Extracting Grammar Rules
from Neural Language Models

Jaap Jumelet & Jelle Zuidema

Language Models and Grammar

What do they ‘know’ about language?

- Syntactic **phenomena**

Language Models and Grammar

What do they 'know' about language?

- Syntactic **phenomena**
  - Subject-verb agreement:
    $P_{\text{LM}}(\textbf{are} \mid \textit{The } \textbf{keys} \textit{ near that } \underline{\textit{table}}) > P_{\text{LM}}(\underline{\textit{is}} \mid \textit{The } \textbf{keys} \textit{ near that } \underline{\textit{table}})$

Language Models and Grammar

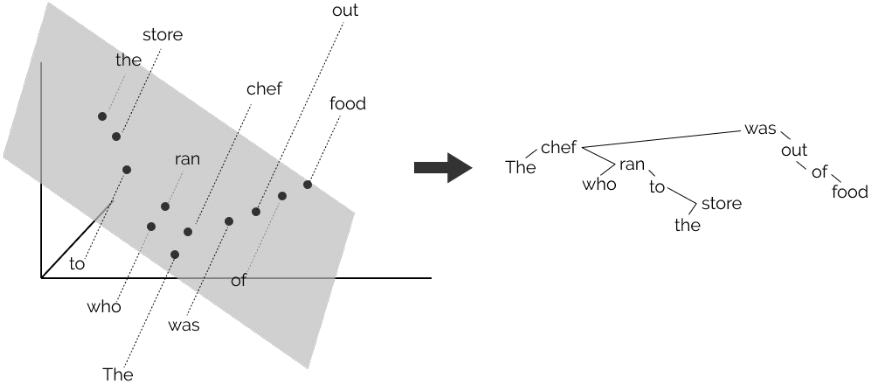
What do they 'know' about language?

- Syntactic **phenomena**
  - Subject-verb agreement:
    $P_{\text{LM}}(\textbf{are} \mid \textit{The } \textbf{keys} \textit{ near that } \underline{\textit{table}}) > P_{\text{LM}}(\underline{\textit{is}} \mid \textit{The } \textbf{keys} \textit{ near that } \underline{\textit{table}})$
  - Determiner-noun agreement:
    $P_{\text{LM}}(\textbf{table} \mid \textit{The keys near } \textbf{that}) > P_{\text{LM}}(\textit{tables} \mid \textit{The keys near } \textbf{that})$
  - *... Many more*

Language Models and Grammar

What do they 'know' about language?

- Syntactic **structure**
  - Structural Probes

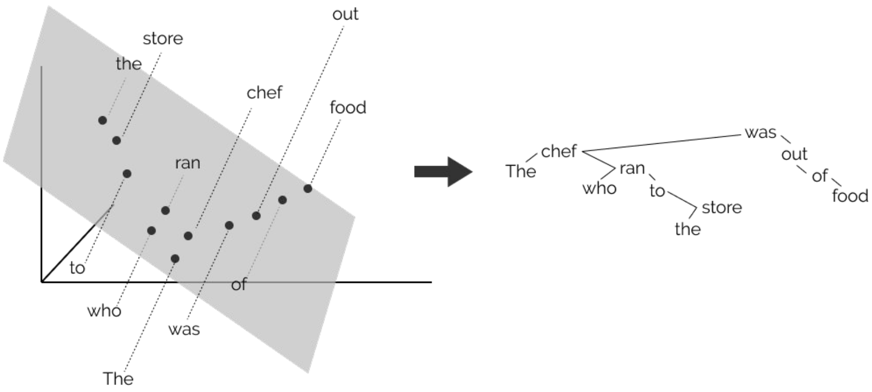


Hewitt et al. (2019)

Language Models and Grammar

What do they 'know' about language?

- Syntactic **structure**
  - Structural Probes



Optimise linear projection B

Predicted distance

$$\min_{B} \sum_{\ell} \frac{1}{|s_\ell|^2} \sum_{i,j} \left(d(w_i, w_j) - \|B(h_i - h_j)\|^2 \right)$$

Gold distance

Hewitt et al. (2019)

Language Models and Grammar

What do they 'know' about language?

- Syntactic **structure**
  - Structural Probes
    - ✔ Easy to train
    - ✔ Applicable to many formalisms

Language Models and Grammar

What do they 'know' about language?

- Syntactic **structure**
  - Structural Probes

    ✔ Easy to train
    ✔ Applicable to many formalisms

    ... but
    ✗ Probing is always **supervised**
    ✗ Did we interpret the model, or did the probe learn the task itself?
    ✗ Is the extracted structure even used for model **predictions**?

