Can You Really Reason: A Novel Framework for Assessing Natural Language Reasoning Datasets and Models

Shanshan Huang(✉)

Shanghai Jiao Tong University, Shanghai, China
huangss_33@sjtu.edu.cn

Abstract. Recent studies have illuminated a pressing issue in the domain of natural language understanding (NLU) and reasoning: many of these datasets are imbued with subtle statistical cues. These cues, often unnoticed, provide sophisticated models an unintended edge, allowing them to exploit these patterns, leading to a potentially misleading overestimation of their genuine capabilities. While the existence of these cues has been noted, a precise and systematic identification has remained elusive in existing literature. Addressing this gap, our paper presents a novel lightweight framework. This framework is meticulously designed to not only detect these hidden biases in multiple-choice NLU datasets but also rigorously evaluate the robustness of models that are developed based on these datasets. By unveiling these biases and assessing model integrity, we aim to pave the way for more genuine and transparent advancements in NLU research.

Keywords: Dataset bias · Model bias · Model robustness

1 Introduction

The advancements in neural network models have yielded significant enhancements in a plethora of tasks, including natural language inference [1,20], argumentation [11], commonsense reasoning [9,14,23], reading comprehension [6], question answering [19], and dialogue analysis [7]. However, recent studies [4,12,15] have unveiled that superficial statistical patterns, including sentiment, word repetition, and shallow n-gram tokens in benchmark datasets, can forecast the correct answer. These patterns or features, termed as spurious **cues** when appearing in both training and test datasets with similar distributions. When these cues are neutralized, leading to a "stress test" [8,10,13], models exhibit reduced performance, suggesting an overestimation of their capabilities when evaluated on these datasets.

Several natural language reasoning tasks, exemplified by those in the Stanford Natural Language Inference (SNLI) dataset, can be cast as multiple-choice questions. A typical question can be structured as follows:

B. Luo et al. (Eds.): ICONIP 2023, CCIS 1969, pp. 54–66, 2024.
https://doi.org/10.1007/978-981-99-8184-7_5

Example 1. An instance from SNLI. **Premise**: A swimmer playing in the surf watches a low flying airplane headed inland.
Hypothesis: Someone is swimming in the sea.
Label: a) Entail. b) Contradict. c) Neutral.

Humans approach these questions by examining the logical relations between the premise and the hypothesis. Yet, previous work [10,16] has unveiled that several NLP models can correctly answer these questions by only considering the hypothesis. This observation often traces back to the presence of artifacts in the manually crafted hypotheses within many datasets. Although identifying problematic questions with a "hypothesis-only" test is theoretically sound, this approach often i) relies on specific models like BERT [3], which require costly retraining, and ii) fails to explain why a question is problematic.

This paper puts forth a lightweight framework aimed at identifying simple yet impactful cues in multiple-choice natural language reasoning datasets, enabling the detection of problematic questions. While not all multiple-choice questions in these datasets include a premise, a hypothesis, and a label, we detail a method to standardize them in Sect. 2. We leverage words as fundamental features in crafting spurious cues, since they serve as the foundational units in modeling natural language across most contemporary machine learning methods. Even complex linguistic features, such as sentiment, style, and opinions, are anchored on word features. Subsequent experimental sections will demonstrate that word-based cues can detect statistical bias in datasets as effectively as the more resource-demanding hypothesis-only method.

2 Approach

We evaluate the information leak in the datasets using only statistical features. First, we formulate a number of natural language reasoning (NLR) tasks in a general form. Then, based on the frequency of words associated with each label, we design a number of metrics to measure the correlation between words and labels. Such correlation scores are called "cue scores" because they are indicative of potential cue patterns. Afterward, we aggregate the scores using a number of simple statistical models to make predictions.

