BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance

R. Thomas McCoy,[1] Junghyun Min,[1] and Tal Linzen[2]
[1]Department of Cognitive Science, Johns Hopkins University
[2]Department of Linguistics and Center for Data Science, New York University
tom.mccoy@jhu.edu, jmin10@jhu.edu, linzen@nyu.edu

Abstract

If the same neural network architecture is trained multiple times on the same dataset, will it make similar linguistic generalizations across runs? To study this question, we fine-tuned 100 instances of BERT on the Multi-genre Natural Language Inference (MNLI) dataset and evaluated them on the HANS dataset, which evaluates syntactic generalization in natural language inference. On the MNLI development set, the behavior of all instances was remarkably consistent, with accuracy ranging between 83.6% and 84.8%. In stark contrast, the same models varied widely in their generalization performance. For example, on the simple case of subject-object swap (e.g., determining that *the doctor visited the lawyer* does not entail *the lawyer visited the doctor*), accuracy ranged from 0.0% to 66.2%. Such variation is likely due to the presence of many local minima in the loss surface that are equally attractive to a low-bias learner such as a neural network; decreasing the variability may therefore require models with stronger inductive biases.

1 Introduction

Generalization is a crucial component of learning a language. No training set can contain all possible sentences, so learners must be able to generalize to sentences that they have never encountered before. We differentiate two types of generalization:

1. **In-distribution generalization:** Generalization to examples which are novel but which are drawn from the same distribution as the training set.

2. **Out-of-distribution generalization:** Generalization to examples drawn from a different distribution than the training set.

Standard test sets in natural language processing are generated in the same way as the corresponding training set, therefore testing only in-distribution generalization. Current neural architectures perform very well at this type of generalization. For example, on the natural language understanding tasks included in the GLUE benchmark (Wang et al., 2019), several Transformer-based models (Liu et al., 2019b,a; Raffel et al., 2020) have surpassed the human baselines from Nangia and Bowman (2019).

However, this strong performance does not necessarily indicate mastery of language. Because of biases in training distributions, it is often possible for a model to achieve strong in-distribution generalization by using shallow heuristics rather than deeper linguistic knowledge. Therefore, evaluating only on standard test sets cannot reveal whether a model has learned abstract properties of language or if it has only learned shallow heuristics.

An alternative evaluation approach addresses this flaw by testing how the model handles particular linguistic phenomena, using datasets designed to be impossible to solve using shallow heuristics. In this line of investigation, which tests out-of-distribution generalization, the results are more mixed. Some works have found successful handling of phenomena such as subject-verb agreement (Gulordava et al., 2018) and filler-gap dependencies (Wilcox et al., 2018). Other works, however, have illuminated surprising failures even on seemingly simple types of examples (Marvin and Linzen, 2018; McCoy et al., 2019). Such results make it clear that there is still much room for improvement in how neural models perform on syntactic structures that are rare in training corpora.

In this work, we investigate whether the linguistic generalization behavior of a given neural architecture is consistent across multiple instances of that architecture. This question is important because, in order to tell which types of architectures generalize best, we need to know whether suc-

Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 217–227
Online, November 20, 2020.

cesses and failures of generalization should be attributed to aspects of the architecture or to random luck in the choice of the model's initial weights.

We investigate this question using the task of natural language inference (NLI). We fine-tuned 100 instances of BERT (Devlin et al., 2019) on the MNLI dataset (Williams et al., 2018).[1] These 100 instances differed only in (i) the initial weights of the classifier trained on top of BERT, and (ii) the order in which training examples were presented. All other aspects of training, including the initial weights of BERT, were held constant. We evaluated these 100 instances on both the in-distribution MNLI development set and the out-of-distribution HANS evaluation set (McCoy et al., 2019), which tests syntactic generalization in NLI models.

We found that these 100 instances were remarkably consistent in their in-distribution generalization accuracy, with all accuracies on the MNLI development set falling in the range 83.6% to 84.8%, and with a high level of consistency on labels for specific examples (e.g., we identified 526 examples that all 100 instances labeled incorrectly). In contrast, these 100 instances varied dramatically in their out-of-distribution generalization performance; for example, on one of the thirty categories of examples in the HANS dataset, accuracy ranged from 4% to 76%. These results show that, when assessing the linguistic generalization of neural models, it is important to consider multiple training runs of each architecture, since models can differ vastly in how they perform on examples drawn from a different distribution than the training set, even when they perform similarly on an in-distribution test set.

