Do Current Natural Language Inference Models Truly Understand Sentences? Insights from Simple Examples

Anonymous ACL submission

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 complexity of the task, especially the complexity of the sentences used for model evaluations, it remains controversial whether these models can truly infer the meaning of sentences or they simply guess the answer via non-humanlike heuristics. Here, we reduce the complexity of the task using two approaches. The first approach simplifies the relationship between the premise and hypothesis by making them unrelated. A test set, referred to as *Random Pair*, is constructed by randomly pairing premises and hypotheses in MNLI/SNLI. Models fine-tuned on MNLI/SNLI identify a large proportion (up to 77.6%) of these unrelated statements as being contradictory. Models fine-tuned on SICK, a dataset that included unrelated premise-hypothesis pairs, perform well on *Random Pair*. The second approach simplifies the task by constraining the premises/hypotheses to be syntactically/semantically simple sentences. A new test set, referred to as *Simple Pair*, is constructed using simple sentences, such as short SVO sentences, and basic conjunction sentences. We find that models fine-tuned on MNLI/SNLI generally fail to understand these simple sentences, but their performance can be boosted by re-fine-tuning the models using only a few hundreds of samples from SICK. All models tested here, however, fail to understand the fundamental compositional binding relation between a subject and a predicate (up to ∼100% error rate) for basic conjunction sentences. Taken together, the results show that models achieving high accuracy on mainstream datasets can still lack basic sentence comprehension capacity, and datasets discouraging non-humanlike heuristics are required to build more robust NLI models.

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; Bowman et al., 2015; 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 to perform the NLI task, such as SICK (Marelli et al., 2014), 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 seems to suggest that these models already have the ability to interpret the meanings of sentences and generate semantic inference. Nevertheless, recent evidence shows that NLI models may have just guessed the answer based on statistical biases in the datasets (Gururangan et al., 2018; Clark et al., 2019). Furthermore, 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 a sentence or simply guess the answer via non-humanlike heuristics (Naik et al., 2018).

The goal of the current paper is two-folded. First, we tackle the issue of potential statistical biases in the current large-scale NLI datasets by designing testing conditions that factor out the effect of statistical biases. To achieve the goal, we extend the current mainstream datasets, such as MNLI and SNLI, to create test conditions in which any heuristics originated in the original datasets (if any) are rendered useless under the new test conditions. This is

done by breaking the original premise-hypothesis pairs and randomly pairing a premises with a hypothesis. Consider a situation in which an annotator designs a hypothesis that is expected to stay in a contradiction relation with a premise. The annotator may use words that are highly suggestive of a contradiction relationship, for example, a higher likelihood of negation in the hypothesis sentence. Instead of truly evaluating the relationship between the premise and the hypothesis, a model may simply exploit the hypothesis-internal regularities to solve the NLI task (Naik et al., 2018; Gururangan et al., 2018; Rudinger et al., 2017). But when the same hypothesis is paired with a random premise, the hypothesis-internal bias remains the same while the relationship between the premise and the hypothesis has been (most likely) changed to a neutral relation. We therefore reason that if a model truly understands the semantic relation between a premise-hypothesis pair, it should answer "neutral" for most of our newly constructed testing conditions; deviations from such a result (i.e. identification of entailment or contradiction to a non-trivial extent) would indicate the model does not truly rely on semantic relations to perform the task.

Second, to probe deeper into the semantic capabilities of the current NLI models, we constructed a large number of simple and conjunction sentences following a set of systematic design features (see more details in Method), and tested whether NLI models can make correct inferences on these sentences. The sentences in the current mainstream datasets are generally highly sophisticated. Training and testing models on difficult and challenging material is 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). But the complexity of the datasets could also potentially hinder an explicit description as to what specific features of the linguistic system the models can learn and more importantly what they can not learn. Furthermore, a focus on complex material implicitly assumes that the current NLP models have the capacity to understand simple sentences and consequently perform the NLI task accurately. The current study, however, shows that models fine-tuned on highly challenging datasets in fact fail on very basic sentences once we systematically probe the semantic knowledge of these models.

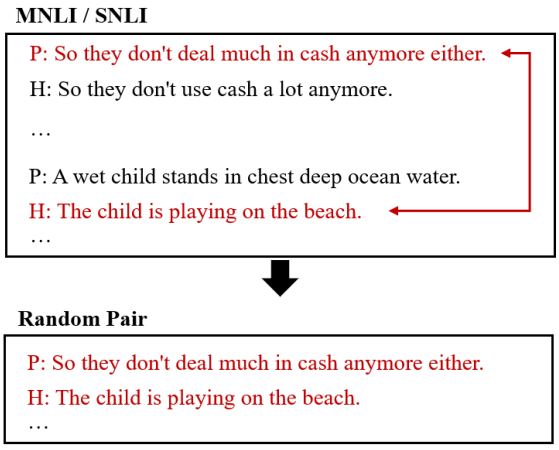

Figure 1: Construction of the *Random Pair* set.

