Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference

R. Thomas McCoy,[1] Ellie Pavlick,[2] & Tal Linzen[1]
[1]Department of Cognitive Science, Johns Hopkins University
[2]Department of Computer Science, Brown University
tom.mccoy@jhu.edu, ellie_pavlick@brown.edu, tal.linzen@jhu.edu

Abstract

A machine learning system can score well on a given test set by relying on heuristics that are effective for frequent example types but break down in more challenging cases. We study this issue within natural language inference (NLI), the task of determining whether one sentence entails another. We hypothesize that statistical NLI models may adopt three fallible syntactic heuristics: the lexical overlap heuristic, the subsequence heuristic, and the constituent heuristic. To determine whether models have adopted these heuristics, we introduce a controlled evaluation set called HANS (Heuristic Analysis for NLI Systems), which contains many examples where the heuristics fail. We find that models trained on MNLI, including BERT, a state-of-the-art model, perform very poorly on HANS, suggesting that they have indeed adopted these heuristics. We conclude that there is substantial room for improvement in NLI systems, and that the HANS dataset can motivate and measure progress in this area.

1 Introduction

Neural networks excel at learning the statistical patterns in a training set and applying them to test cases drawn from the same distribution as the training examples. This strength can also be a weakness: statistical learners such as standard neural network architectures are prone to adopting shallow heuristics that succeed for the majority of training examples, instead of learning the underlying generalizations that they are intended to capture. If such heuristics often yield correct outputs, the loss function provides little incentive for the model to learn to generalize to more challenging cases as a human performing the task would.

This issue has been documented across domains in artificial intelligence. In computer vision, for example, neural networks trained to recognize objects are misled by contextual heuristics: a network that is able to recognize monkeys in a typical context with high accuracy may nevertheless label a monkey holding a guitar as a human, since in the training set guitars tend to co-occur with humans but not monkeys (Wang et al., 2018). Similar heuristics arise in visual question answering systems (Agrawal et al., 2016).

The current paper addresses this issue in the domain of natural language inference (NLI), the task of determining whether a **premise** sentence entails (i.e., implies the truth of) a **hypothesis** sentence (Condoravdi et al., 2003; Dagan et al., 2006; Bowman et al., 2015). As in other domains, neural NLI models have been shown to learn shallow heuristics, in this case based on the presence of specific words (Naik et al., 2018; Sanchez et al., 2018). For example, a model might assign a label of *contradiction* to any input containing the word *not*, since *not* often appears in the examples of contradiction in standard NLI training sets.

The focus of our work is on heuristics that are based on superficial **syntactic** properties. Consider the following sentence pair, which has the target label *entailment*:

(1) *Premise:* The judge was paid by the actor.
Hypothesis: The actor paid the judge.

An NLI system that labels this example correctly might do so not by reasoning about the meanings of these sentences, but rather by assuming that the premise entails any hypothesis whose words all appear in the premise (Dasgupta et al., 2018; Naik et al., 2018). Crucially, if the model is using this heuristic, it will predict *entailment* for (2) as well, even though that label is incorrect in this case:

(2) *Premise:* The actor was paid by the judge.
Hypothesis: The actor paid the judge.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3428–3448
Florence, Italy, July 28 - August 2, 2019.

Heuristic	Definition	Example
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise	**The doctor** was **paid** by **the actor**. $\xrightarrow[\text{WRONG}]{}$ The doctor paid the actor.
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The doctor near **the actor danced**. $\xrightarrow[\text{WRONG}]{}$ The actor danced.
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	If **the artist slept**, the actor ran. $\xrightarrow[\text{WRONG}]{}$ The artist slept.

Table 1: The heuristics targeted by the HANS dataset, along with examples of incorrect entailment predictions that these heuristics would lead to.

We introduce a new evaluation set called HANS (Heuristic Analysis for NLI Systems), designed to diagnose the use of such fallible structural heuristics.[1] We target three heuristics, defined in Table 1. While these heuristics often yield correct labels, they are not valid inference strategies because they fail on many examples. We design our dataset around such examples, so that models that employ these heuristics are guaranteed to fail on particular subsets of the dataset, rather than simply show lower overall accuracy.

We evaluate four popular NLI models, including BERT, a state-of-the-art model (Devlin et al., 2019), on the HANS dataset. All models performed substantially below chance on this dataset, barely exceeding 0% accuracy in most cases. We conclude that their behavior is consistent with the hypothesis that they have adopted these heuristics.

Contributions: This paper has three main contributions. First, we introduce the HANS dataset, an NLI evaluation set that tests specific hypotheses about invalid heuristics that NLI models are likely to learn. Second, we use this dataset to illuminate interpretable shortcomings in state-of-the-art models trained on MNLI (Williams et al., 2018b); these shortcoming may arise from inappropriate model inductive biases, from insufficient signal provided by training datasets, or both. Third, we show that these shortcomings can be made less severe by augmenting a model's training set with the types of examples present in HANS. These results indicate that there is substantial room for improvement for current NLI models and datasets, and that HANS can serve as a tool for motivating and measuring progress in this area.

[1]GitHub repository with data and code: https://github.com/tommccoy1/hans

2 Syntactic Heuristics

We focus on three heuristics: the lexical overlap heuristic, the subsequence heuristic, and the constituent heuristic, all defined in Table 1. These heuristics form a hierarchy: the constituent heuristic is a special case of the subsequence heuristic, which in turn is a special case of the lexical overlap heuristic. Table 2 in the next page gives examples where each heuristic succeeds and fails.

There are two reasons why we expect these heuristics to be adopted by a statistical learner trained on standard NLI training datasets such as SNLI (Bowman et al., 2015) or MNLI (Williams et al., 2018b). First, the MNLI training set contains far more examples that support the heuristics than examples that contradict them:[2]

Heuristic	Supporting Cases	Contradicting Cases
Lexical overlap	2,158	261
Subsequence	1,274	72
Constituent	1,004	58

Even the 261 contradicting cases in MNLI may not provide strong evidence against the heuristics. For example, 133 of these cases contain negation in the premise but not the hypothesis, as in (3). Instead of using these cases to overrule the lexical overlap heuristic, a model might account for them by learning to assume that the label is *contradiction* whenever there is negation in the premise but not the hypothesis (McCoy and Linzen, 2019):

(3) a. I **don't** care. $\nrightarrow$ I care.
b. This is **not** a contradiction. $\nrightarrow$ This is a contradiction.

[2]In this table, the lexical overlap counts include the subsequence counts, which include the constituent counts.

Heuristic	Premise	Hypothesis	Label
Lexical overlap heuristic	The banker near the judge saw the actor.	The banker saw the actor.	E
	The lawyer was advised by the actor.	The actor advised the lawyer.	E
	The doctors visited the lawyer.	The lawyer visited the doctors.	N
	The judge by the actor stopped the banker.	The banker stopped the actor.	N
Subsequence heuristic	The artist and the student called the judge.	The student called the judge.	E
	Angry tourists helped the lawyer.	Tourists helped the lawyer.	E
	The judges heard the actors resigned.	The judges heard the actors.	N
	The senator near the lawyer danced.	The lawyer danced.	N
Constituent heuristic	Before the actor slept, the senator ran.	The actor slept.	E
	The lawyer knew that the judges shouted.	The judges shouted.	E
	If the actor slept, the judge saw the artist.	The actor slept.	N
	The lawyers resigned, or the artist slept.	The artist slept.	N

Table 2: Examples of sentences used to test the three heuristics. The *label* column shows the correct label for the sentence pair; *E* stands for *entailment* and *N* stands for *non-entailment*. A model relying on the heuristics would label all examples as *entailment* (incorrectly for those marked as N).

