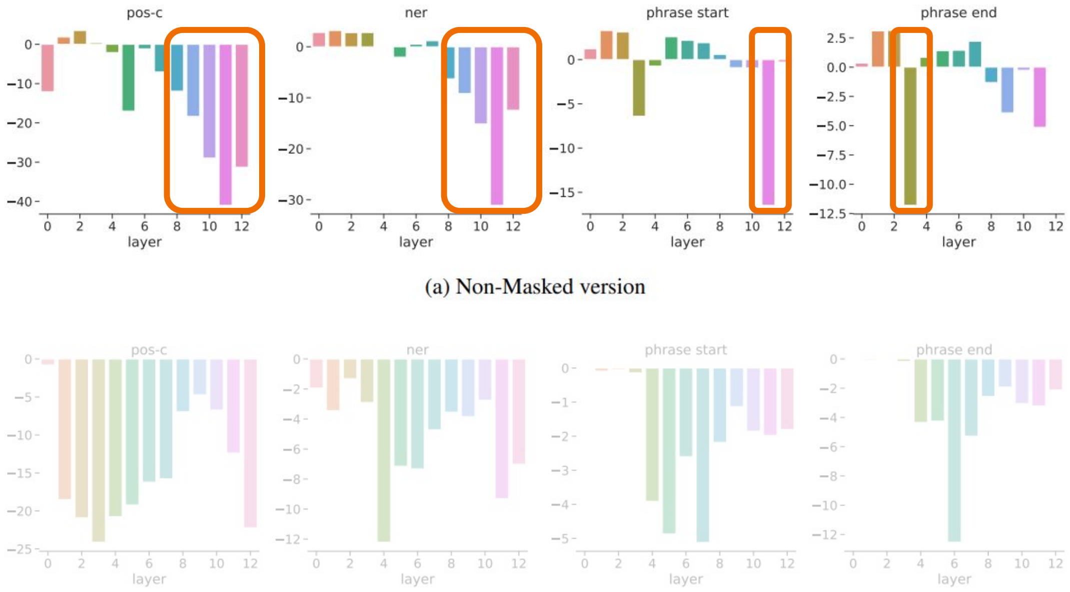
(a) Non-Masked version



(b) Masked version

The Inner Layers: Amnesic Probing

- Removing information from layer *i*, and inspecting the model's predictions

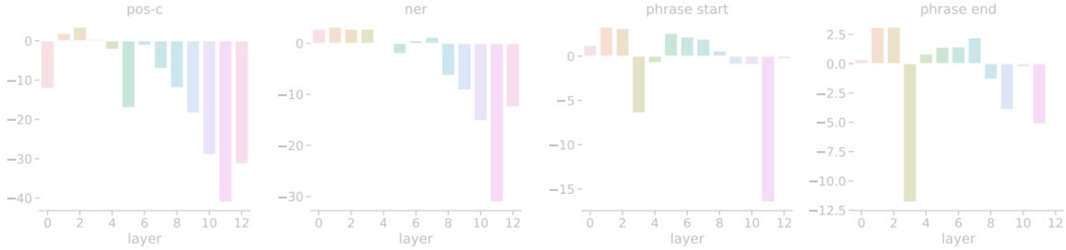


(a) Non-Masked version

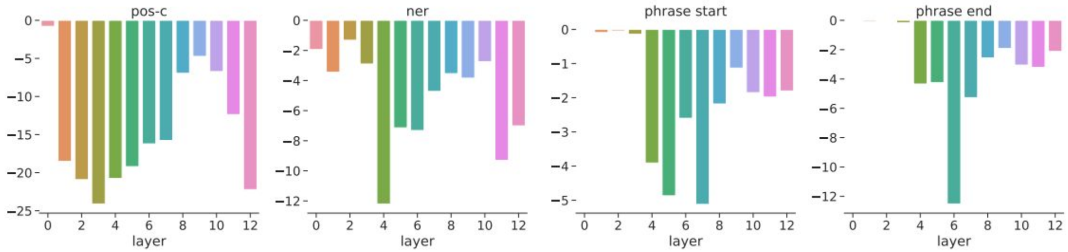
(b) Masked version

The Inner Layers: Amnesic Probing

- Removing information from layer *i*, and inspecting the model's predictions



(a) Non-Masked version