To preview, we tested 3 popular transformer-based models, i.e., BERT (Devlin et al., 2019), ALBERT (Lan et al., 2019), and RoBERTa (Liu et al., 2019), which were respectively fine-tuned on 3 widely-used NLI datasets, i.e., the MNLI, SNLI, and SICK datasets. We found that these models were by and large inaccurate in drawing inference relations on our datasets, and their previously reported success might be due to the inherent biases present in the datasets. More importantly, we also identified a key problem with these models: All the models appeared to fail in the basic semantic composition principles.

2 Method

2.1 NLI Dataset and Pre-trained Models

We employed 3 pre-trained language models, i.e., BERT (Devlin et al., 2019), ALBERT (Lan et al., 2019), and RoBERTa (Liu et al., 2019) to perform the NLI task. For all models, we used the base version. We built our models using Huggingface (Wolf et al., 2020). The models were separately fine-tuned based on 3 datasets, i.e., MNLI (Williams et al., 2018), SNLI (Bowman et al., 2015), and SICK (Marelli et al., 2014). For the 3 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 (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

Templates for the simple-sentence set

N-is-A: The N_1 is A_1. *The apple is expensive.*

SVO: The S_1 V_1 the O_1. *The student saw the dog.*

Templates for the conjunction-sentence set

P1.P2: The N_1 is A_1. The N_2 is A_2.
The apple is expensive. The banana is sweet.

P1.P2: The S_1 V_1 O_1. The S_2 V_2 O_2.
The student saw the dog. The professor lost the key.

$\bar{\text{P}}$1.P2: The N_1 is not A_1. The N_2 is A_2.
The apple is not expensive. The banana is sweet.

$\bar{\text{P}}$1.P2: 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.

P1 and P2: The N_1 is A_1 and the N_2 is A_2.
The apple is expensive and the banana is sweet.

P1 and P2: The S_1 V_1 O_1 and the S_2 V_2 O_2.
The student did not see the dog and the professor lost the key.

P1 and $\bar{\text{P}}$2: The N_1 is A_1 and the N_2 is not A_2.
The apple is expensive and the banana is not sweet.

P1 and $\bar{\text{P}}$2: The S_1 V_1 O_1 and the S_2 did not V_2 O_2.
The student did not see the dog and the professor lost the key.

Figure 2: Template for syntactically simple sentences

softmax classifier. The classifier calculated a score for each class though a linear transformer matrix and softmax function (Devlin et al., 2019).

2.2 Construction of Test Sets

2.2.1 *Random Pair*

We created a *Random Pair* test set by randomly pairing premises and hypotheses in MNLI or SNLI test set, with the constraint that none of the new premise-hypothesis pairs in our test set overlapped with the original pairs in the original datasets (Figure 1). Specifically, 1000 premises were selected from each of the 2 datasets and each premise was paired with 54 hypotheses (18 from MNLI-matched, 18 from MNLI-mismatched, and 18 hypotheses from SNLI). This procedure resulted in 54000 premise-hypothesis pairs (1000 premises × 54 hypotheses) for MNLI and SNLI, respectively. Since the pairing between a premise and a hypothesis is randomized, the relationship between them should generally be neutral. Human annotation was acquired for critical examples to confirm this (see Section 2.3). We did not construct random pairs based on SICK since SICK contained a lot of semantically similar premises/hypotheses (Marelli et al., 2014) and therefore the relationship between random pairs is complex.

2.2.2 *Simple Pair*

We constructed a *Simple Pair* test set using only syntactically simple sentences. The test set were 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 2). One template created N-is-A sentences, where N was a noun and A was an adjective. The noun was 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). As shown in Figure 3, each N-is-A type of premise was paired with 6 hypotheses, and 51000 premise-hypothesis pairs (8500 premises × 6 hypotheses) were created. Each SVO premise was paired with 4 hypotheses, and 48000 premise-hypothesis pairs (12000 premises × 4 hypotheses) were created. Premise-hypothesis pairs containing antonyms or synonyms were excluded in the simple-sentence set and the relationship between all premise-hypothesis pairs was neutral.

For conjunction-sentence set, the premise was constructed by conjoining two simple sentences using one of four possible templates (see Figure 2 for the details). Each premise was paired with 4 hypotheses (see Figure 3). In total, 34000 premise-hypothesis pairs (8500 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 were controlled as being neutral.

