Analytical, Symbolic and First-Order Reasoning within Neural Architectures

Samuel Ryb
Cornell University
shr59@cornell.edu

Marten van Schijndel
Cornell University
mv443@cornell.edu

Abstract

While prior work shows that pre-trained language models (PLMs) primarily emulate knowledge of entailment relations using surface heuristics, this paper examines whether PLMs learn aspects of symbolic and first-order logic relations as a side effect of learning word prediction. We introduce Logic and Knowledge Natural Language Inference (LAKNLI), a new NLI task, and we probe two different PLMs: one fine-tuned on NLI tasks and the other without NLI fine-tuning. Results show that PLMs are sometimes able to use logical knowledge for word prediction, yet they still rely heavily on heuristics. We also examine the conditions under which PLMs succeed and fail at utilizing logical relations.

1 Introduction

State-of-the-art language models often rely on surface level heuristics (McCoy et al., 2019). This is problematic when the heuristics make incorrect predictions involving logical properties (e.g., models may predict that *Either Alice knows Bob or Carl knows Claire* implies *Bob or Carl knows Claire*). This paper examines whether language models, can, in addition to surface level heuristics, infer symbolic and first-order logic relations from textual data. We introduce LAKNLI (Logic and Knowledge Natural Language Inference), a new probing dataset which assesses whether language models can reason by using logical patterns to predict entailment relationships (without being explicitly trained on them). We also examine the conditions under which BERT (Devlin et al., 2019), a widely used pre-trained language model, succeeds and fails at utilizing these logical relations.

2 LAKNLI

2.1 Overview of the Task and Dataset

In order to probe pre-trained language models (PLMs) to examine their symbolic and first-order logic reasoning abilities, we create a new probing task and dataset: LAKNLI (Logic and Knowledge Natural Language Inference). When solving LAKNLI, a language model needs to exploit the logical connective in the premise to predict whether a logical entailment exists between it and the hypothesis.

The dataset is divided according to 7 logical connectives (such as *and*, *or*, etc; full list given in Appendix A). 20 premises are attributed to each logical connective, where each premise is followed by 4 different hypotheses:

- **Premise (P)**: Some statement which is assumed as true. The premise is structured according to one of the deductive schemas given in Appendix A. **Example:** Alice got home by 2PM and met Bob then.
- **Direct Deduction (DD)** The (word-for-word) logical deduction, subject to one of the seven deductive schemas, which logically follows from the premise. A model should always judge a DD hypothesis to be entailed from the premise even if it solely relies on one of our distractor heuristics (LO or SS; defined below). **Example:** Alice got home by 2PM. Alice met Bob at 2PM.
- **Lexical Overlap (LO)** Some (possibly nonsense) bag-of-words reiterated from the premise. The hypothesis does not logically follow from the premise (Parikh et al., 2016). **Example:** Alice met home by Bob.
- **Subsequence Overlap (SS)** A random sequence of consecutive words reiterated from

Proceedings of the ESSLLI 2021 Workshop on Computing Semantics with Types, Frames and Related Structures, pages 61–72
Utrecht, The Netherlands (online), 26–27 July, 2021. ©2021 Association for Computational Linguistics

the premise. The hypothesis does not logically follow from the premise (otherwise SS and DD would be indistinguishable). **Example:** 2PM and met Bob.

- **Knowledge (K)** A hypothesis which significantly restricts lexical and subsequence overlap, yet still logically follows from the premise. A model will only judge K to be entailed by the premise if the training text statistics encode some of the logical subtleties of natural language. **Example:** Someone saw someone else in the afternoon.

Premise-Hypothesis (P-H)	**Entailment**
Premise - Direct Deduction (P-DD)	✓
Premise - Lexical Overlap (P-LO)	✗
Premise - Subsequence (P-SS)	✗
Premise - Knowledge (P-K)	✓

Table 1: The entailment relations that a non-heuristic based statistical learner should predict when probed on LAKNLI.

Note that LO and SS hypotheses do not logically contradict their corresponding premises. Rather, it is not possible to derive a logical entailment relation between an LO or SS hypothesis and the corresponding premise.[1]

2.2 Distinguishing LAKNLI From Other NLI Tasks and Datasets

While other resources provide related benefits to LAKNLI, the structure of LAKNLI differs from existing NLI tasks and datasets. HANS (McCoy et al., 2019) contains LO and SS sentences which are grammatically correct, while in LAKNLI there is no requirement for the LO and SS sentences to be grammatical nor make sense. In LAKNLI, grammatical correctness is an additional heuristic that models can use to determine the entailment relation between a premise and its hypothesis. If a model fails to classify LO and SS hypotheses as being non-entailed from their premise, one should question the model's ability to accurately encode formal properties of syntax.

SuperGLUE (Wang et al., 2019) is another resource which enables probing of a model's sensitivity to logical relations. However, SuperGLUE is not as narrowly targeted for the probing of logical information as LAKNLI. In SuperGLUE, logical connectives can appear in the premise for some items and in the hypothesis for other items. In LAKNLI, the main logical connective always occurs in the premise. LAKNLI also attempts to avoid any sentence ambiguity in terms of the main logical connective and in terms of referent binding.

