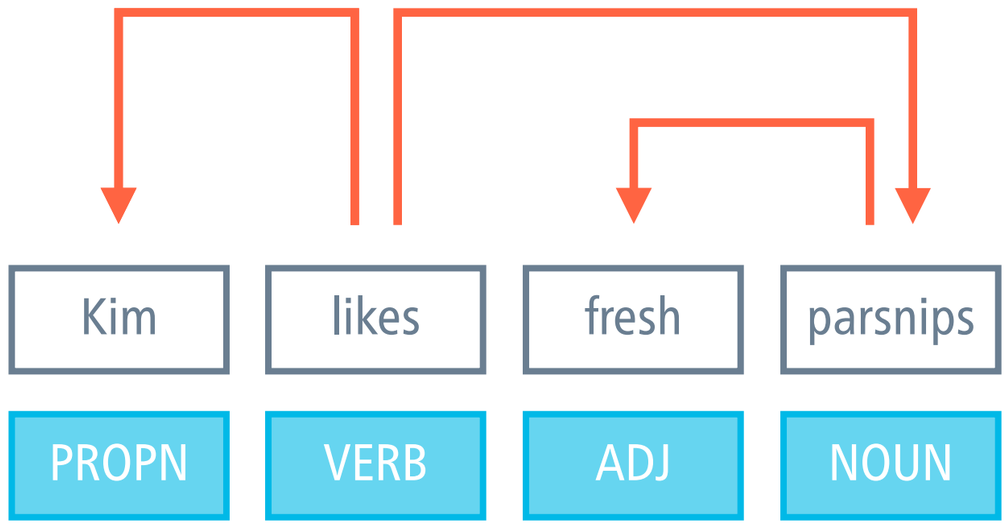
Probing and Explaining Neural Language Models