2.1 Task Formulation

Given a question instance x of an NLR task dataset X, we formulate it as

$$x = (p, h, l) \in X, \tag{1}$$

where p is the context against which to do the reasoning, and p corresponds to the "premise" in example 1; h is the hypothesis given the context p. $l \in \mathcal{L}$ is the label that depicts the type of relation between p and h. The size of the relation set $\mathcal{L}$ varies between tasks. We argue that most of the discriminative NLR tasks can be formulated into this general form. For example, an NLI question consists of a *premise*, a *hypothesis*, and a *label* on the relation between the premise and hypothesis. $|\mathcal{L}| = 3$ for three different relations: *entailment*, *contradiction*, and *neutral*. We will discuss how to transform into this form in Sect. 2.4.

2.2 Cue Metric

For a dataset X, we collect a set of all words $\mathcal{N}$ that exist in X. The cue metric for a word measures the disparity of the word's appearance under a specific label. Let w be a word in $\mathcal{N}$, we compute a scalar statistic metric called *cue score*, $f_{\mathcal{F}}^{(w,l)}$, in one of the following eight ways. We categorized the metrics into two genres: the first four use only statistics, and the last four use a notion of angles in the Euclidean space. Let $\mathcal{L}' = \mathcal{L} - L \setminus l$, and we define

$$\#(w, \mathcal{L}') = \sum_{l' \in \mathcal{L}'} \#(w, l'). \tag{2}$$

Frequency (Freq)
The simplest measurement is the co-occurrence of words and labels, where $\#()$ denotes naive counting. This metric aims to capture the raw frequency of words appearing in a particular label.

$$f_{Freq}^{(w,l)} = \#(w, l) \tag{3}$$

Relative Frequency (RF)
Relative Frequency extends the Frequency metric by accounting for the total frequency of the word across all labels. It's defined as follows:

$$f_{RF}^{(w,l)} = \frac{\#(w, l)}{\#(w)} \tag{4}$$

Conditional Probability (CP)
The Conditional Probability of label l given word w is another way to capture the association between a word and a label. This metric is essentially the Relative Frequency as defined above.

$$f_{CP}^{(w,l)} = p(l|w) = \frac{\#(w, l)}{\#(w)} \tag{5}$$

Point-wise Mutual Information (PMI)
PMI is a popular metric used in information theory and statistics. It measures the strength of association between a word and a label. PMI is higher when the word and label co-occur more often than would be expected if they were independent. We define the PMI of word w and label l as follows, where $p(w)$ and $p(l)$ are the probabilities of w and l respectively, and $p(w, l)$ is the joint probability of w and l.

$$f_{PMI}^{(w,l)} = \log \frac{p(w, l)}{p(w)p(l)} \tag{6}$$

Local Mutual Information (LMI)

The LMI is a variant of PMI that weighs the PMI by the joint probability of the word and label. This has the effect of giving more importance to word-label pairs that occur frequently. The LMI of word w with respect to label l is defined as follows.

$$f_{LMI}^{(w,l)} = p(w,l) \log \frac{p(w,l)}{p(w)p(l)}. \tag{7}$$

Ratio Difference (RD)

The Ratio Difference metric measures the absolute difference between the word-label ratio and the overall label ratio. This metric helps identify words that are disproportionately associated with a specific label.

$$f_{RD}^{(w,l)} = \left| \frac{\#(w,l)}{\#(w,\mathcal{L}')} - \frac{\#(l)}{\#(\mathcal{L}')} \right| \tag{8}$$

Angle Difference (AD)

Angle Difference is similar to *Ratio Difference* but accounts for the non-linear relationship between the ratios by taking the arc-tangent function. This metric can be more robust to outliers.

$$f_{AD}^{(w,l)} = \left| \arctan \frac{\#(w,l)}{\#(w,\mathcal{L}')} - \arctan \frac{\#(l)}{\#(\mathcal{L}')} \right| \tag{9}$$

Cosine (Cos)

The Cosine metric considers $v_w = [\#(w,l), \#(w,\mathcal{L}')]$ and $v_l = [(\#(l), \#(\mathcal{L}')]$ as two vectors on a 2D plane. Intuitively, if v_w and v_l are co-linear, w leaks no spurious information. Otherwise, w is suspected to be a spurious cue as it tends to appear more with a specific label l. This metric quantifies the similarity of the word-label relationship in a geometric manner.

