Testing Pre-trained Language Models' Understanding of Distributivity via Causal Mediation Analysis

Pangbo Ban[†], **Yifan Jiang**[†], **Tianran Liu**[†], **Shane Steinert-Threlkeld**
Department of Linguistics, University of Washington
`{pbban, yfjiang, tianranl, shanest}@uw.edu`

Abstract

To what extent do pre-trained language models grasp semantic knowledge regarding the phenomenon of distributivity? In this paper, we introduce DistNLI, a new diagnostic dataset for natural language inference that targets the semantic difference arising from distributivity, and employ the causal mediation analysis framework to quantify the model behavior and explore the underlying mechanism in this semantically-related task. We find that the extent of models' understanding is associated with model size and vocabulary size. We also provide insights into how models encode such high-level semantic knowledge. Our dataset and code are available on GitHub.

1 Introduction

The ability to understand and utilize semantic knowledge (consciously or unconsciously) is essential to human reasoning process. Although significant progress has been made by large-scale pre-trained language models on many reasoning-required tasks, it is still unclear whether these models have reached a considerable level of competence in discerning and processing semantic knowledge. To break into the black box, recent studies employ various analysis methods and bring evidence that semantic knowledge (Bowman et al., 2015b; Ettinger, 2020; Jumelet et al., 2021) is encoded by pre-trained models. However, some issues still remain. First, due to the difficulty in being analyzed and probed for, many semantic phenomena are not touched on by NLP researchers, even though they have been studied by linguists for decades. Second, current analysis methods are not flawless. For example, Belinkov (2021) reviews some limitations of the probing classifier paradigm such as the spurious correlation between the probing classifier and the original model.

To address these issues, in this paper, we leverage causal mediation analysis (CMA), a new analysis framework introduced by Vig et al. (2020a,b), to test pre-trained language models' understanding of semantic knowledge, with a specific focus on predicative distributivity. As a complex linguistic phenomenon, predicative distributivity involves semantics, pragmatics, and psycholinguistics. With minimal pairs differing in distributivity, we look into whether pre-trained language models grasp the semantic difference, and how much a model component plays a role in such extent of understanding. Our contributions are as follows:

- We introduce DistNLI, a diagnostic NLI dataset which targets testing pre-trained language models' ability of discerning the property of predicative distributivity via minimal pairs of coordinated sentences (Section 3).

- We refine metrics used in the CMA framework, namely total effect (TE), natural indirect effect (NIE) and natural direct effect (NDE) to guarantee the effect decomposition. We apply the framework to the ternary NLI task (Sections 5.1 and 6.1).

- We find that pre-trained models with either more parameters or richer vocabulary show some understanding of distributivity. We also find that knowledge of distributivity is concentrated in middle layers and the level of concentration is patterned with the degree of understanding (Sections 4, 5.2 and 6.2).

2 Related Work

Distributivity A sentence with the verb predication applying to a group, e.g. *Sumon and Frank built a boat.*, can be interpreted into at least two readings: the distributive and the collective ones (Scha, 1981). In the distributive reading, the predication applies to each individual in the group (e.g.

---

both Sumon and Frank individually/separately built a boat), whereas in the collective reading, the predication only applies to the group as a whole (e.g. Sumon and Frank built a boat *jointly*).

To theorize this linguistic phenomenon, Scha (1981) introduces the property of distributivity as a tool to formalize the collective and distributive senses of a predication. While Scha analyzes distributivity as a pure lexical property of different predicates, Link et al. (2002) propose that distributivity is a semantic operator comparable to *each*, which he defines as the 'D-operator'. Some later semanticists settle their analyses on the middle ground where the D-operator and the direct predication (based solely on the lexical property) theory are useful under different circumstances (Dowty et al., 1987; Roberts, 1987; Hoeksema, 1988; Verkuyl, 1993; Winter, 1997; De Vries, 2017; Champollion, 2017). Arguing that the predication of a simple sentence can be directly interpreted based on its lexical meaning while the analysis of a complex one will need a D-operator, Winter (1997) further defines P-distributivity and Q-distributivity, which correspond to the distributive sense under the direct predication and the D-operator respectively. For instance, *Azul and Marsha hold three balloons* falls in the category of Q-distributivity, since it is necessary to introduce the D-operator to analyze the distributivity given the group *Azul and Marsha*. The example *Yu and Vivian laughed*, on the other hand, is a case of P-distributivity, since it is apparent that the predicate *laughed* entails the distributivity originated from its lexical meaning.

Adopting the terminology proposed by De Vries (2017), we investigate the predicative distributivity in this paper by following the approach advocated by Winter (1997) and Champollion (2017) that collectivity and P-distributivity are categorized under the direct predication, which is paralleled to the D-operator that gives the Q-distributivity. We will call predicates which can be perceived in both distributive and collective senses ambiguous predicates in this paper, which corresponds to the mixed predicates in the previous literature. (De Vries, 2017; Champollion, 2017).