Lastly, the construction and use of knowledge sentences (K) which minimize the utility of LO and SS heuristics are unique to LAKNLI and, to our knowledge, have not previously been used within NLI tasks and datasets.

3 Probing BERT architectures

In this paper, we focus on probing the BERT (Devlin et al., 2019) architecture's facility with logical relations. Given BERT's extensive pre-training and success on many NLP tasks, the model can provide an example of the kinds of analyses that can be done with LAKNLI as well as providing a solid baseline measure for other how well neural language models in general would perform on LAKNLI. That is, if BERT can solve LAKNLI, other neural language models may also be able to solve this task. However, if BERT fails on this task, perhaps other neural language models will not be able to solve this task. We first use a BERT-NLI model, which is fine-tuned on SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018), to see whether BERT has the ability to use symbolic logic to infer entailment even when fine-tuned on the task. We then probe a non-fine tuned BERT-base model on LAKNLI, to examine its capacity to reason about and deduce textual inferences correctly without explicit NLI training.

4 BERT-NLI

4.1 Preprocessing

Given a premise and a hypothesis, BERT-NLI outputs the corresponding logical relations: *entailment*, *non-entailment*, and *neutral*. We passed all of the P-{DD,LO,SS,K} sentence pairs from LAKNLI through BERT-NLI, but we coded neutral outputs as non-entailment.

Since exact lexical and syntactic overlap occurs between P and DD, any NLI-competent model should mark the relationship of all P-DD sentence pairs as entailment. However, BERT-NLI did not mark all P-DD pairs as entailment (see Appendix B), indicating a total failure to process

[1] In this paper, we use 'P-X' to indicate a premise-hypothesis pair. We use '{X,Y}_Z' to indicate that a model was trained on hypotheses of types X and Y and tested on hypotheses of type Z.

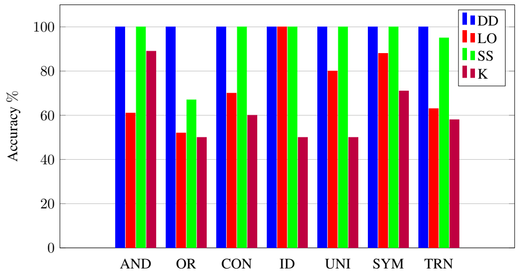


Figure 1: BERT-NLI entailment relation prediction accuracy when evaluated on a subset (492 P-H sentence pairs) of LAKNLI.

Connective	Accuracy
AND (conjunction)	89%
OR (disjunction)	50%
CON (conditional)	60%
ID (identity)	50%
UNI (universal)	50%
SYM (symmetry)	71%
TRN (transitivity)	58%

Table 2: BERT-NLI's accuracy in predicting the entailment relations of LAKNLI's P-K sentences.

those items. Therefore, we removed all premises and associated hypotheses in cases where P-DD was predicted as non-entailment. This resulted in removing 68 premise-hypothesis sets and preserving 492 premise-hypothesis sets.

4.2 Results and Discussion

The results (see Figure 1) parallel those of McCoy and Linzen (2018) and indicate that BERT-NLI primarily encodes entailment relations according to subsequence overlap. BERT-NLI has the most success at correctly predicting the entailment relation between AND P-K sentences, achieving an accuracy of 89% (see Table 2). Above chance performance suggests that BERT-NLI encodes some of the logical relations included in LAKNLI (AND, CON, SYM, TRN).

5 BERT-base (uncased)

Since BERT-NLI only outputs one of three NLI categories, we used BERT-base to conduct a more thorough error analysis of the kinds of logical relations that can be inferred from text statistics.

We used a support-vector machine to probe BERT. The [CLS] token in BERT encodes the overall meaning of each sentence, so we used each layer's encoding of the [CLS] token as the input to our SVM. As only 560 P-H pairs are available in LAKNLI, we used 5-fold cross-validation to train 3 different SVM probes (see Table 3).

Probe Name	Trained On	Tested On
{DD,LO}_LO	P-DD, P-LO	P-LO
{DD,SS}_SS	P-DD, P-SS	P-SS
{DD,K}_K	P-DD, P-K	P-K

Table 3: Trained probes and their descriptions.

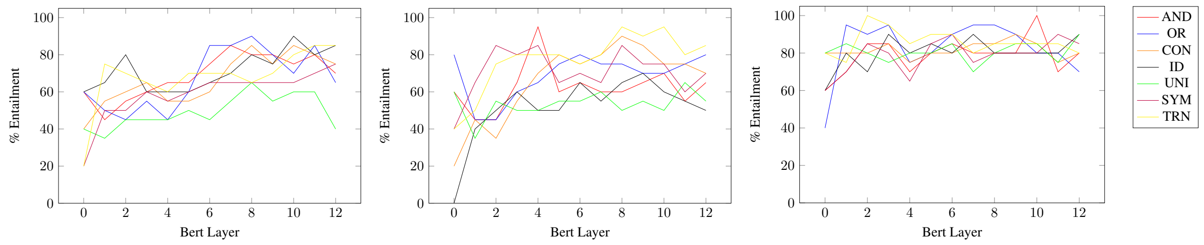
Remark Consider a probe of the form {A,B}_C, where A, B and C are some hypothesis types from LAKNLI. If C=A or C=B (i.e. the test sentences are of type A or B), the probe should predict 1 (an entailment relation), otherwise, the probe should predict 0 (a non-entailment relation). For example, consider the {DD,LO}_LO probe, which was trained on P-DD and P-LO sentence pairs. Since it was tested on P-LO sentence pairs, the probe should output 1 for all of those items. If either P-SS or P-K sentences (the item classes which were not observed during probe training) are used during testing of a {DD,LO} probe, it should output 0 for all of those items.