2 Background

2.1 In-distribution generalization

Several works have noted that the same architecture can have very different in-distribution generalization across restarts of the same training process (Reimers and Gurevych, 2017, 2018; Madhyastha and Jain, 2019). Most relevantly for our work, fine-tuning of BERT is unstable for some datasets, such that some runs achieve state-of-the-art results while others perform poorly (Devlin et al., 2019; Phang et al., 2018). Unlike these past works, we focus on *out-of-distribution* generalization, rather than in-distribution generalization.

[1]The weights for all 100 fine-tuned models are publicly available at https://github.com/tommccoy1/hans.

2.2 Out-of-distribution generalization

Several other works have noted variation in out-of-distribution syntactic generalization. Weber et al. (2018) trained 50 instances of a sequence-to-sequence model on a symbol replacement task. These instances consistently had above 99% accuracy on the in-distribution test set but varied on out-of-distribution generalization sets; in the most variable case, accuracy ranged from close to 0% to over 90%. Similarly, McCoy et al. (2018) trained 100 instances for each of six types of networks, using a synthetic training set that was ambiguous between two generalizations. Some models consistently made the same generalization across runs, but others varied considerably, with some instances of a given architecture strongly preferring one of the two generalizations that were plausible given the training set, while other instances strongly preferred the other generalization. Finally, Liška et al. (2018) trained 5000 instances of recurrent neural networks on the lookup tables task. Most of these instances failed on compositional generalization, but a small number generalized well.

These works on variation in out-of-distribution generalization all used simple, synthetic tasks with training sets designed to exclude certain types of examples. Our work tests if models are still as variable when trained on a natural-language training set that is not adversarially designed. In concurrent work, Zhou et al. (2020) also measured variability in out-of-distribution performance for 3 models (including BERT) on 12 datasets (including HANS). Their work has impressive breadth, whereas we instead aim for depth: We analyze the particular categories within HANS to give a fine-grained investigation of syntactic generalization, while Zhou et al. only report overall accuracy averaged across categories. In addition, we fine-tuned 100 instances of BERT, while Zhou et al. only fine-tuned 10 instances. The larger number of instances allows us to investigate the extent of the variability in more detail.

2.3 Linguistic analysis of BERT

Many recent papers have sought a deeper understanding of BERT, whether to assess its encoding of sentence structure (Lin et al., 2019; Hewitt and Manning, 2019; Chrupała and Alishahi, 2019; Jawahar et al., 2019; Tenney et al., 2019b); its representational structure more generally (Abnar et al., 2019); its handling of specific linguistic

phenomena such as subject-verb agreement (Goldberg, 2019), negative polarity items (Warstadt et al., 2019), function words (Kim et al., 2019), or a variety of psycholinguistic phenomena (Ettinger, 2020); its internal workings (Coenen et al., 2019; Tenney et al., 2019a; Clark et al., 2019); or its inductive biases (Warstadt and Bowman, 2020). The novel contribution of this work is the focus on variability across a large number of fine-tuning runs; previous works have generally used models without fine-tuning or have used only a small number of fine-tuning runs (usually only one fine-tuning run, or at most ten fine-tuning runs).

3 Method

3.1 Task and datasets

We used the task of natural language inference (NLI, also known as Recognizing Textual Entailment; Condoravdi et al., 2003; Dagan et al., 2006, 2013), which involves giving a model two sentences, called the *premise* and the *hypothesis*. The model must then output *entailment* if the premise entails (i.e., implies the truth of) the hypothesis, *contradiction* if the premise contradicts the hypothesis, or *neutral* otherwise. For training, we used the training set of the MNLI dataset (Williams et al., 2018), examples from which are given below:

(1) a. **Premise:** Finally she turned back to him.
b. **Hypothesis:** She turned to him.
c. **Label:** Entailment

(2) a. **Premise:** You outwitted me.
b. **Hypothesis:** You have never outwitted me.
c. **Label:** Contradiction

(3) a. **Premise:** okay well i live in Carrollton
b. **Hypothesis:** I have a house in Carrollton.
c. **Label:** Neutral

To test in-distribution generalization, we used the MNLI `matched` development set, which was generated in the same way as the MNLI training set. We used the development set rather than the test set because the test set labels are not available to the public. This development set was not used in any way during training, making it effectively a test set. To test out-of-distribution generalization, we used the HANS dataset (McCoy et al., 2019), which contains NLI examples designed to require understanding of syntactic structure. More specifically, HANS targets three structural heuristics that models trained on MNLI are likely to learn (for definitions and examples, see Figure 1).