$$f_{Cos}^{(w,l)} = \cos(v_w, v_l) \tag{10}$$

Weighted Power (WP)

The Weighted Power metric combines the Cosine metric with a frequency-based weighting, emphasizing the importance of words with higher frequencies. This metric can help prioritize cues that are more likely to impact the model.

$$f_{WP}^{(w,l)} = (1 - f_{Cos}^{l})\#(w)^{f_{Cos}^{l}} \tag{11}$$

In general, we can denote the *cue score* of a word w w.r.t. label l as $f^{(w,l)}$, by dropping the method subscript $\mathcal{F}$.

These metrics provide different perspectives on the association between words and labels, which can help identify potential spurious correlations.

2.3 Aggregation Methods

We can use simple methods $\mathcal{G}$ to aggregate the cue scores of words within a question instance x to make a prediction. These methods are designed to be easily implemented and computationally efficient, given the low-dimensional cue features.

Average and Max
The most straightforward way to predict a label is to select the label with the highest average or maximum *cue score* in an instance.

$$\mathcal{G}average = \arg\max l \frac{\sum_w f^{w,l}}{|x|}, l \in \mathcal{L}, w \in \mathcal{N} \tag{12}$$

$$\mathcal{G}max = \arg\max l \max_w (f^{w,l}), l \in \mathcal{L}, w \in \mathcal{N} \tag{13}$$

Linear Models
To better utilize the *cue score* in making predictions, we employ two simple linear models: SGDClassifier and logistic regression. The input for the models is a concatenated vector of *cue scores* for each label in instance x:

$$\begin{aligned} input(x) =& [f^{w_1,l_1},, ..., f^{w_d,l_1}, f^{w_1,l_2}, ..., f^{w_d,l_2}, \\ & ..., f^{w_1,l_t}, ..., f^{w_d,l_t}]. \end{aligned} \tag{14}$$

Here, d denotes the length of x. In practice, input vectors are padded to the same length. The training loss for the linear model is:

$$\hat{\phi n} = \arg\min \phi_n loss(\mathcal{G}_{linear}(input(x); \phi_n)) \tag{15}$$

The loss is calculated between the gold label l_g and the predicted label $\mathcal{G}linear(input(x); \phi_n)$. ϕ_n represents the optimal parameters in $\mathcal{G}linear$ that minimize the loss for label l_g.

2.4 Transformation of MCQs with Dynamic Choices

Until now, we have focused on multiple-choice questions (MCQs) that are classification problems with a fixed set of choices. However, some language reasoning tasks involve MCQs with non-fixed choices, such as the ROCStory dataset. In these cases, we can separate the original story into two unified instances, $u_1 = (context, ending1, false)$ and $u_2 = (context, ending2, true)$. We predict the label probability for each instance, $\mathcal{G}(input(u_1); \phi)$ and $\mathcal{G}(input(u_2); \phi)$, and choose the ending with the higher probability as the prediction.

3 Experiment

We proceed to demonstrate the effectiveness of our framework in this section. We apply our method to detect cues and measure the amount of information leakage in 12 datasets from 6 different tasks, as shown in Table 1. Our experimental findings are segmented into five sub-sections: Datasets, Quantifying Information Leakage, Bias Evaluation Methods, Comparison with Hypothesis-only Models, and Identifying Problematic Datasets.

Table 1. Dataset examples and normalized version.