Natural Language Inference   Natural Language Inference (NLI) is the task of determining whether one sentence (premise) entails, contradicts, or is neutral to another sentence (hypothesis). Early attempts include Chen et al. (2017) and Ghaeini et al. (2018) which are LSTM-based. Recent progress

in pre-trained language models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021) and NLI datasets like SNLI (Bowman et al., 2015a), XNLI (Conneau et al., 2018), and MNLI (Williams et al., 2018) provides more opportunities to tackle this task. Pre-trained models fine-tuned on NLI datasets often achieve satisfying performance.

However, diagnostic studies show that high scores achieved by neural models do not mean they truly understand the relationship between sentences. For example, McCoy et al. (2019) find BERT trained on MNLI instead leverages shallow heuristics to make predictions.

Causal Mediation Analysis   Recently, several analysis methods have been proposed to reveal information that are learned and utilized by language models. One of them is causal mediation analysis (Robins and Greenland, 1992; Pearl, 2001), a statistical framework to identify direct and indirect effects of an intervention on an outcome of interest. It is first introduced as an analysis method to the NLP field by Vig et al. (2020a,b) to scrutinize gender bias in language models. More recent studies use it to explore linguistic phenomena such as syntactic agreement (Finlayson et al., 2021) and negation (Dobreva and Keller, 2021). The framework consists of three metrics: TE, NIE and NDE. TE is used to quantify how an *input intervention* (e.g., text edits) would affect a *response variable* (e.g., predicted probabilities). NIE and NDE are used to measure the mediated influence realized through an intermediate variable, or *mediator*, which can be a neuron, a whole layer, or an attention head. TE is usually decomposed into the sum of NIE and NDE. As Vig et al. (2020a,b) suggest, the CMA framework has great potential for extensions. Motivated by their work, in this paper, we apply the framework, which was limited to binary classification, to the NLI task and propose alternative definitions of the metrics for a more robust effect decomposition.

3   The DistNLI Dataset

3.1   Capturing distributivity with NLI

Although semanticists differ in the treatment of lexical distributivity (i.e., P-distributivity) in their theories, the evaluation of predicative distributivity generally relies on the validity of the inference from a predication of a group and the part with the *each* operator in the case of conjunction (e.g.

Premise	Hypothesis	NLI Label	Distributivity
Mia and Lin *laughed*.	Mia *laughed*./Lin *laughed*.	Entailment	Distributive
Mia and Lin *pushed a rock*.	Mia *pushed a rock*./Lin *pushed a rock*.	Entailment non-Entailment	Distributive Collective
Mia and Lin *gathered*.	*Mia *gathered*./*Lin *gathered*.	N/A	Collective

Table 1: The relationship between distributivity and NLI labels illustrated by examples from DistNLI. The NLI label is determined by the type of the predicate in the premise, and *non-entailment* means that both *neutral* and *contradiction* are acceptable as true labels. Here, *pushed a rock* is an ambiguous predicate, so it can be assigned with both NLI labels.

Sumon and Frank each built a boat) (Dowty et al., 1987; Lasersohn, 1995; Winter, 1997; Champollion, 2017; De Vries, 2017).

Rooted in this definition, distributive predicates sanction the entailment relation between the plurality and its part, whereas collective or ambiguous predicates do not. Consequently, we use the NLI task to evaluate models' understanding of predicative distributivity. In our approach, the model's ability to discern distributivity is evaluated by the divergent prediction of models between predicates with differing distributivity. Table 1 demonstrates the relationship between distributivity and NLI labels. For instance, given the premise *Mia and Lin laughed* and the hypothesis *Lin laughed*, the model should predict the label as entailment based on the distributive predicate *laughed*. On the other hand, given the premise *Mia and Lin pushed a rock* and the hypothesis *Mia pushed a rock*, since *pushed a rock* is an ambiguous predicate, the model will exhibit a completely different performance if it grasps the disparity in semantics the distributivity exerts.

3.2 Data Generation

We generate a synthetic NLI dataset consisting of premise-hypothesis pairs with `[DP1] and [DP2] [Pred]` as premise and `[DP1]/[DP2] [Pred]` as hypothesis. `[DP1]` and `[DP2]` denote determiner phrases and `[Pred]` denotes a predicate. For example:

> **Premise**: Mia and Lin wore a mask.
> **Hypothesis**: Mia (Lin) wore a mask.

None of determiner phrase contains quantifiers in its structure, and both lexical and phrasal predicates are included (Champollion, 2017). Three kinds of noun phrases have been formulated, namely person, animal, and object, and it is guaranteed that no group nouns like *the committee* and

conventionalized conjunctions like *Simon and Garfunkel* are included. The template is further instantiated with distributive and ambiguous predicates. We scrape existing categorized predicates from past publications on distributivity, and augment the list with predicates of similar pattern and characteristics regarding the semantic ambiguity in the information structure (Kroch, 1974; Taub, 1989; De Vries, 2017; Champollion, 2017; Coppock and Champollion, 2019). The augmented list aligns with the report of Safir and Stowell (1987) that most predications with indefinite cardinal as the determiner manifest the distributive feature.