To assess whether a model has learned these heuristics, HANS contains examples where each heuristic makes the right predictions (i.e., where the correct label is *entailment*) and examples where each heuristic makes the wrong predictions (i.e., where the correct label is *non-entailment*). A model that has adopted one of the heuristics will output *entailment* for all examples targeting that heuristic, even when the correct answer is *non-entailment*.

3.2 Models and training

All of our models consisted of BERT with a linear classifier on top of it outputting labels of *entailment*, *contradiction*, or *neutral*. We fine-tuned 100 instances of this model on MNLI using the fine-tuning code from the BERT GitHub repository.[2] The BERT component of each instance was initialized with the pre-trained `bert-base-uncased` weights. For evaluation on HANS, we translated outputs of *contradiction* and *neutral* into a single *non-entailment* label, following McCoy et al. (2019). The fine-tuning process proceeded for 3 epochs and modified the weights of both the BERT component and the classifier. Following Devlin et al. (2019), across fine-tuning runs we varied only (i) the random initial weights of the classifier and (ii) the order in which training examples were presented. All other aspects, including the initial pre-trained weights of the BERT component, were held constant.

4 Results

4.1 In-distribution generalization

The 100 instances were remarkably consistent on in-distribution generalization, with all models scoring between 83.6% and 84.8% on the MNLI development set (Figure 2, left). Numerical statistics for the performance of our 100 instances of BERT on MNLI and HANS can be found in Figure 7, and statistics for HANS broken down by linguistic construction can be found in Figures 3 and 4. Finally, to see model-by-model results, see `https://github.com/tommccoy1/hans`.

The instances were also highly consistent in their choice of labels for particular examples (Figure 2, right); in the rest of this subsection, we provide some quantitative and qualitative analysis of consistency of performance on individual examples.

[2] `github.com/google-research/bert`

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

Figure 1: The heuristics targeted by the HANS dataset, along with examples of incorrect entailment predictions that these heuristics would lead to. (Figure from McCoy et al. 2019.)

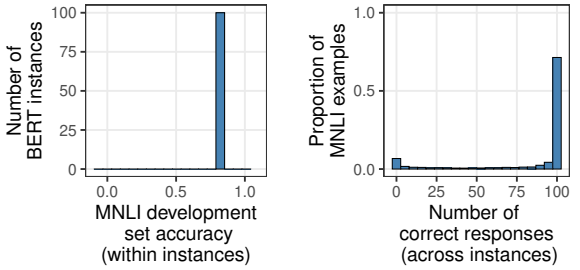

Figure 2: In-distribution generalization. Left: Within-instance accuracy on the MNLI development set; all BERT instances had scores near 84%. Right: Across-instance accuracy on individual examples in the MNLI development set; e.g., 66% of the examples were answered correctly by all 100 instances. For numerical results, see Figure 7.

On average, among any pair of fine-tuned BERT instances, the two members of the pair agreed on the labels of 93.1% of the examples (when considering all three labels of *entailment*, *contradiction*, and *neutral*, rather than the collapsed labels of *entailment* and *non-entailment*). To give a sense of consistency across all 100 instances (rather than only among pairs of instances), Figure 2 (right) illustrates how consistent our 100 instances were on their answers to individual examples in the MNLI development set. Of the 9815 examples in the set, there were 6526 that all 100 instances labeled correctly, and 526 that all instances labeled incorrectly. Thus, the consistent score of about 84% on the MNLI development set can be partially explained by the fact that there are certain examples that all models answered correctly or that all models answered incorrectly, as models were consistently correct or incorrect on 72% of the examples.

