Decomposing Natural Logic Inferences for Neural NLI

Julia Rozanova[1], **Deborah Ferreira**[1], **Mokanarangan Thayaparan**[1], **Marco Valentino**[1] and **André Freitas**[1,2]
Department of Computer Science, University of Manchester, United Kingdom[1]
Idiap Research Institute, Switzerland[2]
{firstname.lastname}@manchester.ac.uk

Abstract

In the interest of interpreting neural NLI models and their reasoning strategies, we carry out a systematic probing study which investigates whether these models capture the crucial semantic features central to natural logic: *monotonicity* and *concept inclusion*. Correctly identifying valid inferences in *downward-monotone contexts* is a known stumbling block for NLI performance, subsuming linguistic phenomena such as negation scope and generalized quantifiers. To understand this difficulty, we emphasize monotonicity as a property of a *context* and examine the extent to which models capture relevant monotonicity information in the vector representations which are intermediate to their decision making process. Drawing on the recent advancement of the probing paradigm, we compare the presence of monotonicity features across various models. We find that monotonicity information is notably weak in the representations of popular NLI models which achieve high scores on benchmarks, and observe that previous improvements to these models based on fine-tuning strategies have introduced stronger monotonicity features together with their improved performance on challenge sets.

1 Introduction

Large, black box neural models which achieve high scores on benchmark datasets designed for testing *natural language understanding* are the subject of much scrutiny and investigation.

It is often investigated whether models are able to capture specific semantic phenomena which mimic human reasoning and/or logical formalism, as there is evidence that they sometimes exploit simple heuristics and dataset artifacts instead (McCoy et al., 2019; Herlihy and Rudinger, 2021).

In this work, we consider the rigorous setting of *natural logic* (MacCartney and Manning, 2007). This is a highly systematic reasoning principle relying on only two abstract features, each of which is in itself linguistically complex: *monotonicity* and *concept inclusion word-pair relations*. It underlies the majority of symbolic/rule-based and hybrid approaches to NLI and is an important baseline reasoning phenomenon to look for in a robust and principled NLI model.

Downward monotone operators such as negation markers and generalized quantifiers result in the kinds of natural logic inferences which are often known to stump neural NLI models that demonstrate high performance on large benchmark sets such as MNLI (Williams et al., 2018): this has been identified in behavioural studies based on targeted challenge test sets, such as in Yanaka et al. (2019a) and Geiger et al. (2020).

In this work, however, we present a **structural** study: we investigate the extent to which the features relevant for identifying natural logic inferences, especially context monotonicty itself, are captured in the model's internal representations. To this end, we carry out a systematic *probing* study.

Our contributions are may be summarized as follows:

1. We perform a structural investigation as to whether the behaviour of *natural logic* formalisms are mimicked within popular transformer-based NLI models.

2. For this purpose, we present a joint NLI and semantic probing dataset format (and dataset) which we call NLI-XY: it is a unique probing dataset in that the probed features relate to the NLI task output in a very systematic way.

3. We employ thorough probing techniques to determine whether the abstract semantic features of *context monotonicity* and *concept inclusion relations* are captured in the models' internal representations.

4. We observe that some well-known NLI models demonstrate a systematic failure to model

Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 394 - 403
December 8, 2022 ©2022 Association for Computational Linguistics

context monotonicity, a behaviour we observe to correspond to poor performance on natural logic reasoning in downward-monotone contexts. However, we show that the existing HELP dataset improves this behaviour.

5. We support the observations in the probing study with several *qualitative analyses*, including decomposed error-breakdowns on the **NLI-XY** dataset, representation visualizations, and evaluations on existing challenge sets.

2 Related Work

Natural logic dates back to the formalisms of Sanchez (1991), but has been received more recent treatments and reformulations in MacCartney and Manning (2007) and Hu and Moss (2018). Symbolic and hybrid neuro-symbolic implementations of the natural logic paradigm have been explored in Chen et al. (2021); Kalouli et al. (2020); Abzianidze (2017) and Hu et al. (2020).

