Understanding the Influence of Question Types for Question Answering

Ashima Arora

Abstract

Pre-trained models have often shown to achieve high dataset-specific performance. However, recent studies have proposed that these high performances can often be attributed to these models' exploring biases in the dataset without learning the intended task. In this paper, I modified a randomly sampled subset of the adversarial SQuAD dataset (Jia et al., 2017) by changing the answer types in the adversarial sentences of the adversarial examples. The modification yielded a 6-point boost in the EM and 4-point boost in the F1-score in the dataset showing evidence towards the models' reliance on the question types along with lexical biases for its predictions instead of performing the intended reading comprehensions. Then, I used an ensemble based self-debiasing framework (Utama et al., 2020) to train a biased model in ensemble with a debiased model to prevent the model from mainly utilizing biases without specifically targeting a certain bias. The self- debiasing model showed close to a 1-point boost in the F1-Score in the modified adversarial examples by down-weighing biased examples in the debiased training objective.

1 Introduction

Pre-trained models such as ELECTRA often achieve remarkable performance on many reading comprehension tasks such as Question Answering (QA). On a benchmark dataset for QA such as Stanford Question Answering Dataset (SQuAD) by (Rajpurkar et al., 2016), it achieves 86% F1 score. However, when the same model was evaluated on adversarial SQuAD (Jia et al., 2017) where adversarial sentences (semantics altering perturbations) were added to the existing development set, the accuracy of the model dropped down to 69% F1 score. In my experiment, the same adversarial sentences were modified to change their answer types (Figure 1) such as adding a date value in the adversarial answer for a

Article: Nikola Tesla
Paragraph: *"In 1875, Tesla enrolled at Austrian Polytechnic in Graz, Austria, on a Military Frontier scholarship. During his first year, Tesla never missed a lecture, earned the highest grades possible, passed nine exams (nearly twice as many required), started a Serbian culture club, and even received a letter of commendation from the dean of the technical faculty to his father, which stated, \"Your son is a star of first rank.\" Tesla claimed that he worked from 3 a.m. to 11 p.m., no Sundays or holidays excepted. He was \"mortified when [his] father made light of [those] hard won honors.\" After his father's death in 1879, Tesla found a package of letters from his professors to his father, warning that unless he were removed from the school, Tesla would be killed through overwork. During his second year, Tesla came into conflict with Professor Poeschl over the Gramme dynamo, when Tesla suggested that commutators weren't necessary. At the end of his second year, Tesla lost his scholarship and became addicted to gambling. During his third year, Tesla gambled away his allowance and his tuition money, later gambling back his initial losses and returning the balance to his family. Tesla said that he \"conquered [his] passion then and there,\" but later he was known to play billiards in the US. When exam time came, Tesla was unprepared and asked for an extension to study, but was denied. He never graduated from the university and did not receive grades for the last semester. Tadakatsu enrolled at an engineering school in the year 1850."*
Question: *"What year did Tesla enroll at an engineering school?"*
Original Prediction: 1875
Prediction under adversary: 1850

Modified adversarial sentence: *Tadakatsu enrolled at an engineering school in the year Thomas Charles*.
Prediction with modified adversary: Thomas Charles

Figure 1: Modified adversarial dataset (in pink) was created by changing the answer type in the adversarial example (in blue) to an incompatible answer type under the given the question type.

'where' question asking about location. The modified answer type in the adversarial example would not contradict the correct answer or confuse humans. It is expected to improve performance for pre-trained models relying on question types to make their predictions, since the answers in the modified adversarial sentences would no longer be appropriate for the respective question types. The modified adversarial dataset showed 6-point increase in the EM score and 4-point increase in the F1-score. Although, there is an increase in the EM score from the baseline ADDONESENT adversarial sentences which is indicative of the model's reliance on question types for its predictions, it is also important to note that there are greater number of examples where the model continued to rely on a lexical overlap between the question and context to makes its final prediction despite a change in the answer type. An analysis of the model's predictions on the discussed dataset highlights that for *'what'* type questions, the model is unable to distinguish a sentence that actually answers the question from one that merely has words in common with it (Jia et al., 2017). It relies on biased features primarily the lexical overlap between n-grams in the question and the given context.