3.3 Annotation

The annotation on predicative distributivity in this dataset consists of two stages. During the first stage, we recruit three graduate students, who are native speakers of American English with both linguistics and NLP background, to annotate grammatical sentences with predicates for whether they are distributive, collective or ambiguous. An example with pictures and explanation is given prior to the task. Considering the subtle nature of distributivity, we synthesize their judgements and discard highly controversial data points (i.e. predicates which received three distinct labels). During the second stage, the dataset is further confirmed by an expert in both Semantics and NLP to validate the result and guarantee a trustworthy dataset. The post-annotated data is split into the control group and the intervention group, with the former containing 164 pairs with distributive predicates and the latter containing 164 pairs with ambiguous predicates. The construction of the groups is explained in Section 5.1. They together form the final DistNLI dataset of 328 premise-hypothesis pairs.

Model	ConjNLI		HANS		DistNLI	
	Total Acc	AND Acc	Ent. Acc	Non-Ent. Acc	Dis. Acc	Amb. Acc
DeBERTa-base	65.81	64.84	99.3	53.25	100.00	0.00
DeBERTa-large	66.61	65.99	99.89	54.81	100.00	0.00
DeBERTa-xlarge	65.01	64.84	100	42.33	100.00	0.00
DeBERTa-v2-xlarge	66.29	65.99	99.96	49.87	100.00	0.00
DeBERTa-v2-xxlarge	66.45	65.42	99.92	43.35	100.00	0.00
RoBERTa-large	64.53	64.55	99.65	46.61	100.00	0.61

Table 2: Pre-examination results on ConjNLI, HANS and DistNLI. For ConjNLI, **Total Acc** is the accuracy on the full ConjNLI data, and **AND Acc** is the accuracy on cases with *and* in the premise, hypothesis, or both. For HANS, **Ent. Acc** and **Non-Ent. Acc** stand for **Entailed Acc** and **Non-Entailed Acc**, which are the accuracy on entailed cases and non-entailed cases respectively. For DistNLI, **Dis. Acc** and **Amb. Acc** stand for **Distributive Acc** and **Ambiguous Acc**, which are the accuracy on the control group and intervention group respectively.

4 Models

4.1 Model Selection

We choose recent pre-trained models that are fine-tuned on MNLI as our target models. We try to control conditions as much as possible, such as model size and training setup. We use six models: DeBERTa (base, large, xlarge), DeBERTa-v2 (xlarge, xxlarge), and RoBERTa-large. RoBERTa-large shares the same vocabulary of size 50K as DeBERTa variants while DeBERTa-v2 variants increase their vocabulary size to 128K.

4.2 Pre-examination

Pre-trained language models can leverage various types of information learned from the training corpus to tackle the downstream tasks. The one we want to test in this paper is distributivity, but there is other related information models may use to make predictions, such as coordination, which is used to generate our dataset, or lexical overlap between the hypothesis and premise. To minimize the effect of these confounders, we run selected models over existing diagnostic NLI datasets, namely ConjNLI (Saha et al., 2020) and HANS (McCoy et al., 2019). In addition, we also report each model's accuracy on DistNLI as a preliminary evaluation of models' ability to recognize distributivity. Results on these datasets are shown in Table 2.

ConjNLI ConjNLI is an NLI dataset testing both Boolean and non-Boolean usage of conjuncts including *and*, *or*, *but*, and their combination with quantifiers and negation (Saha et al., 2020). We report each model's accuracy on the whole development set of ConjNLI as well as the subset where *and* exists in the premise, the hypothesis, or both.

We find that all selected models perform reasonably well on ConjNLI, suggesting that they can handle diverse Boolean and non-Boolean coordinated sentences to a certain extent.

HANS HANS is an NLI dataset testing whether models have adopted syntactic heuristics, e.g., the lexical overlap heuristic, the subsequence heuristic, the constituent heuristic during pre-training (McCoy et al., 2019). HANS is annotated with *entailment* on which the heuristics make correct predictions, and *non-entailment* (*neutral* or *contradiction*) on which the heuristics make incorrect predictions. Therefore, if a model can perform perfectly on the entailed cases but fails on the non-entailed cases, it may exploit the heuristics. For selected models, we report their accuracy on both entailed and non-entailed cases. As expected, we find that all models achieve nearly perfect performance on entailed cases. Nevertheless, they reach much higher scores on the non-entailed cases compared to the experiment with BERT by McCoy et al. (2019). This indicates that these models rely less on syntactic heuristics to solve the NLI task.

DistNLI We report each model's accuracy on the control (distributive) group and the intervention (ambiguous) group. Because the control group is labeled with *entailment* and the intervention group is labeled with *non-entailment*, unsurprisingly results on DistNLI match the same pattern as that on HANS. However, the accuracy gap between the two groups is much bigger. Although this seems to suggest selected models heavily rely on syntactic heuristics and hence fail to recognize distributivity, results on ConjNLI and HANS show that selected models can handle at lease some of the

non-Boolean coordinated sentences on which the heuristics provide no help. It is also worth noting that even for our recruited annotators, distributivity is a challenging phenomenon. Their responses demonstrate more variance in ambiguous predicates labeling. This implies accuracy may not be a good metric to evaluate models' understanding of distributivity since it does not account for the change of predicted probability when the decision is not flipped. Thus, we leverage the CMA framework for further investigation.