We trained all probes on DD hypotheses in addition to another set of hypotheses, as DD hypotheses have a common property with all LO, SS and K hypotheses. That is, DD hypotheses include lexical and syntactic overlap with P (like LO and SS hypotheses), yet are still logically entailed from P (like K sentences, which do not include overlap). Since P-K pairs contain minimal lexical overlap, training probes on only P-K pairs could make the probe negatively correlate lexical overlap with entailment. That is, the probe could learn that lexical overlap indicates non-entailment and vice-versa. Our aim in training the probe on P-DD as well as on P-K was to push the probe to identify more generalizable knowledge within BERT (i.e. lexical overlap can produce entailment under some conditions).

5.1 Experiment #1

Training probes on contextualized embeddings can yield high test accuracy even when embeddings do not necessarily encode relevant information (Hewitt and Liang, 2019). Therefore, we defined two tasks:

- **The linguistic task** tracked whether BERT solved LAKNLI using logical relations. For this task we used the {DD,K}_K probe.



(a) {DD,LO}_LO

(b) {DD,SS}_SS

(c) {DD,K}_K

Figure 2: Percent of test items classified as containing an entailment relation across the linguistic and control task probes.

- **The control task** tracked whether BERT solved LAKNLI using surface heuristics. For this task we used the {DD,LO}_LO and {DD,SS}_SS probes.

Each of these probes was trained on the P-H [CLS] embeddings from each of the 13 layers of the model,[2] and we tracked the flow of information throughout each layer of BERT. Figure 2c shows the number of linguistic task items classified as having an entailment relation while Figures 2a and 2b show the percent of control task items classified as having an entailment relation, for each logical connective, across each layer of BERT-base.

5.1.1 Results and Discussion

All of the probes classified several items as exhibiting entailment across all non-embedding layers for both the linguistic and control tasks (see Figure 2). These results replicate previous findings showing that BERT relies on surface level heuristics (McCoy and Linzen, 2018; McCoy et al., 2019), but we found that information about logical relations were decodable from BERT's internal representations as well.

In order to determine whether BERT's contextualized embeddings encoded the semantics of logical connectives or whether the trained probes learned the logical connectives separately from BERT, we measured the *selectivity* of the probing task (Hewitt and Liang, 2019). However, we modified the original definition to fit our experiment:

Definition 1. *selectivity = percent entailment of {DD,K}_K probe on layer$_i$ - max(percent entailment of {DD,LO}_LO probe on layer$_i$, percent entailment of {DD,SS}_SS probe on layer$_i$).*

A positive selectivity score represents the degree to which the probe tracked entailment relations using logic rather than surface heuristics, while the absolute value of a negative selectivity score represents the degree to which the probe tracked entailment relations using surface heuristics rather than logic (see Figure 3 and Appendix C).

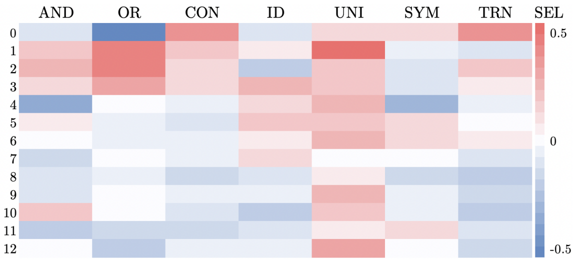


Figure 3: Probe selectivity computed by Definition 1.

A selectivity score of 1 would indicate that BERT solved LAKNLI using *only* logical understanding and knowledge. While this was not what we observed, our results still confirm that BERT was able to track entailment relations and encode some symbolic and first-order reasoning properties when solving LAKNLI (represented by a positive selectively score, with some exceptions where the selectivity scores are negative).

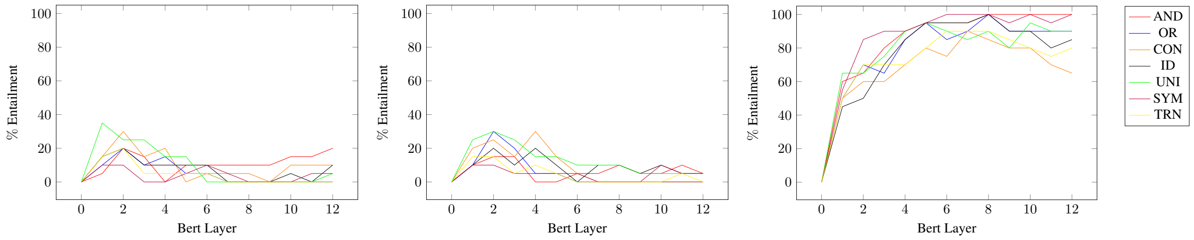
Our results offer a counterpoint to McCoy et al. (2019) who claimed that BERT primarily encodes entailment relations according to surface heuristics. We confirmed that BERT uses surface heuristics but sometimes also encodes knowledge and logical reasoning, though it was trained solely on text data.