The shortcomings of natural logic handling in various neural NLI models have been shown with several *behavioural* studies, where NLI challenge sets exhibiting examples of downward monotone reasoning are used to evaluate performance of models with respect to these reasoning patterns (Richardson et al., 2019; Yanaka et al., 2019b,a; Goodwin et al., 2020; Geiger et al., 2020).

In an attempt to better identify linguistic features that neural models manage or fail to capture, researchers have employed *probing* strategies: namely, the *diagnostic classification* (Alain and Bengio, 2018) of auxiliary feature labels from internal model representations. Most probing studies in natural language processing focus on the *syntactic* features captured in transformer-based language models (Hewitt and Manning, 2019), but calls have been made for more sophisticated probing tasks which rely more on contextual information (Pimentel et al., 2020).

In the realm of semantics, probing studies have focused more on *lexical* semantics (Vulić et al., 2020): word pair relations are central to monotonicity reasoning, and thus form part of our probing study as well, but the novelty of our work is the task of classifying context monotonicity from intermediate contextual embeddings.

3 Problem Formulation: Decomposing Natural Logic

Natural logic inferences (as formalized in Sanchez 1991; MacCartney and Manning 2007) are usually described with respect to *substitution* operations. Certain word substitutions result in either forward or reverse entailment, while others result in neither. This is the basis for a calculus of determining entailment from substitution sequences (MacCartney and Manning, 2007; Hu et al., 2020; Hu and Moss, 2018).

Broadly speaking, we wish to determine whether well-known transformer-based NLI models mimic the reasoning strategies of natural logic. However, as neural NLI models are black box classifiers that only see a premise/hypothesis sentence pair as its input, it is not immediate how to compare its process to a rule-based system.

To this end, we consider a formulation of natural logic which describes its rules in terms of concept pair relations and *context monotonicity* (similar to Rozanova et al. 2021).

3.1 Inferences From Concepts and Contexts

Consider the following example of a single step natural logic inference, which we will decompose into semantic components relevant to its entailment label:

		NLI Label
Premise	I did not eat any **fruit** for breakfast.	Entailment
Hypothesis	I did not eat any **raspberries** for breakfast.	

The hyponym/hypernym pair (raspberries, fruit) exemplifies a more general relation which we will refer to as the *concept inclusion* relation $\sqsubset$, (and dually, *reverse concept inclusion* $\sqsupset$). This mimics the subset relation of the set-based interpretations of the predicates *raspberry* and *fruit*.

In the above example, they occur in a shared **context**, namely the sentence template

"I did not eat any _____ for breakfast".

Such a context may be treated as a term substitution function f

$$f : (\mathcal{X}, \sqsubseteq) \rightarrow (\mathcal{S}, \Rightarrow)$$

between a set of concepts $\mathcal{X}$ (ordered by the concept inclusion relation) and the set $\mathcal{S}$ of full sentences ordered by entailment - we demonstrate this substitution in table 1.

X	$f(X)$
raspberries	I did not eat any **raspberries** for breakfast
fruit	I did not eat any **fruit** for breakfast

Table 1: In the example, the substitution function f behaves on the concept inputs as shown.

3.2 Context Monotonicity

We say that f is *upward monotone* ($\uparrow$) if it is order *preserving*, i.e.

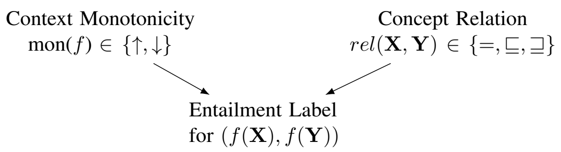
$$\forall_{X,Y}(X \sqsubseteq Y \text{ implies } f(X) \Rightarrow f(Y))$$

and that f is *downward monotone* ($\downarrow$) if it is order *reversing*, i.e.

$$\forall_{X,Y}(X \sqsupseteq Y \text{ implies } f(X) \Rightarrow f(Y)).$$

Given a natural language context f, any pair of grammatically valid insertions (X, Y) (e.g. ("raspberries", "fruit")) yields a sentence pair $(f(X), f(Y))$. Treating $f(X)$ as a *premise* sentence and $f(Y)$ as a *hypothesis* sentence, a trained neural NLI model can provide a classification of whether $f(X)$ entails $f(Y)$.