There are some examples in MNLI that contradict the heuristics in ways that are not easily explained away by other heuristics; see Appendix A for examples. However, such cases are likely too rare to discourage a model from learning these heuristics. MNLI contains data from multiple genres, so we conjecture that the scarcity of contradicting examples is not just a property of one genre, but rather a general property of NLI data generated in the crowdsourcing approach used for MNLI. We thus hypothesize that any crowdsourced NLI dataset would make our syntactic heuristics attractive to statistical learners without strong linguistic priors.

The second reason we might expect current NLI models to adopt these heuristics is that their input representations may make them susceptible to these heuristics. The lexical overlap heuristic disregards the order of the words in the sentence and considers only their identity, so it is likely to be adopted by bag-of-words NLI models (e.g., Parikh et al. 2016). The subsequence heuristic considers linearly adjacent chunks of words, so one might expect it to be adopted by standard RNNs, which process sentences in linear order. Finally, the constituent heuristic appeals to components of the parse tree, so one might expect to see it adopted by tree-based NLI models (Bowman et al., 2016).

3 Dataset Construction

For each heuristic, we generated five templates for examples that support the heuristic and five templates for examples that contradict it. Below is one template for the subsequence heuristic; see Appendix B for a full list of templates.

(4) The N_1 P the N_2 V. $\nrightarrow$ The N_2 V.
The lawyer by the actor ran. $\nrightarrow$ *The actor ran.*

We generated 1,000 examples from each template, for a total of 10,000 examples per heuristic. Some heuristics are special cases of others, but we made sure that the examples for one heuristic did not also fall under a more narrowly defined heuristic. That is, for lexical overlap cases, the hypothesis was not a subsequence or constituent of the premise; for subsequence cases, the hypothesis was not a constituent of the premise.

3.1 Dataset Controls

Plausibility: One advantage of generating examples from templates—instead of, e.g., modifying naturally-occurring examples—is that we can ensure the plausibility of all generated sentences. For example, we do not generate cases such as *The student read the book* $\nrightarrow$ *The book read the student*, which could ostensibly be solved using a hypothesis-plausibility heuristic. To achieve this, we drew our core vocabulary from Ettinger et al. (2018), where every noun was a plausible subject of every verb or a plausible object of every transitive verb. Some templates required expanding this core vocabulary; in those cases, we manually curated the additions to ensure plausibility.

Selectional criteria: Some of our example types depend on the availability of lexically-specific verb frames. For example, (5) requires awareness of the fact that *believed* can take a clause (*the lawyer saw the officer*) as its complement:

(5) The doctor believed the lawyer saw the officer. $\nrightarrow$ The doctor believed the lawyer.

It is arguably unfair to expect a model to understand this example if it had only ever encountered *believe* with a noun phrase object (e.g., *I believed* ***the man***). To control for this issue, we only chose verbs that appeared at least 50 times in the MNLI training set in all relevant frames.

4 Experimental Setup

Since HANS is designed to probe for structural heuristics, we selected three models that exemplify popular strategies for representing the input sentence: DA, a bag-of-words model; ESIM, which uses a sequential structure; and SPINN, which uses a syntactic parse tree. In addition to these three models, we included BERT, a state-of-the-art model for MNLI. The following paragraphs provide more details on these models.

DA: The Decomposable Attention model (DA; Parikh et al., 2016) uses a form of attention to align words in the premise and hypothesis and to make predictions based on the aggregation of this alignment. It uses no word order information and can thus be viewed as a bag-of-words model.

ESIM: The Enhanced Sequential Inference Model (ESIM; Chen et al., 2017) uses a modified bidirectional LSTM to encode sentences. We use the variant with a sequential encoder, rather than the tree-based Hybrid Inference Model (HIM).

SPINN: The Stack-augmented Parser-Interpreter Neural Network (SPINN; Bowman et al., 2016) is tree-based: it encodes sentences by combining phrases based on a syntactic parse. We use the SPINN-PI-NT variant, which takes a parse tree as an input (rather than learning to parse). For MNLI, we used the parses provided in the MNLI release; for HANS, we used parse templates that we created based on parses from the Stanford PCFG Parser 3.5.2 (Klein and Manning, 2003), the same parser used to parse MNLI. Based on manual inspection, this parser generally provided correct parses for HANS examples.

BERT: The Bidirectional Encoder Representations from Transformers model (BERT; Devlin et al., 2019) is a Transformer model that uses attention, rather than recurrence, to process sentences. We use the `bert-base-uncased` pretrained model and fine-tune it on MNLI.

Implementation and evaluation: For DA and ESIM, we used the implementations from AllenNLP (Gardner et al., 2017). For SPINN[3] and BERT,[4] we used code from the GitHub repositories for the papers introducing those models.

We trained all models on MNLI. MNLI uses three labels (*entailment*, *contradiction*, and *neutral*). We chose to annotate HANS with two labels only (*entailment* and *non-entailment*) because the distinction between *contradiction* and *neutral* was often unclear for our cases.[5] For evaluating a model on HANS, we took the highest-scoring label out of *entailment*, *contradiction*, and *neutral*; we then translated *contradiction* or *neutral* labels to *non-entailment*. An alternate approach would have been to add the *contradiction* and *neutral* scores to determine a score for *non-entailment*; we found little difference between these approaches, since the models almost always assigned more than 50% of the label probability to a single label.[6]

5 Results

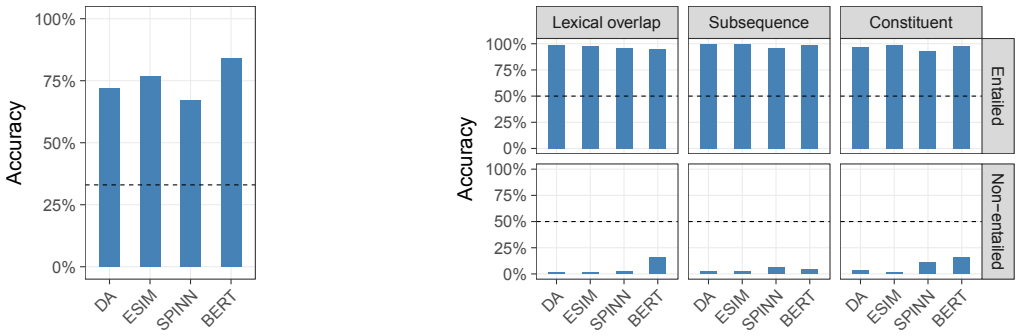
All models achieved high scores on the MNLI test set (Figure 1a), replicating the accuracies found in past work (DA: Gururangan et al. 2018; ESIM: Williams et al. 2018b; SPINN: Williams et al. 2018a; BERT: Devlin et al. 2019). On the HANS dataset, all models almost always assigned the correct label in the cases where the label is *entailment*, i.e., where the correct answer is in line with the hypothesized heuristics. However, they all performed poorly—with accuracies less than 10% in most cases, when chance is 50%—on the cases where the heuristics make incorrect predictions

[3] `https://github.com/stanfordnlp/spinn`; we used the NYU fork at `https://github.com/nyu-mll/spinn`.

[4] `https://github.com/google-research/bert`

[5] For example, with *The actor was helped by the judge* $\nrightarrow$ *The actor helped the judge*, it is possible that the actor did help the judge, pointing to a label of *neutral*; yet the premise does pragmatically imply that the actor did not help the judge, meaning that this pair could also fit the non-strict definition of *contradiction* used in NLI annotation.