5.2 Experiment #2

In order to examine how BERT distinguishes between knowledge and surface heuristic hypotheses, we trained three more probes (see Table 4). The percent of test items the probes classified as entailment can be seen in Figure 4.

The goal of this analysis was to identify the specific features encoded by BERT that can distinguish the hypotheses from one another. Lexical overlap is one obvious difference, which we controlled for

[2] We denote the embedding layer as *layer 0*.



(a) {K}_LO

(b) {K}_SS

(c) {K}_K

Figure 4: Percent of test sentence pairs that are classified as having an entailment relation when trained on K sentences and tested on LO, SS and K.

Probe Name	Trained On	Tested On
{K}_LO	P-K	P-LO
{K}_SS	P-K	P-SS
{K}_K	P-K	P-K

Table 4: Trained probes and their descriptions

in the previous analysis by including DD hypotheses in each probe training set. By removing DD hypotheses from the probe training sets in this analysis, we anticipated that the probes would learn a negative correlation between lexical overlap and logical entailment. However, we also expected that the probes would help us identify other encoded features that distinguish between knowledge-based properties and surface layer heuristics.

5.2.1 Results and Discussion

The {K}_K probe (Figure 4c) performed substantially above chance after layer 4. The high percentage of items classified as containing an entailment relation in the intermediate and upper layers suggests that BERT was able to exploit logical connectives to correctly determine entailment relations. This result is consistent with previous observations that BERT's intermediate and upper layers can encode semantic meaning (Jawahar et al., 2019), though it may also indicate the availability of input features at higher layers of BERT (we explore this more at the end of this section).

The {K}_LO (Figure 4a) and {K}_SS (Figure 4b) probes should have achieved 0 percent entailment classification across all layers, for each logical connective. When trained on P-K pairs, the probe should not have been able to deduce an entailment relation between the P-LO and P-SS sentence pairs. While only a small percentage of items were classified as containing entailment relations in layers 6-12, the lower 6 layers still classified 30% of test items as entailment relations. That the probes marked entailment relations between premises and (sometimes ungrammatical) hypotheses suggests that BERT does not use syntax and grammar as a heuristic to encode logical entailment relations. While BERT has the ability to handle subject-verb agreement, learn rich syntactic features from middle layers and encode the most information regarding linear word order in lower layers (Goldberg, 2019; Jawahar et al., 2019; Rogers et al., 2020) our results should cause one to question BERT's understanding of natural language grammar properties.

In order to determine how much more BERT was able to distinguish knowledge hypotheses (K) from their corresponding lexical and subsequence overlap hypotheses (LO, SS), we again used selectivity (although modified slightly from Definition 1):

Definition 2. *selectivity = percent entailment classification of {K}_K probe on layer$_i$ - max(percent entailment classification of {K}_LO probe on layer$_i$, percent entailment classification of {K}_SS probe on layer$_i$).*

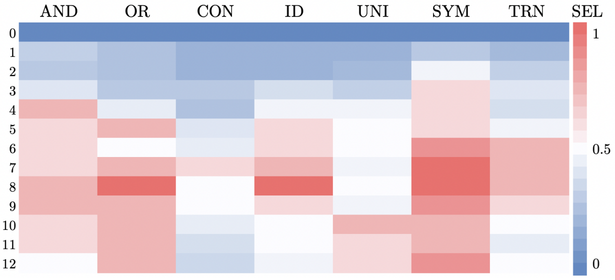


Figure 5: Probe selectivity scores computed by Definition 2.

Per Figure 5, the probe was best able to distinguish between knowledge and surface heuristics within layers 5 - 10 (see Figure 5 and Appendix C). This result aligns with our previous selectivity results (Figure 3 and Figure 8), which indicated that the probe was able to decode logical relations from BERT's representations mainly in the intermediate (and some upper) levels. We hypothesize that this was most likely due to BERT encoding some properties of syntactic structures primarily in lower layers (used for LO and SS sentences) and encoding semantic and pragmatic information in intermediate and upper layers (used for K sentences).

As noted earlier, the high selectivity scores may have been due to a lack of lexical overlap between the premise and knowledge hypotheses. To explore the influence of lexical overlap on the probe, we trained an additional {K}_DD probe (see Figure 6).

Probe Name	Trained On	Tested On
{K}_DD	P-K	P-DD

Table 5: Trained probe and its description.

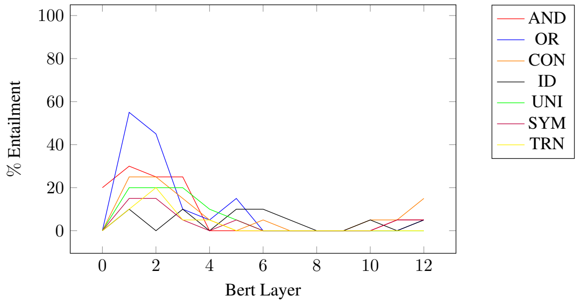


Figure 6: Percent of test items classified as containing an entailment relation by a {K}_DD probe.

