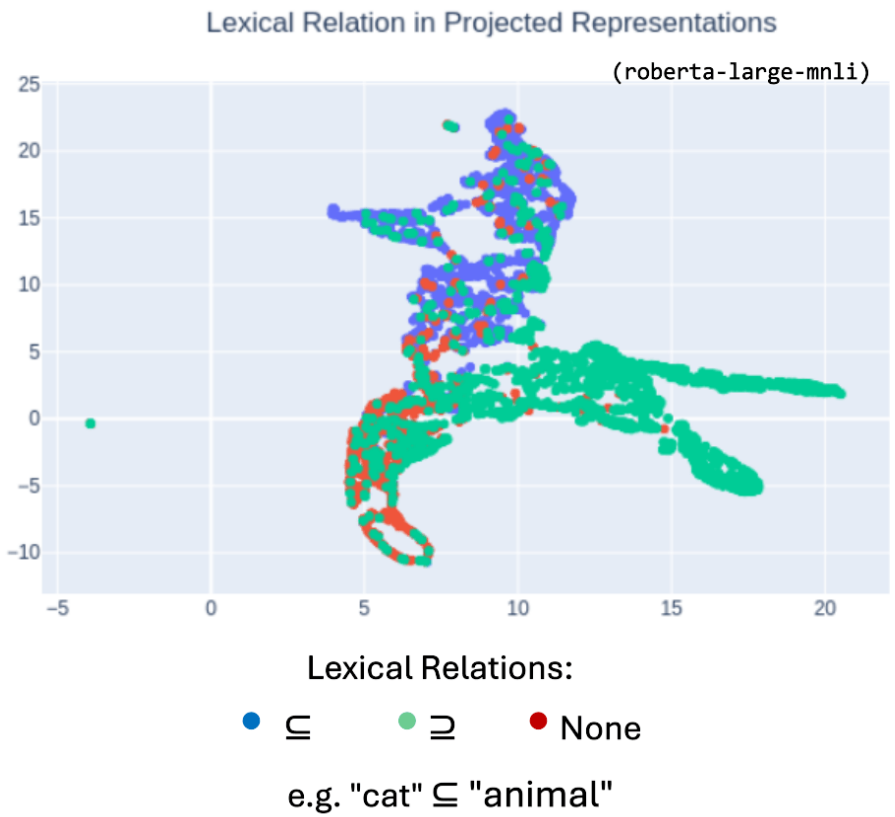
Estimating the Causal Effects of Natural Logic Features in Transformer-Based NLI Models

Julia Rozanova, Marco Valentino, André Freitas

MANCHESTER 1824
The University of Manchester

Introduction: Interpretable Features

- Transformer-based NLP Models fine-tuned for a specific task such as Natural Language Inference *(NLI)* can learn to encode linguistic features required to perform the task.
- Interpretability methods such as *probing* can help demonstrate that given linguistic feature, such as POS or *lexical relations,* are encoded in the intermediate representations of a model.



MANCHESTER 1824
The University of Manchester

Structured Class of NLI Examples

Construct an NLI example as follows:

A "context" ***C***:

I ate some [ ] for breakfast.

A word pair ***W*** with relation ***R***
(or "none):

peaches ⊂ fruit

Combine into a Premise/Hypothesis pair:

Premise:	Hypothesis:
I ate some peaches for breakfast.	I ate some fruit for breakfast.

Causal Diagram: Expected Behaviour

A "context" ***C***:

I ate some ☐ for breakfast.

The context "monotonicity" ***M***:

Upward (↑)

I ate some ☐ for breakfast.

Downward (↓)

I did not eat any ☐ for breakfast.

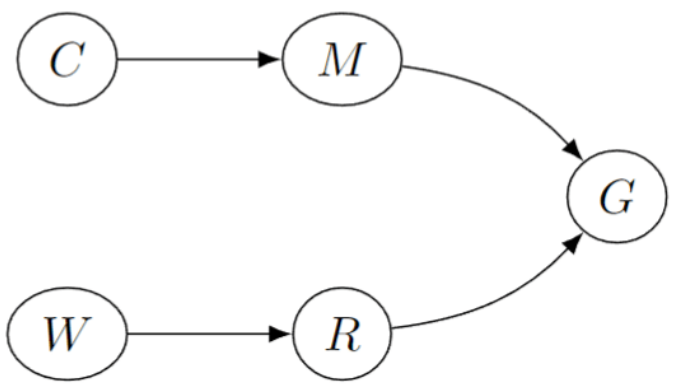
Effect on the entailment label:

M \ R	$\sqsubseteq$	$\sqsupseteq$	$\#$
↑	Entailment	Non-Entailment	Non-Entailment
↓	Non-Entailment	Entailment	Non-Entailment