In this paper, I used a self-debiasing framework within which two models – a shallow model f_b and main model f_d – of the same architecture are lined to address the discussed lexical bias and other unknown biases (Utama et al., 2020). In this framework, a shallow biased model f_b is first obtained automatically with no prior knowledge of existing biases unlike other debiasing frameworks (Clark et al., 2019) that require a knowledge of the existing biases *a-priori*. The predictions of the shallow model p_b are then utilized by an example reweighing method to down-weigh the model's confidence on biased examples in the training objective.

2 Approach

2.1 Data

In this paper, the model was trained on SQuAD (Rajpurkar et al., 2016) training dataset. For evaluation, the baseline model was evaluated with the ADDONESENT adversarial SQuAD dataset which is a model independent dataset (Jia et al., 2017) that adds a random human-approved sentence to the context paragraph of the original SQuAD dataset. A sample of 500 random examples from the ADDONESENT dataset were chosen. These were then used to create a new dataset where the answer type in each adversarial sentence was changed to an inappropriate answer type with respect to the given question type (Figure 1). Like (Jia et al., 2017) this is a semantic altering perturbation which should not be able to confuse humans on the correct answer.

Article: Hugenot
Paragraph: "French Huguenots made two attempts to establish a haven in North America. In 1562, naval officer Jean Ribault led an expedition that explored Florida and the present-day Southeastern U.S., and founded the outpost of Charlesfort on Parris Island, South Carolina. The Wars of Religion precluded a return voyage, and the outpost was abandoned. In 1564, Ribault's former lieutenant Ren\u00e9 Goulaine de Laudonni\u00e8re launched a second voyage to build a colony; he established Fort Caroline in what is now Jacksonville, Florida. War at home again precluded a resupply mission, and the colony struggled. In 1565 the Spanish decided to enforce their claim to La Florida, and sent Pedro Men\u00e9ndez de Avil\u00e9s, who established the settlement of St. Augustine near Fort Caroline. Men\u00e9ndez' forces routed the French and executed most of the Protestant captives. The German event of Central Park caused the Bretons to abandon Acme."

Question: *"What European event caused the Huguenots to abandon Charlesfort?"*
Original Prediction: The Wars of Religion
Prediction under adversary: German event of Central Park

Figure 2: After changing the question into declarative form, the n-grams and syntactic dependency structure of the adversarial example (in blue) matches the question more closely than the actual answer in the original context (in green).

2.2 Motivation and Analysis

On examining the model's predictions on the original ADDONESENT examples, I observed that the model is more likely to assign a higher probability to an adversarial answer (answer in the adversarial example) if the syntactic variability between the question and the adversarial sentence was lower than between the question and the sentence in original context that actually answers the question. In other words, after a question is paraphrased into declarative form, whenever the adversarial sentence had a higher number of

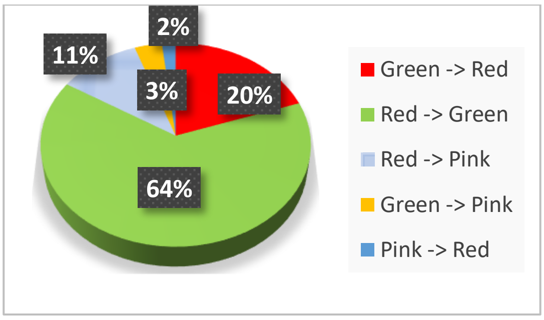


Figure 3: Analysis of the adversarial examples whose predictions were changed in the modified adversarial examples. '->' marks a transition from sampled adversarial examples to their modified version. Green indicates correctly predicted examples, Red indicates incorrectly predicted examples. Pink is to indicate examples that gave a correct answer but picked longer contexts than the list of valid predictions.