5 Total Effect

5.1 Experiment Design

Response Variable To quantify the model behavior, we follow the general idea proposed by Finlayson et al. (2021). We define the response variable y as the odds that a model (parameterized by θ) predicts *non-entailment* for a premise-hypothesis pair S_I, where I is the set of possible readings of S which may contain a distributive reading, a collective reading or both:

$$y(S_I) = \text{Odds}(\textit{non-entailment}|S_I)$$
$$= \frac{P_\theta(\textit{non-entailment}|S_I)}{P_\theta(\textit{entailment}|S_I)} \quad (1)$$

The larger the $y(S_I)$, the more likely the model predicts *non-entailment* on the input pair S_I. With this definition, we can transform the original question into a binary task: whether models have a stronger preference over *non-entailment* given a premise-hypothesis pair with a particular predicate.

We hypothesize that if a model has some understanding of distributivity, *ceteris paribus*, an ambiguous pair should result in a larger predicted probability of *non-entailment* than a distributive pair, even if the model predicts *entailment* for both. In other words, $y(S_I)$ should be small when I only contains a distributive reading but relatively large when I contains both readings, provided that all other aspects of S are equal.

Input Intervention To isolate distributivity, we need a class of interventions to change possible readings of a given premise-hypothesis pair S while keeping everything else the same. Since it is intractable to directly modify the possible readings I, we choose to modify the surface form. For our data templates, I is determined by the predicate. Therefore, we define a do-operator swap-pred, which replaces the predicate in the given pair with a
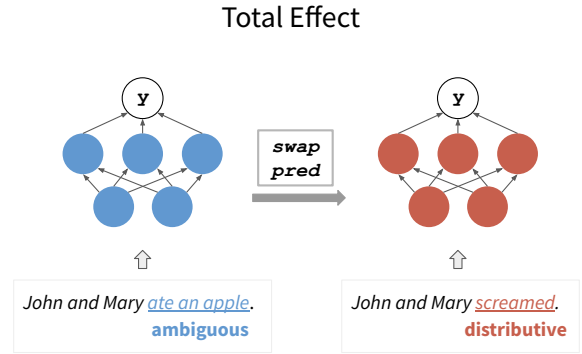

Total Effect

Figure 1: Total effect measures the relative change in y from the input intervention which alters the distributivity of the predicate in the sentence.

random sampled predicate of a different type, as illustrated in Figure 1.[1] We also define the null operator which preserves the original predicate. The response variable $y(S_I)$ can be redefined as $y_i(S_p)$ where i is a do-operator, S is a premise-hypothesis pair, and p indicates the type of the predicate in S:

$$y_{\texttt{null}}(S_p) = \text{Odds}(\textit{non-entailment}|S_p) \quad (2)$$
$$y_{\texttt{swap-pred}}(S_p) = \text{Odds}(\textit{non-entailment}|S_{p'}) \quad (3)$$

Leveraging these interventions, we can split the input dataset into two groups: the control group and the intervention group. Pairs in the control group have distributive predicates with the null operator applied. Pairs in the intervention group are the same sentences but with the swap-pred operator applied. Each pair in the control group can have one or more matches in the intervention group. For example, if *John and Mark smiled.* (premise) and *John smiled.* (hypothesis) is in the control group, *John and Mark built a house.* (premise) and *John built a house.* (hypothesis) could be its potential match in the intervention group.

Metric TE is used to measure how much a response variable y would change if we apply the swap-pred operator rather than the null operator. Instead of the odds difference definition used by previous studies (Vig et al., 2020b; Finlayson et al., 2021; Dobreva and Keller, 2021; Jeoung and Diesner, 2022), we adopt the odds ratio definition proposed by VanderWeele and Vansteelandt (2010). To make the scale more symmetric, we take the

---

[1] The illustration of TE and the following NIE figure are inspired by Vig et al. (2020a); Finlayson et al. (2021).

equivalent logarithmic version:

$$\mathrm{TE}(\mathtt{swap\text{-}pred}, \mathtt{null}; y, S_p)$$

$$= \log\left(\frac{y_{\mathtt{swap\text{-}pred}}(S_p)}{y_{\mathtt{null}}(S_p)}\right) \qquad (4)$$

$$= \log\left(\frac{\mathrm{Odds}(\textit{non-entailment}|S_{p'})}{\mathrm{Odds}(\textit{non-entailment}|S_p)}\right)$$

We calculate the sample average total effect over DistNLI to estimate the average total effect over the population of all possible matched pairs:

$$\overline{\mathrm{TE}}(\mathtt{swap\text{-}pred}, \mathtt{null}; y) \,\widehat{=}\, \mu_{\mathrm{TE}} \qquad (5)$$

One benefit of this definition is that the total effect now inherits the interpretation of odds ratio, in addition to its own causal interpretation. Odds ratio is used to measure the strength of association between a response variable and an intervention. It compares the relative odds of the occurrence of an outcome of interest, given whether a particular intervention is performed (Szumilas, 2010). Therefore, by analogy with the odds ratio, we can interpret $\overline{\mathrm{TE}}$ in three cases: (i) If $\overline{\mathrm{TE}} > 0$, then the presence of ambiguous predicate in S causes higher odds of *non-entailment*; (ii) If $\overline{\mathrm{TE}} = 0$, then there is no causal relationship between the type of the predicate in S and the model prediction; (iii) If $\overline{\mathrm{TE}} < 0$, then the presence of ambiguous predicate in S causes lower odds of *non-entailment*.

