Simple but Challenging: Natural Language Inference Models Fail on Simple Sentences

Cheng Luo[1], **Wei Liu**[2], **Jieyu Lin**[2], **Jiajie Zou**[2], **Ming Xiang**[3] and **Nai Ding**[1,2*]
[1]Zhejiang Lab / Hangzhou, China
[2]Zhejiang University / Hangzhou, China
[3]The University of Chicago / Chicago, United States
luo_cheng@zhejianglab.com, {liuweizju, jieyu_lin}@zju.edu.cn,
{jiajiezou, ding_nai}@zju.edu.cn, mxiang@uchicago.edu

Abstract

Natural language inference (NLI) is a task to infer the relationship between a premise and a hypothesis (e.g., entailment, neutral, or contradiction), and transformer-based models perform well on current NLI datasets such as MNLI and SNLI. Nevertheless, given the linguistic complexity of the large-scale datasets, it remains controversial whether these models can truly infer the relationship between sentences or they simply guess the answer via shallow heuristics. Here, we introduce a controlled evaluation set called *Simple Pair* to test the basic sentence inference ability of NLI models using sentences with syntactically simple structures. Three popular transformer-based models, i.e., BERT, RoBERTa, and DeBERTa, are employed. We find that these models fine-tuned on MNLI or SNLI perform very poorly on *Simple Pair* (< 35.4% accuracy). Further analyses reveal event coreference and compositional binding problems in these models. To improve the model performance, we augment the training set, i.e., MNLI or SNLI, with a few examples constructed based on *Simple Pair* ($\sim$ 1% of the size of the original SNLI/MNLI training sets). Models fine-tuned on the augmented training set maintain high performance on MNLI/SNLI and perform very well on *Simple Pair* ($\sim$100% accuracy). Furthermore, the positive performance of the augmented training models can transfer to more complex examples constructed based on sentences from MNLI and SNLI. Taken together, the current work shows that (1) models achieving high accuracy on mainstream large-scale datasets still lack the capacity to draw accurate inferences on simple sentences, and (2) augmenting mainstream datasets with a small number of target simple sentences can effectively improve model performance.

* Corresponding author: Nai Ding

1 Introduction

Natural language inference (NLI), also known as recognizing textual entailment (RTE), is a basic task to test the semantic inference ability of natural language processing (NLP) models (Cooper et al., 1996; Dagan et al., 2005; Poliak, 2020). The NLI task concerns the relationship between a pair of sentences, i.e., a premise and a hypothesis (Naik et al., 2018; Ravichander et al., 2019; Richardson et al., 2020; Jeretic et al., 2020). In recent years, a number of datasets have been developed to train models for the NLI task, such as Stanford NLI (SNLI) (Bowman et al., 2015) and Multi-genre NLI (MNLI) (Williams et al., 2018), and transformer-based deep neural network models have achieved high accuracy on these datasets (Nangia and Bowman, 2019; Poliak, 2020). The high accuracy of NLI models could be taken to suggest that these models already have the ability to interpret the meaning of sentences and generate semantic inference. Nevertheless, recent evidence shows that NLI models may have just guessed the answer based on statistical biases hidden in the datasets (Gururangan et al., 2018; Clark et al., 2019). It also has been shown that models can achieve high accuracy even when the words in premise/hypothesis are shuffled (Sinha et al., 2021), casting further doubts on whether the NLI models can truly infer the meaning of sentence pairs or simply guess the answer via shallow heuristics (Naik et al., 2018).

To understand the true capacity of the current models, one reasonable approach is to generate more complex cases to break the shallow heuristics and accordingly identify the model defects. There is a growing body of recent NLI work that constructs syntactically/semantically sophisticated material for NLI datasets (Welleck et al., 2019; Nie et al., 2020; Liu et al., 2021). Training and testing models on difficult and challenging material are valuable since this exercise pushes the boundaries

of how much NLI models can cope with linguistic complexity (Nie et al., 2020; Ravichander et al., 2019). However, the complexity of the datasets could also potentially hinder an explicit picture of what specific linguistic features the models can learn and more importantly what they cannot learn. Furthermore, the focus on complex material implicitly assumes that the current NLI models have the capacity to understand simple sentences and consequently perform the NLI task accurately on simple sentences.

In this work, departing from the common practice of constructing complex material, we introduce a controlled evaluation set called *Simple Pair*, which includes a large number of syntactically/semantically simple sentences following a set of systematic design features. The goal of the current study is two-fold. First, we ask whether the current NLI models have the ability to correctly infer the relationship between simple sentences in *Simple Pair*. If not, the failure patterns on these simple cases can more effectively help us identify the basic linguistic operation(s) that the current models fail to capture, and illuminate shortcomings from inappropriate model biases. Second, we ask whether the weakness of the models can be overcome using simple training sentences constructed based on *Simple Pair*. If so, the seemingly basic linguistic information provided by these simple cases can serve as an important supplement for the existent datasets, and robustly improve the model performance on NLI tasks.

To preview, we tested three popular transformer-based models, i.e., BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021), which were respectively fine-tuned on 2 widely-used datasets, i.e., the MNLI and the SNLI datasets. We found that these models were by and large inaccurate in drawing inference relations on our datasets, indicating severe model problems such as event coreference biases and compositional binding failures. To address these problems, we fine-tuned each model on the MNLI or SNLI augmented with a few samples constructed based on sentences in *Simple Pair*. The small number of samples can indeed significantly improve model performance on *Simple Pair*, and the positive improvement can extend to more complex and challenging cases.

2 Methods

2.1 NLI Dataset and Pre-trained Models

We employed three pre-trained language models, i.e., BERT, RoBERTa, and DeBERTa to perform the NLI task. For all models, we used both the base (b) and large (l) versions. We built our models using Huggingface (Wolf et al., 2020). The models were separately fine-tuned based on 2 mainstream datasets, i.e., MNLI and SNLI. For the 2 datasets we used, the relationship between a premise and a hypothesis could be entailment, contradiction, or neutral. The accuracy was evaluated by the proportion of premise-hypothesis pairs for which the inference relation was correctly identified. The parameters for fine-tuning were adopted from previous studies, and the test accuracy was higher than 83.9% (shown in Appendix Table 1). For each sentence pair, the input to the models was [CLS, premise, SEP, hypothesis, SEP]. The concatenated sequence was encoded through the models and the output embedding of CLS was fed into a 3-way softmax classifier. The classifier calculated a score for each class through a linear transformer matrix and softmax function (Devlin et al., 2019).