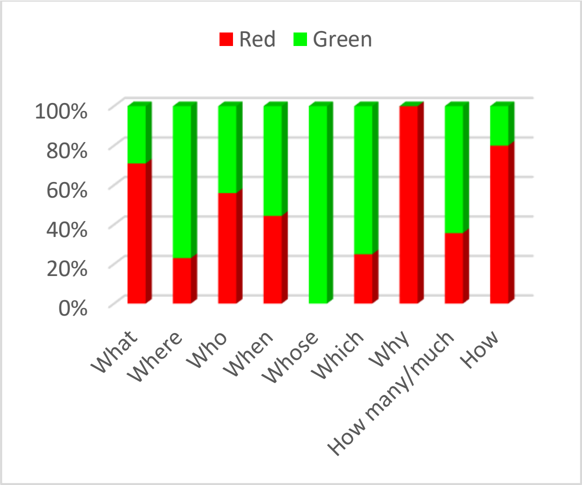


Figure 4: A categorical view of the percentage of each question type that were predicted correctly (green) in the modified adversarial examples vs those predicted incorrectly (red).

matching n-grams or a closer match to the syntactic structure with the said declarative form, the model would predict the answer in the adversary as opposed to answer in the context which signifies that the model has a **label bias** (Shah et al., 2020). More specifically, it relies on a lexical overlap between the question and the sentences in the context and chooses its answer from a sentence with the highest lexical overlap and not on an understanding of the question itself (Figure 2).

To further investigate whether the model actually performs reading comprehension, a random sample of 500 examples from the ADDONESENT dataset were modified to change the answer types in the adversarial sentences to be different from and incompatible with the actual answer types for the respective questions (Figure 1). This was done with the motivation of understanding whether the model truly comprehends the questions or whether it solely relies on the question type to answer questions and make its final predictions.

The original ADDONESENT dataset's adversarial sentences do not explicitly change the answer type corresponding to the given question. For example, a 'where' question asking about a location will have the adversarial sentence replace that location name for another location name like *Quebec* in the original context changed for *Australia* in the adversarial sentence. In this scenario, for a model relying on the question type for its predictions, the adversarial answer is then an equally likely answer. Under this impression, the modified adversarial dataset was changed such that the answer type in the adversarial sentence would now go from a location to another answer type, ex. *Australia* changed to *helping*, such that a model relying on the question type to answer its questions would now consider the adversarial answers as incompatible and continues to predict with answers picked from the original context in most cases.

On comparing the scores between the original adversarial ADDONESENT dataset and the modified version of it, there is roughly a 6% increase in Exact Match score and a 4% increase in the F1-score. Out of 230 adversarial examples, there were 55 examples in the modified version where the prediction was changed from the predictions in original adversarial ADDONESENT dataset. A breakdown of the observed changes can be seen in Figure 3. Out of the 36 examples where the prediction was corrected in the modified version, 21 of them were *where*, *how many*, *when*, *who* question types and the other 15 were "simpler" *what* questions with numerical answers ex: what distance. These examples were then predicted correctly in the modified adversary because the model is seeking for a numeric answer that is no longer found in the modified adversarial example which contains a common noun or location answer type in its place, providing evidence for the models' reliance on question types during prediction.