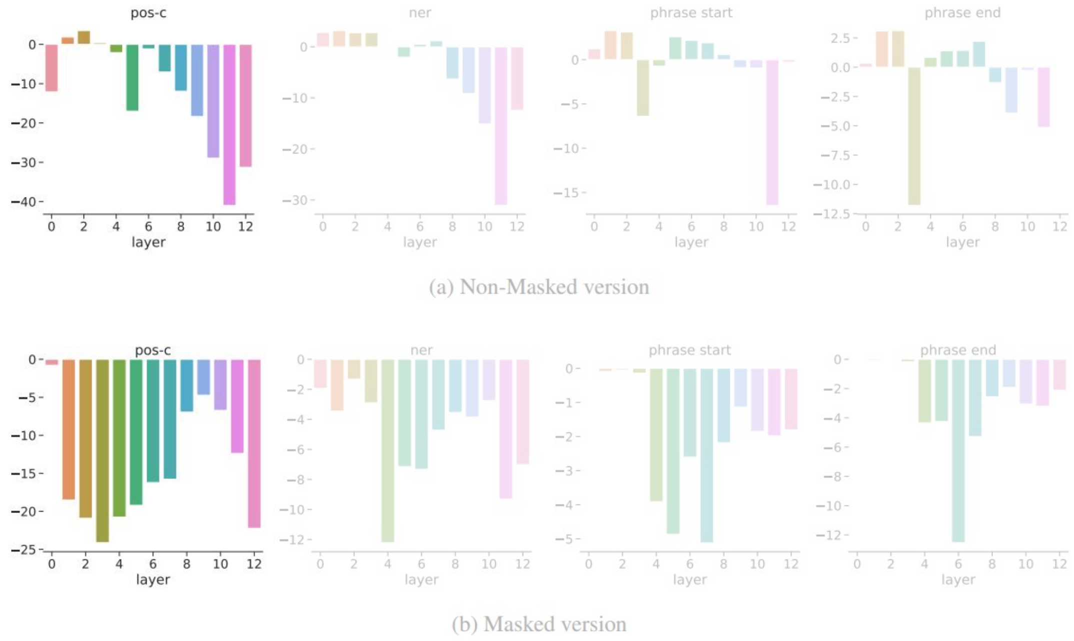


(b) Masked version

The Inner Layers: Amnesic Probing

- Removing information from layer *i*, and inspecting the model's predictions

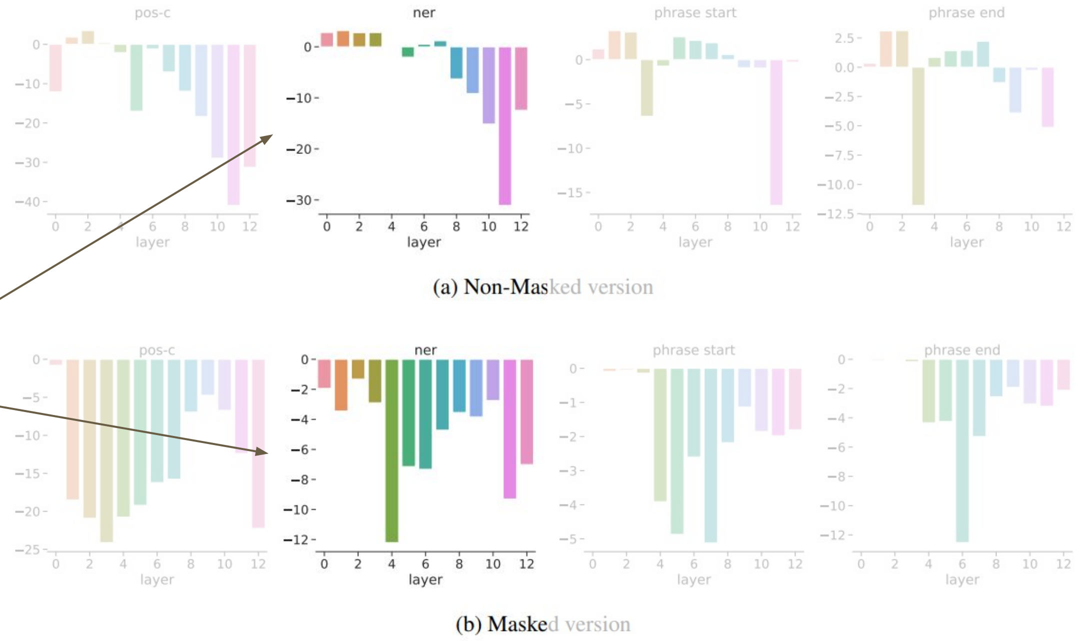
Masked vs Non-Masked



The Inner Layers: Amnesic Probing

- Removing information from layer *i*, and inspecting the model's predictions

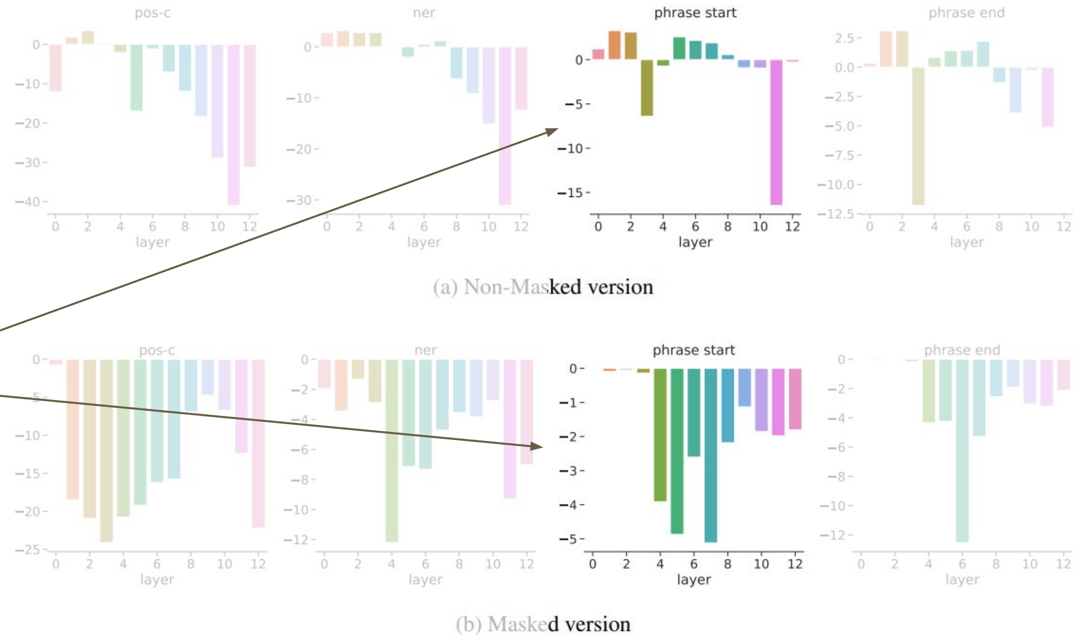
Masked vs Non-Masked



The Inner Layers: Amnesic Probing