In summary, these two abstract linguistic features, *context montonicity* and *concept inclusion relation*, jointly determine the final gold entailment label of this type of NLI example.



4 NLI-XY Dataset

We follow this formalism as the basis for the **NLI-**XY dataset. This is the first probing dataset in NLP where the auxiliary labels for intermediate semantic features influence the final task label in a rigid and deterministic (yet simple) way, with these features being themselves linguistically complex. As such, it is a "decomposed" natural logic dataset, where the positive entailment labels are further enriched with labels for the monotonicity and relational properties which gave rise to them. This allows for informative qualitative and structural analyses into natural logic handling strategies in neural NLI models.

The NLI-XY dataset is comprised of the following:

			Auxilliary Label
Context	f	I did not eat any ▁ for breakfast.	$\downarrow$ (downward monotone)
Insertion Pair	(X,Y)	(fruit, raspberries)	$\sqsupset$ (reverse concept inclusion)
			NLI Label
Premise	$f(X)$	I did not eat any fruit for breakfast.	Entailment
Hypothesis	$f(Y)$	I did not eat any raspberries for breakfast.	

Table 2: A typical NLI-XY example with labels for context monotonicity, lexical relation and the final entailment label.

1. A set of *contexts* f with a blank position (indicated with a lower case 'x' or an underscore), annotated with the context monotonicity label.

2. A set of *insertion pairs* (X, Y), which are either nouns or noun phrases, annotated with the concept inclusion word-pair relation.

3. A derived set of premise and hypothesis pairs $(f(X), f(Y)$ made up of permutations of (X, Y) insertion pairs through contexts f, controlled for grammaticality as far as possible.

We present examples of the component parts and their composition in table 2.The premise/hypothesis pairs may thus be used as input to any NLI model, while the context monotonicity and insertion relation information can be used as the targets of an auxiliary probing task on top of the model's representations.

We make the NLI-XY dataset and all the experimental code used in this work is publically available [1]. We constructed the NLI-XY dataset used here as follows:

Context Extraction We extract context examples from two NLI datasets which were designed for the behavioural analysis of NLI model performance on monotonicity reasoning. In particular, we use the manually curated evaluation set MED (Yanaka et al., 2019a) and the automatically generated HELP training set (Yanaka et al., 2019b). By design, as they are collections of NLI examples exhibiting monotonicity reasoning, these datasets mostly follow our required $(f(X), f(Y))$ structure, and are labeled as instances of upward or downward monotonicity reasoning (although the contexts are not explicitly identified).

[1] Anonymized github link.

We extract the common context f from these examples after manually removing a few which do not follow this structure (differing, for example, in pronoun number agreement or prepositional phrases). We choose to treat determiners and quantifiers as part of the context, as these are the kinds of closed-class linguistic operators whose monotonicity profiles we are interested in. To ensure grammatically valid insertions, we manually identify whether each context as suitable either for a singular noun, mass noun or plural noun in the blank/“x" position.

Insertion Pairs Our (X, Y) insertion phrase pairs come from two sources: Firstly, the labeled word pairs from the MoNLI dataset (Geiger et al., 2020), which features only single-word noun phrases. Secondly, we include an additional hand-curated dataset which has a small number of *phrase-pair* examples, which includes intersective modifiers (e.g. ("brown sugar", "sugar")) and prepositional phrases (e.g. ("sentence", "sentence about oranges")). Several of these examples were drawn from the MED dataset. Each word in the pair is labelled as a singular, plural or mass noun, so that they may be permuted through the contexts in a reasonably grammatical way.

Premise/Hypothesis Pairs Premise/Hypothesis pairs are constructed by permuting insertion pairs through the set of contexts within the grammatical constraints. Such a permutation strategy may generate examples which are not consistently *meaningful*, but we see the monotonicity reasoning pattern as sufficiently rigid and syntactic that it is of interest to observe how models treat less "meaningful" entailment examples that still hold with respect to the natural logic formalism: for example, "I did not swim in a person” entails "I did not swim in an Irishman" at a systematic level. This does raise a question of whether we do (or even should) observe certain systematic behaviours on out-of-distribution examples: we leave the further investigation of this matter for future work.