2.3 Human Annotation

A large number of hypotheses in *Random Pair* were identified as entailment or contradiction by the models fine-tuned on MNLI and SNLI (see Section 3.1). To test whether most of these premise-

premise	hypothesis
The N_1 is A_1.	The $\mathbf{N_2}$ is A_1. The N_1 is $\mathbf{A_2}$. The $\mathbf{N_2}$ is $\mathbf{A_2}$. The $\mathbf{N_2}$ is not A_1. The N_1 is not $\mathbf{A_2}$. The $\mathbf{N_2}$ is not $\mathbf{A_2}$.
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.	The $\mathbf{N_2}$ is A_1. The N_1 is $\mathbf{A_2}$. The $\mathbf{N_2}$ is not A_1. The N_1 is not $\mathbf{A_2}$.

premise	hypothesis
The S_1 V_1 the O_1.	The $\mathbf{S_2}$ V_1 the O_1. The S_1 $\mathbf{V_2}$ the O_1. The S_1 V_1 the $\mathbf{O_2}$. The $\mathbf{O_1}$ V_1 the $\mathbf{S_1}$.
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.	The $\mathbf{S_2}$ V_1 the O_1. The $\mathbf{S_1}$ V_2 the O_2. The $\mathbf{S_2}$ did not V_1 the O_1. The $\mathbf{S_1}$ did not V_2 the O_2.

Figure 3: Construction of the *Simple Pair* set. More examples are shown in Appendix Table 2.

hypothesis pairs were truly neutral, we collected human annotation for part of the data. For each model, we selected 100 premise-hypothesis pairs that received the highest scores for entailment or contradiction. Within these pairs, 40 was randomly selected for human annotation. These premise-hypothesis pairs were listed in Appendix File1. In the annotation process, human annotators (N = 5) were presented with pairs of sentences and asked to label the relationship between the two sentences, i.e., entailment, contradiction, or neutral. The ground truth label was obtained using a majority vote from the 5 annotators. These sentences are largely more frequently classified as neutral by humans. Appendix table 3 shows the summary statistics of ground truth labels.

3 Results

3.1 Model performance on *Random Pair*

For the *Random Pair* set (see Section 2.2.1), a premise was paired with a set of randomly chosen hypotheses, and we expected the relationship for most of these premise-hypothesis pairs to be neutral. Table 1 showed model performance on *Random Pair*. It appeared that only the models fine-tuned on SICK identified the majority of premise-hypothesis pairs, i.e., more than 95%, as being neutral. The models fine-tuned on MNLI or SNLI, however, identified a large proportion of premise-hypothesis pairs, i.e., more than 32.8% and 69.3% respectively, as contradiction. To further evaluate the model performance, we acquired human annotation of the premise-hypothesis pairs that received the highest scores for contradiction under each model, and more than 90% of these premise-hypothesis pairs were manually annotated as neutral (Appendix Table 3). These results suggested that the transformer-based models fine-tuned on MNLI or SNLI were inaccurate when a hypothesis was unrelated to the premise.

Models		acc (%)	*Random Pair* MNLI	*Random Pair* SNLI
MNLI	BERT	61.5		
	ALBERT	56.7		
	RoBERTa	56.0		
SNLI	BERT	22.1		
	ALBERT	27.4		
	RoBERTa	21.2		
SICK	BERT	99.3		
	ALBERT	97.4		
	RoBERTa	99.9		

entailment neutral contradiction

Table 1: Model performance on *Random Pair*. The percent of premise-hypothesis pairs identified as entailment, neutral, and contradiction were shown in blue, red, and yellow, respectively.

3.2 Model performance on *Simple Pair*

Models fine-tuned on MNLI or SNLI performed poorly on *Random Pair*, and one potential reason was that the sentences in *Random Pair* were selected from MNLI or SNLI, which were complex and therefore challenging to interpret. In the following, we tried to tease apart the ability to infer the relationship between sentences and the ability to interpret complex sentences by testing model performance on syntactically/semantically simple sentences. Model performance on *Simple Pair* was