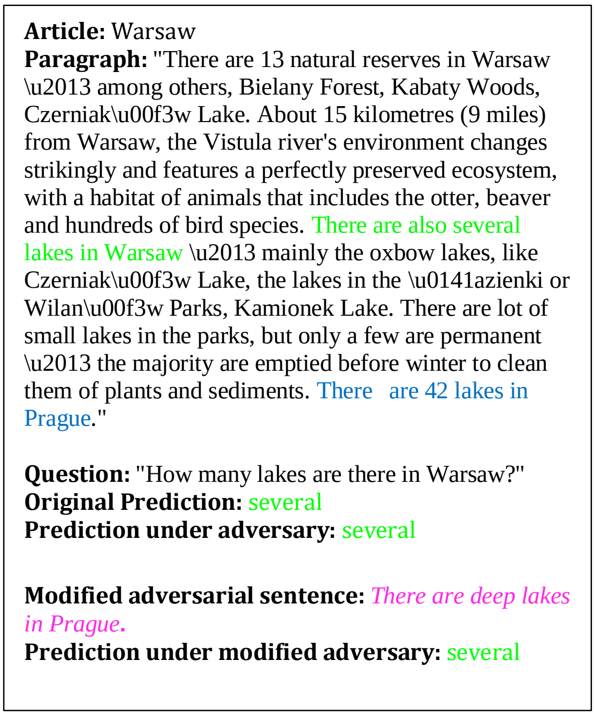
Out of the remaining 175 examples where the predictions remained the same, 69 of them continued to be predicted correctly in the modified adversaries and 106 of them continued to be predicted incorrectly in the modified adversary. Of the 106 that remained incorrect despite a change in the answer type, there were 14 *'who'* type questions and 9 of them were answered by the model by incorrectly picking any person or group name from the context. Similarly, for *'when'* and '*how many'* type questions, the model incorrectly picked any dates or numeric answers from the context. It was observed that 71 (66.9%) of those 106 questions where *'what'* type questions that required a deeper comprehension of the question rather than using lexical overlap patterns to answer. Maximally, in these *'what'* type questions, the model chose an answer from sentences with maximal overlap with the question (lexical bias) without worrying about the expected data type. For example, it spit out date type answer for questions looking for a person.

On the other hand, the 69 examples that continued to remain correct were observed to have a wider distribution of different question types. There seemed to be a wider range of categories covered but it is important to note that most of these required little understanding of the question itself. The *'what'* type questions here were simpler with most of them requiring numerical answers in a context that usually had only one numeric value present. 34 out of 69 of them were *'who'*, *'when'*, *'where'*, '*how many'* type questions with sentences in the original context that had a substantial amount of the lexical overlap with the question. In other words, these examples allowed the model to use its existing lexical biases to correctly answer the questions (Figure 5).

3 Experiment

3.1 Setup

For the experiment, I measured the Exact Match (EM) and F1-score across 500 randomly sampled examples from the ADDONESENT dataset as well as the modified ADDONESENT dataset where the answer types were changed in the adversarial sentences (Table 1). These two datasets formed the baseline for the comparing changes in the debiased model. The same scores were used to evaluate performance against for the ensemble-based debiasing model (Table 2).



Article: Warsaw
Paragraph: "There are 13 natural reserves in Warsaw \u2013 among others, Bielany Forest, Kabaty Woods, Czerniak\u00f3w Lake. About 15 kilometres (9 miles) from Warsaw, the Vistula river's environment changes strikingly and features a perfectly preserved ecosystem, with a habitat of animals that includes the otter, beaver and hundreds of bird species. There are also several lakes in Warsaw \u2013 mainly the oxbow lakes, like Czerniak\u00f3w Lake, the lakes in the \u0141azienki or Wilan\u00f3w Parks, Kamionek Lake. There are lot of small lakes in the parks, but only a few are permanent \u2013 the majority are emptied before winter to clean them of plants and sediments. There are 42 lakes in Prague."

Question: "How many lakes are there in Warsaw?"
Original Prediction: several
Prediction under adversary: several

Modified adversarial sentence: *There are deep lakes in Prague*.
Prediction under modified adversary: several

Figure 5: An example where the prediction does not change from original to ADDONESENT to modified ADDONESENT examples given that there is same amount of lexical overlap in the original context, the adversarial and the modified adversarial sentence.

3.2 Hyperparameters

The experiment requires training a shallow bias model in ensemble with the main debiased model. For the shallow biased model, I followed the suggestion given by (Utama et al., 2020) for parametrizing a biased Quora Question Pairs

(QQP)[1] model. My shallow model was trained on 500 random examples from the SQuAD training dataset. I trained the model for 4 epochs and with a learning rate of 2e-05. For the main model, the training ran over the default setting of 3 epochs and a learning rate of 5e-05.