Lastly, we note that the data is split into train, dev and test partitions *before* this permutation occurs, so that there are **no shared contexts or insertion pairs** between the different data partitions, in an attempt to avoid overlap issues such as those discussed in (Lewis et al., 2021). The full dataset statistics are reported in table 3.

		Context Monotonicity		
Partition	**(X,Y) Relation**	Up $\uparrow$	Down $\downarrow$	Total
train	$\sqsubseteq$	671	543	1214
	$\sqsupseteq$	671	543	1214
	None	244	222	466
	Total	1586	1308	**2894**
dev	$\sqsubseteq$	598	389	987
	$\sqsupseteq$	598	389	987
	None	220	242	462
	Total	1416	1020	**2436**
test	$\sqsubseteq$	1103	1066	2169
	$\sqsupseteq$	1103	1066	2169
	None	502	516	1018
	Total	2708	2648	**5356**

Table 3: Dataset statistics for the NLI-XY dataset. We employ an **aggressive** $30, 20, 50$ train-dev-test split for a more impactful probing result, as probing is meant to demonstrate the *ease of extraction* of features. In particular, higher test accuracy with a smaller training set is a more convincing probing result than one with a large training set and small test set.

5 Experimental Setup

Our experiments are designed to investigate the following questions: Firstly, how do NLI models compare in their learned encoding of context monotonicity and lexical relational features? Secondly, if a model successfully captures these features, to what extent do they correspond with the model's predicted entailment label? We investigate these questions with a detailed probing study and a supporting qualitative analysis, using decomposed error break-downs and representation visualization.

5.1 Model Choices

We consider a selection of neural NLI models based on BERT-like transformer language models (such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and BART (Lewis et al., 2020)) which are fine-tuned on one of two benchmark training sets: either SNLI (Bowman et al., 2015) or MNLI (Williams et al., 2018). Of particular interest, however, is the case where these models are trained on an additional dataset (the HELP dataset from (Yanaka et al., 2019b)) which was designed for improving the overall balance of upward and downward monotone contexts in NLI training data. We use our own random $50-30-20$ train-dev-test split of the HELP dataset (ensuring unique contexts in every split), so that there is no overlap of contexts between the fine-tuning data and the few HELP-test examples we used as part of our NLI-XY dataset[2].

5.2 Probing Tasks

The NLI-XY dataset is equipped with two auxiliary feature labels which are the targets of the probing task: context monotonicity and the relation of the (X, Y) word pair (referred to as concept inclusion relation or lexical relation). We now describe the details of the intermediate representations we choose as inputs to the probing tasks:

Target Representation The standard practice for word-pair relation classification tasks is to concatenate the contextual representation vectors for the (X, Y) word pair (taking the mean vector for multi-token words). We argue that this is a good representation choice for probing context monotonicity as well: as we are considering transformer-based bidirectional encoder architectures, the context (including the order) of each token in the input sequence informs the representation of each token in the final layer. As such, we propose that since contextual information is implicitly encoded, it is feasible to expect that a token's vector representation may encode contextual features such as context monotonicity. As both the X and the Y word occur in the same respective context, we are comfortable probing the concatenated (X, Y) representation for contextual features, and note that it allows for easy comparison with the word pair relation probing results.

Probing Methodology For each auxiliary classification task, we use simple linear models as probes. We train 20 probes of varying complexities using the *probe-ably* framework (Ferreira et al., 2021).