Task Name	Datasets	Example			
		Original	"Premise"	"Hypothesis"	label
Natural Language Inference	SNLI, MNLI, QNLI	(SNLI) **Premise**: A woman and a child holding on to the railing while on trolley. **Hypothesis**: The people are not holding on anything. **Label**: contradiction	A woman and a child holding on to the railing while on trolley .	The people are not holding on anything.	contradiction
Argumentation	ARCT, ARCT_adv	(ARCT) **Reason**: Milk isn't a gateway drug even though most people drink it as children. **Claim**: Marijuana is not a gateway drug. **Warrant 1**: Milk is similar to marijuana. **Warrant 2**: Milk is not marijuana.	Milk isn't a gateway drug even though most people drink it as children. Marijuana is not a gateway drug.	Milk is similar to marijuana.	true
Reading Comprehension	RACE, RECLOR		Milk isn't a gateway drug even though most people drink it as children. Marijuana is not a gateway drug.	Milk is not marijuana.	false
Commonsense Reasoning	ROCStory, COPA, SWAG	(COPA) The woman hummed to herself. What was the cause for this? **Alternative1**: She was in a good mood. **Alternative2**: She was nervous.	The woman hummed to herself. What was the cause for this?	She was in a good mood.	true
Question Answering	CQA		The woman hummed to herself. What was the cause for this?	She was nervous.	false
Dialogue Analysis	Ubuntu				

3.1 Datasets

In this section, we present the results of our experiments conducted on 12 diverse datasets as outlined in Table 1. The datasets can be broadly classified into two categories based on the tasks they present: NLI classification tasks and multiple-choice problems. The NLI classification tasks constitute the first type. They are, in essence, a specialized variant of multiple-choice datasets. The second type includes datasets like ARCT, ARCT_adv [16], RACE [6], and RECLOR [22]. In these, one of the alternatives is the "hypothesis", and the "premise" contains more than a single context role. As an example, in ARCT, **Reason** and **Claim** act as the "premise", requiring the correct warrant to be chosen. Other datasets like Ubuntu [7], COPA [14], ROCStory, SWAG [23], and CQA [19] belong to the second type as well but have only a single context role in the "premise".

Table 2 outlines how hypotheses are gathered in these datasets. Most datasets utilize human-written hypotheses, barring CQA and SWAG.

3.2 Quantifying Information Leakage

In our effort to effectively measure the severity of information leakage or bias in these datasets, we formulated a measurement expressed as $\mathcal{D} = Acc - Majority$. Here, $Majority$ is the accuracy achieved through majority voting and Acc represents the accuracy of a model that bases its prediction solely on spurious cues.

A high absolute value of $\mathcal{D}$ indicates the existence of more cues in a dataset. However, a smaller $\mathcal{D}$ doesn't necessarily mean less bias in the training data, but rather less "leakage" between the training and test data. If $\mathcal{D}$ is positive, it

Table 2. The methods of hypothesis collection for the datasets. AE = Adversarial Experiment, LM = language model, CD = crowdsourcing, Human represents human performance on the datasets.

Datasets	Data Size	Data source	AE	Human(%)
ROCStory	3.9k	CD	No	100.0
COPA	1k	CD	No	100.0
SWAG	113k	LM	Yes	88.0
SNLI	570K	CD	No	80.0
QNLI	11k	CD	No	80.0
MNLI	413k	CD	No	80.0
RACE	100k	CD	No	94.5
RECLOR	6k	CD	No	63.0
CQA	12k	CD	No	88.9
ARCT	2k	CD	No	79.8
ARCT_adv	4k	CD	Yes	-
Ubuntu	100k	Random Selection	No	-

implies the model is utilizing the cues for its prediction. This evaluation method can be universally applied to any multiple-choice dataset.

3.3 Cue Evaluation Methods

The primary technique we use in our analysis is the hypothesis-only method, which we use as a gold standard for examining the existence of spurious cues. This method assumes that the model can only access the hypothesis and has to make its prediction without considering the premise.

To simplify this process and to find a measure that is as close to the hypothesis only method, we employed four simpler methods to make decisions based solely on spurious statistical cues. These methods include the average value classifier (Ave), the maximum value classifier (Max), SGD classifier (SGDC), and logistic regression (LR). These are outlined in detail in Sect. 2.

The main difference between our methods and the hypothesis-only method lies in the type of cues used. While our method uses word-level cues that are interpretable, the hypothesis-only method uses more complex cues, which are not easily interpretable.

3.4 Comparison with Hypothesis-Only Models

Our research aimed to assess and validate our proposed bias detection methods, chiefly by comparing their performance with hypothesis-only models. The goal was to demonstrate the effectiveness of our method in identifying spurious

Table 3. The Pearson Score of $\mathcal{D}$ on 12 datasets, between our methods and hypothesis-only models, fastText and BERT. P is the average Pearson score of BERT and fastText(FT).