Examples (4) through (6) show some of the 6526 cases that all 100 instances answered correctly:

(4) a. **Premise:** The new rights are nice enough
b. **Hypothesis:** Everyone really likes the newest benefits
c. **Label:** Neutral

(5) a. **Premise:** This site includes a list of all award winners and a searchable database of Government Executive articles.
b. **Hypothesis:** The Government Executive articles housed on the website are not able to be searched.
c. **Label:** Contradiction

(6) a. **Premise:** You and your friends are not welcome here, said Severn.
b. **Hypothesis:** Severn said the people were not welcome there.
c. **Label:** Entailment

Examples (7) through (12) show some of the 526 cases that all 100 instances answered incorrectly. Some of these examples arguably have incorrect labels in the dataset, such as (7) (because the hypothesis mentions a report which the premise does not mention), so it is unsurprising that models found such examples difficult. Other consistently difficult examples involve areas that one might intuitively expect to be tricky for models trained on natural language, such as world knowledge (e.g., (8) requires knowledge of how long forearms are, and (9) requires knowledge of what nodding is), the ability to count (e.g., (10)), or fine-grained shades of meaning that might require multiple steps of reasoning (e.g., (11) and (12)). Some of the consistently difficult examples have a high degree of lexical overlap yet are not labeled *entailment* (such as (13)); the difficulty of such examples adds further evidence to the conclusion that these models have adopted the lexical overlap heuristic. Finally, there are some examples, such as (14), for which it

is unclear why models find them so difficult.

(7) a. **Premise:** Indeed, 58 percent of Columbia/HCA's beds lie empty, compared with 35 percent of nonprofit beds.
b. **Hypothesis:** 58% of Columbia/HCA's beds are empty, said the report.
c. **Label:** Entailment

(8) a. **Premise:** One he broke back to about the length of his forearm.
b. **Hypothesis:** He snapped it until it was just a couple of inches long.
c. **Label:** Contradiction

(9) a. **Premise:** The Kal nodded.
b. **Hypothesis:** The Kal then shook its head side to side.
c. **Label:** Contradiction

(10) a. **Premise:** Load time is divided into elemental and coverage related load time.
b. **Hypothesis:** Load time is comprised of three parts.
c. **Label:** Contradiction

(11) a. **Premise:** I thought working on Liddy's campaign would be better than working on Bob's.
b. **Hypothesis:** I thought I would like working on Liddy's campaign the best.
c. **Label:** Neutral

(12) a. **Premise:** Sure enough, there was the chest, a fine old piece, all studded with brass nails, and full to overflowing with every imaginable type of garment.
b. **Hypothesis:** The chest wasn't big enough to completely contain all of the garments.
c. **Label:** Entailment

(13) a. **Premise:** True to his word to his faithful mare, Ca'daan left Whitebelly in Fena Dim and borrowed Gray Cloud from his uncle.
b. **Hypothesis:** Ca'daan kept his word to Gray Cloud and borrowed Whitebelly from his uncle.
c. **Label:** Contradiction

(14) a. **Premise:** Clearly, yes.
b. **Hypothesis:** Obviously, the answer is yes.
c. **Label:** Entailment

Finally, examples (15) through (17) show some of the 8 cases that exactly half of our 100 instances got correct. Plausibly, such examples are the ones that lie close to a decision boundary that is relatively consistent across instances.

(15) a. **Premise:** He bent down to study the tiny little jeweled gears.
b. **Hypothesis:** He bent down to examine the decorated gears.
c. **Label:** Entailment

(16) a. **Premise:** Conversely, an increase in government saving adds to the supply of resources available for investment and may put downward pressure on interest rates.
b. **Hypothesis:** Interest rates should increase to increase saving.
c. **Label:** Contradiction

(17) a. **Premise:** More than 100 judges, lawyers and dignitaries were present for the gathering.
b. **Hypothesis:** 152 judges and lawyers showed up
c. **Label:** Neutral

4.2 Out-of-distribution generalization

On HANS, performance was much more variable than on the MNLI development set. HANS consists of 6 main categories of examples, each of which can be further divided into 5 subcategories. Performance was reasonably consistent on five of these categories, but on the sixth category—lexical overlap examples that are inconsistent with the lexical overlap heuristic—performance varied dramatically, ranging from 5% accuracy to 55% accuracy (Figure 6). Since this is the most variable category, we focus on it for the rest of the analysis.