Jenny Kunz and Marco Kuhlmann
Department of Computer and Information Science

LINKÖPING UNIVERSITY

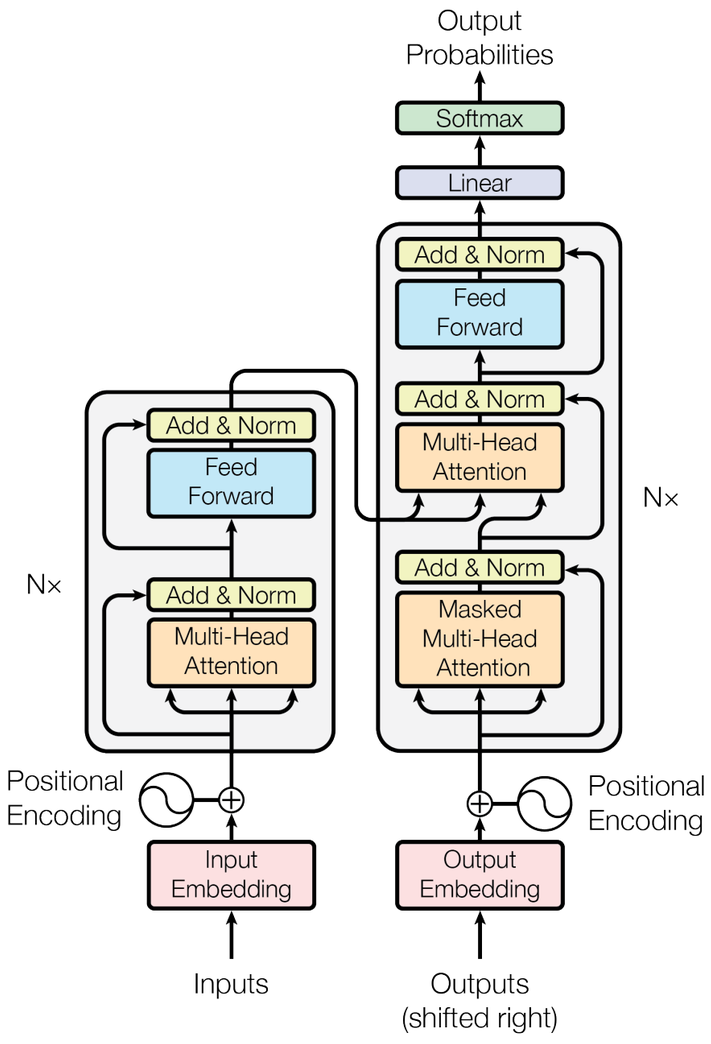
Linguistic structure



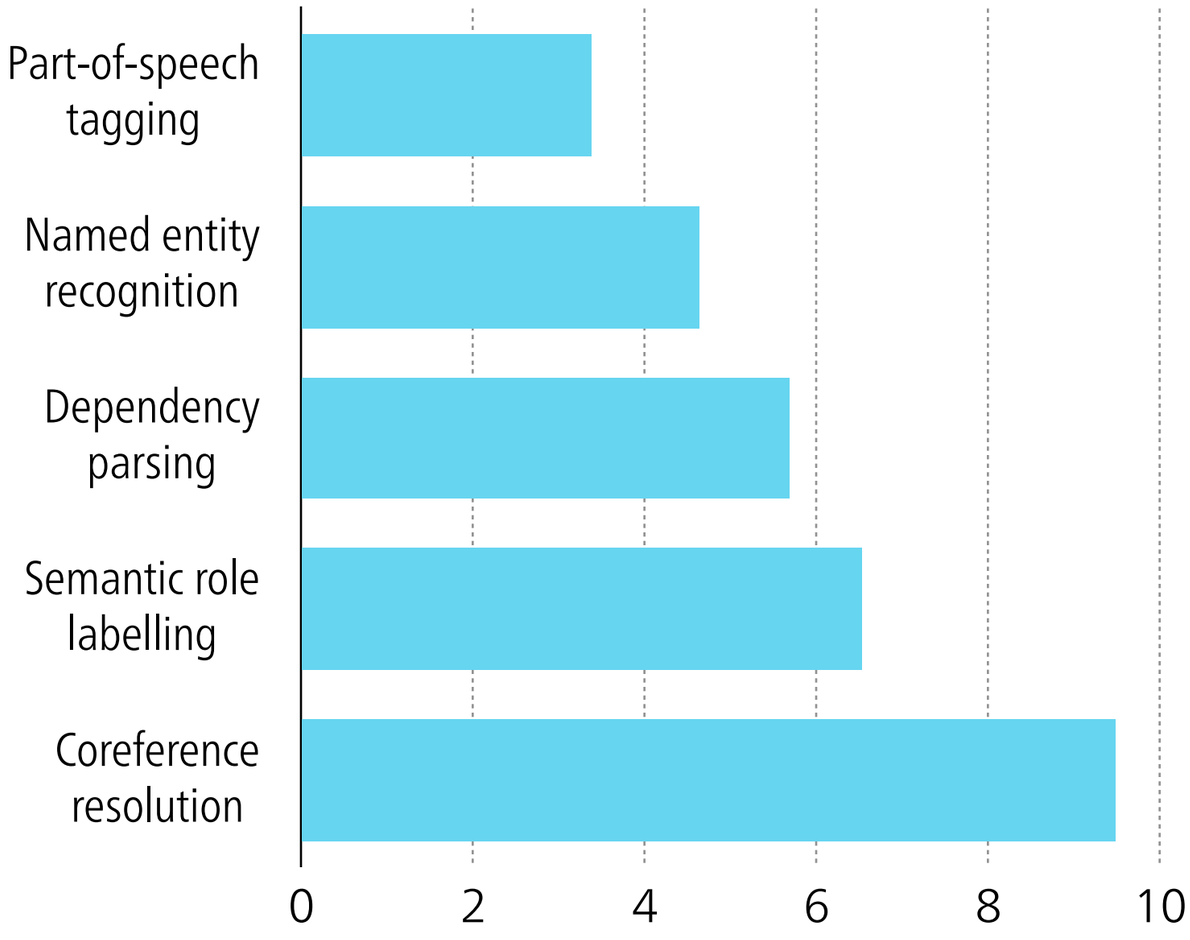
part-of-speech tags

dependency trees

‘Natural language processing from scratch’



Vaswani et al. (2017)



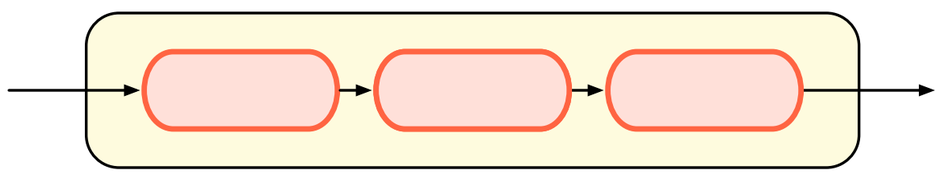
Tenney et al. (2019)

Pre-training & fine-tuning

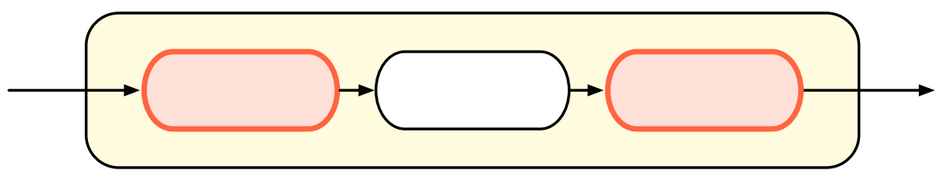
- **Transfer learning** aims to re-use knowledge gained while solving one problem when solving the next problem.