Another benefit is that the lexical overlap heuristic is not a problem to our experiment, because $\overline{\mathrm{TE}}$ measures the difference between sentences with swapped and unswapped predicates. If a model completely depends on the heuristic, $\overline{\mathrm{TE}}$ will be close to 0 since the overlap between the premise and hypothesis remains the same when the input is intervened. In this case, no causal relationship is concluded between distributivity and the model prediction. If, however, we obtain a non-zero $\overline{\mathrm{TE}}$, this should be due to factors other than the heuristic.

Since the size of DistNLI is relatively small, we perform the one-sample t-test with a significance level of 0.05 to infer about the average total effect over the full population. If $\overline{\mathrm{TE}}$ is statistically significantly positive, we can conclude that the model has some understanding of distributivity.

5.2 Result and Discussion

Table 3 presents a one-sample t-test of the average total effect for each model. Except for DeBERTa-base, all models have a significantly positive $\overline{\mathrm{TE}}$

Model	Mean	SD	T	P-value
D-b	0.040	1.091	0.468	0.320
D-l	0.314	0.900	4.452	<7e-06
D-xl	0.351	0.507	8.844	<7e-16
D-v2-xl	0.856	0.796	13.724	<2e-29
D-v2-xxl	0.828	1.088	9.724	<3e-18
R-l	0.779	1.279	7.774	<4e-13

Table 3: One Sample t-test of $\overline{\mathrm{TE}}$ for each model. Here, R stands for RoBERTa, D stands for DeBERTa, b stands for base, and l stands for large.
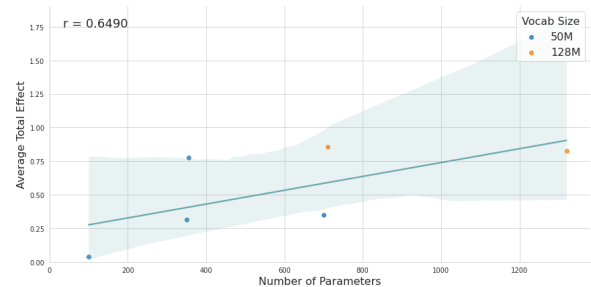

Figure 2: Relationship between $\overline{\mathrm{TE}}$, the number of parameters and vocabulary size. Each point is a model.

with a significant level of 0.05. Based on the interpretation of $\overline{\mathrm{TE}}$, these models are able to discern distributivity to some extent.

Models with more parameters tend to show understanding of distributivity. We find that $\overline{\mathrm{TE}}$ is positively correlated with the number of parameters, as shown by Figure 2 ($r = 0.649$). While there are confounders such as vocabulary size, number of layers, pre-training task, etc., the trend holds when we control for model architecture and only consider DeBERTa variants. This finding may suggest that larger models have a stronger ability to capture linguistic phenomena presented in the training corpus. Vig et al. (2020a) report an analogous result on gender bias. We also observe that the effect on $\overline{\mathrm{TE}}$ vanishes when the number of parameters increases: as shown by Table 3, DeBERTa-large has a $\overline{\mathrm{TE}}$ eight times greater than DeBERTa-base, but merely 0.04 lesser than DeBERTa-xlarge. This observation is in line with the finding of K et al. (2020) that the number of parameters has little effect on model performance after a certain threshold.

Models with richer vocabulary tend to show understanding of distributivity. We find that $\overline{\mathrm{TE}}$ is associated with the size of vocabulary. As illustrated in Figure 2, DeBERTa-v2 variants have
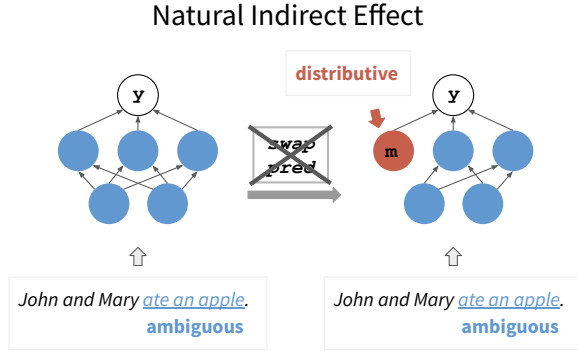
Natural Indirect Effect

Figure 3: Natural indirect effect measures the relative change in y, given the presence of the input intervention, if every path into the mediator m is blocked by setting m to the value it would have been without the input intervention.

a $\overline{\text{TE}}$ of around 0.8, which is considerably larger than other models. We suspect this significant increase is due to their larger vocabulary size. Since distributivity is determined by the type of predicates, a richer vocabulary is expected to lead to a better semantic representation of predicates which in turn boosts $\overline{\text{TE}}$, although the effects of other confounders might not be ruled out.