[6] We also tried training the models on MNLI with *neutral* and *contradiction* collapsed into *non-entailment*; this gave similar results as collapsing after training (Appendix D) .



(a) (b)

Figure 1: (a) Accuracy on the MNLI test set. (b) Accuracies on six sub-components of the HANS evaluation set; each sub-component is defined by its correct label and the heuristic it addresses. The dashed lines indicate chance performance. All models behaved as we would expect them to if they had adopted the heuristics targeted by HANS. That is, they nearly always predicted *entailment* for the examples in HANS, leading to near-perfect accuracy when the true label is *entailment*, and near-zero accuracy when the true label is *non-entailment*.

(Figure 1b). Thus, despite their high scores on the MNLI test set, all four models behaved in a way consistent with the use of the heuristics targeted in HANS, and not with the correct rules of inference.

Comparison of models: Both DA and ESIM had near-zero performance across all three heuristics. These models might therefore make no distinction between the three heuristics, but instead treat them all as the same phenomenon, i.e. lexical overlap. Indeed, for DA, this must be the case, as this model does not have access to word order; ESIM does in theory have access to word order information but does not appear to use it here.

SPINN had the best performance on the subsequence cases. This might be due to the tree-based nature of its input: since the subsequences targeted in these cases were explicitly chosen not to be constituents, they do not form cohesive units in SPINN's input in the way they do for sequential models. SPINN also outperformed DA and ESIM on the constituent cases, suggesting that SPINN's tree-based representations moderately helped it learn how specific constituents contribute to the overall sentence. Finally, SPINN did worse than the other models on constituent cases where the correct answer is *entailment*. This moderately greater balance between accuracy on entailment and non-entailment cases further indicates that SPINN is less likely than the other models to assume that constituents of the premise are entailed; this harms its performance in cases where that assumption happens to lead to the correct answer.

BERT did slightly worse than SPINN on the subsequence cases, but performed noticeably less poorly than all other models at both the constituent and lexical overlap cases (though it was still far below chance). Its performance particularly stood out for the lexical overlap cases, suggesting that some of BERT's success at MNLI may be due to a greater tendency to incorporate word order information compared to other models.

Analysis of particular example types: In the cases where a model's performance on a heuristic was perceptibly above zero, accuracy was not evenly spread across subcases (for case-by-case results, see Appendix C). For example, within the lexical overlap cases, BERT achieved 39% accuracy on conjunction (e.g., *The actor and the doctor saw the artist* $\nrightarrow$ *The actor saw the doctor*) but 0% accuracy on subject/object swap (*The judge called the lawyer* $\nrightarrow$ *The lawyer called the judge*). Within the constituent heuristic cases, BERT achieved 49% accuracy at determining that a clause embedded under *if* and other conditional words is not entailed (*If the doctor resigned, the lawyer danced* $\nrightarrow$ *The doctor resigned*), but 0% accuracy at identifying that the clause outside of the conditional clause is also not entailed (*If the doctor resigned, the lawyer danced* $\nrightarrow$ *The lawyer danced*).

6 Discussion

Independence of heuristics: Though each heuristic is most closely related to one class of model (e.g., the constituent heuristic is related to tree-based models), all models failed on cases illustrating all three heuristics. This finding is unsurprising since these heuristics are closely related

to each other, meaning that an NLI model may adopt all of them, even the ones not specifically targeting that class of model. For example, the subsequence and constituent heuristics are special cases of the lexical overlap heuristic, so all models can fail on cases illustrating all heuristics, because all models have access to individual words.

Though the heuristics form a hierarchy—the constituent heuristic is a subcase of the subsequence heuristic, which is a subcase of the lexical overlap heuristic—this hierarchy does not necessarily predict the performance of our models. For example, BERT performed worse on the subsequence heuristic than on the constituent heuristic, even though the constituent heuristic is a special case of the subsequence heuristic. Such behavior has two possible causes. First, it could be due to the specific cases we chose for each heuristic: the cases chosen for the subsequence heuristic may be inherently more challenging than the cases chosen for the constituent heuristic, even though the constituent heuristic as a whole is a subset of the subsequence one. Alternately, it is possible for a model to adopt a more general heuristic (e.g., the subsequence heuristic) but to make an exception for some special cases (e.g., the cases to which the constituent heuristic could apply).

Do the heuristics arise from the architecture or the training set? The behavior of a trained model depends on both the training set and the model's architecture. The models' poor results on HANS could therefore arise from architectural limitations, from insufficient signal in the MNLI training set, or from both.

The fact that SPINN did markedly better at the constituent and subsequence cases than ESIM and DA, even though the three models were trained on the same dataset, suggests that MNLI does contain some signal that can counteract the appeal of the syntactic heuristics tested by HANS. SPINN's structural inductive biases allow it to leverage this signal, but the other models' biases do not.

Other sources of evidence suggest that the models' failure is due in large part to insufficient signal from the MNLI training set, rather than the models' representational capacities alone. The BERT model we used (`bert-base-uncased`) was found by Goldberg (2019) to achieve strong results in syntactic tasks such as subject-verb agreement prediction, a task that minimally requires a distinction between the subject and direct object of a sentence (Linzen et al., 2016; Gulordava et al., 2018; Marvin and Linzen, 2018). Despite this evidence that BERT has access to relevant syntactic information, its accuracy was 0% on the subject-object swap cases (e.g., *The doctor saw the lawyer* $\nrightarrow$ *The lawyer saw the doctor*). We believe it is unlikely that our fine-tuning step on MNLI, a much smaller corpus than the corpus BERT was trained on, substantially changed the model's representational capabilities. Even though the model most likely had access to information about subjects and objects, then, MNLI did not make it clear how that information applies to inference. Supporting this conclusion, McCoy et al. (2019) found little evidence of compositional structure in the InferSent model, which was trained on SNLI, even though the same model type (an RNN) did learn clear compositional structure when trained on tasks that underscored the need for such structure. These results further suggest that the models' poor compositional behavior arises more because of the training set than because of model architecture.

Finally, our BERT-based model differed from the other models in that it was pretrained on a massive amount of data on a masking task and a next-sentence classification task, followed by fine-tuning on MNLI, while the other models were only trained on MNLI; we therefore cannot rule out the possibility that BERT's comparative success at HANS was due to the greater amount of data it has encountered rather than any architectural features.

Is the dataset too difficult? To assess the difficulty of our dataset, we obtained human judgments on a subset of HANS from 95 participants on Amazon Mechanical Turk as well as 3 expert annotators (linguists who were unfamiliar with HANS: 2 graduate students and 1 postdoctoral researcher). The average accuracy was 76% for Mechanical Turk participants and 97% for expert annotators; further details are in Appendix F.

Our Mechanical Turk results contrast with those of Nangia and Bowman (2019), who report an accuracy of 92% in the same population on examples from MNLI; this indicates that HANS is indeed more challenging for humans than MNLI is. The difficulty of some of our examples is in line with past psycholinguistic work in which humans have been shown to incorrectly answer comprehension questions for some of our subsequence subcases. For example, in an experiment in which participants read the sentence *As Jerry played the violin*

gathered dust in the attic, some participants answered *yes* to the question *Did Jerry play the violin?* (Christianson et al., 2001).

Crucially, although Mechanical Turk annotators found HANS to be harder overall than MNLI, their accuracy was similar whether the correct answer was *entailment* (75% accuracy) or *non-entailment* (77% accuracy). The contrast between the balance in the human errors across labels and the stark imbalance in the models' errors (Figure 1b) indicates that human errors are unlikely to be driven by the heuristics targeted in the current work.