The complexities are represented and controlled as follows: For linear models, we follow Pimentel et al. (2020) in using the nuclear norm of the linear transformation matrix as the approximate measure of complexity, as it is a continuous approximation of the transformation matrix rank. Naively, a strong accuracy on the probing test set may be understood to indicate strong presence of the target features within the learned representations, but there has been much discussion about whether this evidence is compelling on its own. In fact, certain probing experiments have found the same accuracy scores for random representations (Zhang and Bowman, 2018), indicating that high accuracy scores are meaningless in isolation. Hewitt and Liang (2019) describe this as a dichotomy between the representation's encoding of the target features and the probe's capacity for *memorization*, and propose the use of the *selectivity* measure to always place the probe accuracy in the context of a controlled probing task with shuffled labels on the same vector representations. For each fully trained probe, we report both the test accuracy and the *selectivity* measure: tracking the selectivity ensures that we are not using a probe that is complex enough to be *overly expressive* to the point of having the capacity to overfit the randomised control training set. The *selectivity* score is calculated with respect to a *control task*. At its core, this is just a balanced random relabelling of the auxiliary data, but Hewitt and Liang (2019) advocate for more targeted control tasks with respect to the features in question and a hypothesis about the model's possible capacity for *memorization*. For the lexical relation classification control task, we assign a shared random label for all identical insertion pairs, regardless of context. Thus, a probe which is expressive enough to "memorize" individual labels of word pairs would attain high accuracy on this control task. Analogously, the context monotonicity classification control tasks assigns shared random labels to identical contexts.

5.3 NLI Challenge Set Evaluations

As well as the NLI-XY dataset (which can function as an ordinary NLI evaluation set), for completeness we report NLI task evaluation scores on the full MED dataset (Yanaka et al., 2019a), which was designed as a thorough stress-test of monotonicity reasoning performance. Furthermore, we report scores on the HELP-test set (from the dataset split in Rozanova et al. 2021): this data partition was not used in the fine-tuning of models on HELP, but we include the test scores here for insight.

5.4 Decomposed Error Analysis

The compositional structure and auxiliary labels in the NLI-XY dataset allow for qualitative analysis which may enrich the observations. To this end, we construct decomposed error analysis heatmaps which indicate whether a given premise-hypothesis data point $(f(X), f(Y))$ is correctly classified by an entailment model. These are structured with individual (X, Y) insertion pairs on the vertical axis and contexts on the horizontal axis. For brevity (and because this is representative of our observa-

[2] We use the *transformers* library (Wolf et al., 2020) and their available pretrained models for this work.