MANCHESTER 1824
The University of Manchester

Causal Diagram: Expected Behaviour

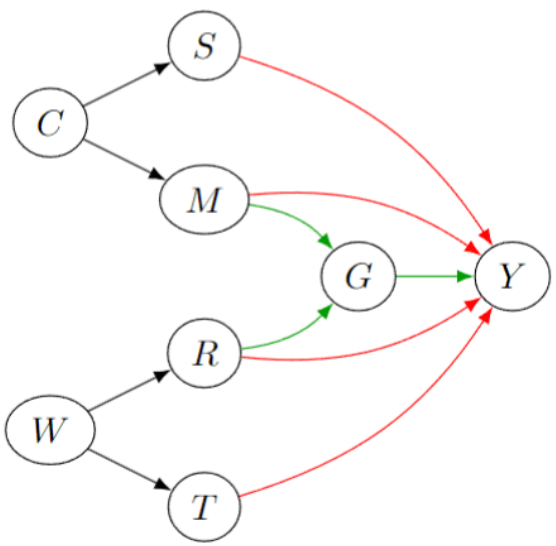
Variable	Description
G	Gold Label
C	Context
M	Context Monotonicity
W	Inserted Word Pair
R	Word-Pair Relation



MANCHESTER 1824
The University of Manchester

Causal Diagram: Model Behaviour

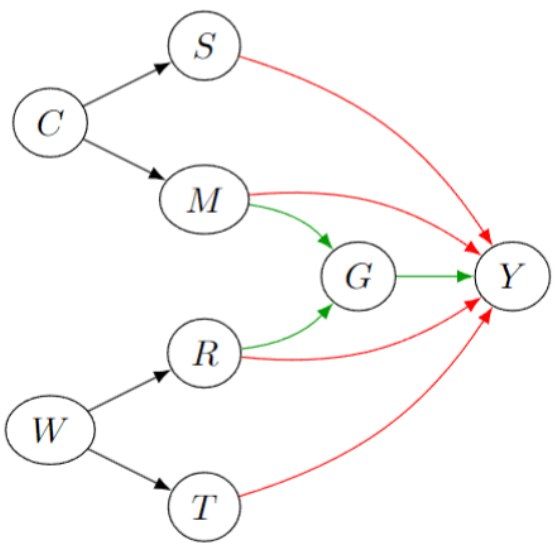
Variable	Description
Y	Model Prediction
G	Gold Label
C	Context
M	Context Monotonicity
S	Context Textual Surface Form
W	Inserted Word Pair
R	Word-Pair Relation
T	Word-Pair Textual Surface Form



MANCHESTER 1824
The University of Manchester

Causal Diagram: Model Behaviour

Variable	Description
Y	Model Prediction
G	Gold Label
C	Context
M	Context Monotonicity
S	Context Textual Surface Form
W	Inserted Word Pair
R	Word-Pair Relation
T	Word-Pair Textual Surface Form



- Not all causal effects can be estimated, but we can calculate:

- DCE S -> Y
- DCE T -> Y
- TCE C on Y
- TCE W on Y

MANCHESTER 1824
The University of Manchester

Interventions

- Calculate the causal effects along different paths by introducing an **intervention** which changes the variable of interest while holding relevant variables constant
- See Stolfo et al (202x)

Intervention Set	Target Measure	C	W	M	R	G	Interventions in Dataset
$\mathcal{I}_0$	TCE $(C \to Y)$	$\neq$	$=$	$\neq$	$=$	$\neq$	14270
$\mathcal{I}_1$	TCE $(W \to Y)$	$=$	$\neq$	$=$	$\neq$	$\neq$	22640
$\mathcal{I}_2$	DCE $(S \to Y)$	$\neq$	$=$	$=$	$=$	$=$	20910
$\mathcal{I}_3$	DCE $(T \to Y)$	$=$	$\neq$	$=$	$=$	$=$	25960

MANCHESTER 1824
The University of Manchester

Intervention Examples

Intervention Set	Target Quantity	Intervention Step	Premise	Hypothesis	M	R	G
$\mathcal{I}_0$	TCE(C on Y)	Before	You can't live without fruit .	You can't live without strawberries .	$\uparrow$	$\sqsupseteq$	Non-Entailment
		After	All fruit study English.	All strawberries study English.	$\downarrow$	$\sqsupseteq$	Entailment
$\mathcal{I}_2$	DCE($S \rightarrow Y$)	Before	He has no interest in seafood .	He has no interest in oysters .	$\downarrow$	$\sqsupseteq$	Entailment
		After	I don't want to argue about this in front of seafood .	I don't want to argue about this in front of oysters .	$\downarrow$	$\sqsupseteq$	Entailment

Table 3: Example context interventions for determining the total causal effect of label-relevant context changes and the direct causal effect of label-irrelevant context changes.