7 Augmenting the training data with HANS-like examples

The failure of the models we tested raises the question of what it would take to do well on HANS. One possibility is that a different type of model would perform better. For example, a model based on hand-coded rules might handle HANS well. However, since most models we tested are in theory capable of handling HANS's examples but failed to do so when trained on MNLI, it is likely that performance could also be improved by training the same architectures on a dataset in which these heuristics are less successful.

To test that hypothesis, we retrained each model on the MNLI training set augmented with a dataset structured exactly like HANS (i.e. using the same thirty subcases) but containing no specific examples that appeared in HANS. Our additions comprised 30,000 examples, roughly 8% of the size of the original MNLI training set (392,702 examples). In general, the models trained on the augmented MNLI performed very well on HANS (Figure 2); the one exception was that the DA model performed poorly on subcases for which a bag-of-words representation was inadequate.[7] This experiment is only an initial exploration and leaves open many questions about the conditions under which a model will successfully avoid a heuristic; for example, how many contradicting examples are required? At the same time, these results do suggest that, to prevent a model from learning a heuristic, one viable approach is to use a training set that does not support this heuristic.

[7]The effect on MNLI test set performance was less clear; the augmentation with HANS-like examples improved MNLI test set performance for BERT (84.4% vs. 84.1%) and ESIM (77.6% vs 77.3%) but hurt performance for DA (66.0% vs. 72.4%) and SPINN (63.9% vs. 67.0%).

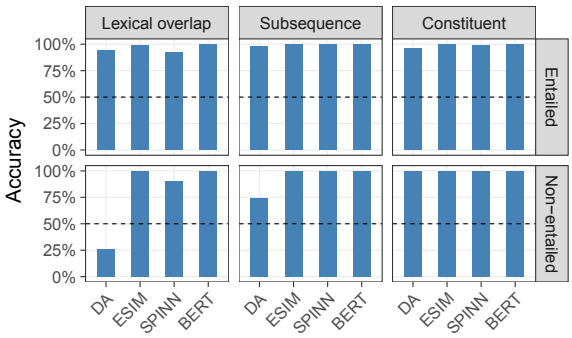


Figure 2: HANS accuracies for models trained on MNLI plus examples of all 30 categories in HANS.

Transfer across HANS subcases: The positive results of the HANS-like augmentation experiment are compatible with the possibility that the models simply memorized the templates that made up HANS's thirty subcases. To address this, we retrained our models on MNLI augmented with *subsets* of the HANS cases (withholding some cases; see Appendix E for details), then tested the models on the withheld cases.

The results of one of the transfer experiments, using BERT, are shown in Table 3. There were some successful cases of transfer; e.g., BERT performed well on the withheld categories with sentence-initial adverbs, regardless of whether the correct label was *non-entailment* or *entailment*. Such successes suggest that BERT is able to learn from some specific subcases that it should rule out the broader heuristics; in this case, the non-withheld cases plausibly informed BERT not to indiscriminately follow the constituent heuristic, encouraging it to instead base its judgments on the specific adverbs in question (e.g., *certainly* vs. *probably*). However, the models did not always transfer successfully; e.g., BERT had 0% accuracy on entailed passive examples when such examples were withheld, likely because the training set still included many non-entailed passive examples, meaning that BERT may have learned to assume that all sentences with passive premises are cases of non-entailment. Thus, though the models do seem to be able to rule out the broadest versions of the heuristics and transfer that knowledge to some new cases, they may still fall back to the heuristics for other cases. For further results involving withheld categories, see Appendix E.

Transfer to an external dataset: Finally, we tested models on the `comp_same_short` and

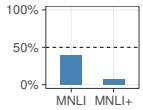
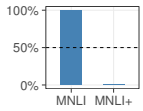
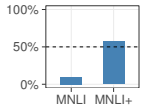
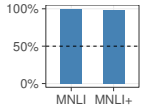
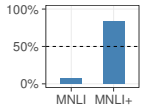
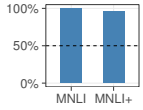
Withheld category	Results
Lexical overlap: Conjunctions (↛) *The doctor saw the author and the tourist.* ↛ *The author saw the tourist.*	
Lexical overlap: Passives (→) *The authors were helped by the actor.* → *The actor helped the authors.*	
Subsequence: NP/Z (↛) *Before the actor moved the doctor arrived.* ↛ *The actor moved the doctor.*	
Subsequence: PP on object (→) *The authors saw the judges by the doctor.* → *The authors saw the judges.*	
Constituent: Adverbs (↛) *Probably the artists helped the authors.* ↛ *The artists helped the authors.*	
Constituent: Adverbs (→) *Certainly the lawyers shouted.* → *The lawyers shouted.*	

Table 3: Accuracies for BERT fine-tuned on basic MNLI and on MNLI+, which is MNLI augmented with most HANS categories except withholding the categories in this table. The two lexical overlap cases shown here are adversarial in that MNLI+ contains cases superficially similar to them but with opposite labels (namely, the *Conjunctions (→)* and *Passives (↛)* cases from Table 4 in the Appendix). The remaining cases in this table are not adversarial in this way.

`comp_same_long` datasets from Dasgupta et al. (2018), which consist of lexical overlap cases:

(6) the famous and arrogant cat is not more nasty than the dog with glasses in a white dress. ↛ the dog with glasses in a white dress is not more nasty than the famous and arrogant cat.

This dataset differs from HANS in at least three important ways: it is based on a phenomenon not present in HANS (namely, comparatives); it uses a different vocabulary from HANS; and many of its sentences are semantically implausible.

We used this dataset to test both BERT fine-tuned on MNLI, and BERT fine-tuned on MNLI augmented with HANS-like examples. The augmentation improved performance modestly for the long examples and dramatically for the short examples, suggesting that training with HANS-like examples has benefits that extend beyond HANS.[8]

[8]We hypothesize that HANS helps more with short examples because most HANS sentences are short.

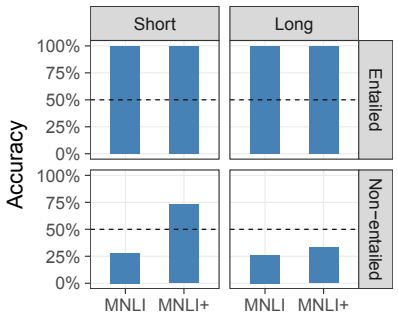


Figure 3: Results on the lexical overlap cases from Dasgupta et al. (2018) for BERT fine-tuned on MNLI or on MNLI augmented with HANS-like examples.

8 Related Work

8.1 Analyzing trained models

This project relates to an extensive body of research on exposing and understanding weaknesses in models' learned behavior and representations. In the NLI literature, Poliak et al. (2018b) and Gururangan et al. (2018) show that, due to biases in NLI datasets, it is possible to achieve far better than chance accuracy on those datasets by only looking at the hypothesis. Other recent works address possible ways in which NLI models might use fallible heuristics, focusing on semantic phenomena, such as lexical inferences (Glockner et al., 2018) or quantifiers (Geiger et al., 2018), or biases based on specific words (Sanchez et al., 2018). Our work focuses instead on *structural* phenomena, following the proof-of-concept work done by Dasgupta et al. (2018). Our focus on using NLI to address how models capture structure follows some older work about using NLI for the evaluation of parsers (Rimell and Clark, 2010; Mehdad et al., 2010).