The {K}_DD probe should have achieved near 100 percent entailment classification across all layers. However, as seen in Figure 6, this did not occur, which indicates that a probe trained solely on K sentences will learn entailment based on the *lack* of surface level heuristics. However, a probe trained on both DD and K sentences (that is, a probe trained equally on sentences with and without surface level overlap), can produce more generalizable entailment predictions, independent of lexical overlap, as seen in Figure 2c.

Therefore, one potential solution to ensure accurate probing results is to train probes on data that contains a balance of surface features and knowledge features. Furthermore, adding contradictory knowledge hypotheses to LAKNLI would enable researchers to calculate precision and recall scores, which would give a better indication as to whether a probe solves LAKNLI using logical relations.

6 Error Analyses

We next conducted qualitative analyses, where, when applicable, we categorized P-K sentence pairs from LAKNLI according to an error type. All analyses used data from the {K}_K probe in Section 5.2.[3] Below are the error types we used, followed by their descriptions and a sample P-K sentence pair:

- **Visual Reasoning (VR)** Sentences which require an analytic understanding of phrases related to spatial reasoning, such as *left of, in front of, behind, between* etc. **Example:** *P: If Alice is next to Bob, Bob is to the right of Carl. Alice is next to Bob → K: Carl is to the left of someone.*
- **Common Knowledge (CK)** Sentences which should be generally understood without any specialized knowledge. **Example:** *P: Alice is friends with Bob → K: Alice and Bob both like each other.*
- **World Knowledge (WK)** Sentences which require general knowledge about entities (such as animals, geographic locations, etc.) in the real world. **Example:** *P: Every boy who likes Alice is in New York City. Bob likes Alice → K: Bob is in North America.*

	AND	OR	CON	ID	UNI	SYM	TRN	TOTAL
VR	4	3	2	1	2	0	7	19
CK	7	3	8	8	9	14	10	59
WK	5	3	5	7	6	1	2	29

Table 6: Number of P-K sentence pairs tagged according to an error type within LAKNLI.

Visual Reasoning Prior visual commonsense reasoning (VCR) tasks such as NLVR (Suhr et al., 2017) have some similarities to LAKNLI, although we only probed PLMs on textual data. Our goal was to examine PLMs' inferential understanding of object relations through the use of non-visually grounded language and analytic consequences (e.g., X is to the right of Y $\iff$ Y is to the left of X).

Sentences which are tagged as VR within LAKNLI can be further classified according to the reasoning phrase involved. Such reasoning phrases and their frequencies are: *next to* (5), *right of / left of* (9), *in front of / behind* (5), *on top of / below* (3), *north of / south of* (1).

[3] Within this section the phrase *the probe* refers to the {K}_K probe.

The probe seemed to struggle with understanding analytic consequences including *right of/left of* (e.g., X is to the right of Y $\iff$ Y is to the left of X), *on top of/below* (e.g., X is on top of Y $\iff$ Y is below X), and *in front of/behind* (e.g., X is in front of Y $\iff$ Y is behind X) in lower layers. However, from layer 4 and above, the degree of error substantially decreased, with only 1 error in layers 4, 5, 6 and 8 and 2 errors in layer 7. This result suggests that BERT was able to perform best at a visual reasoning task between layers 4-8, which aligns with the results from Section 5.2.

The majority of incorrect entailment predictions in the upper layers (9-12) required understanding the relationship between left and right. Even in upper layers, which are supposedly more pragmatically advanced (Tenney et al., 2019), BERT was not fully able to understand such spatial implications.

In terms of logical connectives, it is possible that BERT also encodes the semantics of the biconditional deductive schema (which is not one of the seven deductive schemas included in LAKNLI) from layer 4 upwards, since the probe correctly predicted the majority of visual reasoning phrases which involved analytic consequences. However, due to the small size of LAKNLI, and since BERT did not encode the pragmatic relationship between left and right, future work should probe BERT on larger visual reasoning datasets with an emphasis on analytic consequences.

Common Knowledge Common knowledge reasoning tasks require language models to understand general scenarios that humans intuitively understand (Mostafazadeh et al., 2016; Zellers et al., 2018; LoBue and Yates, 2011).

For the error analysis, we further subcategorized CK P-K sentence pairs into the following types: *Description* (sentences which include a general description about an individual or circumstance), *General Action* (sentences which involve an individual doing an action), *Spatial Relation* (sentences with phrases that impose a spatial relation between at least two entities yet do not require a language model to solve a visual reasoning task), *Time* (sentences which refer to specific and/or general times of the day).

While many errors occurred with no particular pattern in the lower layers (1-3), BERT seems able to encode information regarding *spatial relations* in these lower layers, particularly understanding phrases such as *in proximity, near, adjacent, is in*. BERT seemed to encode general descriptions from layer 6 and above, with the probe classifying 100% of the test items as entailment relations within layers 7-11.

The probe also struggled to correctly predict the entailment relations of some *general action* P-K sentence pairs, particularly in lower layers 1-4 and in upper layers 9, 11-12 despite classifying a large number of items as entailment relations in the intermediate (and one upper) layers 6-8, 10. The probe incorrectly predicted the entailment relations of the following two general action P-K sentence pairs (a) *P: Alice is at home. If Bob walks to the park then Alice walks to the park. Bob walks to the park. → K: Alice leaves her house*, (b) *P: Alice is at home. Alice walks to the park with Bob. Bob walks to the park with Carl. → K: Alice leaves her house* in layers 1-4, 11, 12 and 1-5, 9-12, respectively. Since these were the only P-K sentence pairs in which the displacement of an agent from one location to another was apparent, it is plausible that BERT fails to understand such a relation. This hypothesis is supported by Forbes et al. (2019) who highlighted BERT's struggle to reason about objects and properties in the physical world.