MANCHESTER 1824
The University of Manchester

Causal Effect Calculation

We then calculate:

$$\text{TCE}(C \text{ on } Y) = \frac{1}{|\mathcal{I}_0|} \sum_{(n,n') \in I_0} CP(n, n')$$

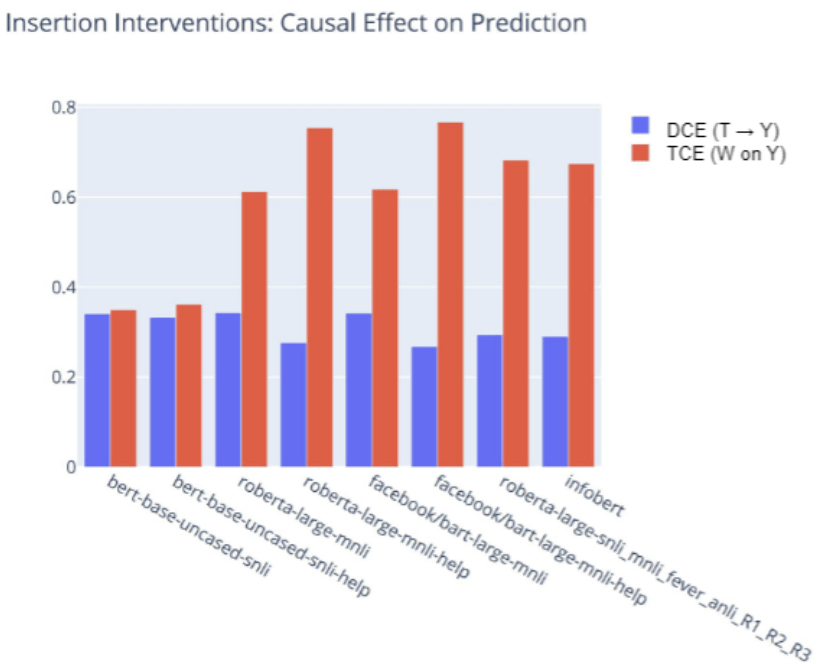
- For each intervention set, we calculate the average "change in prediction":

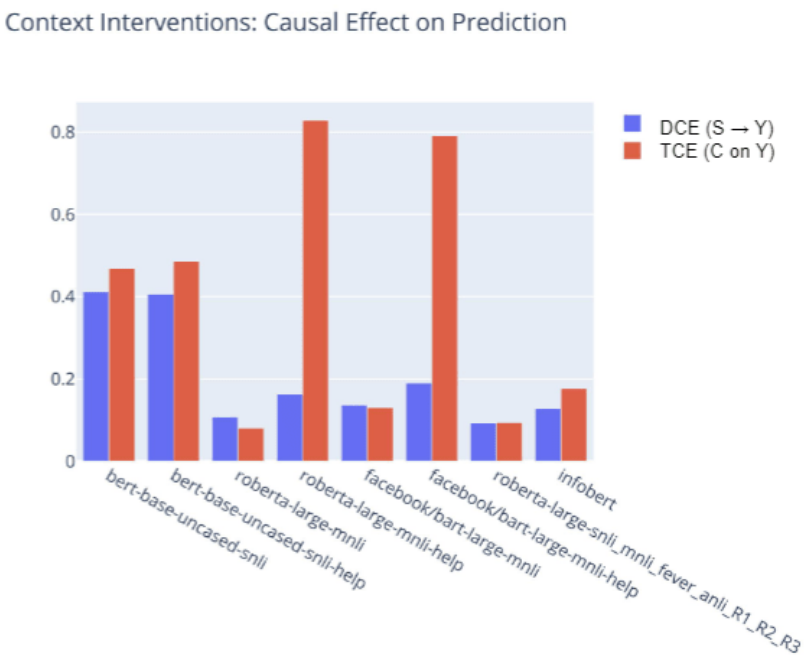
  This is equal to 1 if the prediction changes, 0 if it does not.

MANCHESTER 1824
The University of Manchester

Causal Effect Calculation

The best case scenario is a high TCE (high sensitivity to relevant changes) combined with a low DCE (high robustness to irrelevant changes).





MANCHESTER 1824
The University of Manchester

Interpreting Causal Effects

- Stolfo et al interpret these causal effects in terms of ***robustness*** and ***sensitivity***:

- A high TCE for a relevant variable means high sensitivity to relevant changes,
- while a low DCE for an irrelevant variable means high robustness to irrelevant changes.

MANCHESTER 1824
The University of Manchester

Interpreting Causal Effects

Model	Context Changes		Inserted Word-Pair Changes	
	Robustness	Sensitivity	Robustness	Sensitivity
bert-base-uncased-snli	Mid	Mid	Mid	Low
bert-base-uncased-snli-help	Mid	Mid	Mid	Low
facebook/bart-large-mnli	High	Low	Mid	Mid
facebook/bart-large-mnli-help	Mid/High	**Highest**	**Highest**	**Highest**
roberta-large-mnli	**Highest**	Lowest	Mid	Mid
roberta-large-mnli-help	High	**Highest**	**Highest**	**Highest**
roberta-large-snli_mnli_fever_anli	**Highest**	Lowest	Mid	Mid/High
infobert	High	Low	Mid	Mid/High