  reduce the need for training data

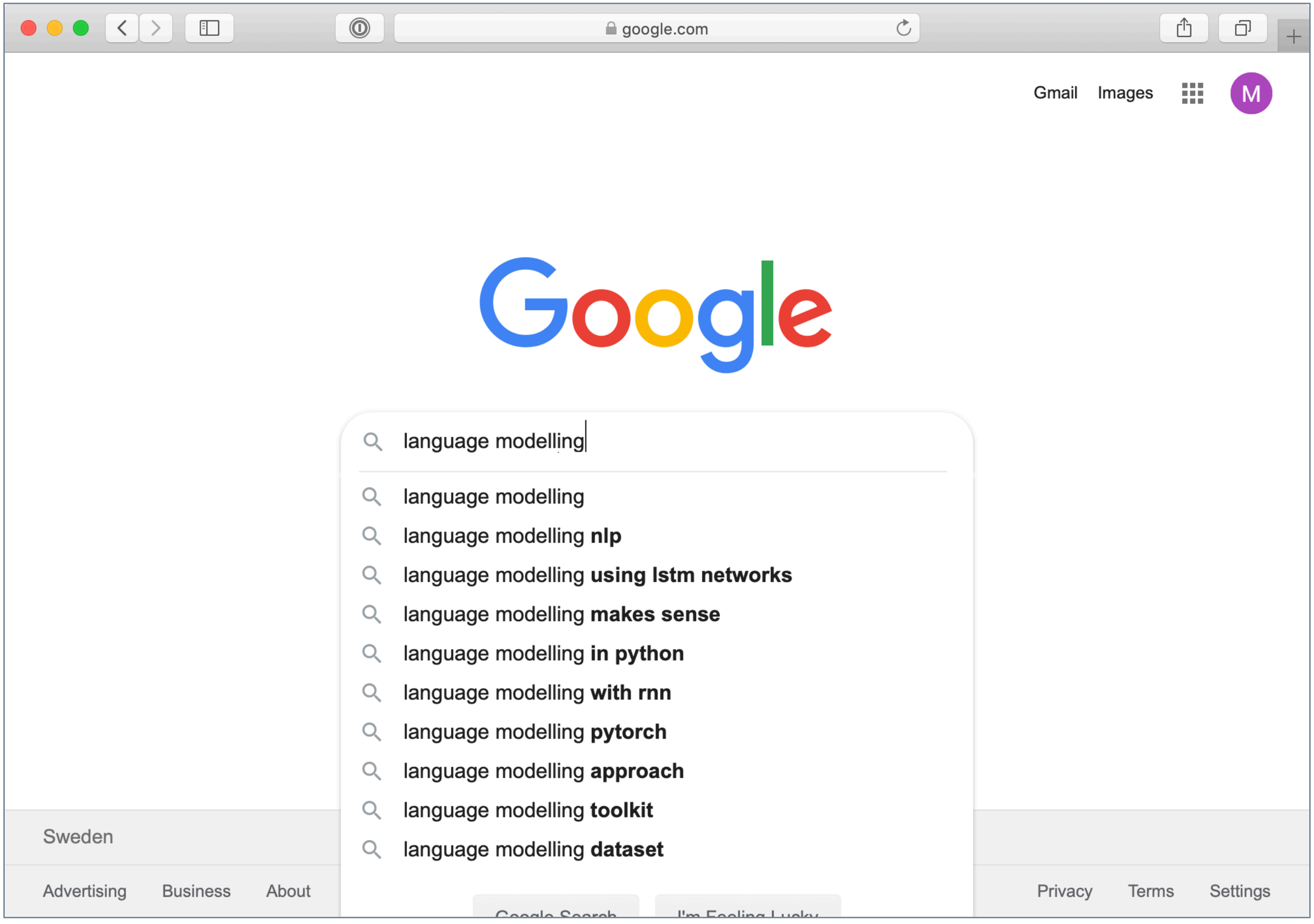
- In contemporary NLP, transfer learning is usually implemented through **pre-training and fine-tuning**.