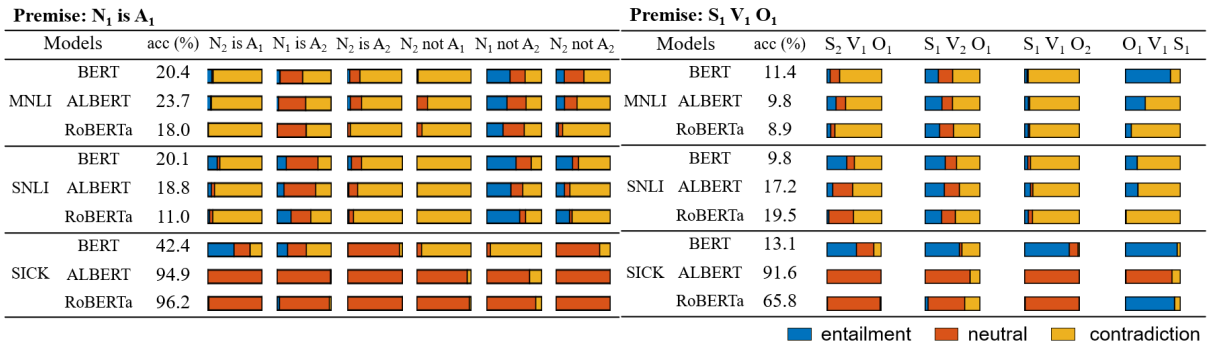

Table 2: Performance on the simple-sentence set in *Simple Pair*. In *Simple Pair*, each premise is paired with a few hypotheses and each hypothesis is shown in a column. The relationship between all premise-hypothesis pairs, when correctly identified, is neutral.

shown in Tables 2 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., [S] or [V] in an SVO sentence, [N] or [A] in a N-is-A sentence) with a different word. It was found that models fine-tuned on MNLI or SNLI 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*".

For the conjunction-sentence set, we constructed neutral hypotheses by breaking the compositional binding relation between a subject and a predicate in the hypothesis (see Figure 3). Model performance was similar for conjunction sentences constructed using SVO or N-is-A sentences. The results showed that all the models tested here failed to understand the fundamental compositional binding relation between a subject and a predicate. The error rate of the models fine-tuned on MNLI or SNLI was near 100%. For example, these 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: after the models have wrongly allowed the composition of "*The apple is sweet*" based on the premise, this inference will 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 models fine-tuned on MNLI or SNLI. 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 "*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. As a result, the models allowed one subject noun in the premise "*the apple*" to be composed with the predicate "*not sweet*", and also allowed the other subject noun in the premise "*the orange*" to be composed with "*expensive*".

3.3 Why do model fine-tuned on SICK perform better?

Compared with large-scale datasets such as MNLI and SNLI, SICK was a small dataset that was automatically generated. However, models fine-tuned on SICK performed better on our test sets compared with models fine-tuned on MNLI or SNLI. Samples in SICK were constructed using eleven expansion rules (Appendix Table 4), and we hypothesized that only some expansion rules were critical to boost model performance on our test sets. In the following, we analyzed the impact of each expansion rule by removing samples from SICK corresponding to the rule. To simplify this analysis, we only considered the ALBERT model which achieved the best performance on most of our test sets. We first removed the samples con-