2.2 Dataset Construction

2.2.1 *Simple Pair* set

To test the basic sentence inference ability of NLI models, we constructed a *Simple Pair* test set using syntactically simple sentences as shown in Figure 1. The test set was further divided into a simple-sentence set and a conjunction-sentence set. For the simple-sentence set, the premise was a short sentence constructed using one of two templates (see Figure 1). One template created N-is-A sentences, where [N] was a noun and [A] was an adjective. The noun was selected from 5 categories, i.e., fruits (N = 40), animals (N = 90), human (N = 100), names (N = 100), and objects (N = 90), and each noun was mapped to a compatible adjective (N = 25, 30, 55, 55, and 28 for nouns from the fruit, animal, human, name, and object categories, respectively). The other template created SVO sentences. The subject and object were selected from the same 5 categories of nouns used in N-is-A sentences, and they were randomly paired with a compatible verb (N = 20). Following the templates in Figure 1, each premise was then paired with a number of hypotheses that all have a *neutral* relationship with the premise. In particular,

N-is-A simple-sentence templates

premise	hypothesis
The N_1 is A_1. *The apple is expensive.*	**The N_2 is (not) A_1.** *The pear is (not) expensive.* **The N_1 is (not) A_2.** *The apple is (not) sweet.* **The N_2 is (not) A_2.** *The pear is (not) sweet.* **premise-relevant/irrelevant** **The N_1 is A_1 + Hypothesis (The N_2 is A_2 for example)** *The apple is expensive and the pear is sweet.* **The N_3 is A_3 + Hypothesis (The N_2 is A_2 for example)** *The banana is popular and the pear is sweet.*

N-is-A conjunction-sentence templates

premise	hypothesis
The N_1 is A_1. The N_2 is A_2. *The apple is expensive. The pear is sweet.*	
The N_1 is not A_1. The N_2 is A_2. *The apple is not expensive. The pear is sweet.*	**The N_2 is (not) A_1.** *The pear is (not) expensive.*
The N_1 is A_1 and the N_2 is A_2. *The apple is expensive and the pear is sweet.*	**The N_1 is (not) A_2.** *The apple is (not) sweet.*
The N_1 is A_1 and the N_2 is not A_2. *The apple is expensive and the pear is not sweet.*	

SVO simple-sentence templates

premise	hypothesis
The S_1 V_1 the O_1. *The student saw the dog.*	**The S_2 (did not) V_1 the O_1.** *The professor saw/did not see the dog.* **The S_1 (did not) V_2 the O_1.** *The student lost/did not lose the dog.* **The S_1 (did not) V_1 the O_2.** *The student saw/did not see the key.* **The O_1 (did not) V_1 the S_1.** *The dog saw/did not see the student.* **premise-relevant/irrelevant** **The S_1 V_1 O_1 + Hypothesis (S_2 V_1 O_1 for example)** *The student saw the dog and the professor saw the dog.* **The S_3 V_3 O_3 + Hypothesis (S_2 V_1 O_1 for example)** *The kid played the ball and the professor saw the dog.*

SVO conjunction-sentence templates

premise	hypothesis
The S_1 V_1 O_1. The S_2 V_2 O_2. *The student saw the dog. The professor lost the key.*	
The S_1 did not V_1 O_1. The S_2 V_2 O_2. *The student did not see the dog. The professor lost the key.*	**The S_2 (did not) V_1 the O_1.** *The professor saw/did not see the dog.*
The S_1 V_1 O_1 and the S_2 V_2 O_2. *The student saw the dog and the professor lost the key.*	**The S_1 (did not) V_2 the O_2.** *The student lost/did not lose the key.*
The S_1 V_1 O_1 and the S_2 did not V_2 O_2. *The student saw the dog and the professor did not lose the key.*	

Figure 1: Construction of the *Simple Pair* set.

each N-is-A type of premise was paired with 6 hypotheses (3 affirmative sentences and 3 negative sentences), and 24000 premise-hypothesis pairs (4000 premises × 6 hypotheses) were created in total. Each SVO premise was paired with 8 hypotheses (4 affirmative sentences and 4 negative sentences), and 32000 premise-hypothesis pairs (4000 premises × 8 hypotheses) were created. No premise-hypothesis pair in the simple-sentence set contained antonyms or synonyms.

In addition to the above neutral premise-hypothesis pairs, to test the event coreference bias of models fine-tuned on MNLI or SNLI, we introduced premise-relevant hypotheses into the simple-sentence set to create a condition where a part of the information of the hypothesis was in line with its premise. As is shown in Figure 1, the premise-relevant hypotheses were constructed by conjoining the original hypothesis with its premise. As a contrary condition, we also created premise-irrelevant hypotheses by conjoining the original hypothesis with a new sentence which was irrelevant with its premise. The two sentences in the premise-relevant/premise-irrelevant hypotheses were conjoined together in a random order, with or without the word "*and*". This procedure resulted in 12000 premise-hypothesis pairs (1000 premise × 6 hypotheses × premise-relevant/premise-irrelevant cases) for N-is-A type and 16000 premise-hypothesis pairs (1000 premise × 8 hypotheses × premise-relevant/premise-irrelevant cases) for SVO type. For all these pairs, the relationship between each premise-hypothesis pair is also in principle *neutral*.

For the conjunction-sentence set, the premise was constructed by conjoining two simple sentences using one of four possible templates (see Figure 1). Each premise was paired with 4 hypotheses (2 affirmative sentences and 2 negative sentences). In total, 16000 premise-hypothesis pairs (4000 premises × 4 hypotheses) were created for the premise constructed using each template. Similar to the simple-sentence set, the relationship between all premise-hypothesis pairs was controlled as being *neutral*. Human annotation was acquired for a part of samples in *Simple Pair* to confirm the neutral relationship between premise-hypothesis pairs (see section 2.3).

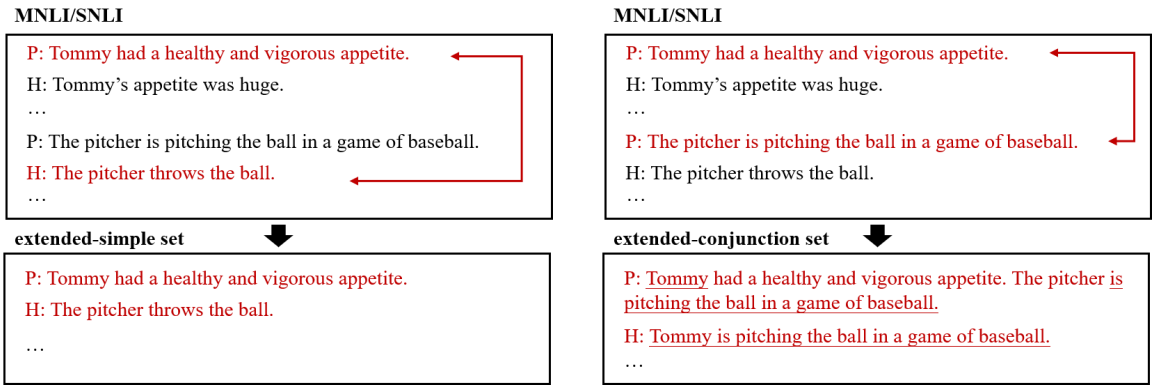


Figure 2: Construction of the *Extended Pair* set.