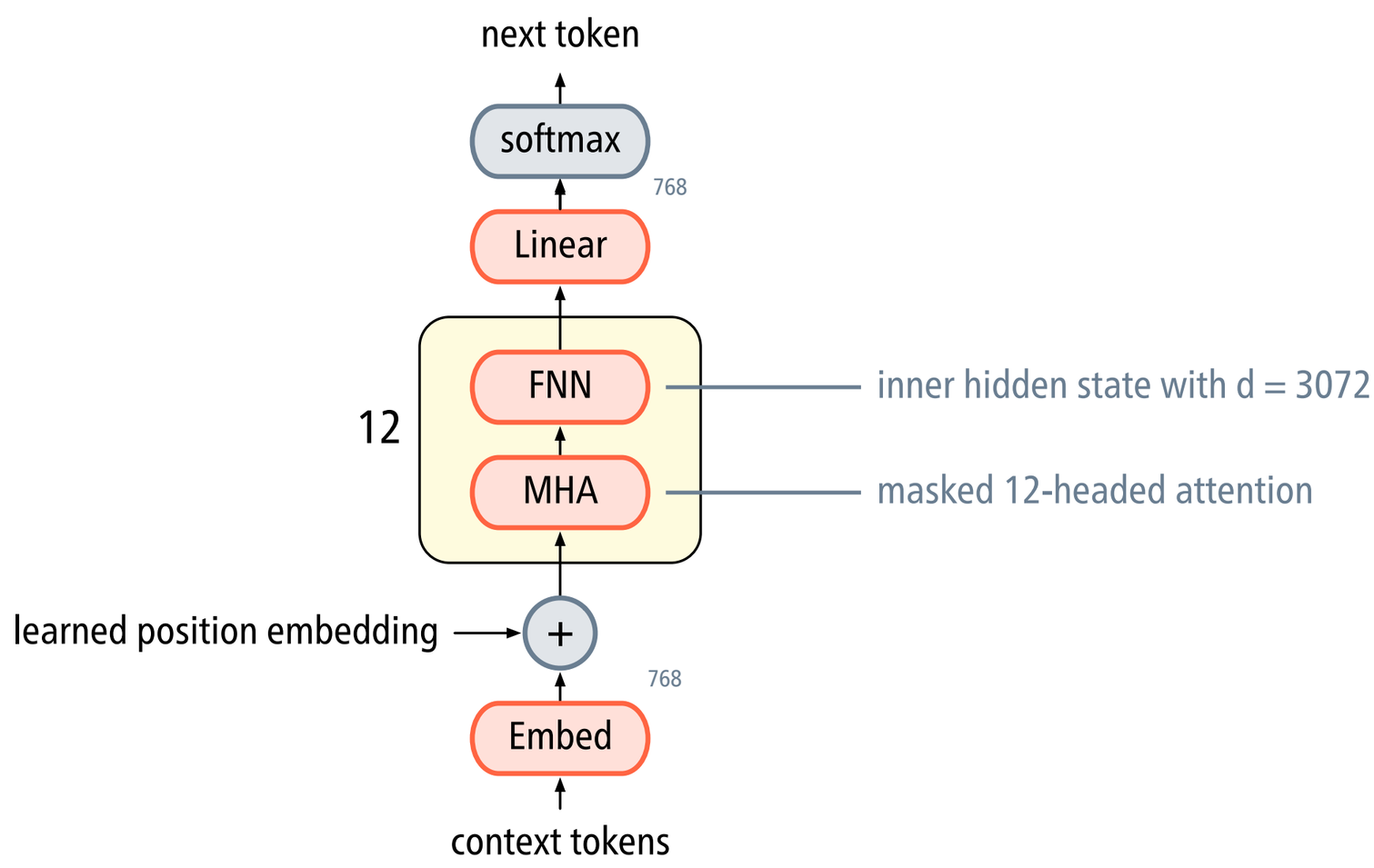
Model trained on task A



Model to be trained on task B


google.com
Gmail
Images
M
Google
language modelling
language modelling
language modelling nlp
language modelling using lstm networks
language modelling makes sense
language modelling in python
language modelling with rnn
language modelling pytorch
language modelling approach
language modelling toolkit
language modelling dataset
Sweden
Advertising
Business
About
Privacy
Terms
Settings

GPT model architecture



Radford et al. (2018)

GPT-3 example output

Model prompt (human-written)	In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.
Model completion (machine-written)	The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow. Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez. Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns. […]

OpenAI Blog (2019)

Larger and larger models, more and more data

	GPT-1	GPT-2	GPT-3
Number of dimensions	768	1,600	12,288
Number of layers	12	48	96
Trainable parameters	**0.117 B**	**1.542 B**	**175 B**
Training data size (tokens)	800 M	(40 GB text)	499 B

Radford et al. (2018), Radford et al. (2019), Brown et al. (2020)

Large language models are zero-shot learners

Sentiment classification

Tweet: I hate it when my battery dies.
Sentiment: Negative

Tweet: My day has been great!
Sentiment: Positive

Tweet: This music video was incredible!
Sentiment: **Positive**

Machine translation

Translate English to French:

sea otter => loutre de mer

peppermint => menthe poivrée

plush giraffe => girafe en peluche

cheese => **fromage**

black text provided by the user, red text generated by GPT-3

Stochastic Parrots 🦜 – Are they worth it?

> How big is too big? What are the possible risks associated with [large pre-trained language models] and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.
>
> Bender et al. (2021)

Overview of this presentation

Question: What do neural language models learn about language, and how can we even know?

- **Method 1: Probing.** Using diagnostic classifiers to draw conclusions about linguistic structure encoded in model representations. (But can we, really?)

- **Method 2: Explanations.** Letting models generate free-form explanations that tell us something about how a model arrived at a prediction. (But do they, really?)

Publications

- Jenny Kunz and Marco Kuhlmann. *Classifier Probes May Just Learn from Linear Context Features*. COLING 2020.

- Jenny Kunz and Marco Kuhlmann. *Test Harder Than You Train: Probing with Extrapolation Splits*. BlackboxNLP 2021.

- Jenny Kunz and Marco Kuhlmann. *Where Does Linguistic Information Emerge in Neural Language Models? Measuring Gains and Contributions Across Layers*. COLING 2022.

- Jenny Kunz, Martin Jirénius, Oskar Holmström, and Marco Kuhlmann. *Human Ratings Do Not Reflect Downstream Utility: A Study of Free-Text Explanations for Model Predictions*. Accepted to BlackboxNLP 2022.

Insights from probing

Probing for linguistic structure

- A **probe** is a classifier trained on a task designed with the intention of revealing what information is present in a model.

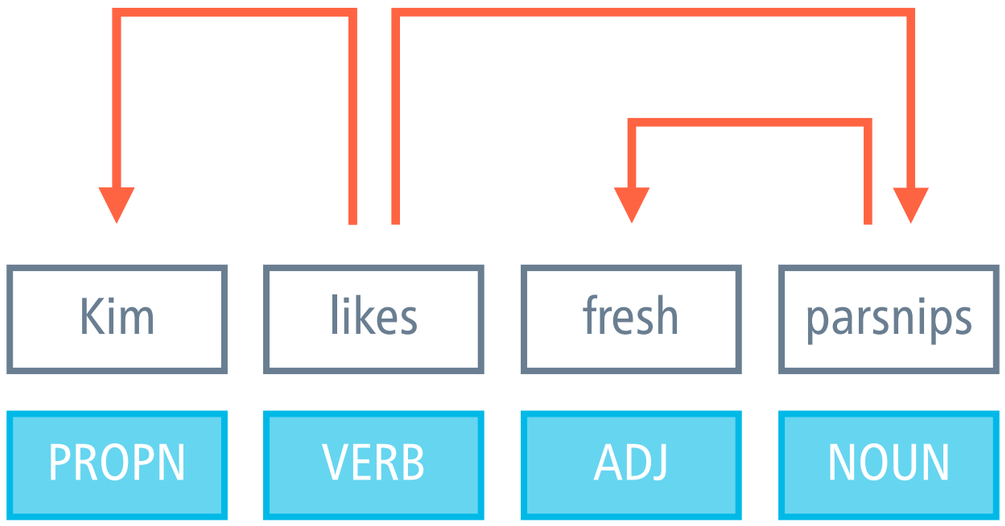
  often a simple linear layer

- The diagnostic task uses data in the form of pairs (x, y) where x is a representation extracted from the model and y is a label.