NLI has been used to investigate many other types of linguistic information besides syntactic structure (Poliak et al., 2018a; White et al., 2017). Outside NLI, multiple projects have used classification tasks to understand what linguistic and/or structural information is present in vector encodings of sentences (e.g., Adi et al., 2017; Ettinger et al., 2018; Conneau et al., 2018). We instead choose the behavioral approach of using task performance on critical cases. Unlike the classification approach, this approach is agnostic to model structure; our dataset could be used to evaluate a symbolic NLI system just as easily as a neural one, whereas typical classification approaches only work for models with vector representations.

8.2 Structural heuristics

Similar to our lexical overlap heuristic, Dasgupta et al. (2018), Nie et al. (2018), and Kim et al. (2018) also tested NLI models on specific phenomena where word order matters; we use a larger set of phenomena to study a more general notion of lexical overlap that is less dependent on the properties of a single phenomenon, such as passives. Naik et al. (2018) also find evidence that NLI models use a lexical overlap heuristic, but our approach is substantially different from theirs.[9]

This work builds on our pilot study in McCoy and Linzen (2019), which studied one of the subcases of the subsequence heuristic. Several of our subsequence subcases are inspired by psycholinguistics research (Bever, 1970; Frazier and Rayner, 1982; Tabor et al., 2004); these works have aims similar to ours but are concerned with the representations used by humans rather than neural networks.

Finally, all of our constituent heuristic subcases depend on the implicational behavior of specific words. Several past works (Pavlick and Callison-Burch, 2016; Rudinger et al., 2018; White et al., 2018; White and Rawlins, 2018) have studied such behavior for verbs (e.g., *He knows it is raining* entails *It is raining*, while *He believes it is raining* does not). We extend that approach by including other types of words with specific implicational behavior, namely conjunctions (*and, or*), prepositions that take clausal arguments (*if, because*), and adverbs (*definitely, supposedly*). MacCartney and Manning (2009) also discuss the implicational behavior of these various types of words within NLI.

8.3 Generalization

Our work suggests that test sets drawn from the same distribution as the training set may be inadequate for assessing whether a model has learned to perform the intended task. Instead, it is also necessary to evaluate on a generalization set that departs from the training distribution. McCoy et al. (2018) found a similar result for the task of question formation; different architectures that all succeeded on the test set failed on the generalization set in different ways, showing that the test set alone was not sufficient to determine what the models had learned. This effect can arise not just from different architectures but also from different initializations of the same architecture (Weber et al., 2018).

[9] Naik et al. (2018) diagnose the lexical overlap heuristic by appending *and true is true* to existing MNLI hypotheses, which decreases lexical overlap but does not change the sentence pair's label. We instead generate new sentence pairs for which the words in the hypothesis all appear in the premise.

9 Conclusions

Statistical learners such as neural networks closely track the statistical regularities in their training sets. This process makes them vulnerable to adopting heuristics that are valid for frequent cases but fail on less frequent ones. We have investigated three such heuristics that we hypothesize NLI models are likely to learn. To evaluate whether NLI models do behave consistently with these heuristics, we have introduced the HANS dataset, on which models using these heuristics are guaranteed to fail. We find that four existing NLI models perform very poorly on HANS, suggesting that their high accuracies on NLI test sets may be due to the exploitation of invalid heuristics rather than deeper understanding of language. However, these models performed significantly better on both HANS and on a separate structure-dependent dataset when their training data was augmented with HANS-like examples. Overall, our results indicate that, despite the impressive accuracies of state-of-the-art models on standard evaluations, there is still much progress to be made and that targeted, challenging datasets, such as HANS, are important for determining whether models are learning what they are intended to learn.

Acknowledgments

We are grateful to Adam Poliak, Benjamin Van Durme, Samuel Bowman, the members of the JSALT General-Purpose Sentence Representation Learning team, and the members of the Johns Hopkins Computation and Psycholinguistics Lab for helpful comments, and to Brian Leonard for assistance with the Mechanical Turk experiment. Any errors remain our own.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1746891 and the 2018 Jelinek Summer Workshop on Speech and Language Technology (JSALT). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the JSALT workshop.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations*.

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960. Association for Computational Linguistics.

Thomas G. Bever. 1970. The cognitive basis for linguistic structures.

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.

Samuel R. Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D. Manning, and Christopher Potts. 2016. A fast unified model for parsing and sentence understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1466–1477. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Kiel Christianson, Andrew Hollingworth, John F Halliwell, and Fernanda Ferreira. 2001. Thematic roles assigned along the garden path linger. *Cognitive Psychology*, 42(4):368–407.

Cleo Condoravdi, Dick Crouch, Valeria de Paiva, Reinhard Stolle, and Daniel G. Bobrow. 2003. Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning*.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, pages 177–190, Berlin, Heidelberg. Springer-Verlag.

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J. Gershman, and Noah D. Goodman. 2018. Evaluating compositionality in sentence embeddings. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pages 1596–1601, Madison, WI.

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.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801. Association for Computational Linguistics.

Lyn Frazier and Keith Rayner. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14(2):178–210.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. AllenNLP: A Deep Semantic Natural Language Processing Platform. In *Proceedings of the Workshop for NLP Open Source Software (NLP-OSS)*.

Atticus Geiger, Ignacio Cases, Lauri Karttunen, and Christopher Potts. 2018. Stress-testing neural models of natural language inference with multiply-quantified sentences. *arXiv preprint arXiv:1810.13033*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. 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 1195–1205. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112. Association for Computational Linguistics.

Juho Kim, Christopher Malon, and Asim Kadav. 2018. Teaching syntax by adversarial distraction. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 79–84. Association for Computational Linguistics.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Bill MacCartney and Christopher D Manning. 2009. *Natural language inference*. Ph.D. thesis, Stanford University.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202. Association for Computational Linguistics.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2018. Revisiting the poverty of the stimulus: Hierarchical generalization without a hierarchical bias in recurrent neural networks. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pages 2093–2098, Madison, WI.

R. Thomas McCoy and Tal Linzen. 2019. Non-entailed subsequences as a challenge for natural language inference. In *Proceedings of the Society for Computation in Linguistics*, volume 2.

R. Thomas McCoy, Tal Linzen, Ewan Dunbar, and Paul Smolensky. 2019. RNNs implicitly implement tensor-product representations. In *International Conference on Learning Representations*.

Yashar Mehdad, Alessandro Moschitti, and Fabio Massimo Zanzotto. 2010. Syntactic/semantic structures for textual entailment recognition. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1020–1028. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353. Association for Computational Linguistics.

Nikita Nangia and Samuel R. Bowman. 2019. Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark.

Yixin Nie, Yicheng Wang, and Mohit Bansal. 2018. Analyzing compositionality-sensitivity of NLI models. *arXiv preprint arXiv:1811.07033*.

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. Association for Computational Linguistics.

Ellie Pavlick and Chris Callison-Burch. 2016. Tense manages to predict implicative behavior in verbs. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2225–2229. Association for Computational Linguistics.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018a. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018b. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191. Association for Computational Linguistics.

Laura Rimell and Stephen Clark. 2010. Cambridge: Parser evaluation using textual entailment by grammatical relation comparison. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 268–271. Association for Computational Linguistics.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural models of factuality. 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 731–744. Association for Computational Linguistics.

Ivan Sanchez, Jeff Mitchell, and Sebastian Riedel. 2018. Behavior analysis of NLI models: Uncovering the influence of three factors on robustness. 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 1975–1985. Association for Computational Linguistics.

Whitney Tabor, Bruno Galantucci, and Daniel Richardson. 2004. Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, 50(4):355–370.

Jianyu Wang, Zhishuai Zhang, Cihang Xie, Yuyin Zhou, Vittal Premachandran, Jun Zhu, Lingxi Xie, and Alan Yuille. 2018. Visual concepts and compositional voting. *Annals of Mathematical Sciences and Applications*, 3(1):151–188.