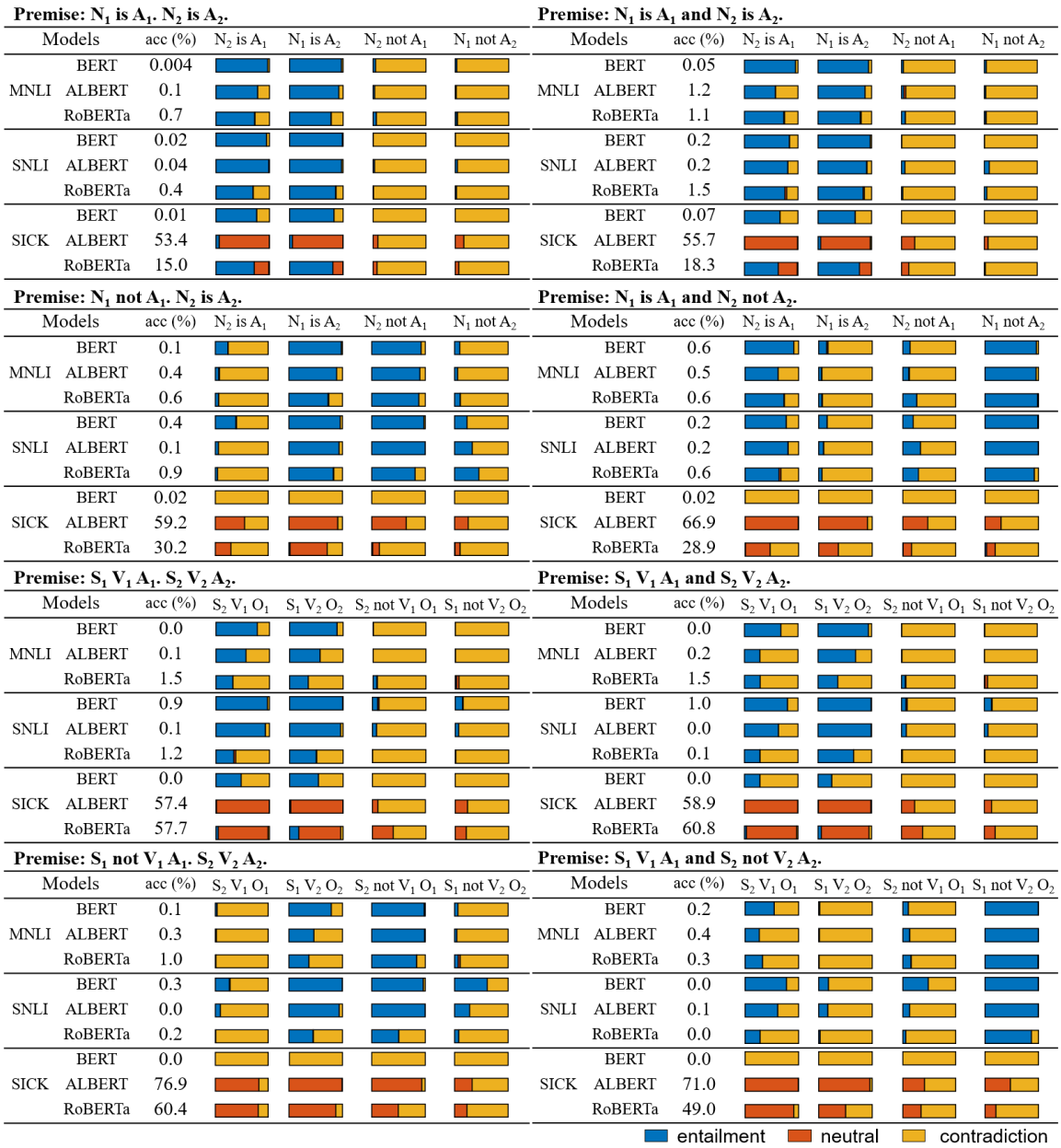

Table 3: Performance on the conjunction-sentence set in *Simple Pair*. Each hypothesis is shown in a column. The relationship between all premise-hypothesis pairs, when correctly identified, is neutral.

sisting of unrelated premise-hypothesis pairs from SICK, since the *Random Pair* set was constructed using a similar rule. We fine-tuned ALBERT on the remaining samples in SICK, and the model fine-tuned this way performed worse on both *Random Pair* and *Simple Pair* (Table 4, first row). Nevertheless, the performance of this model was still better than models fine-tuned on MNLI or SNLI. We then removed each of the remaining 10 expansion rules at a time and fine-tune models using the rest 9 rules. It was found that the scramble-words rule and the replace-words rule were also important to maintain model performance (Table 4).

We next investigated whether the 3 rules critical to maintaining model performance, i.e., unrelated-sentences rule, scramble-words rule, and replace-words rule, could be used to improve the performance of models fine-tuned on MNLI or SNLI. We randomly selected samples corresponding to the 3 rules with the constraint that the textual entailment labels of these samples were balanced. In total, 450 premise-hypothesis pairs (150 entailment + 150 neutral + 150 contradiction) were selected, and we used these 450 samples to re-fine-tune the models that were already fine-tuned on MNLI or SNLI. The parameters were shown in Appendix Table 5, and the re-fine-tuning process did not significantly decrease model performance on MNLI/SNLI (com-

Removed type	*Random Pair*		*Simple Pair*					
	MNLI	SNLI	SVO	A-is-B	P1.P2	P1-and-P2	P1.P2	P1-and-P2
Unrelated sentences	76.4(↓**18.6**)	92.4(↓**7.4**)	76.6(↓15.0)	93.5(↓1.4)	26.4(↓27.0)	48.0(↓7.7)	38.3(↓20.9)	45.6(↓21.3)
Replace words	92.5(↓2.5)	97.7↓2.1)	14.9(↓**76.7**)	38.4(↓**56.5**)	1.3(↓52.1)	1.5(↓54.2)	0.5(↓**58.7**)	0.4(↓**66.5**)
Scramble words	99.6(↑4.6)	99.7(↓0.1)	49.0(↓42.6)	90.4(↓4.5)	0.6(↓**52.8**)	0.3(↓**55.4**)	1.6(↓57.6)	0.6(↓66.3)
Turn adjectives into relative clauses	99.5(↑4.5)	99.9(↑0.1)	70.6(↓21.0)	92.9(↓2.0)	5.8(↓47.6)	10.9(↓44.8)	11.8(↓47.4)	19.8(↓47.1)
Replace quantifiers	89.8(↓5.2)	97.5(↓2.3)	85.4(↓6.2)	84.5(↓10.4)	17.9(↓35.5)	28.0(↓27.7)	18.1(↓41.1)	19.8(↓47.1)
Change determiners with opposites	90.9(↓4.1)	97.9(1.9)	77.4(↓14.2)	99.0(↑4.1)	14.2(↓39.2)	26.8(↓28.9)	28.2(↓31.0)	45.4(↓21.5)
Add modifiers	99.9(↑4.9)	100.0(↑0.2)	89. 9(↓1.7)	93.9(↓1.0)	12.6(↓50.8)	21.5(↓34.2)	26.9(↓32.3)	37.2(↓29.7)
Turn active sentences into passive	99.9(↑4.9)	100.0(↑0.2)	84.8(↓6.6)	87.5(↓7.4)	8.2(↓45.2)	45.2(↓10.5)	25.7(↓33.5)	46.5(↓20.4)
Turn compounds into relative clauses	100.0(↑5.0)	100.0(↑0.2)	85.2(↓6.4)	99.5(↑4.6)	43.2(↓10.2)	56.3(↑0.6)	44.2(↓15.0)	25.3(↓44.6)
Turn passive sentences into active	99.8(↑4.8)	100.0(↑0.2)	86.9(↓25.0)	98.2(↑3.3)	11.9(↓41.5)	20.1(↓35.6)	65.1(↑5.9)	86.7(↓10.2)
Insert a negation	97.4(↑2.4)	99.1(↓0.7)	77.4(↓24.2)	99.6(↑4.7)	59.1(↑5.7)	61.4(↑5.7)	85.8(↑26.6)	97.8(↑30.9)