BERT's understanding of *time* was inconsistent, with low entailment accuracy in the higher layers of the model. The probe incorrectly predicted the entailment relations of two P-K sentence pairs in layers 7, 9-12 and layers 9-10, which required the model to understand that eating at 2PM is associated with lunch time. These initial results correlate with those of (Han et al., 2019) that BERT embeddings do not achieve high accuracy at temporal relation extraction tasks. While their testing sentences included phrases which were indicative of temporal relations, the sentences in LAKNLI make use of numerical symbols to denote time (e.g., *eats lunch at **2**PM*). Therefore, it may be that the probe incorrectly predicted some entailment relations due to BERT struggling to encode numerical symbols (Wallace et al., 2019).

World Knowledge For this analysis, world knowledge P-K sentence pairs were further subcategorized into three types: *Location* (sentences which include relations between cities, countries and continents), *Eco-Systems* (sentences which require an understanding of the basic facts of nature), *Languages* (sentences which require an understanding of the common facts about languages and where they are typically spoken).

The probe correctly predicted the entailment relations of *language* P-K sentence pairs such as *P: If Alice studies French, Bob studies Italian and no other languages. Alice studies French and English and Spanish* → *K: Bob has knowledge of the language spoken in Venice* throughout all of BERT's layers, suggesting that BERT encodes where certain languages are commonly spoken. This was supported by the probe correctly predicting the entailment relation of *P: Alice studies either English or French and Bob either studies English or Spanish. Neither Alice nor Bob study English* → *K: Alice and Bob both learn a Romance language* in all layers, which indicates that perhaps BERT knows that French and Spanish are both considered Romance languages. This suggests that BERT may be encoding properties of set-theoretic membership (e.g., Spanish $\in$ Romance languages, French $\in$ Romance languages).

Half of the *eco-system* P-K sentence pairs required understanding common features of sea creatures, such as their abilities to swim or live in water. The probe correctly predicted the entailment relations of all those P-K sentence pairs in layers 1-8, yet struggled in upper layers (9, 11-12). This result aligns with recent work by Singh et al. (2020) which stressed that many of the intermediate layers contain knowledge-based information that is not included in the final layer.

7 Conclusions

Our work shows that BERT encodes some symbolic and first-order logic relations after training on only textual information. Despite lexical overlap having a large effect, we find that SVM probes of BERT trained on LAKNLI's knowledge sentences still achieve high NLI accuracy. It is therefore possible that the text statistics encode enough about symbolic logic relations for BERT to use these relations to solve NLI tasks. However, since text-trained models are likely unable to effectively learn reference (Merrill et al., 2021), future work must be done to determine the extent of symbolic understanding possible in models that lack reference. We hope that the LAKNLI dataset can help further investigate this question.

Acknowledgements

We thank the reviewers for their feedback on this work. Furthermore, we thank John Whitman and Claire Cardie for earlier feedback and suggestions pertaining to this work.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. 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.

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.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? *CoRR*, abs/1908.02899.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *CoRR*, abs/1901.05287.

Rujun Han, Mengyue Liang, Bashar Alhafni, and Nanyun Peng. 2019. Contextualized word embeddings enhanced event temporal relation extraction for story understanding. *CoRR*, abs/1904.11942.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. *CoRR*, abs/1909.03368.

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.

Peter LoBue and Alexander Yates. 2011. Types of common-sense knowledge needed for recognizing textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 329–334, Portland, Oregon, USA. Association for Computational Linguistics.

R. Thomas McCoy and Tal Linzen. 2018. Non-entailed subsequences as a challenge for natural language inference. *CoRR*, abs/1811.12112.

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.

Will Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021. Provable limitations of acquiring meaning from ungrounded form:what will future language models understand? *TACL*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works.

Jaspreet Singh, Jonas Wallat, and Avishek Anand. 2020. BERTnesia: Investigating the capture and forgetting of knowledge in BERT. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 174–183, Online. Association for Computational Linguistics.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP models know numbers? probing numeracy in embeddings. 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 5307–5315, Hong Kong, China. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *CoRR*, abs/1905.00537.