Project Goal

Can we extract grammatical structure from a model in a way that is:

- Unsupervised
- Reflective of model predictions

... to gain insights into a model’s comprehension of the structural patterns that underlie the task it was trained on.

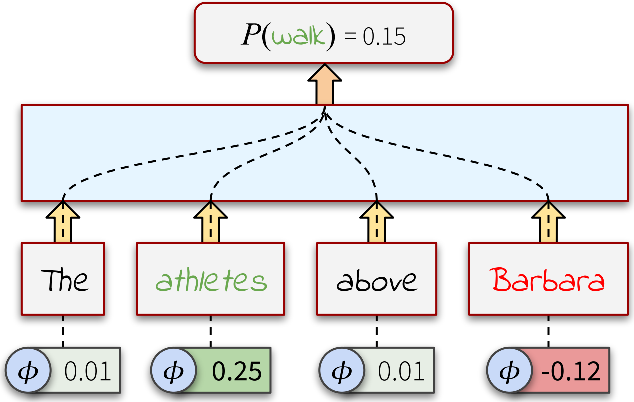
Language Models and Explanations

What do we ‘know’ about language models?

Language Models and Explanations

What do we 'know' about language models?

- Feature attributions
    - Explain model behaviour as a **sum** of **contributions**:



Language Models and Explanations

What do we 'know' about language models?

- Feature attributions
  - Explain model behaviour as a **sum** of **contributions**
  - Often explained in relation to a **baseline:**
    - `<pad>`-token
    - zero-valued
    - random input

Language Models and Explanations

What do we 'know' about language models?

- Feature attributions
  - Explain model behaviour as a **sum** of **contributions**
  - Often explained in relation to a **baseline**
  - ✗ **Faithfulness** is hard to guarantee:
    - Are odd explanations indicative of odd model behaviour, or of a faulty explanation method?
    - How can we know the ***true explanation*** of a model?

Language Models and Explanations

What do we 'know' about language models?

- Feature attributions
  - Explain model behaviour as a **sum** of **contributions**
  - Often explained in relation to a **baseline**
  - ✗ **Faithfulness** is hard to guarantee
  - ✗ 'Flat' contributions represent a limited view of model behaviour:
    - What are the contributions of a sentiment classifier for

      "*This movie was **not bad***" ?
    - **Feature interactions** can provide more fine-grained insights

Language Models and Interactions

Integrated Hessians

Janizek et al. (2021)

Language Models and Interactions

Integrated Hessians

Integral along linear path

- Integrated Gradients:

$$\phi_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$

Baseline

Gradients wrt interpolations

Janizek et al. (2021)

Language Models and Interactions

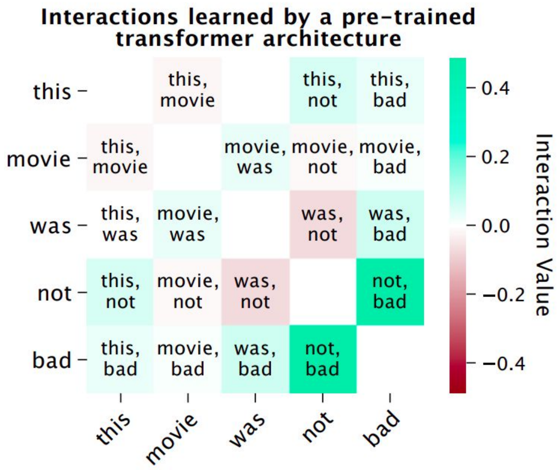
Integrated Hessians

- Integrated Gradients
- Apply **IG** to itself:

$$\Gamma_{i,j}(x) = \phi_j(\phi_i(x))$$

$\Gamma_{i,j}$ represents how much feature j contributed to *the contribution of* feature i

Language Models and Interactions



Integrated Hessians

- Integrated Gradients
- Apply **IG** to itself:

$$\Gamma_{i,j}(x) = \phi_j(\phi_i(x))$$

$\Gamma_{i,j}$ represents how much feature j contributed to *the contribution of* feature i

Janizek et al. (2021)

The Plan

Use Integrated Hessians to gain insights into the grammatical knowledge of a LM.