6 Natural Indirect Effect

6.1 Experiment Design

Neuron Intervention To study the causal contribution of neurons, the hypothesized mediator in our experiment will be a single neuron or a group of neurons. Rather than investigating neurons for each input token independently (Vig et al., 2020a,b; Finlayson et al., 2021), we intervene on neurons for all input tokens simultaneously. This approach is computationally cheaper but still comprehensive enough to give a full picture of the underlying causal mechanism.[2]

Metric Natural Indirect Effect is used to measure how much the response variable y would change with the swap-pred operator applied, if we set the hypothesized mediator m to the value it would have been without rather than with the input intervention (demonstrated in Figure 3). Similar to TE, we use the log odds ratio and estimate the popula-

tion average natural indirect effect:

$$
\begin{aligned}
&\text{NIE}(\texttt{swap-pred}, \texttt{null}; y, m, S_p) \\
&= \log \left(\frac{y_{\texttt{swap-pred}}(S_p)}{y_{\texttt{swap-pred},m_{\texttt{null}}}(S_p)} \right) \quad (6)
\end{aligned}
$$

where m is a hypothesized mediator and $m_{\texttt{null}}$ means that m is set to the value it would have been in the absence of the input intervention. We can also define NDE in a similar way:

$$
\begin{aligned}
&\text{NDE}(\texttt{swap-pred}, \texttt{null}; y, m, S_p) \\
&= \log \left(\frac{y_{\texttt{swap-pred},m_{\texttt{null}}}(S_p)}{y_{\texttt{null}}(S_p)} \right) \quad (7)
\end{aligned}
$$

VanderWeele and Vansteelandt (2010) prove that the log odds ratio definition of causal effects holds a decomposition property: TE = NIE + NDE even when there are interactions and nonlinearities.

The NIE and NDE defined above are in principle an implementation of what Robins and Greenland (1992) refer to as "total indirect effect" and "pure direct effect". The NDE given by Vig et al. (2020a,b) follows the same idea, but their NIE instead formulates the "pure indirect effect". A consequence is that the decomposition property is only guaranteed for linear models (Pearl, 2001; Vig et al., 2020a), which is a potential shortcoming as an analysis method for neural networks.

Given the decomposition property, $\overline{\text{NIE}}$ allows us to measure the magnitude of causal contribution a model component makes to the model behavior, which is quantified by $\overline{\text{TE}}$. In this respect, it potentially solves the problem of spurious correlation between the probing classifier and the original model (Belinkov, 2021). In our experiment, we use it to verify the causal relationship between the semantic information encoded in the original model and the prediction given by the NLI classifier. We can interpret the values of $\overline{\text{NIE}}$ similarly to $\overline{\text{TE}}$.

6.2 Result and Discussion

We experiment on models which pass the significance threshold.[3] Figure 4 illustrates the neuronwise $\overline{\text{NIE}}$: for all models, most neurons have $\overline{\text{NIE}}$s around zero, but a few outliers can also be identified. In order to determine which neurons are responsible the most for the model behaviour, we also select the top 1% of neurons with highest individual $\overline{\text{NIE}}$s from each layer and evaluate the layer-wise $\overline{\text{NIE}}$. Figure 5 illustrates the layer-wise $\overline{\text{NIE}}$ obtained from selected neurons.

---

[2]Pilot experiments focusing on the `[CLS]` token, an approach used by Dobreva and Keller (2021), yielded mixed results, which is consistent with the findings of Reimers and Gurevych (2019) that the `[CLS]` token is not an ideal representation of sentence level meaning.

[3]Due to limited computational resources, DeBERTa-v2-xxlarge is excluded.

(a) DeBERTa-large

(b) DeBERTa-xlarge
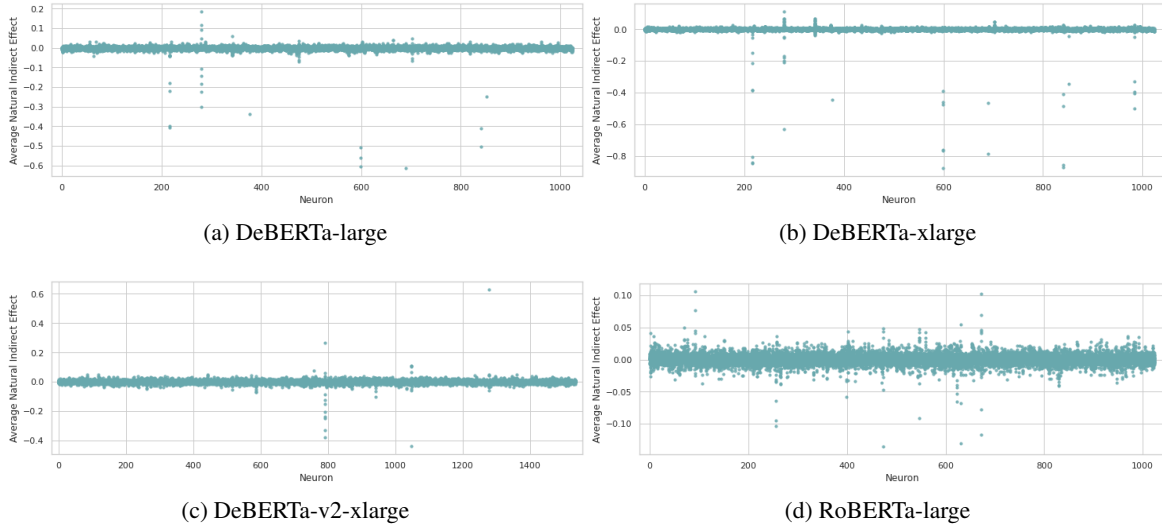
(c) DeBERTa-v2-xlarge