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.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

A LAKNLI's Logical Deductive Schema Templates

The seven logical connectives and their deductive schemas are defined below in the Fitch format. This demonstrates (1) the crossover between zeroth/first-order logic and equivalence relations within natural language (the main logical connective is bolded within P) and (2) the relationship between premises and the 4 different hypotheses (DD, LO, SS, K) subject to each logical connective:

And (conjunction) Elim (AND)

1	$A \wedge B$	
2	A	$\wedge Elim(1)$
3	B	$\wedge Elim(1)$

Sample AND Sentences

P *Alice is friends with Bob* ***and*** *Alice is not friends with Carl.*

DD *Alice is friends with Bob. Alice is not friends with Carl.*

LO *Bob is friends with Carl.*

SS *Bob and Alice is not friends with Carl.*

K *Alice knows two people.*

Or (disjunction) Elim (OR)

1	$A \lor B$	
2	$\neg A$	
3	A	
4	$\bot$	$\bot Intro(2, 3)$
5	B	$\bot Elim(4)$
6	B	
7	B	$Reit(6)$
8	B	$Or - Elim(1, 3 - 5, 6 - 7)$

Sample OR Sentences

P *Alice is at home.* ***Either*** *Alice walks to the park* ***or*** *Bob walks to the park. Alice does not walk to the park.*
DD *Bob walks to the park.*
LO *Home walks to the park.*
SS *to the park or Bob walks.*
K *A man does not remain in his current state.*

Conditional Elim (CON)

1	$A \to B$	
2	A	
3	B	$\to Elim(1, 2)$

Sample CON Sentences

P ***If*** *Alice attends the party Bob attends the party. Alice attends the party.*
DD *Bob attends the party.*
LO *Attends Alice the party.*
SS *The party Bob attends.*
K *Two people attend an event.*

Identity Elim (ID)

1	$b = f(a)$	
2	$P(f(a))$	
3	$P(b)$	$= Elim(1, 2)$

Sample ID Sentences

P *Bob* ***is*** *Alice's uncle. Alice's uncle* ***is*** *in New York City.*
DD *Bob is in New York City.*
LO *New York City is Alice's uncle.*
SS *Uncle is in New York City.*
K *Somebody is in North America.*

Universal Elim (UNI)

1	$\forall x (P(x, a) \to Q(x))$	
2	$P(b, a)$	
3	$P(b, a) \to Q(b)$	$\forall Elim(1)$
4	$Q(b)$	$\to Elim(2, 3)$

Sample UNI Sentences

P ***Every*** *boy who likes Alice is in New York City. Bob likes Alice.*
DD *Bob is in New York City.*
LO *New York City is Alice.*
SS *Alice is in New.*
K *Somebody is in North America.*

Symmetry (SYM)

1	$A \sim B$
2	$B \sim A$

Sample SYM Sentences

P *Alice's party outfit* ***is similar*** *to Bob's shirt.*
DD *Bob's shirt is similar to Alice's party outfit.*
LO *Bob's outfit is similar to Alice's party.*
SS *Party outfit is similar to Bob's.*
K *Two people have comparable clothing.*

Transitivity (TRN)

1	$A \sim B$
2	$B \sim C$
3	$A \sim C$

Sample TRN Sentences

P *Alice's party* ***is north of*** *Bob's party. Bob's party* ***is north of*** *Carl's party.*
DD *Alice's party is north of Carl's party.*
LO *Alice's party is Bob's party.*
SS *Bob's party is north.*
K *Carl's event is located in a southern location compared to Alice's event.*

We avoided using negation (¬), except when necessary (e.g. in the OR deductive schema), as it is often used as a heuristic by PLMs to infer a non-entailment relation between a premise and a hypothesis (McCoy and Linzen, 2018). However, considering that negation changes the semantics of universal quantifiers (e.g. $\neg\forall x P(x) \iff \exists x \neg P(x)$), conditionals (e.g. Modus Tollens $P \rightarrow Q, \neg Q \therefore \neg P$) and conjunctions/dijunctions $(\neg(A \wedge B) \iff \neg A \vee \neg B)$ amongst other logical connectives, we leave it for future work to determine a more sophisticated approach for probing PLM's abilities to understand first-order equivalences involving negation.

B BERT-NLI Preprocessing results

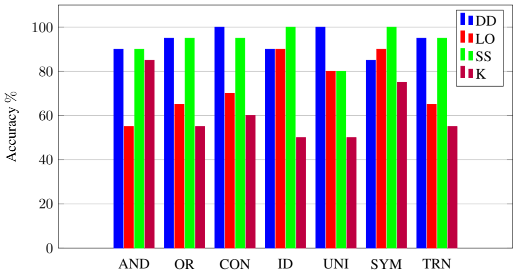
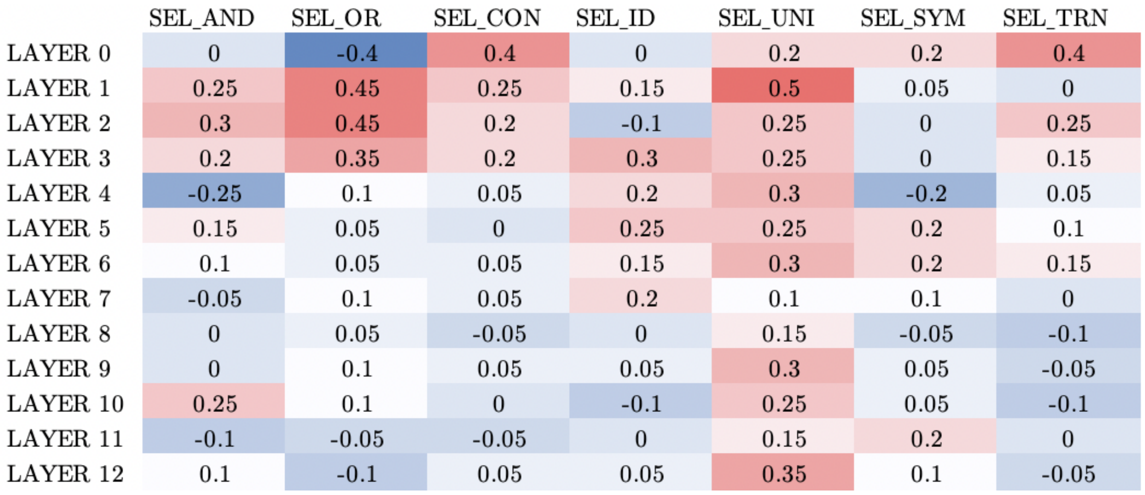