- Removing information from layer *i*, and inspecting the model's predictions

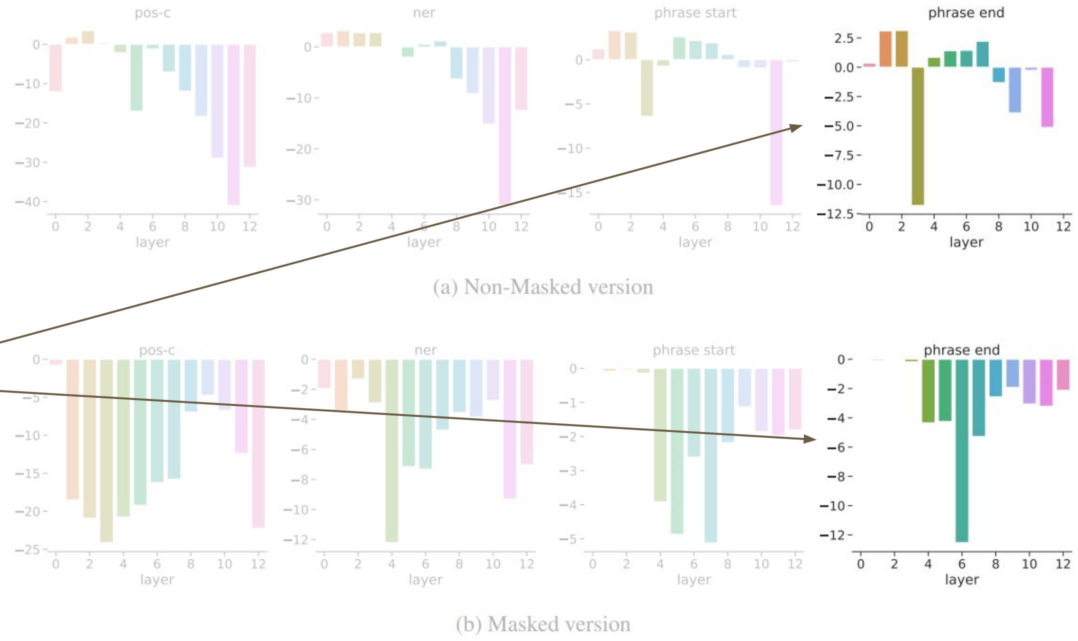
Masked vs Non-Masked



The Inner Layers: Amnesic Probing

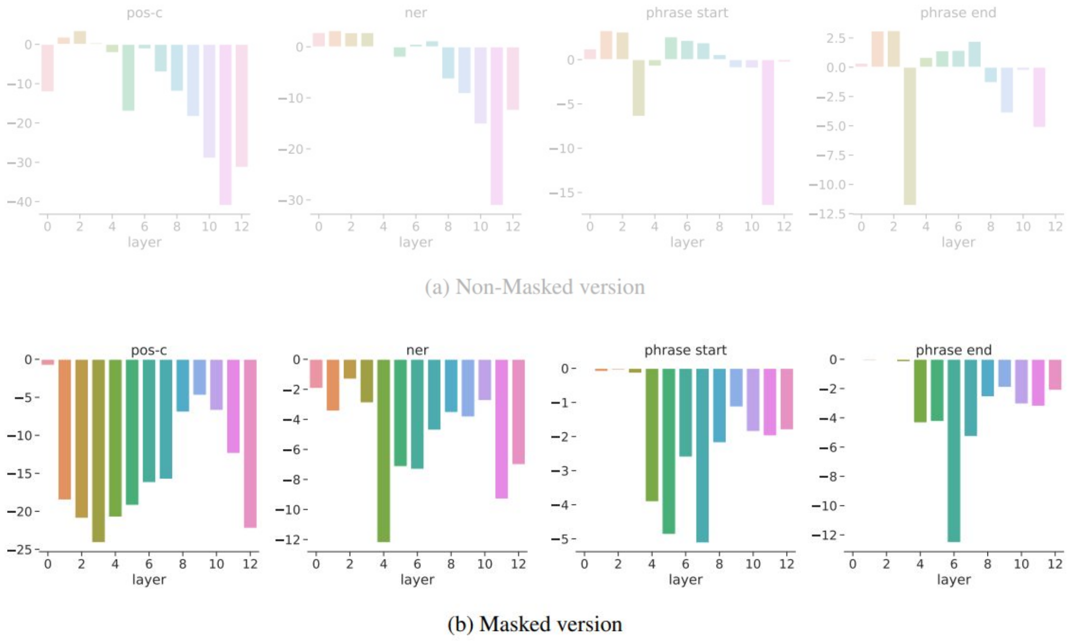
- Removing information from layer *i*, and inspecting the model's predictions

Masked vs Non-Masked



The Inner Layers: Amnesic Probing

- Removing information from layer *i*, and inspecting the model's predictions



Strong impact in the first few layers!!

Summary - Amnesic Probing

- New method for answering questions about what properties are being used by models
- Analysis of different linguistic properties and how they are being used by the popular BERT MLM
- Structural Analysis != Behavioral Analysis

Going Forward

- What **does** it mean that some information is extractable?
- ... or, why is it there from the first place?
- Algorithms that remove also non-linear information

Thanks!

arxiv paper

Yanai Elazar

@yanaiela

yanaiela.github.io