2.2.2 *Extended Pair* set

To test the generalization ability of models fine-tuned on augmented MNLI/SNLI, we created an *Extended Pair* test set using more complex sentences originating from MNLI and SNLI. The test set was also divided into an extended-simple set and an extended-conjunction set (see Figure 2). For the extended-simple set, we randomly paired premises and hypotheses in MNLI and SNLI test sets, with the constraint that none of the new premise-hypothesis pairs in our test set overlapped with the pairs in the original datasets. Specifically, 2000 premises were selected (500 from the MNLI-matched, 500 from MNLI-mismatched, and 1000 from SNLI), and each premise was paired with 3 hypotheses (1 from MNLI-matched, 1 from MNLI-mismatched, and 1 from SNLI). This procedure resulted in 6000 premise-hypothesis pairs (2000 premises × 3 hypotheses) in total. Since the pairing between a premise and a hypothesis is random, the relationship between them should be *neutral*.

For the extended-conjunction set, we randomly selected 60 irrelevant premise sentences from MNLI and SNLI test sets (15 from MNLI-matched, 15 from MNLI-mismatched, and 30 from SNLI), with the constraint that the subject was not a pronoun in each sentence. Following the conjunction templates of *Simple Pair*, the premise was constructed by randomly conjoining 2 of the 60 sentences, and the hypotheses were created by breaking the compositional binding relation between a subject and a predicate in the premise (see Figure 2). This procedure resulted in 6000 premise-hypotheses pairs (375 premises × 4 hypotheses × 4 templates) in total. Like the extended-simple set, we expected the relationship for the premise-hypothesis pairs in the extended-conjunction set to be *neutral* as well. Human annotation was also acquired for a part of samples in *Extended Pair* to confirm the neutral relationship between premise-hypothesis pairs (see section 2.3).

2.3 Human Annotation

A large number of hypotheses in our datasets were identified as entailment or contradiction by the models fine-tuned on MNLI and SNLI (see Results). To test whether most of these premise-hypothesis pairs were truly neutral as we expected, we collected human annotations for part of the data. In total, we randomly selected 200 premise-hypothesis pairs from *Simple Pair* (50 from the simple-sentence set, 50 from the conjunction-sentence set, 50 from the premise-relevant set, and 50 from the premise-irrelevant set), and 100 premise-hypothesis pairs from *Extended Pair* (50 from the extended-simple set, and 50 from the extended-conjunction set). These premise-hypothesis pairs were listed in Supplementary Materials.

Five human annotators were presented with the pairs of sentences and asked to label the relationship between the two sentences, i.e., entailment, contradiction, or neutral. Since the annotation guideline might affect annotators' decisions in the annotation process (Bowman et al., 2015; Glockner et al., 2018; Gururangan et al., 2018), we directly used premise-hypothesis pairs from MNLI and SNLI as examples for the annotators. The examples presented 9 premise-hypothesis pairs (3 premises × E/N/C hypotheses) randomly selected from MNLI and SNLI sets, respectively. For quality control, we also mixed 4 non-neutral examples (2 entailment and 2 contradiction) into the samples of each test set. All five annotators correctly identified these samples. After the annotation, the ground truth label was obtained using a majority vote from

the five annotators. The premise-hypothesis pairs from *Simple Pair* and *Extended Pair* were more frequently classified as neutral by human annotators. Appendix Table 2 shows the summary statistics of ground truth labels.

3 Results

3.1 Model Performance on *Simple Pair*

For the *Simple Pair* set, we constructed premise-hypothesis pairs using syntactically simple sentences and simplified the relationship between the premise and hypothesis by making them *neutral*. However, all models fine-tuned on MNLI or SNLI failed to correctly infer the relationship between the premise-hypothesis pairs. The model performance on *Simple Pair* is shown in Tables 1 and 3, for the simple-sentence set and the conjunction-sentence set, respectively.

For the simple-sentence set, we constructed neutral hypotheses by replacing at least one constituent in the premise (e.g., [N] or [A] in a N-is-A sentence) with a different word. As is shown in Table 1, models fine-tuned on MNLI or SNLI performed poorly on the simple-sentence set (< 28.3% accuracy). It was found that these models identified the relationship between a large proportion of premise-hypothesis pairs as contradiction, especially when the subjects were different between the hypothesis and the premise. For example, the models judged that "*The apple is expensive*" contradicts "*The banana is expensive*". Similarly, the model judged that "*The professor saw the dog*" contradicts "*The student saw the dog*".

Previous works concerning SNLI and MNLI datasets consistently mentioned the issue of event coreference, which could confound neutral and contradictory relationships between premise-hypothesis pairs (Bowman et al., 2015; Williams et al., 2018). It is possible that the consistent model bias for "contradiction" on our simple-sentence set might be attributed to the bias of event coreference originating from SNLI and MNLI. To test this possibility, we introduced premise-relevant hypotheses and premise-irrelevant hypotheses into the simple-sentence set (see Methods for details). For the premise-relevant hypothesis, the sentence described the same event as its premise, with the addition of irrelevant information from the original hypothesis, e.g., the premise "*The apple is expensive*" was paired with the hypothesis "*The apple is expensive and the orange is juicy*". For the premise-irrelevant hypothesis, in contrast, the sentence described an event totally irrelevant with the premise, e.g., the premise "*The apple is expensive*" was paired with the hypothesis "*The banana is sweet and the orange is juicy*". As is shown in Table 2, it was found that the model accuracy was significantly increased when the original hypothesis was replaced by a premise-relevant hypothesis rather than a premise-irrelevant hypothesis. The results indicated that models fine-tuned on SNLI or MNLI had a severe event coreference bias: Only when the premise and hypothesis contained the same event could the neutral hypotheses be correctly identified.