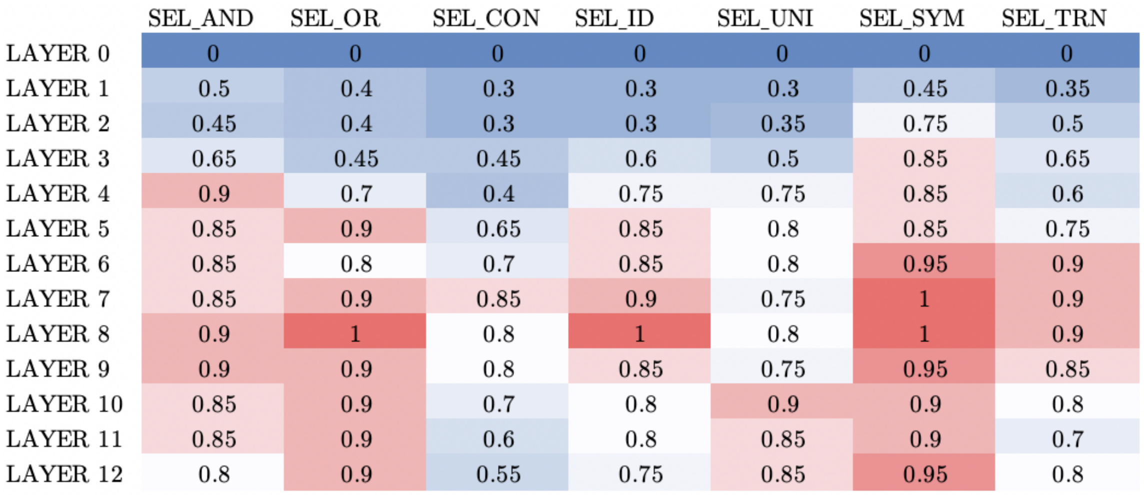
Figure 7: BERT-NLI entailment relation prediction accuracy across all 560 P-H sentence pairs from LAKNLI.

C Selectivity Scores



	SEL_AND	SEL_OR	SEL_CON	SEL_ID	SEL_UNI	SEL_SYM	SEL_TRN
LAYER 0	0	-0.4	0.4	0	0.2	0.2	0.4
LAYER 1	0.25	0.45	0.25	0.15	0.5	0.05	0
LAYER 2	0.3	0.45	0.2	-0.1	0.25	0	0.25
LAYER 3	0.2	0.35	0.2	0.3	0.25	0	0.15
LAYER 4	-0.25	0.1	0.05	0.2	0.3	-0.2	0.05
LAYER 5	0.15	0.05	0	0.25	0.25	0.2	0.1
LAYER 6	0.1	0.05	0.05	0.15	0.3	0.2	0.15
LAYER 7	-0.05	0.1	0.05	0.2	0.1	0.1	0
LAYER 8	0	0.05	-0.05	0	0.15	-0.05	-0.1
LAYER 9	0	0.1	0.05	0.05	0.3	0.05	-0.05
LAYER 10	0.25	0.1	0	-0.1	0.25	0.05	-0.1
LAYER 11	-0.1	-0.05	-0.05	0	0.15	0.2	0
LAYER 12	0.1	-0.1	0.05	0.05	0.35	0.1	-0.05

Figure 8: Probe selectivity scores computed by Definition 1.



	SEL_AND	SEL_OR	SEL_CON	SEL_ID	SEL_UNI	SEL_SYM	SEL_TRN
LAYER 0	0	0	0	0	0	0	0
LAYER 1	0.5	0.4	0.3	0.3	0.3	0.45	0.35
LAYER 2	0.45	0.4	0.3	0.3	0.35	0.75	0.5
LAYER 3	0.65	0.45	0.45	0.6	0.5	0.85	0.65
LAYER 4	0.9	0.7	0.4	0.75	0.75	0.85	0.6
LAYER 5	0.85	0.9	0.65	0.85	0.8	0.85	0.75
LAYER 6	0.85	0.8	0.7	0.85	0.8	0.95	0.9
LAYER 7	0.85	0.9	0.85	0.9	0.75	1	0.9
LAYER 8	0.9	1	0.8	1	0.8	1	0.9
LAYER 9	0.9	0.9	0.8	0.85	0.75	0.95	0.85
LAYER 10	0.85	0.9	0.7	0.8	0.9	0.9	0.8
LAYER 11	0.85	0.9	0.6	0.8	0.85	0.9	0.7
LAYER 12	0.8	0.9	0.55	0.75	0.85	0.95	0.8

Figure 9: Probe selectivity scores computed by Definition 2.