	Ave			Max			SGDC			LR		
	FT	BERT	P	FT	BERT	P	FT	BERT	P	FT	BERT	P
PMI	90.87	96.23	93.55	95.37	79.82	87.59	97.81	91.01	94.41	97.14	96.05	96.6
LMI	65.13	49.18	57.16	34.52	30.71	32.62	69.88	79.06	74.47	77.46	81.21	79.33
AD	84.62	72.49	78.56	90.75	73.02	81.89	93.73	76.24	84.98	97.56	86.91	92.24
WP	86.87	73.09	79.98	92.47	79.87	86.17	94.0	22.59	56.53	61.28	75.55	65.86
RD	96.59	93.82	95.21	98.23	91.04	94.63	94.30	93.98	94.14	94.21	95.59	94.90
Cos	94.84	82.94	88.89	92.73	75.40	84.07	98.08	87.86	92.97	87.38	78.44	82.91
Freq	68.00	50.02	59.01	34.45	30.67	32.56	64.08	67.11	65.60	74.58	88.64	81.61
CP	93.09	96.61	94.85	95.29	79.80	87.54	97.19	96.16	96.67	97.17	97.34	**97.26**

statistical cues in multiple-choice datasets, underpinning the contribution we introduced.

In the context of this experimental comparison, we utilized the Pearson Correlation Coefficient (PCC) to measure the similarity between our method and the established hypothesis-only models, specifically fastText and BERT. The analysis encompassed a range of twelve datasets, making use of eight distinct cue score metrics and four aggregation algorithms.

The outcomes of this analysis, as depicted in Table 3, highlight that the CP cue score coupled with the logistic regression model achieved high correlations across all twelve datasets when compared to the gold standard hypothesis-only models. The PCC scores obtained were 97.17% with fastText and 97.34% with BERT. These remarkable results led us to conclude that the combination of CP and logistic regression forms a robust method for evaluating all datasets in subsequent experiments. The detailed data behind this study is comprehensively presented in the Appendix.

Given these findings, we are confident in asserting that our CP based approach is a powerful tool in identifying problematic word features within datasets, through the calculation of a "cueness score" described in Sect. 2. Furthermore, the coupling of CP and logistic regression offers a compelling measure to determine the extent to which multiple-choice datasets are affected by information leakage, a significant contribution to this field of research.

Further, we visualized our findings by plotting $\mathcal{D}$ for our CP+LR method and two hypothesis-only models (fastText and BERT) on 12 datasets in Fig. 1. The close tracking lines in the plot clearly indicate the strong correlation between our method and the hypothesis-only models.

Overall, our method effectively identifies and quantifies biases in the datasets, and the strong correlation with hypothesis-only models demonstrates the validity and effectiveness of our approach.

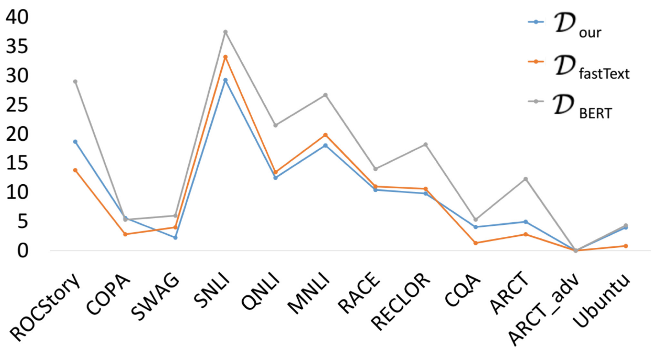

Fig. 1. Deviation scores for three prediction models on all 12 datasets.

3.5 Identifying Problematic Datasets

To better discern problematic datasets, we developed a criterion based on our experiment findings. According to this criterion, if a model's $\mathcal{D}$ exceeds 10% on any cue feature, the dataset is deemed problematic. This straightforward criterion allows for a quick identification of datasets with severe statistical cue issues.