  Example: BERT representation at some layer k

- Basic assumption: The probe can only learn the task well if the necessary information is already encoded in the representation.

Simple probing tasks



Given a word, what is its part-of-speech (POS) tag?

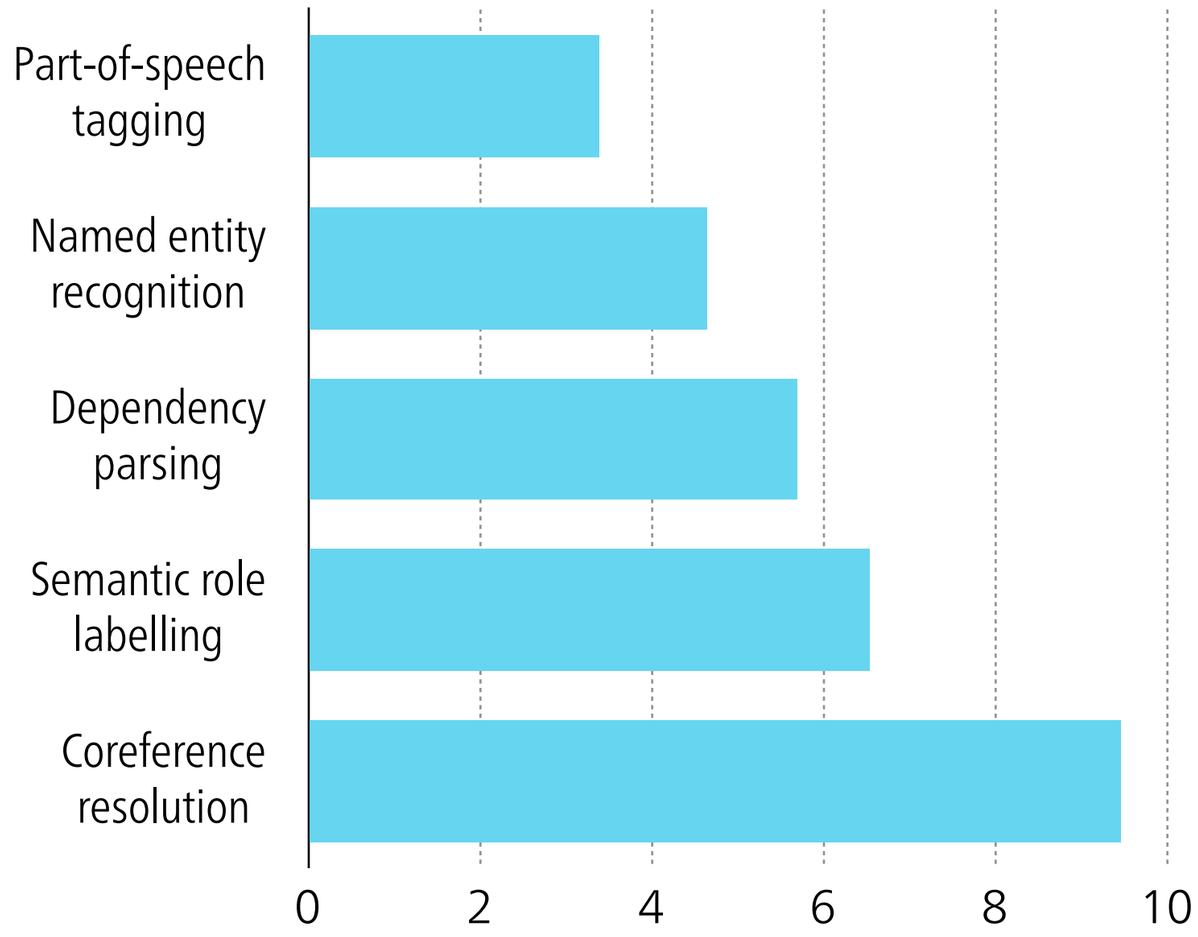
Given a word, is its POS tag the most frequent tag?

Given a word, what is the position of its syntactic head?

Given two words, is one the syntactic head of the other?

Kunz and Kuhlmann (2020); Kunz and Kuhlmann (2022)