Noah Weber, Leena Shekhar, and Niranjan Balasubramanian. 2018. The fine line between linguistic generalization and failure in seq2seq-attention models. In *Proceedings of the Workshop on Generalization in the Age of Deep Learning*, pages 24–27. Association for Computational Linguistics.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005. Asian Federation of Natural Language Processing.

Aaron Steven White and Kyle Rawlins. 2018. The role of veridicality and factivity in clause selection. In *Proceedings of the 48th Annual Meeting of the North East Linguistic Society*.

Aaron Steven White, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2018. Lexicosyntactic inference in neural models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4717–4724. Association for Computational Linguistics.

Adina Williams, Andrew Drozdov, and Samuel R. Bowman. 2018a. Do latent tree learning models identify meaningful structure in sentences? *Transactions of the Association of Computational Linguistics*, 6:253–267.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018b. 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.

A MNLI examples that contradict the HANS heuristics

The sentences in (7) show examples from the MNLI training set that contradict the lexical overlap, subsequence, and constituent heuristics. The full set of all 261 contradicting examples in the MNLI training set may be viewed at https://github.com/tommccoy1/hans/blob/master/mnli_contradicting_examples.

(7) a. A subcategory of accuracy is consistency. $\nrightarrow$ Accuracy is a subcategory of consistency.
b. At the same time, top Enron executives were free to exercise their stock options, and some did. $\nrightarrow$ Top Enron executives were free to exercise.
c. She was chagrined at The Nation's recent publication of a column by conservative education activist Ron Unz arguing that liberal education reform has been an unmitigated failure. $\nrightarrow$ Liberal education reform has been an unmitigated failure.

B Templates

Tables 4, 5, and 6 contain the templates for the lexical overlap heuristic, the subsequence heuristic, and the constituent heuristic, respectively.

In some cases, a given template has multiple versions, such as one version where a noun phrase modifier attaches to the subject and another where the modifier attaches to the object. For clarity, we have only listed one version of each template here. The full list of templates can be viewed in the code on GitHub.[10]

C Fine-grained results

Table 7 shows the results by subcase for models trained on MNLI for the subcases where the correct answer is *entailment*. Table 8 shows the results by subcase for these models for the subcases where the correct answer is *non-entailment*.

D Results for models trained on MNLI with *neutral* and *contradiction* merged

Table 9 shows the results on HANS for models trained on MNLI with the labels *neutral* and *contradiction* merged in the training set into the single label *non-entailment*. The results are similar to the results obtained by merging the labels after training, with the models generally outputting *entailment* for all HANS examples, whether that was the correct answer or not.

[10] https://github.com/tommccoy1/hans

Subcase	Template	Example
Entailment: Untangling relative clauses	The N_1 who the N_2 V_1 V_2 the N_3 $\rightarrow$ The N_2 V_1 the N_1.	The athlete who the judges admired called the manager. $\rightarrow$ The judges admired the athlete.
Entailment: Sentences with PPs	The N_1 P the N_2 V the N_3 $\rightarrow$ The N_1 V the N_3	The tourists by the actor recommended the authors. $\rightarrow$ The tourists recommended the authors.
Entailment: Sentences with relative clauses	The N_1 that V_2 V_1 the N_2 $\rightarrow$ The N_1 V_1 the N_2	The actors that danced saw the author. $\rightarrow$ The actors saw the author.
Entailment: Conjunctions	The N_1 V the N_2 and the N_3 $\rightarrow$ The N_1 V the N_3	The secretaries encouraged the scientists and the actors. $\rightarrow$ The secretaries encouraged the actors.
Entailment: Passives	The N_1 were V by the N_2 $\rightarrow$ The N_1 V the N_2	The authors were supported by the tourists. $\rightarrow$ The tourists supported the authors.
Non-entailment: Subject-object swap	The N_1 V the N_2. $\nrightarrow$ The N_2 V the N_1.	The senators mentioned the artist. $\nrightarrow$ The artist mentioned the senators.
Non-entailment: Sentences with PPs	The N_1 P the N_2 V the N_3 $\nrightarrow$ The N_3 V the N_2	The judge behind the manager saw the doctors. $\nrightarrow$ The doctors saw the manager.
Non-entailment: Sentences with relative clauses	The N_1 V_1 the N_2 who the N_3 V_2 $\nrightarrow$ The N_2 V_1 the N_3	The actors advised the manager who the tourists saw. $\nrightarrow$ The manager advised the tourists.
Non-entailment: Conjunctions	The N_1 V the N_2 and the N_3 $\nrightarrow$ The N_2 V the N_3	The doctors advised the presidents and the tourists. $\nrightarrow$ The presidents advised the tourists.
Non-entailment: Passives	The N_1 were V by the N_2 $\nrightarrow$ The N_1 V the N_2	The senators were recommended by the managers. $\nrightarrow$ The senators recommended the managers.

Table 4: Templates for the lexical overlap heuristic

Subcase	Template	Example
Entailment: Conjunctions	The N_1 and the N_2 V the N_3 $\rightarrow$ The N_2 V the N_3	The actor and the professor mentioned the lawyer. $\rightarrow$ The professor mentioned the lawyer.
Entailment: Adjectives	Adj N_1 V the N_2 $\rightarrow$ N_1 V the N_2	Happy professors mentioned the lawyer. $\rightarrow$ Professors mentioned the lawyer.
Entailment: Understood argument	The N_1 V the N_2 $\rightarrow$ The N_1 V	The author read the book. $\rightarrow$ The author read.
Entailment: Relative clause on object	The N_1 V_1 the N_2 that V_2 the N_3 $\rightarrow$ The N_1 V_1 the N_2	The artists avoided the senators that thanked the tourists. $\rightarrow$ The artists avoided the senators.
Entailment: PP on object	The N_1 V the N_2 P the N_3 $\rightarrow$ The N_1 V the N_2	The authors supported the judges in front of the doctor. $\rightarrow$ The authors supported the judges.
Non-entailment: NP/S	The N_1 V_1 the N_2 V_2 the N_3 $\nrightarrow$ The N_1 V_1 the N_2	The managers heard the secretary encouraged the author. $\nrightarrow$ The managers heard the secretary.
Non-entailment: PP on subject	The N_1 P the N_2 V $\nrightarrow$ The N_2 V	The managers near the scientist resigned. $\nrightarrow$ The scientist resigned.
Non-entailment: Relative clause on subject	The N_1 that V_1 the N_2 V_2 the N_3 $\nrightarrow$ The N_2 V_2 the N_3	The secretary that admired the senator saw the actor. $\nrightarrow$ The senator saw the actor.
Non-entailment: MV/RR	The N_1 V_1 P the N_2 V_2 $\nrightarrow$ The N_1 V_1 P the N_2	The senators paid in the office danced. $\nrightarrow$ The senators paid in the office.
Non-entailment: NP/Z	P the N_1 V_1 the N_2 V_2 the N_3 $\nrightarrow$ The N_1 V_1 the N_2	Before the actors presented the professors advised the manager. $\nrightarrow$ The actors presented the professors.