Table 4. Highest accuracy of our 4 simple classification models on 12 datasets and the deviations from majority selection.

Datasets	Majority	Word Cues	
	(%)	Acc.(%)	$\mathcal{D}$(%)
ROCStory	50.0	68.68	**18.68**
COPA	50.0	55.60	5.60
SWAG	25.0	27.23	2.23
SNLI	33.33	62.57	**29.24**
QNLI	50.0	62.49	**12.49**
MNLI	33.33	51.37	**18.04**
RACE	25.0	35.42	**10.42**
RECLOR	25.0	34.80	9.80
CQA	20.0	23.42	3.42
ARCT	50.0	54.95	4.95
ARCT_adv	50.0	50	0.0
Ubuntu	1.0	4.96	3.96

As per this criterion, we identified ROCStories, SNLI, MNLI, QNLI, RACE, and RECLOR as datasets with considerable statistical cue problems. These findings are detailed in Table 4, which highlights the selection results using word cue features on several datasets. For some of these datasets, our methods significantly outperform the random selection probability, showcasing the extent of the statistical cues present. For instance, in the case of the ROCStories dataset, the highest accuracy achieved with our methods exceeds the random selection probability by 20.92%, and even higher for the SNLI dataset by 33.59%. This indicates that the datasets contain substantial spurious statistical cues that the models can exploit.

In the case of manually intervened datasets without adversarial filtering, such as ARCT, we found that they contained more spurious statistical cues. For instance, human adjustments to the ARCT dataset(ARCT_adv) have a notable impact on accuracy (from 54.95% to 50%).

Finally, in Table 4, we report the highest accuracy of our four simple classification models on the 12 datasets, along with the deviations from majority selection. Our findings reveal that deviation $\mathcal{D}$ can effectively identify problematic datasets. We can thus use $\mathcal{D}$ to assess the extent to which a dataset contains word cues.

In conclusion, our analysis and criteria for problematic datasets can help researchers identify datasets with substantial statistical cue issues. This critical insight can improve the development of more robust models that do not rely on superficial cues.

4 Related Work

Our work is related to and, to some extent, comprises elements in three research directions: spurious features analysis, bias calculation.

Spurious Features Analysis has been increasingly studied recently. Much work [17,18,23] has observed that some NLP models can surprisingly get good results on natural language understanding questions in MCQ form without even looking at the stems of the questions. Such tests are called "hypothesis-only" tests in some works. Further, some research [15] discovered that these models suffer from insensitivity to certain small but semantically significant alterations in the hypotheses, leading to speculations that the hypothesis-only performance is due to simple statistical correlations between words in the hypothesis and the labels. Spurious features can be classified into lexicalized and unlexicalized [1]: lexicalized features mainly contain indicators of n-gram tokens and cross-ngram tokens, while unlexicalized features involve word overlap, sentence length, and BLUE score between the premise and the hypothesis. [10] refined the lexicalized classification to Negation, Numerical Reasoning, Spelling Error. [8] refined the word overlap features to Lexical overlap, Subsequence, and Constituent which also considers the syntactical structure overlap. [15] provided unseen tokens an extra lexicalized feature.

Bias Calculation is concerned with methods to quantify the severity of the cues. Some work [2,5,21] attempted to encode the cue feature implicitly by hypothesis-only training or by extracting features associated with a certain label from the embeddings. Other methods compute the bias by statistical metrics. For example, [22] used the probability of seeing a word conditioned on a specific label to rank the words by their biasness. LMI [16] was also used to evaluate cues and re-weight in some models. However, these works did not give the reason to use these metrics, one way or the other. Separately, [13] gave a test data augmentation method, without assessing the degree of bias in those datasets.

5 Conclusion and Future Work

We have addressed the critical issue of statistical biases present in natural language understanding and reasoning datasets. We have proposed a lightweight framework that automatically identifies potential biases in multiple-choice NLU-related datasets and assesses the robustness of models designed for these datasets. Our experimental results have demonstrated the effectiveness of this framework in detecting dataset biases and evaluating model performance.