The pipeline hypothesis



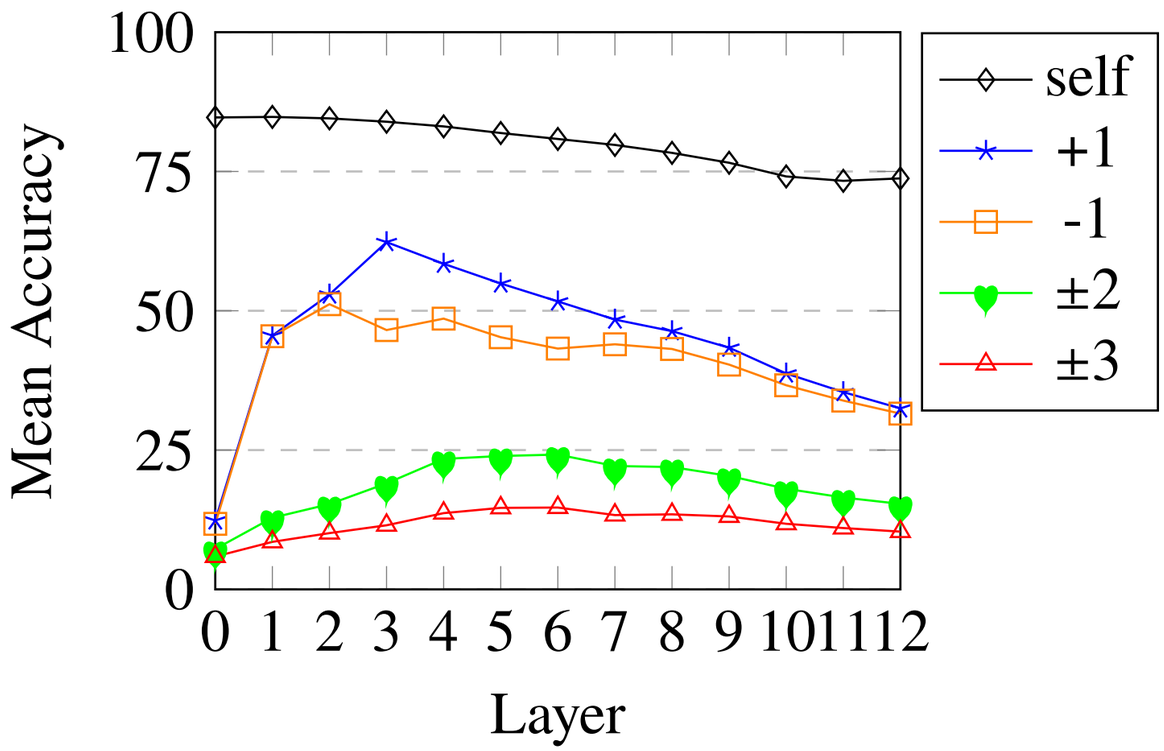
Tenney et al. (2019)

Methodological questions

- What accuracy do we take as sufficient evidence to conclude that a model has ‘learned a task’? What are suitable baselines?
- When a model ‘learns a task’, does that really mean that it encodes linguistic information?
- What is a suitable learning paradigm for asking these questions? What are suitable metrics?

Kunz and Kuhlmann (2020); Kunz and Kuhlmann (2021); Kunz and Kuhlmann (2022)

BERT is very good at predicting neighbouring words



The probe is trained to predict the identity of the word at position ± k.

Vocabulary size = 8,282

A null hypothesis for probing experiments

- The word-level representations learned by BERT contain precise information about the exact linear neighbourhood of the word.
- Because of this, we argue that any study claiming that a model encodes linguistic structure should be able to reject the …
- **Context-only Hypothesis:** The only information that the classifier probe uses to learn the diagnostic task is information about the identities of the neighbouring words of the target word.

Methodological questions

- What accuracy do we take as sufficient evidence to conclude that a model has 'learned a task'? What are suitable baselines?
- When a model 'learns a task', does that really mean that it encodes linguistic information?
- What is a suitable learning paradigm for asking these questions? What are suitable metrics?

Kunz and Kuhlmann (2020); Kunz and Kuhlmann (2021); Kunz and Kuhlmann (2022)

Insights from explanations

Asking language models to explain their predictions

Premise:
A woman in a teal apron prepares a meal at a restaurant.

Hypothesis:
A woman is walking in a park.

Label:
contradiction

Human explanation:
A restaurant is not a park.

Generated explanation:
The woman cannot be walking and preparing a meal at the same time.

Example from e-SNLI; Camburu et al. (2018)

Evaluation of free-form explanations

How similar are generated explanations to human explanations?

- surface-level similarity (BLEU, ROUGE, BERTScore)
- similarity to human reasoning

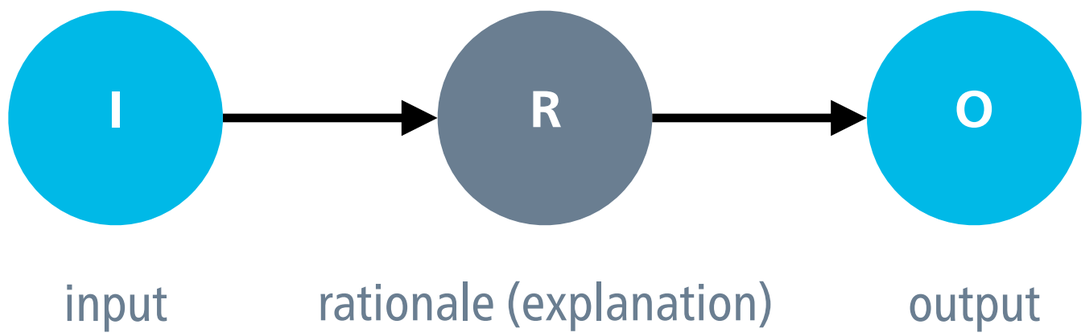
How faithful are generated explanations to model behaviour?