Table 5: Templates for the subsequence heuristic

Subcase	Template	Example
Entailment: Embedded under preposition	P the N_1 V_1, the N_2 V_2 the N_3 $\rightarrow$ The N_1 V_1	Because the banker ran, the doctors saw the professors. $\rightarrow$ The banker ran.
Entailment: Outside embedded clause	P the N_1 V_1 the N_2, the N_3 V_2 the N_4 $\rightarrow$ The N_3 V_2 the N_4	Although the secretaries recommended the managers, the judges supported the scientist. $\rightarrow$ The judges supported the scientist.
Entailment: Embedded under verb	The N_1 V_1 that the N_2 V_2 $\rightarrow$ The N_2 V_2	The president remembered that the actors performed. $\rightarrow$ The actors performed.
Entailment: Conjunction	The N_1 V_1, and the N_2 V_2 the N_3. $\rightarrow$ The N_2 V_2 the N_3	The lawyer danced, and the judge supported the doctors. $\rightarrow$ The judge supported the doctors.
Entailment: Adverbs	Adv the N V $\rightarrow$ The N V	Certainly the lawyers resigned. $\rightarrow$ The lawyers resigned.
Non-entailment: Embedded under preposition	P the N_1 V_1, the N_2 V_2 the N_2 $\nrightarrow$ The N_1 V_1	Unless the senators ran, the professors recommended the doctor. $\nrightarrow$ The senators ran.
Non-entailment: Outside embedded clause	P the N_1 V_1 the N_2, the N_3 V_2 the N_4 $\nrightarrow$ The N_3 V_2 the N_4	Unless the authors saw the students, the doctors helped the bankers. $\nrightarrow$ The doctors helped the bankers.
Non-entailment: Embedded under verb	The N_1 V_1 that the N_2 V_2 the N_3 $\nrightarrow$ The N_2 V_2 the N_3	The tourists said that the lawyer saw the banker. $\nrightarrow$ The lawyer saw the banker.
Non-entailment: Disjunction	The N_1 V_1, or the N_2 V_2 the N_3 $\nrightarrow$ The N_2 V_2 the N_3	The judges resigned, or the athletes mentioned the author. $\nrightarrow$ The athletes mentioned the author.
Non-entailment: Adverbs	Adv the N_1 V the N_2 $\nrightarrow$ The N_1 V the N_2	Probably the artists saw the authors. $\nrightarrow$ The artists saw the authors.

Table 6: Templates for the constituent heuristic

Heuristic	Subcase	DA	ESIM	SPINN	BERT
Lexical overlap	Untangling relative clauses	0.97	0.95	0.88	0.98
	The athlete who the judges saw called the manager. → The judges saw the athlete.				
	Sentences with PPs	1.00	1.00	1.00	1.00
	The tourists by the actor called the authors. → The tourists called the authors.				
	Sentences with relative clauses	0.98	0.97	0.97	0.99
	The actors that danced encouraged the author. → The actors encouraged the author.				
	Conjunctions	1.00	1.00	1.00	0.77
	The secretaries saw the scientists and the actors. → The secretaries saw the actors.				
	Passives	1.00	1.00	0.95	1.00
	The authors were supported by the tourists. → The tourists supported the authors.				
Subsequence	Conjunctions	1.00	1.00	1.00	0.98
	The actor and the professor shouted. → The professor shouted.				
	Adjectives	1.00	1.00	1.00	1.00
	Happy professors mentioned the lawyer. → Professors mentioned the lawyer.				
	Understood argument	1.00	1.00	0.84	1.00
	The author read the book. → The author read.				
	Relative clause on object	0.98	0.99	0.95	0.99
	The artists avoided the actors that performed. → The artists avoided the actors.				
	PP on object	1.00	1.00	1.00	1.00
	The authors called the judges near the doctor. → The authors called the judges.				
Constituent	Embedded under preposition	0.99	0.99	0.85	1.00
	Because the banker ran, the doctors saw the professors. → The banker ran.				
	Outside embedded clause	0.94	1.00	0.95	1.00
	Although the secretaries slept, the judges danced. → The judges danced.				
	Embedded under verb	0.92	0.94	0.99	0.99
	The president remembered that the actors performed. → The actors performed.				
	Conjunction	0.99	1.00	0.89	1.00
	The lawyer danced, and the judge supported the doctors. → The lawyer danced.				
	Adverbs	1.00	1.00	0.98	1.00
	Certainly the lawyers advised the manager. → The lawyers advised the manager.				

Table 7: Results for the subcases where the correct label is *entailment*.

Heuristic	Subcase	DA	ESIM	SPINN	BERT
Lexical overlap	Subject-object swap	0.00	0.00	0.03	0.00
	The senators mentioned the artist. ↛ The artist mentioned the senators.				
	Sentences with PPs	0.00	0.00	0.01	0.25
	The judge behind the manager saw the doctors. ↛ The doctors saw the manager.				
	Sentences with relative clauses	0.04	0.04	0.06	0.18
	The actors called the banker who the tourists saw. ↛ The banker called the tourists.				
	Conjunctions	0.00	0.00	0.01	0.39
	The doctors saw the presidents and the tourists. ↛ The presidents saw the tourists.				
	Passives	0.00	0.00	0.00	0.00
	The senators were helped by the managers. ↛ The senators helped the managers.				
Subsequence	NP/S	0.04	0.02	0.09	0.02
	The managers heard the secretary resigned. ↛ The managers heard the secretary.				
	PP on subject	0.00	0.00	0.00	0.06
	The managers near the scientist shouted. ↛ The scientist shouted.				
	Relative clause on subject	0.03	0.04	0.05	0.01
	The secretary that admired the senator saw the actor. ↛ The senator saw the actor.				
	MV/RR	0.04	0.03	0.03	0.00
	The senators paid in the office danced. ↛ The senators paid in the office.				
	NP/Z	0.02	0.01	0.11	0.10
	Before the actors presented the doctors arrived. ↛ The actors presented the doctors.				
Constituent	Embedded under preposition	0.14	0.02	0.29	0.50
	Unless the senators ran, the professors recommended the doctor. ↛ The senators ran.				
	Outside embedded clause	0.01	0.00	0.02	0.00
	Unless the authors saw the students, the doctors resigned. ↛ The doctors resigned.				
	Embedded under verb	0.00	0.00	0.01	0.22
	The tourists said that the lawyer saw the banker. ↛ The lawyer saw the banker.				
	Disjunction	0.01	0.03	0.20	0.01
	The judges resigned, or the athletes saw the author. ↛ The athletes saw the author.				
	Adverbs	0.00	0.00	0.00	0.08
	Probably the artists saw the authors. ↛ The artists saw the authors.				

Table 8: Results for the subcases where the correct label is *non-entailment*.

		Correct: *Entailment*			Correct: *Non-entailment*		
Model	Model class	Lexical	Subseq.	Const.	Lexical	Subseq.	Const.
DA	Bag-of-words	1.00	1.00	0.98	0.00	0.00	0.03
ESIM	RNN	0.99	1.00	1.00	0.00	0.01	0.00
SPINN	TreeRNN	0.94	0.96	0.93	0.06	0.14	0.11
BERT	Transformer	0.98	1.00	0.99	0.04	0.02	0.20

Table 9: Results for models trained on MNLI with *neutral* and *contradiction* merged into a single label, *non-entailment*.

E Results with augmented training with some subcases withheld

For each model, we ran five experiments, each one having 6 of the 30 subcases withheld. Each trained model was then evaluated on the categories that had been withheld from it. The results of these experiments are in Tables 10, 11, 12, 13 and 14.

F Human experiments

To obtain human results, we used Amazon Mechanical Turk. We subdivided HANS into 114 different categories of examples, covering all possible variations of the template used to generate the example and the specific word around which the template was built. For example, for the constituent heuristic subcase of clauses embedded under verbs (e.g. *The doctor believed the lawyer danced* $\nrightarrow$ *The lawyer danced*), each possible verb under which the clause could be embedded (e.g. *believed*, *thought*, or *assumed*) counted as a different category.