The Hurdle

We can’t just blindly apply this to BERT and see what happens without better guarantees of **faithfulness**

The Solution

First test the setup on *grey*-box LMs:

- Trained on a simple task that is well understood
- Trained to 100% accuracy

Here:

- Simple CFGs:
  - Palindromes: aabcCBAA
  - Dyck: ([(())()])

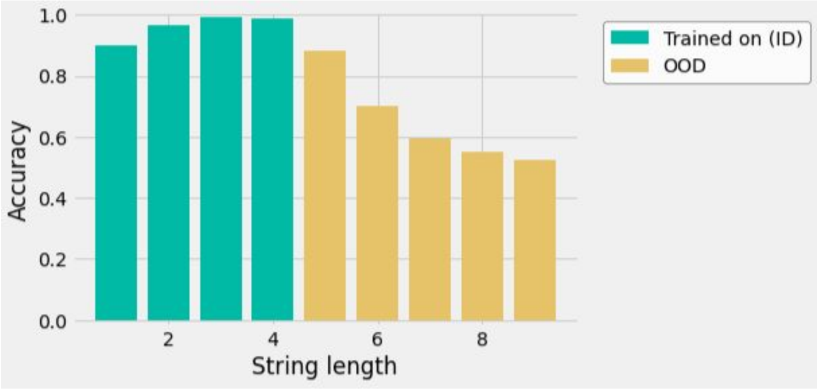
Setup

- Train LSTMs as **string classifiers**: was a string well-formed or not?
    - E.g. for palindromes:

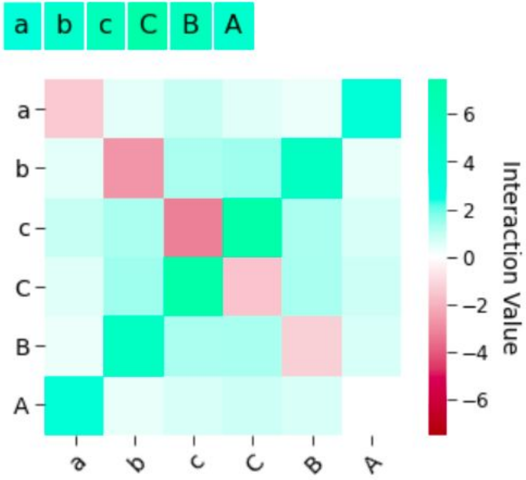
      LSTM(abbBBA) = 1, LSTM(abbB**A**A) = 0
- Apply Integrated Hessians (IH) to the string classification
- Check if the IH interactions reflect the dependencies of the task

Palindromes

Task Performance not perfect yet:

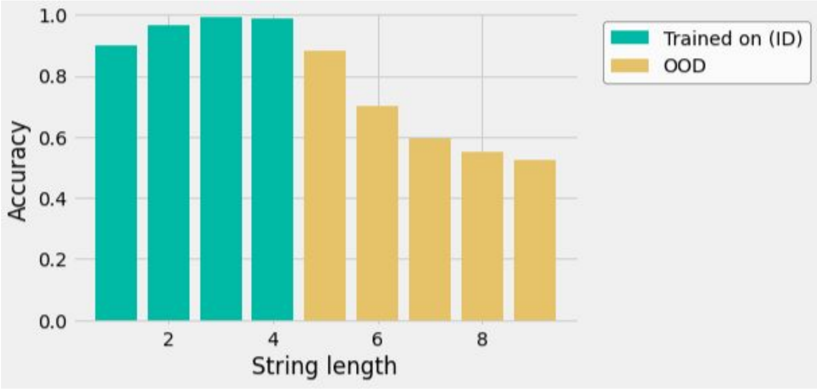


Interaction pattern already insightful:

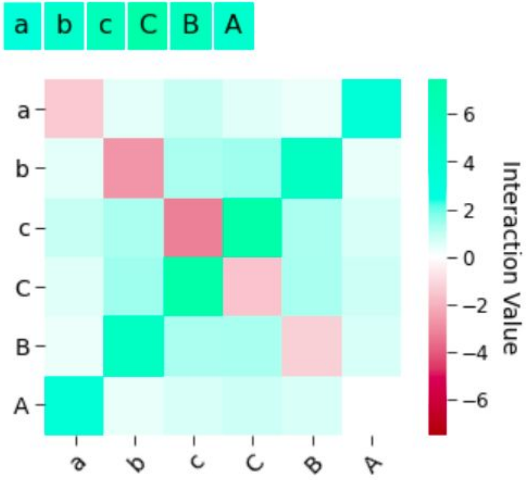


Palindromes

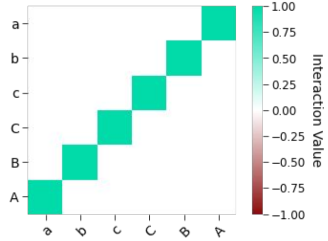
Task Performance not perfect yet:



Interaction pattern already insightful:

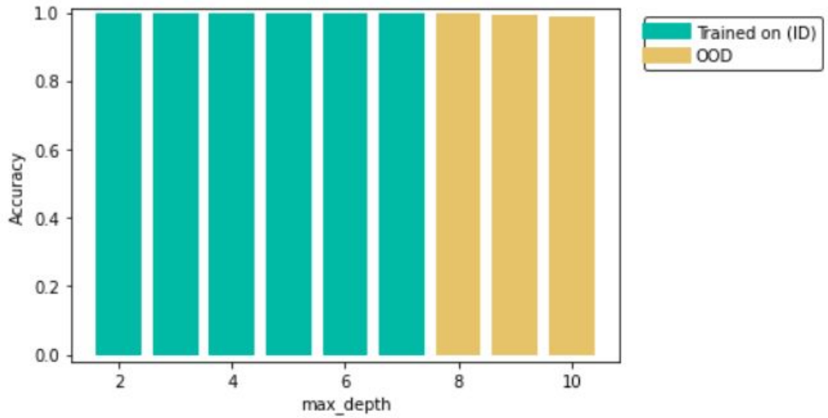


'Optimal':

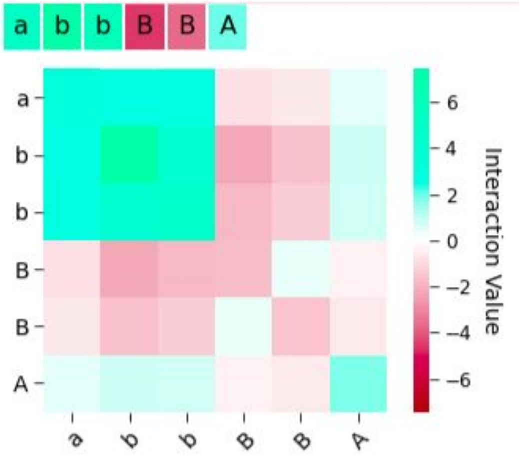


Dyck-2

Task Performance (near) perfect:

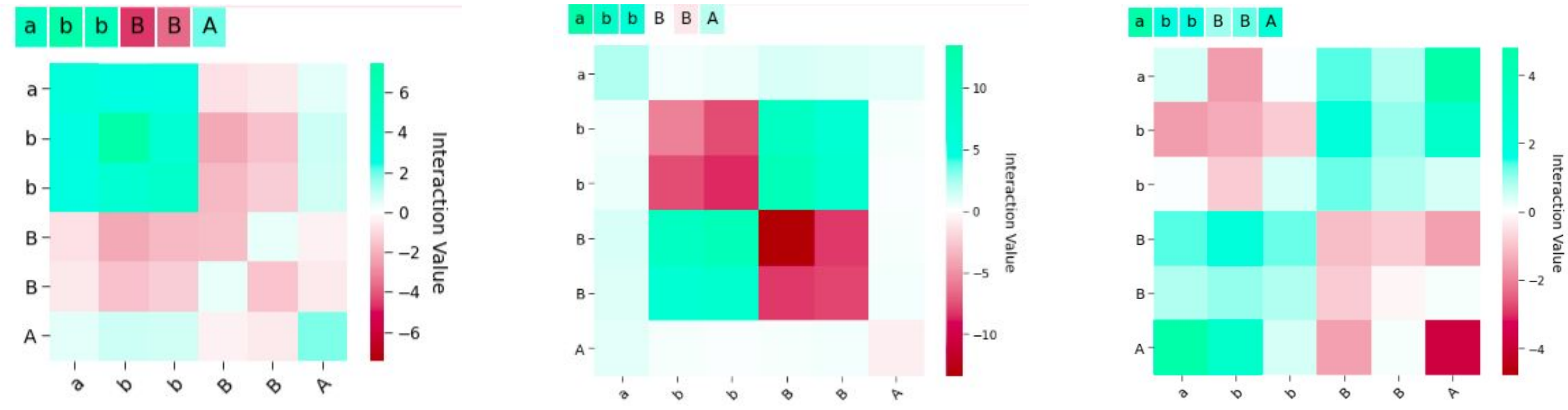


Interaction pattern less insightful:



Instability of IH

When retraining with the same hyperparameters different interactions arise:

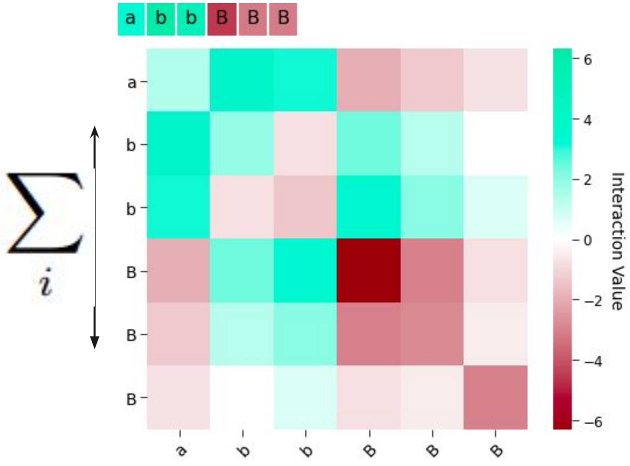


Evaluating Integrated Hessians

We can evaluate the obtained interactions with respect to the attributions of Integrated Gradients, which in turn can be compared to the output of the model

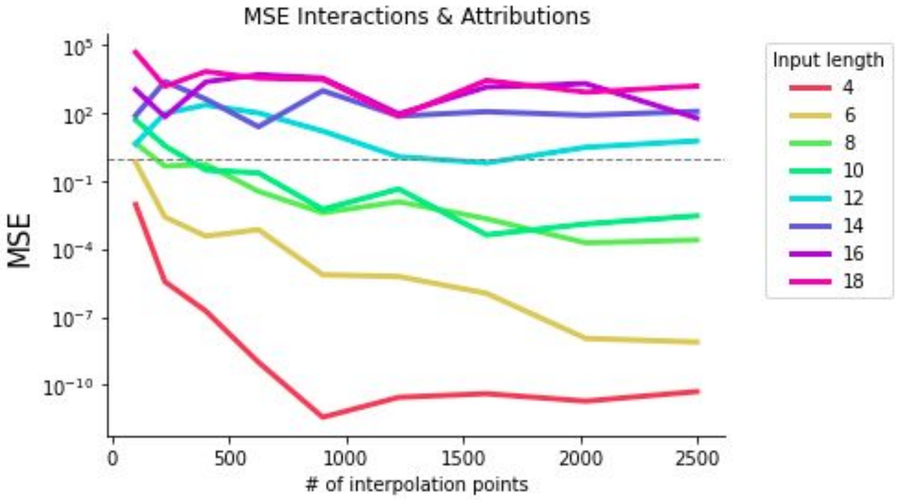
IG: $$\sum_i \phi_i(x) = f(x) - f(x').$$

IH: $$\sum_i \sum_j \Gamma_{i,j}(x) = f(x) - f(x').$$



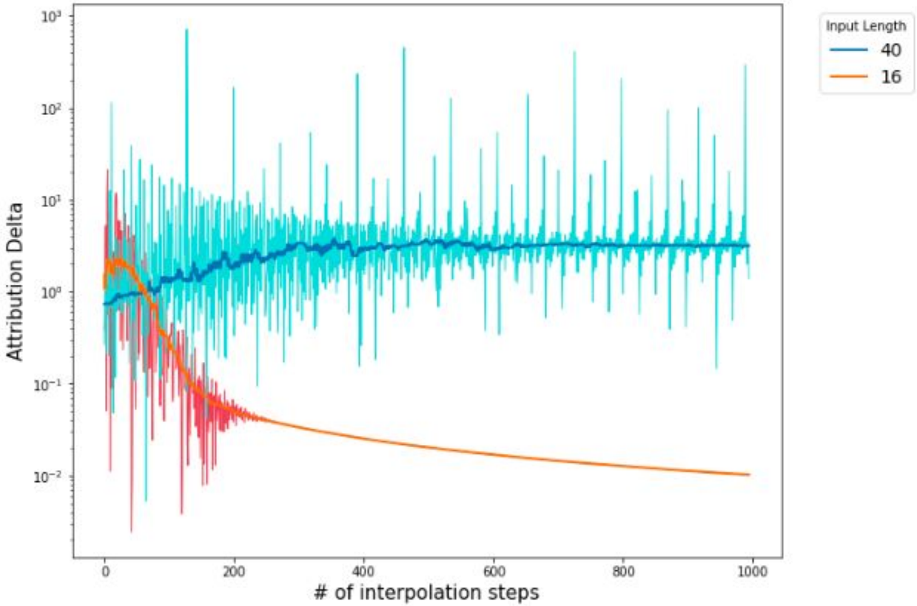
Evaluating Integrated Hessians

No convergence for longer input strings!



Evaluating Integrated Hessians

Integrated Gradients also fails on longer input!



What now?

- It turns out not only the faithfulness of a method to a model is of importance
- Because explanation methods often present an approximation to a complex quantity (Integral, Shapley values, etc.), the output of the method contains uncertainty as well

Future steps

- Reduce instability of IH on longer strings
- Experiment with more baselines
- Test on more tasks (both simpler and more complex -> NL)