The category of lexical overlap examples that are inconsistent with the lexical overlap heuristic encompasses examples for which the correct label is *non-entailment* and for which all the words in the hypothesis also appear in the premise but not as a contiguous subsequence. This category has five subcategories; examples and results for each subcategory are in Figure 5. Chance performance on HANS was 50%; on all subcategories except for passives, accuracies ranged from far below chance to modestly above chance. Models varied considerably even on categories that humans find simple (McCoy et al., 2019). For example, accuracy on the subject-object swap examples, which can be handled with only rudimentary knowledge of syntax (in particular, the distinction between subjects and objects), ranged from 0% to 66%. Overall,

Heuristic	Subcase	Minimum	Maximum	Mean	Std. dev.
Lexical overlap	Untangling relative clauses	0.94	1.00	0.98	0.01
	The athlete who the judges saw called the manager. → *The judges saw the athlete.*				
	Sentences with PPs	0.98	1.00	1.00	0.00
	The tourists by the actor called the authors. → *The tourists called the authors.*				
	Sentences with relative clauses	0.97	1.00	0.99	0.01
	The actors that danced encouraged the author. → *The actors encouraged the author.*				
	Conjunctions	0.72	0.92	0.83	0.05
	The secretaries saw the scientists and the actors. → *The secretaries saw the actors.*				
	Passives	0.99	1.00	1.00	0.00
	The authors were supported by the tourists. → *The tourists supported the authors.*				
Subsequence	Conjunctions	0.93	1.00	0.98	0.02
	The actor and the professor shouted. → *The professor shouted.*				
	Adjectives	1.00	1.00	1.00	0.00
	Happy professors mentioned the lawyer. → *Professors mentioned the lawyer.*				
	Understood argument	0.95	1.00	1.00	0.01
	The author read the book. → *The author read.*				
	Relative clause on object	0.98	1.00	0.99	0.01
	The artists avoided the actors that performed. → *The artists avoided the actors.*				
	PP on object	1.00	1.00	1.00	0.00
	The authors called the judges near the doctor. → *The authors called the judges.*				
Constituent	Embedded under preposition	0.81	1.00	0.96	0.02
	Because the banker ran, the doctors saw the professors. → *The banker ran.*				
	Outside embedded clause	1.00	1.00	1.00	0.00
	Although the secretaries slept, the judges danced. → *The judges danced.*				
	Embedded under verb	0.93	1.00	0.99	0.01
	The president remembered that the actors performed. → *The actors performed.*				
	Conjunction	1.00	1.00	1.00	0.00
	The lawyer danced, and the judge supported the doctors. → *The lawyer danced.*				
	Adverbs	1.00	1.00	1.00	0.00
	Certainly the lawyers advised the manager. → *The lawyers advised the manager.*				

Figure 3: Results for the HANS subcases for which the heuristics make correct predictions (i.e., where the correct label is *entailment*). All statistics are based on 100 runs.

Heuristic	Subcase	Minimum	Maximum	Mean	Std. dev.
Lexical overlap	Subject-object swap	0.00	0.66	0.19	0.17
	The senators mentioned the artist. $\nrightarrow$ *The artist mentioned the senators.*				
	Sentences with PPs	0.04	0.76	0.41	0.18
	The judge behind the manager saw the doctors. $\nrightarrow$ *The doctors saw the manager.*				
	Sentences with relative clauses	0.09	0.67	0.33	0.14
	The actors called the banker who the tourists saw. $\nrightarrow$ *The banker called the tourists.*				
	Conjunctions	0.12	0.72	0.45	0.15
	The doctors saw the presidents and the tourists. $\nrightarrow$ *The presidents saw the tourists.*				
	Passives	0.00	0.04	0.01	0.01
	The senators were helped by the managers. $\nrightarrow$ *The senators helped the managers.*				
Subsequence	NP/S	0.00	0.05	0.02	0.01
	The managers heard the secretary resigned. $\nrightarrow$ *The managers heard the secretary.*				
	PP on subject	0.00	0.35	0.12	0.07
	The managers near the scientist shouted. $\nrightarrow$ *The scientist shouted.*				
	Relative clause on subject	0.00	0.23	0.07	0.04
	The secretary that admired the senator saw the actor. $\nrightarrow$ *The senator saw the actor.*				
	MV/RR	0.00	0.02	0.00	0.00
	The senators paid in the office danced. $\nrightarrow$ *The senators paid in the office.*				
	NP/Z	0.02	0.13	0.06	0.02
	Before the actors presented the doctors arrived. $\nrightarrow$ *The actors presented the doctors.*				
Constituent	Embedded under preposition	0.14	0.70	0.41	0.12
	Unless the senators ran, the professors recommended the doctor. $\nrightarrow$ *The senators ran.*				
	Outside embedded clause	0.00	0.03	0.00	0.01
	Unless the authors saw the students, the doctors resigned. $\nrightarrow$ *The doctors resigned.*				
	Embedded under verb	0.02	0.42	0.17	0.08
	The tourists said that the lawyer saw the banker. $\nrightarrow$ *The lawyer saw the banker.*				
	Disjunction	0.00	0.03	0.00	0.01
	The judges resigned, or the athletes saw the author. $\nrightarrow$ *The athletes saw the author.*				
	Adverbs	0.00	0.17	0.06	0.04
	Probably the artists saw the authors. $\nrightarrow$ *The artists saw the authors.*				

Figure 4: Results for the HANS subcases for which the heuristics make incorrect predictions (i.e., where the correct label is *non-entailment*). All statistics are based on 100 runs.