Table 4: Performance on *Simple Pair* for models fine-tuned on part of the samples in SICK. Each row shows the results when samples corresponding to an expansion rule are removed from the fine-tuning process. The numbers in parenthesis show the change in performance compared with the models fine-tuned on the original SICK dataset. For each test set, shown as a column, the largest decrease is shown in blue. The expansion rules are ordered based on the mean decrease averaged over test sets.

Models		*Random Pair*		*Simple Pair*					
		MNLI	SNLI	SVO	A-is-B	P1&P2	P1-and-P2	$\overline{\text{P1}}$&P2	P1-and $\overline{\text{P2}}$
MNLI (re-fine-tune)	BERT	89.2 (↑**22.1**)	83.4 (↑**59.1**)	16.6(↑**5.2**)	47.0(↑**26.6**)	0.1(↑**0.1**)	0.5(↑**0.5**)	0.9(↑**0.8**)	2.1(↑**1.5**)
	ALBERT	91.6(↑**28.3**)	82.6(↑**55.0**)	20.8(↑**11.0**)	56.4(↑**32.7**)	1.3(↑**1.2**)	7.4(↑**6.2**)	4.8(↑**4.4**)	6.2(↑**5.7**)
	RoBERTa	89.1(↑**30.0**)	88.9(↑**66.8**)	20.3(↑**11.4**)	57.1(↑**39.1**)	8.8(↑**8.1**)	10.1(↑**9.0**)	8.6(↑**8.0**)	4.8(↑**4.2**)
SNLI (re-fine-tune)	BERT	63.4(↑**7.5**)	60.1(↑**40.2**)	21.8(↑**12.0**)	46.3(↑**26.2**)	0.4(↑**0.4**)	1.0(↑**0.8**)	2.5(↑**2.1**)	1.8(**1.6**)
	ALBERT	59.1(↑**9.1**)	54.6(↑**27.5**)	26.1(↑**8.9**)	37.3(↑**18.5**)	0.2(↑**0.2**)	0.7(↑**0.5**)	0.7(↑**0.6**)	1.3(↑**1.1**)
	RoBERTa	55.0(↑**2.1**)	50.0(↑**29.7**)	30.7(↑**11.2**)	26.9(↑**15.9**)	3.4(↑**3.0**)	6.5(↑**5.0**)	1.9(↑**1.0**)	5.4(↑**4.8**)

Table 5: Performance of models re-fine-tuned based on 450 samples in SICK. The numbers in parenthesis show the change in performance compared with the models only fine-tuned on MNLI or SNLI.

paring Appendix Tables 1 and 5). Performance of the models receiving a re-fine-tuning process was shown in Table 5. It was found that the small number of samples selected from SICK can indeed significantly improve model performance on *Random Pair* and the simple-sentence set in *Simple Pair*. Performance on the conjunction-sentence set in *Simple Pair* was not significantly improved since models fine-tuned on the original SICK datasets performance also performed poorly on the conjunction sentences.

4 Related work

Transformer-based models have achieved human-level performance on many NLI datasets such as MNLI, SNLI, and SICK (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 possess the ability to interpret sentences in the current datasets and generate correct inferences. Accordingly, follow-up work aimed 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 work that constructs datasets to test more fine-grained linguistically motivated inference patterns such as pragmatic inferences and numerical reasoning (Ravichander et al., 2019; Jeretic et al., 2020) or correlate model errors with well-defined linguistic phenomena (Naik et al., 2018), with the purpose to identify whether models have trouble making certain types of inferences. Compared with these studies, the current work took a different approach: by intentionally reducing the difficulty of the test material, we aim to uncover whether models can truly infer the meaning of simple sentences. The results show models perform poorly inferring the meaning of basic SVO and N-is-A sentences.