- high feature agreement between labels and explanations
- correlation between noise robustness of labels, explanations

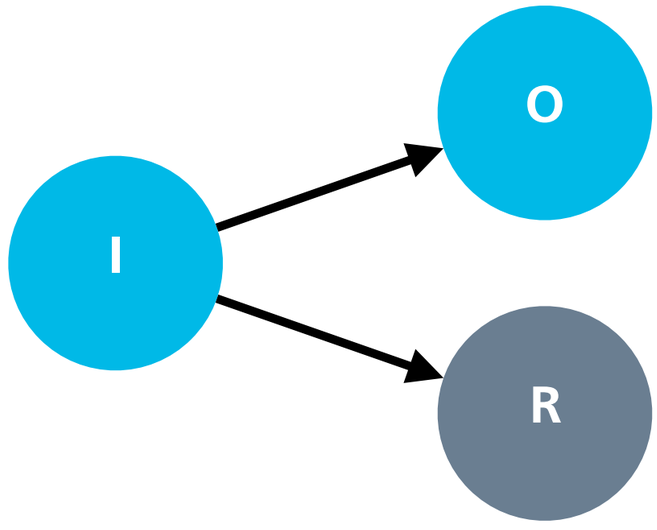
Papineni et al. (2002); Lin (2004); Zhang et al. (2020); Hase et al. (2020); Wiegreffe et al. (2021)

Rationale-augmented model architectures

pipeline architecture

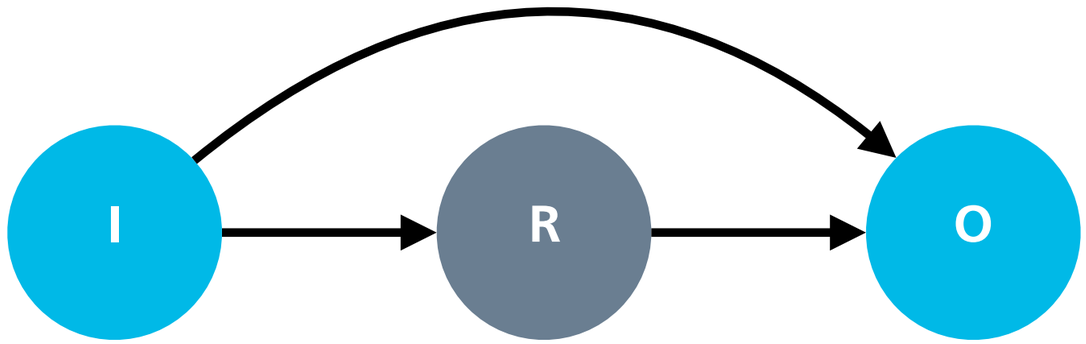


self-rationalizing architecture



Wiegreffe et al. (2021)

Rationale-enriched pipelines



- **pro:** easier to study than self-rationalizing models; higher performance than the R → O component of pipeline models
- **con:** not inherently faithful, as a causal path from input to output remains open

Rajani et al. (2019)

Models

- **Generator:** GPT-2 fine-tuned on the task-specific data set.
  GPT-ST: language modelling, GPT-MT: + label prediction

  Statement: Premise Statement: Hypothesis Explanation: Explanation

- **Classifier:** BERT-base fine-tuned on the task-specific data set.
  Six-different setups depending on the type of explanations.

  [CLS] Premise [SEP] Hypothesis [SEP] Explanation [SEP]

Kunz et al. (2022, forthcoming)

Experimental setups

		Trained with	Tested with
1	None	–	–
2	Gold	Gold explanations	Gold explanations
3	ST –ft	Gold explanations	GPT-ST
4	ST +ft	GPT-ST	GPT-ST
5	MT –ft	Gold explanations	GPT-MT
6	MT +ft	GPT-MT	GPT-MT

Human evaluation

- To assess qualitative properties of the generated explanations, we conduct a human evaluation over 200 samples.
- Each sample is rated yes/no by three persons familiar with the task.
- We report the average score across annotators and Krippendorff's α.

If you disagree with the label or find the example to be non-sense: Flag the example with *N/A*.

Step 1: Look only at e internally:

- Is e a well-formed sentence? (e is grammatical and structurally sensible.)
- Is the content of e factually correct? (e itself is a true statement about the real world. e is factually and logically correct.)

Step 2: Look at e and the label:

- Does e support the label? (Looking at e alone, it is reasonable that the label is correct.)

Step 3: Use all available context:

- Does e provide a valid reasoning path for the label? (e convincingly explains how to get from the context to the label.)
- Does e add new information? (Rather than re-combining information from the context, e comes up with new information.)

Main findings

- Surface similarity, semantic similarity and human ratings do not correlate well with classification accuracy.
- Fine-tuning on generated explanations is crucial for achieving high classification accuracy.
- Existing data sets differ greatly with respect to the quality and uniformity of explanations.

Kunz et al. (2022, forthcoming)

Surface similarity and semantic similarity

Data	ST	MT	Gold
ECQA	7,946	4,436	11,033
e-SNLI	9,398	9,346	14,935

Vocabulary size for generated explanations and gold-standard explanations.

For e-SNLI, explanations generated by ST and MT are similar. Differences are larger for ECQA.

Data	ST	MT
ECQA	0.311	0.250
e-SNLI	0.399	0.401

Semantic similarity as measured by BERTScores (F1)

We see the same trends as in the results about surface similarity.

Classification results

Data set	None	Gold	ST –ft	ST +ft	MT –ft	MT +ft
ECQA	0.378	0.906	0.514	0.631	0.489	**0.634**
e-SNLI	0.898	0.980	0.836	**0.861**	0.836	**0.861**

Classification accuracy measured in terms of macro-averaged F1 scores

Using fine-tuned explanations yields higher scores. ST and MT show similar performance on both data sets.