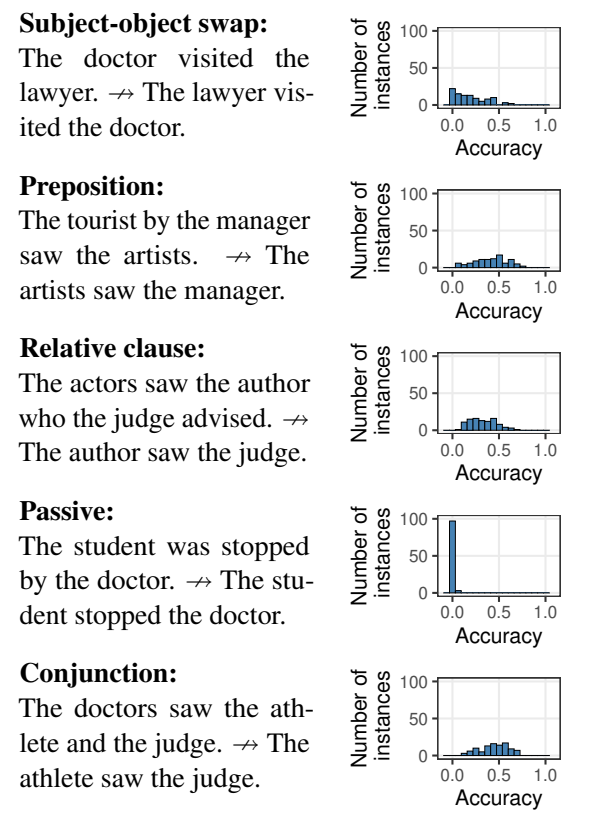

Figure 5: Accuracy distributions on the subcategories of the non-entailed lexical overlap examples of the HANS dataset (i.e., the examples that are inconsistent with the lexical overlap heuristic). For numerical results, and results for the other 25 subcategories of HANS, see Figures 3 and 4.

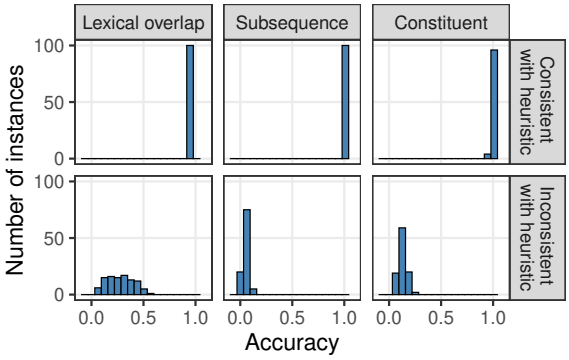

Figure 6: Out-of-distribution generalization: Performance on HANS, broken down into six categories of examples, based on the syntactic heuristic that each example targets and whether the example is consistent with the relevant heuristic (i.e., has a correct label of *entailment*) or inconsistent with the heuristic (i.e., has a correct label of *non-entailment*). The lexical overlap cases that are inconsistent with the heuristic (lower left plot) are highly variable across instances. For numerical results, see Figure 7.

although these models performed consistently on the in-distribution test set, they have nevertheless learned highly variable representations of syntax.

5 Discussion

We have found that models that differ only in their initial weights and the order of training examples can vary substantially in out-of-distribution linguistic generalization. We found this variation even with the vast majority of initial weights held constant (i.e., all the weights in the BERT component of the model). We conjecture that models might be even more variable if the pre-training of BERT were also redone across instances. These results underscore the importance of evaluating models on multiple restarts, as conclusions drawn from a single instance of a model might not hold across instances. Further, these results highlight the importance of evaluating out-of-distribution generalization; since all of our instances displayed similar in-distribution generalization, only their out-of-distribution generalization illuminates the substantial differences in what they have learned.