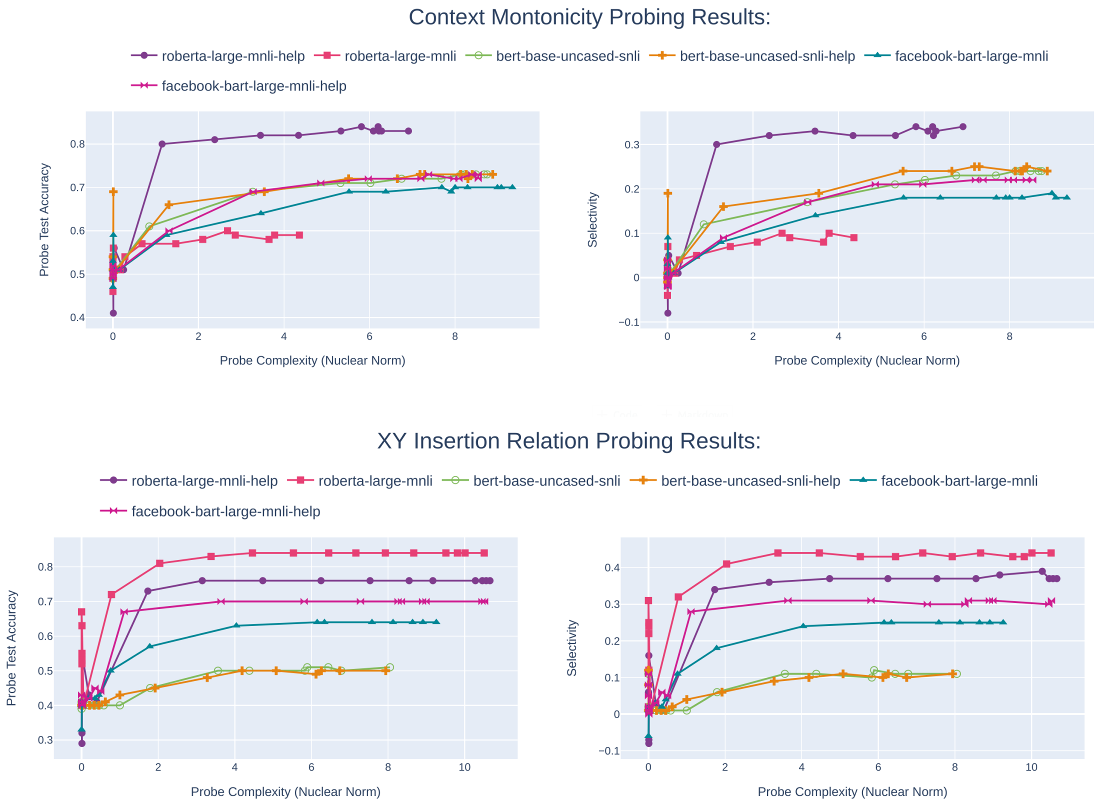


Figure 1: Linear probing results for all examined models.

tions), we include only the error breakdowns for the two sublasses of the positive entailment label: where the context monotonicity is upward and lexical relation is forward incusion, and where the context monotonicity is downward and the lexical relation is reverse inclusion.

6 Results and Discussion

6.1 Probing Results

The results for the linear probing experiments for both the *context monotonicity classification* task and the *lexical relation* classification task may be found in figure 1, with a summary score of accuracy at maximum selectivity visible in table 4. The results of the control tasks are taken into account as part of the selectivity measure, which is represented on the right hand plot for each experiment.

It is particularly notable that large models trained only on the MNLI dataset have inferior performance on context monotonicity classification. This corresponds with the further qualitative observations, suggesting that even in some of the most successful transformer-based NLI models, *there is a poor "understanding" of the logical regularities of contexts and how these are altered with downward monotone operators.*

6.2 Comparison to Challenge Set Performance

A summary of the probing results (presented as accuracy at maximum selectivity) can be compared with challenge set performance in table 4. Evaluation on the challenge test sets is relatively consistent with monotonicity probing performance, in the sense that there is a correspondence between poor/successful modeling of monotonicity features and poor/successful performance on a targeted natural logic test set. As these challenge sets are focused on testing monotonicity reasoning, this is a result which strongly bolsters the suggestion that explicit representation of the context monotonicity feature is crucial, especially for examples involving negation and other downward monotone operators. Furthermore, we generally confirm previous results that additional fine-tuning on the HELP data set has been helpful for these specialized test sets, and add to this that it similarly improves the explicit extractability of relevant context montonicity features from the latent vector representations.