(d) RoBERTa-large

Figure 4: Neuron-wise $\overline{\text{NIE}}$ of the models that pass the $\overline{\text{TE}}$ threshold. The *x*-axis represents the indices of neurons, which range from 0 to hidden size. The indices are not unique: neurons from different layers have the same index.
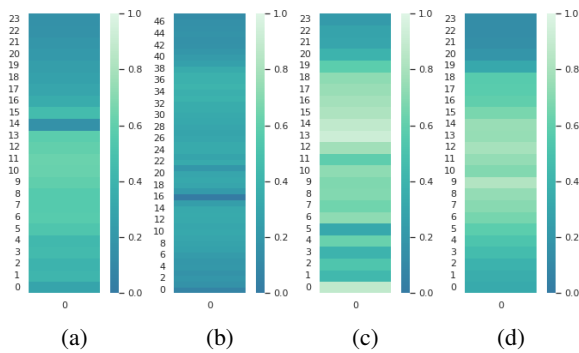

(a) (b) (c) (d)

Figure 5: Layer-wise $\overline{\text{NIE}}$ (top 1% neurons) of the models that pass the $\overline{\text{TE}}$ threshold. From left to right are DeBERTa-large, DeBERTa-xlarge, DeBERTa-v2-xlarge and RoBERTa-large. The y-axis represents layers and the color represents values of the $\overline{\text{NIE}}$.

Knowledge of distributivity is mostly concentrated in middle layers. We define the depth of a layer as its number divided by the total number of layers. Based on this metric, we can divide layers into three groups: early (0 - 0.33), middle (0.33 - 0.67), and final (0.67 - 1). In Figure 5, a concentration pattern of $\overline{\text{NIE}}$ is clearly shown by the color opacity of layers for all models. Specifically, middle layers have higher $\overline{\text{NIEs}}$ than other layers. The exact layer where $\overline{\text{NIE}}$ peaks occur is more idiosyncratic, but still inside or near middle layers: 0.5 (layer #11) for DeBERTa-large, 0.73 (layer #35) for DeBERTa-xlarge, 0.54 (layer #13) for DeBERTa-v2-xlarge, and 0.42 (layer #9) for RoBERTa-large. This finding differs from the conclusion of Tenney

et al. (2019) that semantic information is hardly localized in BERT-like models, although Jawahar et al. (2019) also report that most semantic tasks archive the best performance around middle layers.

Knowledge of distributivity is more concentrated in the models with higher degree of understanding. We find that the level of concentration of $\overline{\text{NIE}}$ patterns with the magnitude of $\overline{\text{TE}}$: $\overline{\text{NIEs}}$ are concentrated in fewer neurons in DeBERTa-v2-xlarge and RoBERTa-large (both have a $\overline{\text{TE}}$ of about 0.8) than in DeBERTa-large and DeBERTa-xlarge (both have a $\overline{\text{TE}}$ of about 0.3). This finding is supported by the following observations: First, as shown by Figure 5, top 1% of neurons is sufficient to achieve the full total effect for the former two models, but not for the latter two models. Second, we notice that a few neurons have extremely higher $\overline{\text{NIE}}$ for the former two models. For example, neuron #1279 at layer #0 (located at the top right of Figure 4c) in DeBERTa-v2-xlarge has a $\overline{\text{NIE}}$ of 0.6735, much higher than most other neurons. According to the interpretation of $\overline{\text{NIE}}$, these neurons are causally and positively responsible for the model behaviour. The pattern is not observed in DeBERTa-large and DeBERTa-xlarge.

7 Conclusion

In this paper, we propose DistNLI, a diagnostic NLI dataset to examine to what extent pre-trained language models can discern the phenomenon of distributivity. By extending the CMA framework,

we show that models including DeBERTa and RoBERTa have some understanding of distributivity, which provides further evidence that models have ability to encode high-level semantic knowledge, and reveal some interesting patterns related to the underlying mechanism of these models.