Results of the human evaluation

	Well-formed	Support	Correctness	Validity	Novelty
ECQA gold	0.603 (+0.22)	0.682 (+0.13)	0.593 (−0.03)	0.490 (+0.18)	0.173 (+0.20)
ECQA GPT-ST	0.573 (+0.25)	0.513 (+0.45)	0.443 (+0.19)	0.285 (+0.48)	0.126 (+0.28)
ECQA GPT-MT	0.607 (+0.32)	0.320 (+0.43)	0.333 (+0.15)	0.107 (+0.43)	0.211 (+0.23)
e-SNLI gold	0.833 (+0.04)	0.873 (+0.06)	0.860 (+0.08)	0.772 (−0.06)	0.052 (−0.02)
e-SNLI GPT-ST	0.868 (+0.10)	0.807 (+0.57)	0.755 (+0.73)	0.670 (+0.65)	0.018 (+0.26)
e-SNLI GPT-MT	0.830 (+0.24)	0.813 (+0.56)	0.813 (+0.56)	0.688 (+0.54)	0.012 (−0.01)

Table 6: Human evaluation: average share of *yes* answers across all samples that were not flagged as invalid. The numbers in parentheses show Krippendorf's α ($n = 3$, interval from -1 to $+1$) for inter-rater agreement.

For ECQA, annotators have a preference for ST. For e-SNLI, there is a slight preference for MT.

Results of the human evaluation

	Well-formed	Support	Correctness	Validity	Novelty
ECQA gold	0.603 (+0.22)	0.682 (+0.13)	0.593 (−0.03)	0.490 (+0.18)	0.173 (+0.20)
ECQA GPT-ST	0.573 (+0.25)	0.513 (+0.45)	0.443 (+0.19)	0.285 (+0.48)	0.126 (+0.28)
ECQA GPT-MT	0.607 (+0.32)	0.320 (+0.43)	0.333 (+0.15)	0.107 (+0.43)	0.211 (+0.23)
e-SNLI gold	0.833 (+0.04)	0.873 (+0.06)	0.860 (+0.08)	0.772 (−0.06)	0.052 (−0.02)
e-SNLI GPT-ST	0.868 (+0.10)	0.807 (+0.57)	0.755 (+0.73)	0.670 (+0.65)	0.018 (+0.26)
e-SNLI GPT-MT	0.830 (+0.24)	0.813 (+0.56)	0.813 (+0.56)	0.688 (+0.54)	0.012 (−0.01)

Table 6: Human evaluation: average share of *yes* answers across all samples that were not flagged as invalid. The numbers in parentheses show Krippendorf's α ($n = 3$, interval from -1 to $+1$) for inter-rater agreement.

On ECQA, MT adds more novel information. (Most of it is factually incorrect.)

Hallucination may be useful for prediction

Question:
The archaeologist was seeing artifacts that he knew were fake, how did he feel?

Target answer:
Painful memories

high rating

ST:
Painful memories is not a feeling.

low rating

MT:
A person who is in fear of being embarrassed is called a bad person.

Results of the human evaluation

	Well-formed	Support	Correctness	Validity	Novelty
ECQA gold	0.603 (+0.22)	0.682 (+0.13)	0.593 (−0.03)	0.490 (+0.18)	0.173 (+0.20)
ECQA GPT-ST	0.573 (+0.25)	0.513 (+0.45)	0.443 (+0.19)	0.285 (+0.48)	0.126 (+0.28)
ECQA GPT-MT	0.607 (+0.32)	0.320 (+0.43)	0.333 (+0.15)	0.107 (+0.43)	0.211 (+0.23)
e-SNLI gold	0.833 (+0.04)	0.873 (+0.06)	0.860 (+0.08)	0.772 (−0.06)	0.052 (−0.02)
e-SNLI GPT-ST	0.868 (+0.10)	0.807 (+0.57)	0.755 (+0.73)	0.670 (+0.65)	0.018 (+0.26)
e-SNLI GPT-MT	0.830 (+0.24)	0.813 (+0.56)	0.813 (+0.56)	0.688 (+0.54)	0.012 (−0.01)

Table 6: Human evaluation: average share of *yes* answers across all samples that were not flagged as invalid. The numbers in parentheses show Krippendorf's α ($n = 3$, interval from -1 to $+1$) for inter-rater agreement.

Overall, scores and inter-rater agreement are low, even for gold explanations.

Main findings

- Surface similarity, semantic similarity and human ratings do not correlate well with classification accuracy.
- Fine-tuning on generated explanations is crucial for achieving high classification accuracy.
- Existing data sets differ greatly with respect to the quality and uniformity of explanations.

Kunz et al. (2022, forthcoming)

Summary

Main conclusions

- Large language models can produce impressive results, but knowing exactly what they have learned is hard.
- Existing methods are often inconclusive and counter-intuitive. There is no comprehensive, well-understood methodology.

  but several interesting ideas, e.g. Voita and Titov (2020)

- We need to get better at clearly formulating the potential and the limitations of different evaluation methods.

Publications

- Jenny Kunz and Marco Kuhlmann. *Classifier Probes May Just Learn from Linear Context Features*. COLING 2020.

- Jenny Kunz and Marco Kuhlmann. *Test Harder Than You Train: Probing with Extrapolation Splits*. BlackboxNLP 2021.

- Jenny Kunz and Marco Kuhlmann. *Where Does Linguistic Information Emerge in Neural Language Models? Measuring Gains and Contributions Across Layers*. COLING 2022.

- Jenny Kunz, Martin Jirénius, Oskar Holmström, and Marco Kuhlmann. *Human Ratings Do Not Reflect Downstream Utility: A Study of Free-Text Explanations for Model Predictions*. Accepted to BlackboxNLP 2022.