For the conjunction-sentence set, we constructed neutral hypotheses by breaking the compositional binding relation between a subject and a predicate in the premise. As is shown in Table 3, the models fine-tuned on MNLI or SNLI performed poorly on the conjunction-sentence set (< 35.4% accuracy). Similar to the simple-sentence set, the DeBERTa models identified a large proportion of these unrelated statements as being contradictory. In addition, the BERT and RoBERTa models also revealed a new problem. They failed to understand the fundamental compositional binding relation between a subject and a predicate. For example, the models consistently made the incorrect judgment that "*The apple is expensive and the orange is sweet*" entails "*The apple is sweet*". This suggests that the models are confused as to which subject should be paired with which predicate (i.e. the compositional binding failure). The models also judged the same premise to contradict "*The apple is not sweet*", again suggesting a composition problem: Once the models had wrongly allowed the composition of "*The apple is sweet*" based on the premise, this inference would now be in contradiction to the hypothesis "*The apple is not sweet*", assuming that the models have the ability to distinguish "*sweet*" and "*not sweet*" as describing two opposite properties.

We also introduced negation into the premises to test if models could bind "*not*" with a positive predicate to form a more complex predicate. These conditions again revealed the composition failure problem on the BERT and RoBERTa models (see Table 3). For example, when the premise was "*The apple is expensive and the orange is not sweet*", the models tended to judge that the premise entailed "*The apple is not sweet*" but contradicted

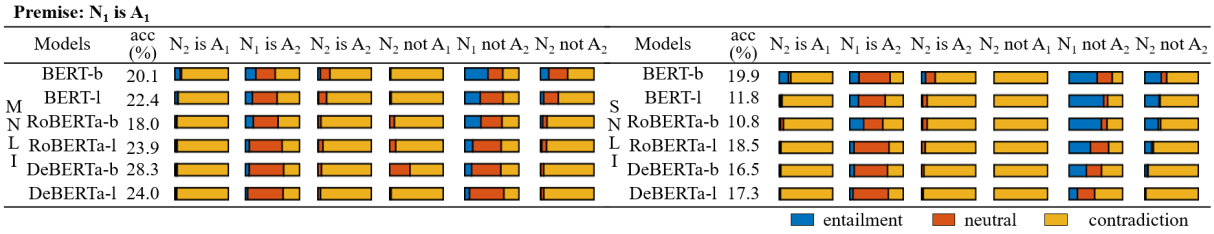


Table 1: Model performance on the simple-sentence set in *Simple Pair*. In *Simple Pair*, each premise is paired with a few hypotheses and each hypothesis is shown in a column. The percent of premise-hypothesis pairs identified as entailment, neutral, and contradiction were shown in blue, red, and yellow, respectively. This table only shows the results for N-is-A sentences and the results for SVO sentences are shown in the Appendix Table 3.

	Models	N-is-A relevant	N-is-A irrelevant	SVO relevant	SVO irrelevant
MNLI	BERT-b	97.1 (77.0)	37.5 (17.4)	68.0 (56.9)	31.5 (20.4)
	BERT-l	97.7 (75.3)	27.7 (5.3)	64.4 (53.9)	24.0 (13.5)
	RoBERTa-b	94.2 (76.2)	23.4 (5.4)	62.0 (50.9)	20.6 (9.5)
	RoBERTa-l	95.8 (71.9)	23.2 (-0.7)	66.2 (52.8)	12.6 (-0.8)
	DeBERTa-b	89.3 (61.0)	21.2 (-7.1)	63.2 (37.9)	13.9 (-11.4)
	DeBERTa-l	86.6 (62.6)	18.6 (-5.4)	66.0 (46.8)	8.7 (-10.5)
SNLI	BERT-b	63.8 (43.9)	25.0 (5.1)	52.7 (43.0)	27.4 (17.7)
	BERT-l	63.8 (52.0)	12.0 (0.2)	53.1 (42.0)	11.3 (0.2)
	RoBERTa-b	50.9 (40.1)	13.7 (2.9)	28.2 (16.9)	8.8 (-2.5)
	RoBERTa-l	49.4 (30.9)	12.4 (-6.1)	38.8 (24.7)	10.3 (-3.8)
	DeBERTa-b	51.5 (35.0)	6.6 (-9.9)	41.8 (26.8)	5.4 (-9.6)
	DeBERTa-l	54.7 (37.4)	5.8 (-11.5)	43.6 (26.5)	3.3 (-13.8)

Table 2: Model performance on simple-sentence set when the original hypothesis was replaced by premise-relevant hypothesis or premise-irrelevant hypothesis. The numbers in the parenthesis show the change in performance compared with the model performance on the original simple-sentence set.

"*The orange is not expensive*".This suggests that the models can correctly combine "*not*" with "*sweet*" to form a new predicate, but they still freely (and wrongly) paired up the subject nouns and the predicates in the premise.

3.2 Improving the Model Performance using *Simple Pair*

To recap, the test on *Simple Pair* identified severe limitations with the models fine-tuned on MNLI or SNLI, i.e., these models demonstrated substantial event coreference bias and compositional binding problem. We next investigated whether the *Simple Pair* set could be used to improve the performance of models fine-tuned on MNLI or SNLI. Specifically, we fine-tuned each model on the MNLI or SNLI training set augmented with a set constructed based on *Simple Pair* but containing no identical samples that appeared in *Simple Pair*. The label distribution of these samples was balanced, i.e., we also created entailment and contradictory hypotheses in the augmented set (see Appendix Table 5). Our additions comprised 6000 examples, roughly 1.5% and 1% of the size of the original MNLI and SNLI training set. The parameters are shown in Appendix Table 1. In general, all models maintained high performance. Some of the models, e.g., RoBERTa-large and DeBERTa-large, even got better performance on MNLI and SNLI test sets (see Appendix Table 1). The performance of the models receiving an augmented fine-tuning process is shown in Table 4. It was found that the small number of samples structured based on *Simple Pair* can significantly improve model performance on *Simple Pair* (close to 100% accuracy).

The positive results of the augmented fine-tuning process are compatible with the possibility that the models simply memorized the template of *Simple Pair*. Therefore, we created an *Extended Pair* set to test the generalization ability for the models fine-tuned on augmented MNLI/SNLI. In *Extended Pair*, all premise-hypothesis pairs were constructed using more complex sentences randomly selected from MNLI and SNLI. The relationship between each premise-hypothesis pair was controlled to be neutral, and they were designed in such as a way to also induce the event coreference bias and compositional binding problem. The model performance on *Extended Pair* is shown in Table 5.