One direction for future improvement would be increasing the diversity of predicates and subjects in DistNLI. At present, we only look at subjects that are two coordinated DPs, but the phenomenon of distributivity applies to all noun phrases which denote groups and even without the utilization of conjuncts. It is possible that pre-trained language models can also differentiate more complicated combinations, as they are trained on large-scale text data. Another direction would be investigating how robust the CMA framework is to the definition of metrics, such as an empirical comparison between alternative definitions.

8 Limitation

Due to the specificity of the linguistic phenomenon involved and its size, this DistNLI dataset should only be used as a diagnostic dataset in the investigation of distributivity of verb predication. Also, occasionally some minimal pairs in the dataset could contradict with the world knowledge considering the nature of artificiality. On the one hand, the creators of this dataset have filtered out pairs that are tremendously deviant from the world knowledge by majority voting. On the other hand, even if there is still any deviating pair against the common-sense (i.e. *The lion and the seal found a habitat*), the distributivity manifested in such examples will not be confounded as long as the grammaticality is guaranteed, since the extent of deviance is constant between the premise and the hypothesis.

References

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48:207–219.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015a. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Samuel R. Bowman, Christopher Potts, and Christopher D. Manning. 2015b. Recursive neural networks can learn logical semantics. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 12–21, Beijing, China. Association for Computational Linguistics.

Lucas Champollion. 2017. *Parts of a whole: Distributivity as a bridge between aspect and measurement*, volume 66. Oxford University Press.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Elizabeth Coppock and Lucas Champollion. 2019. Invitation to formal semantics. *Manuscript, Boston University and New York University. eecoppock. info/semantics-boot-camp. pdf*.

Hanna De Vries. 2017. Two kinds of distributivity. *Natural Language Semantics*, 25(2):173–197.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Radina Dobreva and Frank Keller. 2021. Investigating negation in pre-trained vision-and-language models. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 350–362, Punta Cana, Dominican Republic. Association for Computational Linguistics.

David Dowty et al. 1987. Collective predicates, distributive predicates, and all. In *Proceedings of the 3rd ESCOL*, pages 97–115. (Eastern States Conference on Linguistics), Ohio State University Ohio.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models.

In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843, Online. Association for Computational Linguistics.

Reza Ghaeini, Sadid A. Hasan, Vivek Datla, Joey Liu, Kathy Lee, Ashequl Qadir, Yuan Ling, Aaditya Prakash, Xiaoli Fern, and Oladimeji Farri. 2018. DR-BiLSTM: Dependent reading bidirectional LSTM for natural language inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1460–1469, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Jack Hoeksema. 1988. The semantics of non-boolean "and". *Journal of Semantics*, 6(1):19–40.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Sullam Jeoung and Jana Diesner. 2022. What changed? investigating debiasing methods using causal mediation analysis. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 255–265, Seattle, Washington. Association for Computational Linguistics.

Jaap Jumelet, Milica Denic, Jakub Szymanik, Dieuwke Hupkes, and Shane Steinert-Threlkeld. 2021. Language models use monotonicity to assess NPI licensing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4958–4969, Online. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Anthony S Kroch. 1974. *The semantics of scope in English*. Ph.D. thesis, Massachusetts Institute of Technology.

Peter Lasersohn. 1995. Plurality, conjunction, and events, vol. 55 of. *Studies in Linguistics and Philosophy*.

Godehard Link et al. 2002. The logical analysis of plurals and mass terms: A lattice-theoretical approach. *Formal semantics: The essential readings*, pages 127–146.

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

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Judea Pearl. 2001. Direct and indirect effects. *Probabilistic and Causal Inference*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Craige Roberts. 1987. *Modal subordination, anaphora, and distributivity*. Ph.D. thesis, University of Massachusetts Amherst.

James M Robins and Sander Greenland. 1992. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, pages 143–155.

Ken Safir and Tim Stowell. 1987. Binominal each. *North East Linguistics Society*, 18(3).

Swarnadeep Saha, Yixin Nie, and Mohit Bansal. 2020. ConjNLI: Natural language inference over conjunctive sentences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8240–8252, Online. Association for Computational Linguistics.

Remko Scha. 1981. Collective, distributive and cumulative quantification. *Groenendijk, JAG; Janssen, TMV; and Stokhof, MBJ, editors*, pages 483–512.

Magdalena Szumilas. 2010. Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie canadienne de psychiatrie de l'enfant et de l'adolescent*, 19 3:227–9.

Alison Taub. 1989. Collective predicates, aktionsarten and all. *University of Massachusetts Occasional Papers in Linguistics*, 15(2):16.

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. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

323

Tyler J. VanderWeele and Stijn Vansteelandt. 2010. Odds ratios for mediation analysis for a dichotomous outcome. *American journal of epidemiology*, 172 12:1339–48.

Henk Verkuyl. 1993. *Distributivity and collectivity: a couple at odds*. Research Institute for Language and Speech. Utrech University.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020a. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *ArXiv*, abs/2004.12265.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020b. Investigating gender bias in language models using causal mediation analysis. In *NeurIPS*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yoad Winter. 1997. Choice functions and the scopal semantics of indefinites. *Linguistics and philosophy*, pages 399–467.