As future work, we plan to further investigate the nature of biases in NLU datasets and explore more sophisticated techniques to detect and mitigate these biases. Additionally, we aim to extend our framework to other types of NLU tasks beyond multiple-choice settings. By continuing to refine our understanding of dataset biases and their impact on model performance, we hope to contribute to the development of more robust, accurate, and reliable NLU models that can better generalize to real-world applications.

References

1. Bowman, S., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 632–642 (2015)
2. Clark, C., Yatskar, M., Zettlemoyer, L.: Don't take the easy way out: ensemble based methods for avoiding known dataset biases. 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), pp. 4060–4073 (2019)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., Smith, N.A.: 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, vol. 2 (Short Papers), pp. 107–112 (2018)
5. He, H., Zha, S., Wang, H.: Unlearn dataset bias in natural language inference by fitting the residual. EMNLP-IJCNLP **2019**, 132 (2019)
6. Lai, G., Xie, Q., Liu, H., Yang, Y., Hovy, E.: RACE: large-scale reading comprehension dataset from examinations. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 785–794 (2017)

7. Lowe, R., Pow, N., Serban, I.V., Pineau, J.: The ubuntu dialogue corpus: a large dataset for research in unstructured multi-turn dialogue systems. In: Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 285–294 (2015)
8. McCoy, T., Pavlick, E., Linzen, T.: 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, pp. 3428–3448 (2019)
9. Mostafazadeh, N., et al.: A corpus and cloze evaluation for deeper understanding of commonsense stories. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 839–849 (2016)
10. Naik, A., Ravichander, A., Sadeh, N., Rose, C., Neubig, G.: Stress test evaluation for natural language inference. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 2340–2353 (2018)
11. Niven, T., Kao, H.Y.: Probing neural network comprehension of natural language arguments. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4658–4664 (2019)
12. Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., Van Durme, B.: Hypothesis only baselines in natural language inference. In: Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, pp. 180–191 (2018)
13. Ribeiro, M.T., Wu, T., Guestrin, C., Singh, S.: Beyond accuracy: Behavioral testing of NLP models with checklist. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020, pp. 4902–4912 (2020)
14. Roemmele, M., Bejan, C.A., Gordon, A.S.: Choice of plausible alternatives: an evaluation of commonsense causal reasoning. In: 2011 AAAI Spring Symposium Series (2011)
15. Sanchez, I., Mitchell, J., Riedel, S.: Behavior analysis of NLI models: Uncovering the influence of three factors on robustness. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long Papers), pp. 1975–1985 (2018)
16. Schuster, T., Shah, D.J., Yeo, Y.J.S., Filizzola, D., Santus, E., Barzilay, R.: Towards debiasing fact verification models. arXiv preprint arXiv:1908.05267 (2019)
17. Sharma, R., Allen, J., Bakhshandeh, O., Mostafazadeh, N.: Tackling the story ending biases in the story cloze test. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 752–757 (2018)
18. Srinivasan, S., Arora, R., Riedl, M.: A simple and effective approach to the story cloze test. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 2 (Short Papers), pp. 92–96 (2018)
19. Talmor, A., Herzig, J., Lourie, N., Berant, J.: CommonsenseQA: a question answering challenge targeting commonsense knowledge. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers), pp. 4149–4158 (2019)
20. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: GLUE: a multi-task benchmark and analysis platform for natural language understanding. EMNLP **2018**, 353 (2018)
21. Yaghoobzadeh, Y., Tachet, R., Hazen, T., Sordoni, A.: Robust natural language inference models with example forgetting. arXiv preprint arXiv:1911.03861 (2019)

22. Yu, W., Jiang, Z., Dong, Y., Feng, J.: Reclor: A reading comprehension dataset requiring logical reasoning. arXiv preprint arXiv:2002.04326 (2020)
23. Zellers, R., Bisk, Y., Schwartz, R., Choi, Y.: Swag: A large-scale adversarial dataset for grounded commonsense inference. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 93–104 (2018)