For the extended-simple set, a premise was paired with a randomly chosen hypothesis, and therefore most of these premise-hypothesis pairs would not describe the same entity or event. As expected, the models fine-tuned on MNLI or SNLI inaccurately identified a large proportion of premise-hypothesis pairs as contradiction. The error rate of these models was over 32.1%. The performance of the models fine-tuned on augmented MNLI or SNLI was generally improved. The exceptions were that the DeBERTa-base model fine-tuned

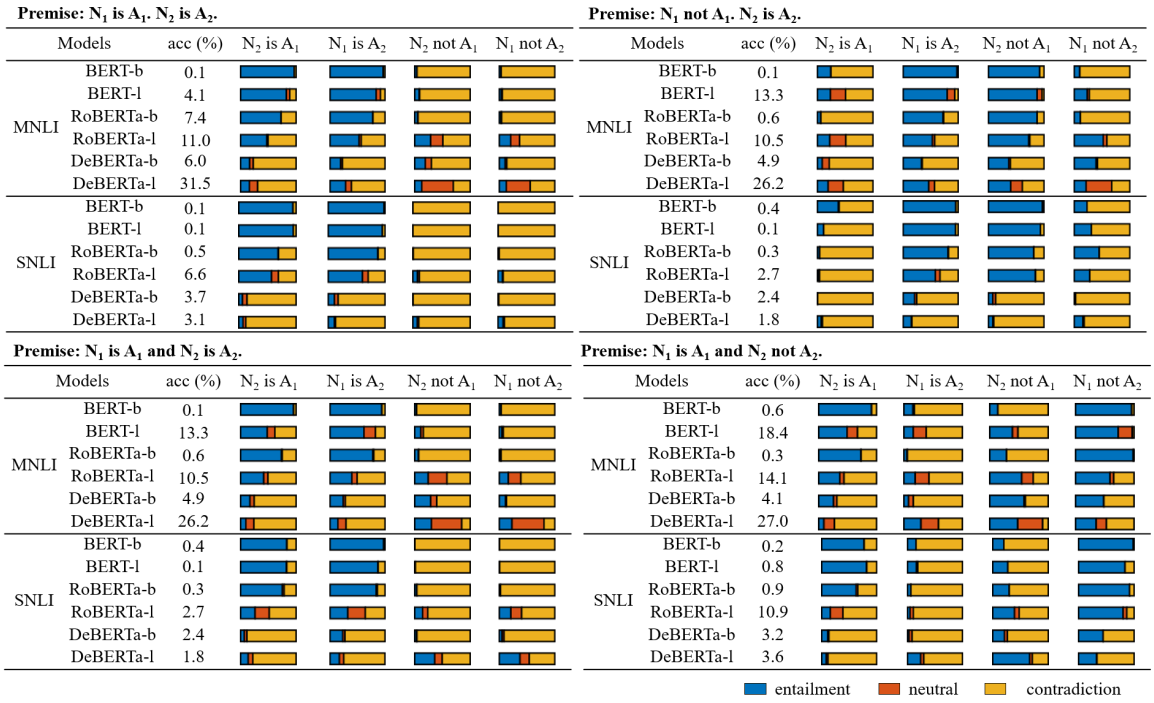


Table 3: Model performance on the conjunction-sentence set of *Simple Pair*. This table only shows the results for N-is-A sentences and the results for SVO sentences are shown in the Appendix Table 4.

	Models	simple-sentence set N-is-A	simple-sentence set SVO	conjunction-sentence set N-is-A	conjunction-sentence set SVO
MNLI	BERT-b	99.5 (79.4)	84.3 (73.2)	99.6 (99.4)	99.6 (99.5)
	BERT-l	98.5 (76.1)	87.3 (76.8)	99.7 (88.1)	99.9 (98.1)
	RoBERTa-b	97.3 (79.4)	90.4 (79.3)	99.7 (99.0)	99.9 (98.7)
	RoBERTa-l	94.8 (70.9)	86.0 (72.6)	99.4 (85.9)	99.9 (90.3)
	DeBERTa-b	97.4 (69.1)	92.1 (66.8)	99.8 (94.5)	99.9 (97.7)
	DeBERTa-l	98.4 (74.5)	92.8 (73.6)	99.6 (69.6)	99.9 (79.9)
SNLI	BERT-b	96.5 (76.7)	67.0 (57.3)	99.5 (99.3)	99.8 (99.2)
	BERT-l	98.9 (87.1)	83.0 (71.9)	99.7 (99.4)	99.9 (99.8)
	RoBERTa-b	97.5 (86.7)	96.3 (85.0)	99.6 (98.8)	99.9 (99.5)
	RoBERTa-l	96.8 (78.3)	94.5 (80.4)	99.6 (89.2)	99.9 (93.3)
	DeBERTa-b	98.7 (82.2)	97.0 (82.0)	99.7 (96.6)	99.9 (98.0)
	DeBERTa-l	98.8 (81.5)	97.9 (80.8)	99.7 (94.7)	99.9 (96.3)

Table 4: Performance of models fine-tuned on MNLI or SNLI augmented with *Simple Pair*. The numbers in the parenthesis show the change in performance compared with the models only fine-tuned on MNLI or SNLI.

on augmented MNLI, and BERT/RoBERTa-base models fine-tuned on augmented SNLI performed worse on the extended-simple set.

For the extended-conjunction set, each premise was created by randomly conjoining two sentences from MNLI and SNLI, and the neutral hypothesis was created by breaking the compositional binding relation between a subject and a predicate in the premise. The models fine-tuned on MNLI or SNLI failed on the extended-conjunction set. The error rate of these models was over 85.1%. However, all models fine-tuned on augmented MNLI or SNLI significantly improved in performance. Compared with the models fine-tuned on MNLI and SNLI, the improvement of accuracy rate was up to 56.9% and 47.7% for the models fine-tuned on augmented MNLI and SNLI respectively.

We also expected that the augmented fine-tuning process could enhance the basic inference capacity of the NLI models and generalize to samples with other syntactically simple structures. To further evaluate the generalization ability of the augmented training models, we used an NLI diagnostic dataset, called HANS (McCoy et al., 2019). The HANS dataset probed various syntactic heuristics from the superficial similarity (i.e., word overlap) between the premise and hypothesis. Therefore, the HANS dataset was similar to the *Simple Pair* and *Extended Pair* sets in the property of word overlap, and its samples with diverse syntactic structures were appropriate to evaluate the generalization ability of the augmented training models. Three nested heuristics, i.e., the lexical overlap, the subsequence, and the constituent heuristics, were measured in HANS. Given that the models fine-tuned on MNLI or SNLI had achieved high accuracy on the lexical overlap set (up to 98.5% accuracy), we employed the subsequence and constituent sets to evaluate the augmented training models. In the evaluation process, we collapsed the model outputs of neutral and contradiction labels into a single non-entailment label, following McCoy et al. (2019).