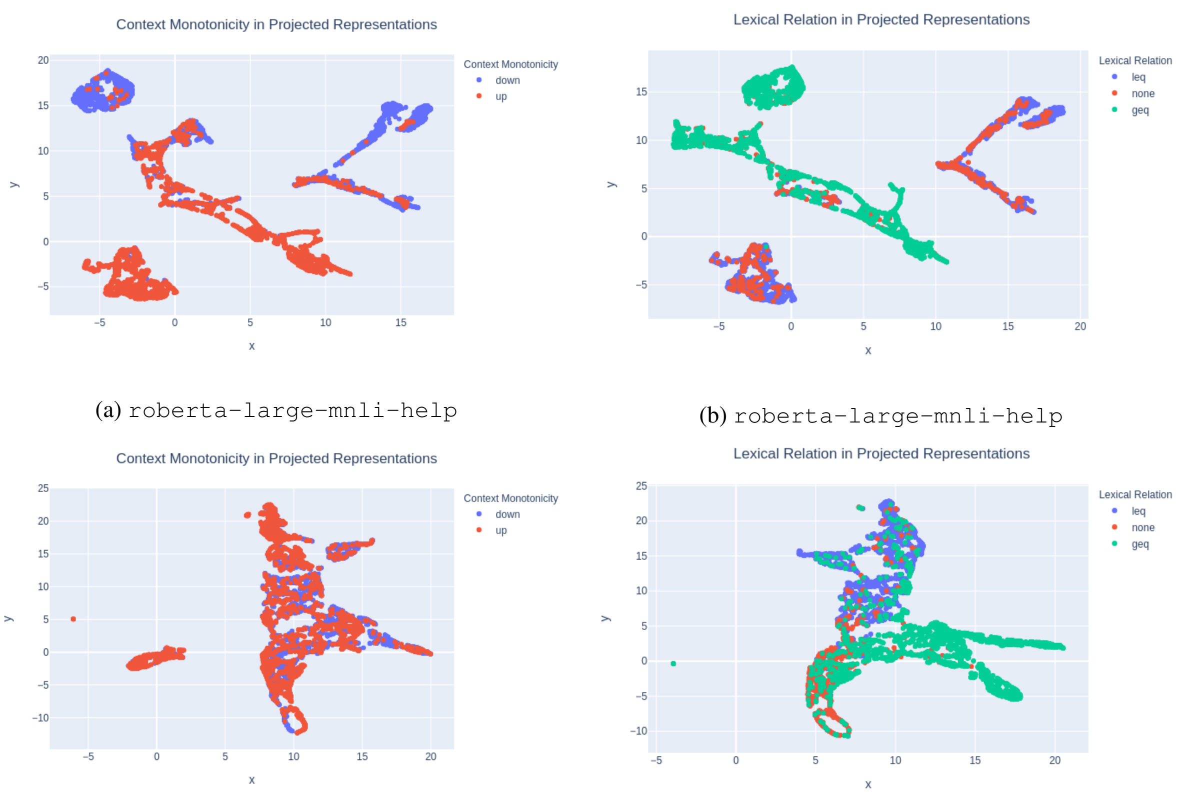
NLI Models	Fine-Tuning Data	Feature Probing: Context Monotonicity (%*)	Feature Probing: XY Insertion Relation (%*)	NLI Monotonicity Challenge Sets: HELP-Test (%)	NLI Monotonicity Challenge Sets: MED (%)	NLI Monotonicity Challenge Sets: NLI-XY (%)
roberta-large-mnli	-	59.00	84.00	36.69	46.10	59.01
roberta-large-mnli	HELP	84.00	76.00	**97.63**	**78.22**	**80.68**
facebook/bart-large-mnli		70.00	64.00	43.61	46.54	60.59
facebook/bart-large-mnli	HELP	73.00	70.00	88.99	77.16	79.34
bert-base-uncased-snli		73.00	51.00	63.55	49.38	49.09
bert-base-uncased-snli	HELP	73.00	51.00	66.80	46.13	44.79

Table 4: Summary NLI challenge test set and probing results for all considered models. *Probing results are summarized with the *accuracy at max selectivity*.

Figure 2: Decomposed error heat maps for portions of the NLI-XY dataset corresponding to the indicated context monotonicity and insertion relations (blank positions are present as only grammatical insertions were included in the dataset.)

6.3 Qualitative Analyses

Error Break-Downs An error heat map according to decomposed context monotonicity and word-pair insertion relation can be seen in figure 2. We are less concerned with the accuracy score (on NLI challenge sets) of a given model as with the behavioural *systematicity* visible in the errors, as we are not interested in noisy errors which may be due to words or phrases from outside the training domain. Consistent mis-classification for all examples derived from a fixed context or insertion pair are actually *also* strongly suggestive of a regularity in reasoning. The decomposed error analyses paint a striking picture: we generally see that models trained on MNLI routinely fail to distinguish between the expected behaviour of upward and downward monotone contexts, despite generally achieving high accuracies on large benchmark sets. This is in accordance with observations in Yanaka



(a) `roberta-large-mnli-help` (b) `roberta-large-mnli-help`

(c) `roberta-large-mnli` (d) `roberta-large-mnli`

Figure 3: UMAP projections of selected classification token representations comparing `roberta-large-mnli` and the improved `roberta-large-mnli-help`, which shows greater distinction between context monotonicity features.

et al. (2019b) and Yanaka et al. (2019a), where low accuracy on the downward-monotone reasoning sections of challenge sets points to this possibility. Howver, they show consistently show strong behavioural regularity with respect to concept inclusion. Even when the contexts are downward monotone, they still treat them systematically as if they were *upward* monotone, echoing the concept insertion pair relation *only*: they completely fail to discriminate between upward/downward monotone contexts and their opposite behaviours.