In the meantime, many studies have discussed the potential danger 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 the accuracy higher than chance level, even when just considering the hypothesis (Gururangan et al., 2018). A recent study has also shown that the current models fail to generalize across different datasets (Talman and Chatzikyriakidis, 2019). The concern of overfitting also arises for other NLP tasks (Jia and Liang, 2017; Wallace et al., 2019; Sugawara et al., 2020; Lin et al., 2021). For example, it has been found that models can give the correct answer reading comprehension questions even when crucial information is removed so that the questions are no longer answerable (Poliak et al., 2018; Gururangan et al., 2018; Si et al., 2019; Berzak et al., 2020; Kaushik and Lipton, 2018). Here, by randomly shuffling the premises and hypotheses in MNLI and SNLI, we provide additional evidence that existing models are severely over-fitted and tend to judge the relationship between two randomly paired sentences to be contradictory.

The current work differs from previous studies in two major aspects. First, we propose a new method to extend existing datasets to create new samples that are minimally affected by non-humanlike heuristics. Relatedly, the study in Wang et al. (2019) switched the premise and hypothesis and used the switched pairs to test NLI models; but by randomly pairing premises and hypotheses, we were able to generate a much larger dataset. Our method can be combined with the method by Wang et al. (2019) to further increase the size of datasets and reduce the inherent statistical biases. Second, we constructed a large set of simple sentences to test 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.

Related to the *Simple Pair* set, a recent study constructs a test set, i.e., HANS, based on 3 rules so that the hypothesis overlaps with the premise in terms of words, subsequences, or constituents (McCoy et al., 2019). All our test samples belong to the word overlapping case defined by McCoy et al. (2019), but the results clearly differ, e.g., between simple and conjunction sentences, suggesting that word overlap cannot fully explain model performance. Some samples in HANS are similar to the samples in *Simple Pair*, e.g., when the subject and object of a sentence are swapped. McCoy et al. (2019) found that BERT fine-tuned on MNLI predicts that the original sentence entails the subject-object swapped sentence, and the current study replicates this phenomenon (Table 2, upper right), but we also show that the results strongly depend on the model and training set. Furthermore, we show that incorporating a small number of samples from SICK can improve model performance. Finally, a new and important finding of the current results is that current models have substantial difficulty solving the compositional binding problem for conjunction sentences.

5 Conclusion

In summary, since existing models seem to have shown good performance on mainstream NLI datasets such as MNLI and SNLI, 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 datasets (Nie et al., 2020; Liu et al., 2021). The current study, however, shows that models achieving good performance on mainstream datasets do not necessarily generalize to simpler datasets. In fact, models fine-tuned on MNLI/SNLI generally have lower than chance level performance when predicting the relationship between simple sentences. The results here show that combining part of the automatically generated samples with large-scale human-created datasets such as MNLI and SNLI can potentially increase model's ability to generalize to simpler test samples while largely maintaining the performance on challenging samples. Nevertheless, even with the current approach to combine datasets, all the models still could not bind the subject noun with the correct predicate in conjunctive sentences. Future studies should develop both the design properties of these models and the properties of the training datasets to achieve more robust NLI performance.

References

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. STARC: Structured annotations for reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5726–5735, Online. Association for Computational Linguistics.

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.

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.

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.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. 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*.

Jieyu Lin, Jiajie Zou, and Nai Ding. 2021. Using adversarial attacks to reveal the statistical bias in machine reading comprehension models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 333–342, Online. Association for Computational Linguistics.

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*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

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.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Chenglei Si, Shuohang Wang, Min-Yen Kan, and Jing Jiang. 2019. What does bert learn from multiple-choice reading comprehension datasets? *arXiv preprint arXiv:1910.12391*.

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*.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the benchmarking capacity of machine reading comprehension datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8918–8927.

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.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China. 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.

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.

A Appendices

Train data	MNLI/SNLI/SICK	MNLI/SNLI/SICK	MNLI/SNLI/SICK
Train model	BERT-base	ALBERT-base	RoBERTa-base
Learning rate	2e-5/3e-5/2e-5	2e-5/2e-5/2e-5	2e-5/2e-5/2e-5
Train epochs	3/2/5	3/2/5	3/3/5
Batch size	32/32/32	32/32/32	32/32/32
Weight decay	0.01/0.1/0.1	0/0.01/0.1	0.1/0.01/0.1
Test accuracy	84.04/83.93 (-m/-mm) 89.44/86.63	85.04/85.12 (-m/-mm) 90.05/89.45	87.84/87.62 (-m/-mm) 91.03/89.35