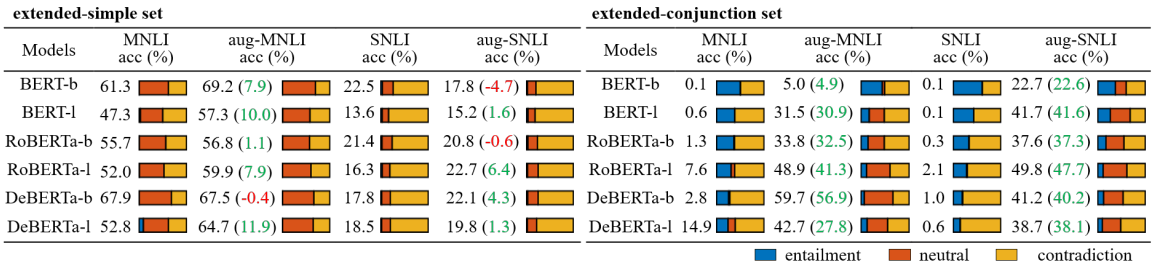


Table 5: Model Performance on *Extended Pair* set. The numbers in the parenthesis show the change in performance comparing the models fine-tuned on augmented MNLI or SNLI with the models only fine-tuned on MNLI or SNLI.

The model performance is shown in Appendix Table 6. Through the augmented fine-tuning process, the model performance was generally improved on the subsequence and constituent sets of HANS (up to a 13.3% increase).

4 Related work

Transformer-based models have achieved human-level performance on many NLI datasets such as MNLI and SNLI (Devlin et al., 2019; Lan et al., 2019; Liu et al., 2019; Nangia and Bowman, 2019). The good performance seems to suggest that these models have the ability to interpret sentences in the current datasets and generate correct inferences. Accordingly, follow-up works aim at constructing even more challenging datasets to train and test the models (Nie et al., 2020; Liu et al., 2021). There is also a growing body of works that constructs datasets to test more fine-grained linguistically motivated inference patterns such as pragmatic inferences and numerical reasoning (Jeretic et al., 2020; Ravichander et al., 2019) or correlates model errors with well-defined linguistic phenomena (Yanaka et al., 2019; Geiger et al., 2020; Yanaka et al., 2020; Hossain et al., 2022), with the purpose to identify whether models have trouble making certain types of inferences. Compared with these studies, the current work take a different approach: By intentionally reducing the difficulty of the test material, we aim to uncover whether models can truly infer the relation between simple sentences. The results show models perform poorly inferring the relation between basic N-is-A and SVO sentences.

The current work differs from previous studies in two major aspects. First, we constructed a large set of simple sentences, i.e., *Simple Pair*, to test and enhance models. Most current datasets are composed of syntactically complicated sentences and it is usually difficult to isolate specific linguistic constructs from these sentences (Naik et al., 2018). In our study, the sentences are simple enough so that the mechanisms to understand (or fail to understand) them are relatively transparent. Second, we extended the current mainstream datasets, i.e., MNLI and SNLI, to test the generalization ability of models. In *Extended Pair*, the original premise-hypothesis pairs in MNLI/SNLI are broken and recombined in a random way. It is an effective method to tackle the issue of potential statistical biases in NLI datasets, since most heuristics originating in the original datasets are rendered useless under the new test conditions where all the sentences are unrelated. Relatedly, the study of Wang et al. (2019) switched the premise and hypothesis, and used the switched pairs to test NLI models. Our method can be combined with the method by Wang et al. (2019) to further reduce the inherent statistical biases in NLI datasets.

Many studies have discussed the potential risk of overfitting on benchmark datasets, and emphasized the need to more accurately evaluate the true language capacity of various models (Smith, 2012; Talman and Chatzikyriakidis, 2019; Sinha et al., 2021; Poliak, 2020). For example, it has been shown that models can guess the relationship between a premise and a hypothesis with an accuracy higher than the chance level, even when just considering the hypothesis (Gururangan et al., 2018; Poliak et al., 2018). Here, by creating premise-hypothesis pairs characterized by neutral relationship, we provide additional evidence that existing models are severely over-fitted: (1) All models tend to judge the relationship between two unrelated simple sentences to be contradictory, which suggests the event coreference bias, and (2) some of them have substantial difficulty solving the compositional binding relations for conjunction sentences.

Regarding the event coreference bias, many studies have mentioned the event coreference problem in NLI tasks (Bowman et al., 2015; Williams et al., 2018; Glockner et al., 2018; Storks et al., 2019). Consider the sentence pair "*A boat sank in the*

Pacific Ocean" and "*A boat sank in the Atlantic Ocean*" as an example. The pair could be labeled as a contradiction if one assumes that the two sentences refer to the same single event, but could also be reasonably labeled as neutral if they are two independent events. For the SNLI set, the human annotators were instructed to judge the relation between sentences given that the two sentences describe the same scenario (Bowman et al., 2015). Hence, sentences that described different entities or events should be considered as contradiction by human annotators. For the MNLI set, despite no strict restrictions for a specific scenario between premise and hypothesis in each sample, it is still possible that the annotators adopted a similar annotation strategy in MNLI (Williams et al., 2018). Therefore, the coreference bias is regarded as an inherent problem in models fine-tuned on SNLI or MNLI, and no studies, to our knowledge, have tried to address this problem. In this work, we show that augmenting SNLI or MNLI with a few samples from *Simple Pair* can attenuate the coreference bias in these models. Regarding the compositional binding problem, it is surprising that large pre-trained models, e.g., BERT and RoBERTa, failed to deal with the fundamental compositional binding relation between a subject and a predicate. It is possible that the compositional failures we observed are also attributed to the inherent biases originating from MNLI and SNLI, given the model performance on conjunction sentences can be significantly improved by the augmented fine-tuning process.

5 Conclusion

In summary, since existing models have shown good performance on large-scale NLI datasets, the received wisdom is that these models are capable of doing at least some sophisticated inferences, and more progress can be made by evaluating them on even more challenging and complex datasets. The current work, however, shows that models achieving good performance on large-scale datasets do not necessarily generalize to simpler datasets. In fact, models fine-tuned on MNLI or SNLI generally have lower than chance level performance when inferring the relationship between simple sentences. Nevertheless, the results here show that combining a few simple examples with large-scale datasets, e.g., MNLI and SNLI, can significantly increase the model's ability to deal with simple test samples while largely maintaining the performance on original test samples. The positive results on simple test samples can also robustly transfer to improving the model accuracy on more complex samples. These results indicated that, in addition to more complex material, simple and transparent material, such as *Simple Pair*, can also serve as a tool for motivating and measuring progress in NLI tasks.