Visualization In figure 3, each data point corresponds to an embedded example (contextual XY word pair representation) in the NLI-XY dataset, with the left and right columns colored with the *gold* auxiliary labels for context monotonicity and concept inclusion relations respectively. These illustrate the probing observations: in the well-known `roberta-large-mnli` model, concept inclusion relation features are distinguishable, whereas context monotonicity is very randomly scattered, with no emergent clustering. However, the `roberta-large-mnli-help` model shows an improvement in this behaviour, demonstrating a stronger context monotonicity distinction.

7 Conclusion

In summary, the NLI-XY has enabled us to present evidence that explicit context monotonicity feature clustering in neural model representations seems to correspond to better performance on natural logic challenge sets which test downward-monotone reasoning. In particular, many popular models trained on MNLI seem to lack this behaviour, accounting for previous observations that they systematically fail in downward-monotone contexts.

Furthermore, the probes' labels also have some explanatory value: both entailment and non-entailment labels can each further be broken down into sub-regions. This qualifies the classification with the observations that the data point occurs in a cluster of examples with a) upward (resp. downward) contexts and b) a forward (resp. backward) containment relation between the substituted noun phrases. In this sense, the analyses in this work can thus be interpreted as an explainable "decomposition" of the treatment natural logic examples in neural models.

References

Lasha Abzianidze. 2017. LangPro: Natural language theorem prover. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 115–120, Copenhagen, Denmark. Association for Computational Linguistics.

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Zeming Chen, Qiyue Gao, and Lawrence S. Moss. 2021. Neurallog: Natural language inference with joint neural and logical reasoning.

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.

Deborah Ferreira, Julia Rozanova, Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2021. Does my representation capture X? probe-ably. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 194–201, Online. Association for Computational Linguistics.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 163–173, Online. Association for Computational Linguistics.

Emily Goodwin, Koustuv Sinha, and Timothy J. O'Donnell. 2020. Probing linguistic systematicity. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1958–1969, Online. Association for Computational Linguistics.

Christine Herlihy and Rachel Rudinger. 2021. MedNLI is not immune: Natural language inference artifacts in the clinical domain. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1020–1027, Online. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Hai Hu, Qi Chen, Kyle Richardson, Atreyee Mukherjee, Lawrence S. Moss, and Sandra Kuebler. 2020. MonaLog: a lightweight system for natural language inference based on monotonicity. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 334–344, New York, New York. Association for Computational Linguistics.

Hai Hu and Larry Moss. 2018. Polarity computations in flexible categorial grammar. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 124–129, New Orleans, Louisiana. Association for Computational Linguistics.

Aikaterini-Lida Kalouli, Richard Crouch, and Valeria de Paiva. 2020. Hy-NLI: a hybrid system for natural language inference. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5235–5249, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008, Online. Association for Computational Linguistics.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Bill MacCartney and Christopher D. Manning. 2007. Natural logic for textul inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 193–200, Prague. Association for Computational Linguistics.

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.

Tiago Pimentel, Naomi Saphra, Adina Williams, and Ryan Cotterell. 2020. Pareto probing: Trading off accuracy for complexity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3138–3153, Online. Association for Computational Linguistics.

Kyle Richardson, Hai Hu, Lawrence S. Moss, and Ashish Sabharwal. 2019. Probing natural language inference models through semantic fragments. *CoRR*, abs/1909.07521.

Julia Rozanova, Deborah Ferreira, Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2021. Supporting context monotonicity abstractions in neural nli models.

V. Sanchez. 1991. Studies on natural logic and categorial grammar.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019a. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40, Florence, Italy. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019b. HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 250–255, Minneapolis, Minnesota. Association for Computational Linguistics.

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361, Brussels, Belgium. Association for Computational Linguistics.