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

Premise	The N_1 is A_1. *The apple is expensive.*				
Hypothesis					
The $\mathbf{N_2}$ is A_1. *The pear is expensive.*	The N_1 is $\mathbf{A_2}$. *The apple is sweet.*	The $\mathbf{N_2}$ is $\mathbf{A_2}$. *The pear is sweet.*	The $\mathbf{N_2}$ is not A_1. *The pear is not expensive.*	The N_1 is not $\mathbf{A_2}$. *The apple is not sweet.*	The $\mathbf{N_2}$ is not $\mathbf{A_2}$. *The pear is not sweet.*

Premise	The S_1 V_1 the O_1. *The student saw the dog.*			
Hypothesis				
The $\mathbf{S_2}$ V_1 the O_1. *The professor saw the dog.*	The S_1 $\mathbf{V_2}$ the O_1. *The student lost the dog.*	The S_1 V_1 the $\mathbf{O_2}$. *The student lost the key.*	The $\mathbf{O_1}$ V_1 the $\mathbf{S_1}$. *The dog saw the student.*	

Premise	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 and the pear is sweet.*		
	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_1 and the N_2 is not A_2.	*The apple is expensive and the pear is not sweet.*		
Hypothesis				
The $\mathbf{N_2}$ is A_1. *The pear is expensive.*	The N_1 is $\mathbf{A_2}$. *The apple is sweet.*	The $\mathbf{N_2}$ is not A_1. *The pear is not expensive.*	The N_1 is not $\mathbf{A_2}$. *The apple is not sweet.*	

Premise	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_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 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.*		
Hypothesis				
The $\mathbf{S_2}$ V_1 the O_1. *The professor saw the dog.*	The $\mathbf{S_1}$ V_2 the O_2. *The student lost the key.*	The $\mathbf{S_2}$ did not V_1 the O_1. *The professor did not see the dog.*	The $\mathbf{S_1}$ did not V_2 the O_2. *The student did not lose the key.*	

Appendix Table 2: Examples of the *Simple Pair* set.

Samples	*Random Pair*	
	MNLI	SNLI
samples receive high contradiction scores	7.5 / 90.0 / 2.5	2.5 / 97.5 / 0
samples receive high entailment scores	10.0 / 90.0 / 0	42.5 / 57.5 / 0

entailment neutral contradiction

Appendix Table 3: Human classification for premise-hypothesis pairs that received the highest scores for entailment or contradiction under each model. The percent of premise-hypothesis pairs classified as entailment, neutral, and contradiction were shown in blue, red, and yellow, respectively.

Rule	Size	Example
Unrelated sentences	1197	Premise: *The man is dancing.* Hypothesis: *Three women are standing still.*
Add modifiers	290	Premise: *An **old** man is sitting in a field.* Hypothesis: *A man is sitting in a field.*
Scramble words	377	Premise: *A **pan** is being dropped over the **meat**.* Hypothesis: *The **meat** is being dropped into a **pan**.*
Insert a negation	419	Premise: *The person is **not** slicing onions.* Hypothesis: *The person is slicing onions.*
Replace words	1791	Premise: *A man in a **cap** is playing a harp.* Hypothesis: *A man in a **hat** is playing a harp.*
Replace quantifiers	267	Premise: ***A few** people are dancing.* Hypothesis: ***A group of** people are dancing.*
Change determiners with opposites	608	Premise: *There is **no group of** people dancing.* Hypothesis: ***A group of** people are dancing.*
Turn passive sentences into active	34	Premise: *A woman is peeling the potato.* Hypothesis: *The potato is **being peeled** by a woman.*
Turn active sentences into passive	209	Premise: *An interview is **being granted** by the man.* Hypothesis: *The man is granting an interview.*
Turn compounds into relative clauses	56	Premise: *A woman is using a **machine made for sewing**.* Hypothesis: *A woman is using a sewing machine.*
Turn adjectives into relative clauses	189	Premise: *A girl, **who is little**, is playing the piano.* Hypothesis: *A little girl is playing the piano.*

Appendix Table 4: The types of premise-hypothesis pairs in SICK.

Train model	MNLI / SNLI BERT-base	MNLI / SNLI ALBERT-base	MNLI / SNLI RoBERTa-base
Learning rate	2e-6/2e-6	2e-6/2e-6	2e-6/2e-6
Train epochs	3/3	3/3	3/3
Batch size	16/16	16/16	16/16
Weight decay	0.01/0.01	0.01/0.01	0.01/0.01
Test accuracy	80.67 /81.02 (-m/-mm) /88.97	82.02/82.60 (-m/-mm) /88.63	84.65/84.54 (-m/-mm) /89.88

Appendix Table 5: Hyper-parameters for re-fine-tuning models. The performance of models on MNLI, and SNLI test sets is shown in the last row.