6 Limitations

In our test sets, we tried to ensure each premise-hypothesis pair has a neutral relation. One caveat is that the results of human classification (Appendix Table 2) showed that the current manipulation did not completely exclude entailment or contradictory samples in the *Simple Pair* and *Extended Pair* sets. But we note that it is unlikely that the small amount of entailment and contradictory samples in the test sets could account for the severe inaccuracy of NLI models, and we therefore did not employ more controls on the *Simple Pair* or *Extended Pair* sets. Overall, the current work mainly revealed the effect of some general biases when the NLI models were applied to deal with simple premise-hypothesis pairs characterized by neutral relationships. Future work could focus on the NLI model performance on simple sentences characterized by entailment or contradictory relationships.

Through our augmented fine-tuning process, the model performance was generally improved on the *Simple Pair* and *Extended Pair* sets. However, the performance improvement on the *Extended Pair* set was smaller than that on the *Simple Pair* set (Table 4 vs. Table 5). We argued that augmenting MNLI/SNLI with samples from *Simple Pair* was an effective way to attenuate shallow heuristics, but it may not have successfully dealt with deeper biases (for instance the event coreference bias) originated from the MNLI/SNLI. To achieve more robust performance on NLI tasks, future work could pursue more effective examples to augment the existent large-scale datasets.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful suggestions and comments. This work is supported by the National Key Research and Development Program of China (No. 2021ZD0204105), and the Exploratory Research Project of Zhejiang Lab (No.2022RC0AN01). Ming Xiang is supported by the University of Chicago Humanities Division Council.

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.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. 1996. Using the framework. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

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.

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

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, Melbourne, Australia. 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, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Md Mosharaf Hossain, Dhivya Chinnappa, and Eduardo Blanco. 2022. An analysis of negation in natural language understanding corpora. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 716–723, Dublin, Ireland. Association for Computational Linguistics.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Hanmeng Liu, Leyang Cui, Jian Liu, and Yue Zhang. 2021. Natural language inference in context-investigating contextual reasoning over long texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13388–13396.

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

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

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nikita Nangia and Samuel R. Bowman. 2019. Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4566–4575, Florence, Italy. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*,

pages 4885–4901, Online. Association for Computational Linguistics.

Adam Poliak. 2020. A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109, Online. Association for Computational Linguistics.

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

Abhilasha Ravichander, Aakanksha Naik, Carolyn Rose, and Eduard Hovy. 2019. EQUATE: A benchmark evaluation framework for quantitative reasoning in natural language inference. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 349–361, Hong Kong, China. Association for Computational Linguistics.

Kyle Richardson, Hai Hu, Lawrence Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8713–8721.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021. UnNatural Language Inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7329–7346, Online. Association for Computational Linguistics.

Noah A Smith. 2012. Adversarial evaluation for models of natural language. *arXiv preprint arXiv:1207.0245*.

Shane Storks, Qiaozi Gao, and Joyce Y Chai. 2019. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*.

Aarne Talman and Stergios Chatzikyriakidis. 2019. Testing the generalization power of neural network models across NLI benchmarks. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 85–94, Florence, Italy. Association for Computational Linguistics.

Haohan Wang, Da Sun, and Eric P Xing. 2019. What if we simply swap the two text fragments? a straightforward yet effective way to test the robustness of methods to confounding signals in nature language inference tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7136–7143.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

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

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

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, and Kentaro Inui. 2020. Do neural models learn systematicity of monotonicity inference in natural language? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6105–6117, Online. Association for Computational Linguistics.

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

A Appendices

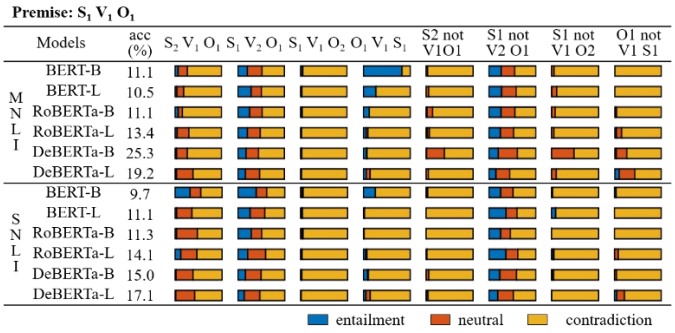
Train model	BERT-base	RoBERTa-base	DeBERTa-v3-base
Learning rate	2e-5/3e-5	2e-5/2e-5	2e-5/2e-5
Train epochs	3/2	3/3	3/2
Batch size	32/32	32/32	64/64
Weight decay	0.01/0.1	0.1/0.01	0.0/0.0
Train data	MNLI / SNLI		
Test accuracy	84.04/83.93(-m/-mm) 89.44	87.84/87.62(-m/-mm) 91.03	90.21/90.73(-m/-mm) 93.33
Train data	aug-MNLI / aug-SNLI		
Test accuracy	82.72/83.12(-m/-mm) 90.98	87.63/87.53(-m/-mm) 91.98	90.37/90.44(-m/-mm) 93.40

Train model	BERT-large	RoBERTa-large	DeBERTa-v3-large
Learning rate	2e-5/3e-5	6e-6/6e-6	6e-6/5e-6
Train epochs	2/2	2/2	2/2
Batch size	32/32	64/64	32/32
Weight decay	0.01/0.1	0.0/0.0	0.0/0.0
Train data	MNLI / SNLI		
Test accuracy	86.24/86.51(-m/-mm) 91.85	90.34/90.13(-m/-mm) 92.66	91.18/91.10(-m/-mm) 93.92
Train data	aug-MNLI / aug-SNLI		
Test accuracy	85.77/86.01(-m/-mm) 91.53	90.80/90.53(-m/-mm) 93.18	91.69/91.67(-m/-mm) 93.94

Appendix Table 1: Hyper-parameters for fine-tuning models. The performance of models on MNLI, and SNLI test sets is shown in the last row.

Human annotation		
	Simple Pair	
	simple-sentence set	conjunction-sentence set
E/N/C rate (%)	4.0 / **94.0** / 2.0	0 / **100** / 0
	premise-related set	premise-irrelevant set
E/N/C rate (%)	6.0 / **88.0** / 6.0	0 / **100** /0
	Extended Pair	
	extended-simple set	extended-conjunction set
E/N/C rate (%)	0 / **100** / 0	6.0 / **88.0** / 6.0

Appendix Table 2: Human classification for premise-hypothesis pairs randomly selected from the *Simple Pair* and *Extended Pair* sets.