For each of these 114 categories, we chose 20 examples from HANS and obtained judgments from 5 human participants for each of those 20 examples. Each participant provided judgments for 57 examples plus 10 controls (67 stimuli total) and was paid $2.00. The controls consisted of 5 examples where the premise and hypothesis were the same (e.g. *The doctor saw the lawyer* $\rightarrow$ *The doctor saw the lawyer*) and 5 examples of simple negation (e.g. *The doctor saw the lawyer* $\nrightarrow$ *The doctor did not see the lawyer*). For analyzing the data, we discarded any participants who answered any of these controls incorrectly; this led to 95 participants being retained and 105 being rejected (participants were still paid regardless of whether they were retained or filtered out). On average, each participant spent 6.5 seconds per example; the participants we retained spent 8.9 seconds per example, while the participants we discarded spent 4.2 seconds per example. The total amount of time from a participant accepting the experiment to completing the experiment averaged 17.6 minutes. This included 9.1 minutes answering the prompts (6.4 minutes for discarded participants and 12.1 minutes for retained participants) and roughly one minute spent between prompts (1 second after each prompt). The remaining time was spent reading the consent form, reading the instructions, or waiting to start (Mechanical Turk participants often wait several minutes between accepting an experiment and beginning the experiment).

The expert annotators were three native English speakers who had a background in linguistics but who had not heard about this project before providing judgments. Two of them were graduate students and one was a postdoctoral researcher. Each expert annotator labeled 124 examples (one example from each of the 114 categories, plus 10 controls).

Heuristic	Subcase	DA	ESIM	SPINN	BERT
Lexical overlap	Subject-object swap *The senators mentioned the artist.* $\nrightarrow$ *The artist mentioned the senators.*	0.01	1.00	1.00	1.00
Lexical overlap	Untangling relative clauses *The athlete who the judges saw called the manager.* $\rightarrow$ *The judges saw the athlete.*	0.34	0.23	0.23	0.20
Subsequence	NP/S *The managers heard the secretary resigned.* $\nrightarrow$ *The managers heard the secretary.*	0.27	0.00	0.00	0.10
Subsequence	Conjunctions *The actor and the professor shouted.* $\rightarrow$ *The professor shouted.*	0.49	0.38	0.38	0.38
Constituent	Embedded under preposition *Unless the senators ran, the professors recommended the doctor.* $\nrightarrow$ *The senators ran.*	0.51	0.51	0.51	1.00
Constituent	Embedded under preposition *Because the banker ran, the doctors saw the professors.* $\rightarrow$ *The banker ran.*	1.00	0.06	1.00	0.03

Table 10: Accuracies for models trained on MNLI augmented with most HANS example categories except withholding the categories in this table (experiment 1/5 for the withheld category investigation).

Heuristic	Subcase	DA	ESIM	SPINN	BERT
Lexical overlap	Sentences with PPs *The judge behind the manager saw the doctors.* $\nrightarrow$ *The doctors saw the manager.*	0.00	0.96	0.71	0.97
Lexical overlap	Sentences with PPs *The tourists by the actor called the authors.* $\rightarrow$ *The tourists called the authors.*	1.00	1.00	0.94	1.00
Subsequence	PP on subject *The managers near the scientist shouted.* $\nrightarrow$ *The scientist shouted.*	0.00	0.07	0.57	0.39
Subsequence	Adjectives *Happy professors mentioned the lawyer.* $\rightarrow$ *Professors mentioned the lawyer.*	0.71	0.99	0.64	1.00
Constituent	Outside embedded clause *Unless the authors saw the students, the doctors resigned.* $\nrightarrow$ *The doctors resigned.*	0.78	1.00	1.00	0.17
Constituent	Outside embedded clause *Although the secretaries slept, the judges danced.* $\rightarrow$ *The judges danced.*	0.78	0.78	0.78	0.97

Table 11: Accuracies for models trained on MNLI augmented with most HANS example categories except withholding the categories in this table (experiment 2/5 for the withheld category investigation).

Heuristic	Subcase	DA	ESIM	SPINN	BERT
Lexical overlap	Sentences with relative clauses	0.00	0.04	0.02	0.84
	The actors called the banker who the tourists saw. $\nrightarrow$ *The banker called the tourists.*				
Lexical overlap	Sentences with relative clauses	1.00	0.97	1.00	1.00
	The actors that danced encouraged the author. $\rightarrow$ *The actors encouraged the author.*				
Subsequence	Relative clause on subject	0.00	0.04	0.00	0.93
	The secretary that admired the senator saw the actor. $\nrightarrow$ *The senator saw the actor.*				
Subsequence	Understood argument	0.28	1.00	0.81	0.94
	The author read the book. $\rightarrow$ *The author read.*				
Constituent	Embedded under verb	0.00	0.00	0.05	0.98
	The tourists said that the lawyer saw the banker. $\nrightarrow$ *The lawyer saw the banker.*				
Constituent	Embedded under verb	1.00	0.94	0.98	0.43
	The president remembered that the actors performed. $\rightarrow$ *The actors performed.*				

Table 12: Accuracies for models trained on MNLI augmented with most HANS example categories except withholding the categories in this table (experiment 3/5 for the withheld category investigation).

Heuristic	Subcase	DA	ESIM	SPINN	BERT
Lexical overlap	Passives	0.00	0.00	0.00	0.00
	The senators were helped by the managers. $\nrightarrow$ *The senators helped the managers.*				
Lexical overlap	Conjunctions	0.05	0.51	0.52	1.00
	The secretaries saw the scientists and the actors. $\rightarrow$ *The secretaries saw the actors.*				
Subsequence	MV/RR	0.76	0.44	0.32	0.07
	The senators paid in the office danced. $\nrightarrow$ *The senators paid in the office.*				
Subsequence	Relative clause on object	0.72	1.00	0.99	0.99
	The artists avoided the actors that performed. $\rightarrow$ *The artists avoided the actors.*				
Constituent	Disjunction	0.11	0.29	0.51	0.44
	The judges resigned, or the athletes saw the author. $\nrightarrow$ *The athletes saw the author.*				
Constituent	Conjunction	0.99	1.00	0.74	1.00
	The lawyer danced, and the judge supported the doctors. $\rightarrow$ *The lawyer danced.*				

Table 13: Accuracies for models trained on MNLI augmented with most HANS example categories except withholding the categories in this table (experiment 4/5 for the withheld category investigation).

Heuristic	Subcase	DA	ESIM	SPINN	BERT
Lexical overlap	Conjunctions	0.00	0.44	0.00	0.08
	The doctors saw the presidents and the tourists. $\nrightarrow$ *The presidents saw the tourists.*				
Lexical overlap	Passives	0.00	0.00	0.00	0.00
	The authors were supported by the tourists. $\rightarrow$ *The tourists supported the authors.*				
Subsequence	NP/Z	0.00	0.10	0.18	0.57
	Before the actors presented the doctors arrived. $\nrightarrow$ *The actors presented the doctors.*				
Subsequence	PP on object	0.04	0.76	0.04	0.98
	The authors called the judges near the doctor. $\rightarrow$ *The authors called the judges.*				
Constituent	Adverbs	0.76	0.33	0.20	0.84
	Probably the artists saw the authors. $\nrightarrow$ *The artists saw the authors.*				
Constituent	Adverbs	0.66	1.00	0.59	0.96
	Certainly the lawyers advised the manager. $\rightarrow$ *The lawyers advised the manager.*				

Table 14: Accuracies for models trained on MNLI augmented with most HANS example categories except withholding the categories in this table (experiment 5/5 for the withheld category investigation).