In stark contrast to the models we have looked at—which generalized in highly variable ways despite being trained on the same set of examples—humans tend to converge to similar linguistic generalizations despite major differences in the linguistic input that they encounter as children (Chomsky, 1965, 1980). This suggests that reducing the generalization variability of NLP models may help bring them closer to human performance in one major area where they still dramatically lag behind humans, namely in out-of-distribution generalization.

How could the out-of-distribution generalization of models be made more consistent? The variability that we have observed likely reflects the presence of many local minima in the loss surface, all of which are equally attractive to our models. This makes the model's choice of a minimum essentially arbitrary and easily affected by the initial weights and the order of training examples. To reduce this variability, then, one approach would be to use models with stronger inductive biases, which can help distinguish between the many local minima. An alternate approach would be to use training sets that better represent a large set of linguistic phenomena, to decrease the probability of there being local minima that ignore certain phenomena.

		HANS: Consistent with heuristic			HANS: Inconsistent with heuristic		
	MNLI	Lexical	Subseq.	Const.	Lexical	Subseq.	Const.
Minimum	0.84	0.93	0.98	0.96	0.05	0.01	0.03
Maximum	0.85	0.98	1.00	1.00	0.55	0.14	0.24
Mean	0.84	0.96	0.99	0.99	0.28	0.05	0.13
Standard deviation	0.00	0.01	0.00	0.01	0.12	0.02	0.04

Figure 7: Results for models trained on MNLI. The MNLI column reports accuracy on the MNLI `matched` development set, where there are three possible labels (*entailment*, *contradiction*, and *neutral*). The remaining columns are subsets of the HANS dataset, with *neutral* and *contradiction* merged into a single label, *non-entailment*, such that there are only two possible labels: *entailment* and *non-entailment*. The examples that are consistent with the heuristics are those that have a correct label of *entailment*, while the examples that are inconsistent with the heuristics are those with a correct label of *non-entailment*. All statistics are based on 100 runs.

Acknowledgments

We are grateful to Emily Pitler, Dipanjan Das, and the members of the Johns Hopkins Computation and Psycholinguistics lab group for helpful comments. Any errors are our own.

This project is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1746891 and by a gift to TL from Google, and it was conducted using computational resources from the Maryland Advanced Research Computing Center (MARCC). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, Google, or MARCC.

References

Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. 2019. Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 191–203, Florence, Italy. Association for Computational Linguistics.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Noam Chomsky. 1980. Rules and representations. *Behavioral and Brain Sciences*, 3(1):1–15.

Grzegorz Chrupała and Afra Alishahi. 2019. Correlating neural and symbolic representations of language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2952–2962, Florence, Italy. 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.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of BERT. *33rd Conference on Neural Information Processing Systems*.

Cleo Condoravdi, Dick Crouch, Valeria de Paiva, Reinhard Stolle, and Daniel G. Bobrow. 2003. Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning*, pages 38–45.

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

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing Textual Entailment: Models and Applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for

language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

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

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Adam Liška, Germán Kruszewski, and Marco Baroni. 2018. Memorize or generalize? Searching for a compositional RNN in a haystack. In *Proceedings of the 2018 workshop on Architectures and Evaluation for Generality, Autonomy, and Progress in AI (AEGAP)*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

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

Pranava Madhyastha and Rishabh Jain. 2019. On model stability as a function of random seed. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 929–939, Hong Kong, China. Association for Computational Linguistics.

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

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

R. Thomas McCoy, 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.

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.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on STILTs: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2018. Why comparing single performance scores does not allow to draw conclusions about machine learning approaches. *arXiv preprint arXiv:1803.09578*.

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

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Alex Warstadt and Samuel R Bowman. 2020. Can neural networks acquire a structural bias from raw linguistic data? *Proceedings of the 42nd Annual Conference of the Cognitive Science Society.*

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. 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 2870–2880, Hong Kong, China. Association for Computational Linguistics.

Noah Weber, Leena Shekhar, and Niranjan Balasubramanian. 2018. The fine line between linguistic generalization and failure in Seq2Seq-attention models. In *Proceedings of the Workshop on Generalization in the Age of Deep Learning*, pages 24–27, New Orleans, Louisiana. Association for Computational Linguistics.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

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

Xiang Zhou, Yixin Nie, Hao Tan, and Mohit Bansal. 2020. The curse of performance instability in analysis datasets: Consequences, source, and suggestions. *arXiv preprint arXiv:2004.13606*.