Appendix Table 3: Performance on SVO sentences in the simple-sentence set of *Simple Pair*.

Premise: S_1 V_1 O_1. S_2 V_2 O_2.

	Models	acc (%)	S_2 V_1 O_1	S_1 V_2 O_2	S_2 not V_1 O_1	S_1 not V_2 O_2
MNLI	BERT-b	0				
	BERT-l	1.1				
	RoBERTa-b	1.8				
	RoBERTa-l	8.6				
	DeBERTa-b	1.6				
	DeBERTa-l	18.1				
SNLI	BERT-b	0.9				
	BERT-l	0.1				
	RoBERTa-b	1.0				
	RoBERTa-l	2.2				
	DeBERTa-b	2.4				
	DeBERTa-l	1.3				

Premise: S_1 not V_1 O_1. S_2 V_2 O_2.

	Models	acc (%)	S_2 V_1 O_1	S_1 V_2 O_2	S_2 not V_1 O_1	S_1 not V_2 O_2
MNLI	BERT-b	0.1				
	BERT-l	2.0				
	RoBERTa-b	1.3				
	RoBERTa-l	6.7				
	DeBERTa-b	3.8				
	DeBERTa-l	11.5				
SNLI	BERT-b	0.4				
	BERT-l	0.1				
	RoBERTa-b	0.3				
	RoBERTa-l	2.3				
	DeBERTa-b	1.6				
	DeBERTa-l	0.4				

Premise: S_1 V_1 O_1 and S_2 V_2 O_2.

	Models	acc (%)	S_2 V_1 O_1	S_1 V_2 O_2	S_2 not V_1 O_1	S_1 not V_2 O_2
MNLI	BERT-b	0.1				
	BERT-l	1.9				
	RoBERTa-b	1.4				
	RoBERTa-l	13.7				
	DeBERTa-b	1.6				
	DeBERTa-l	31.9				
SNLI	BERT-b	0.8				
	BERT-l	0.1				
	RoBERTa-b	0.1				
	RoBERTa-l	11.4				
	DeBERTa-b	1.7				
	DeBERTa-l	8.7				

Premise: S_1 V_1 O_1 and S_2 not V_2 O_2.

	Models	acc (%)	S_2 V_1 O_1	S_1 V_2 O_2	S_2 not V_1 O_1	S_1 not V_2 O_2
MNLI	BERT-b	0.2				
	BERT-l	2.3				
	RoBERTa-b	0.2				
	RoBERTa-l	9.3				
	DeBERTa-b	1.9				
	DeBERTa-l	18.4				
SNLI	BERT-b	0.1				
	BERT-l	0.4				
	RoBERTa-b	0.1				
	RoBERTa-l	10.4				
	DeBERTa-b	1.7				
	DeBERTa-l	4.1				

entailment neutral contradiction

Appendix Table 4: Performance on SVO sentences in the conjunction-sentence set of *Simple Pair*.

200 premise × E/N/C= 600 pairs

Premise:	The N_1 is A_1.
E-hypo:	The N_1 is A_1.
C-hypo:	The N_1 is not A_1.
N-hypo:	The **N_2** is(not) A_1.
	The N_1 is (not) **A_2**.
	The **N_2** is (not) **A_2**.

200 × 4 premise × E/N/C= 2400 pairs

Premise:	The N_1 is A_1. The N_2 is A_2.
	The N_1 is A_1 and the N_2 is A_2.
	The N_1 is not A_1. The N_2 is A_2.
	The N_1 is A_1 and the N_2 is not A_2.
E-hypo:	The N_1 is A_1.
	The N_2 is A_2.
C-hypo:	The N_1 is not A_1.
	The N_2 is not A_2.
N-hypo:	The **N_2** is (not) A_1.
	The N_1 is (not) **A_2**.

200 premise × E/N/C= 600 pairs

Premise:	The S_1 V_1 O_1.
E-hypo:	The S_1 V_1 O_1.
C-hypo:	The S_1 did notV_1 O_1.
N-hypo:	The **S_2** (did not) V_1 the O_1.
	The S_1 (did not) **V_2** the O_1.
	The S_1 (did not) V_1 the **O_2**.
	The **O_1** (did not) V_1 the **S_1**.

200 × 4 premise × E/N/C= 2400 pairs

Premise:	The S_1 V_1 O_1. The S_2 V_2 O_2.
	The S_1 did not V_1 O_1. The S_2 V_2 O_2.
	The S_1 V_1 O_1 and the S_2 V_2 O_2.
	The S_1 did not V_1 O_1 and the S_2 V_2 O_2.
E-hypo:	The S_1 V_1 O_1.
	The S_2 V_2 O_2.
C-hypo:	The S_1 did not V_1 O_1.
	The S_2 did not V_2 O_2.
N-hypo:	The **S_2** (did not)V_1 the O_1.
	The **S_1** (did not) V_2 the O_2.

Appendix Table 5: Construction of the augmented examples based on the *Simple Pair* set.

Models	subsequence acc(%)				constituent acc(%)			
	MNLI	aug-MNLI	SNLI	aug-SNLI	MNLI	aug-MNLI	SNLI	aug-SNLI
BERT-b	52.3	60.5 (8.2)	50.2	55.7 (5.5)	53.3	59.9 (6.6)	50.2	52.6 (2.4)
BERT-l	56.4	61.5 (5.1)	53.4	62.4 (9.0)	66.3	67.6 (1.3)	52.8	56.1 (3.3)
RoBERTa-b	67.8	69.4 (1.6)	56.6	61.6 (5.0)	72.0	72.0 (0.0)	53.2	66.5 (13.3)
RoBERTa-l	67.5	70.5 (3.0)	59.2	71.4 (12.2)	69.7	65.8 (-3.9)	56.0	61.0 (5.0)
DeBERTa-b	64.9	68.6 (3.7)	64.8	66.3 (1.5)	66.2	68.7 (2.5)	61.7	61.8 (0.1)
DeBERTa-l	65.4	67.8 (2.4)	65.6	69.7 (4.1)	69.0	60.6 (-8.4)	63.4	69.0 (5.6)

Appendix Table 6: Model performance on the HANS dataset. The numbers in the parenthesis show the change in performance comparing the models fine-tuned on augmented MNLI or SNLI with the models only fine-tuned on MNLI or SNLI.