3.3 Training a shallow model

As per (Utama et al., 2020), the self-debiasing framework requires training a shallow model f_b by training a copy of the main model on a small random subset of the main training dataset for several epochs. The model f_b is then used to make predictions on the remaining unseen examples. The probabilities p_b assigned by the shallow model $f_b(x^{(i)}) = p_b$ to each training instance $x^{(i)}$ is indicative of the likelihood of that instance containing bias. More specifically, the probability $p_b^{(i,c)}$ assigned by the model to the index of correct label c can be used to interpret the model's bias in the following manner: when the model predicts $x^{(i)}$ correctly with high confidence, i.e., it assigns a probability of 1 or close to 1 to $x^{(i)}$, then $x^{(i)}$ is potentially biased. On the other hand, if it makes an overconfident error, i.e., it assigns a probability of 0 or close to 0 to $x^{(i)}$, then $x^{(i)}$ is likely a harder example for the model from which the model should learn.

For training a shallow biased model, I obtained 500 random samples from the SQuAD training dataset. The model was trained with the discussed hyperparameters settings. Through this training, I obtained the start and end logits assigned by this biased model to each training example. Then, the remaining unseen training examples were put through an evaluation loop to obtain the start and end logits assigned by this shallow model to these examples during evaluation.

3.4 Debiased training objective

The main debias model f_d is then trained by utilizing the shallow model f_b. Specifically, p_b from the shallow model is used by an existing model agnostic debiasing method, i.e., example reweighing (Clark et al., 2019) to lower the importance of the biased instances in the training objective. Example reweighing adjusts the importance of training instance by assigning a scalar weight value to the log probability p_d of the instances. The scalar weight is computed using the following: $1\text{-}p_b^{(i,c)}$. The loss is then computed as follows:

$$L(\theta_d) = -(1 - p_b^{(i,c)})\, y^{(i)} \cdot \log p_d$$

Where $y^{(i)}$ is the one hot vector associated with the index of the correct label and p_d is the softmax output of f_d.

3.5 Main (debiased) model

The main debiased model f_d is trained on the entire SQuAD training dataset. The start_logits and end_logits obtained from the shallow model f_b were used to compute the start and end probability values for each training instance by taking a softmax of the start and end logits. Then, the probability value associated with the correct start and end index was obtained by using the start_positions and end_positions from inputs parameter passed to the compute_loss method. Then, the scalar weights for the start and end positions in the debias model were obtained by using $1\text{-}p_b^{(i,s)}$ and $1\text{-}p_b^{(i,e)}$ respectively, where s is index of the correct start position and e is the index of the correct end position. The start_loss and end_loss values are then computed by multiplying the scalar start and end weight value with the dot product of the one hot vector of the start and end positions with the log softmax of their corresponding start and end logits computed by the main model f_d. An average of the start and end loss formed the total loss returned by the compute_loss method.

4 Results

First, a baseline was established by running the default model with a learning rate of 5e-05, an embedding size of 128, 12 hidden layers and a hidden size of 256 for 3 epochs. This model was evaluated for Exact Match (EM) and F1-score on the three datasets: the original SQuAD

[1] The dataset is available at https://www.kaggle.com/c/quora-question-pairs

development dataset, 500 randomly sampled ADDONESENT examples, and the modified version of the 500 sampled adversarial examples (Table 1).

Then, the model was trained under a self-debiasing framework (Utama et al., 2020) as described in section 3. The debiased (or main) model under this framework was trained on the default settings like the baseline version (3 epochs, 128 embedding size, 5e-05 learning rate, 12 hidden layers and 256 hidden size). It was then evaluated similarly on the three datasets in question (Table 2).

	SQuAD Dev	Adversarial	Modified Adversarial
EM	78.38	56.4	61.2
F1-Score	86.17	65.34	69.33

Table 1: Metrics – Baseline (before debiasing).

	SQuAD Dev	Adversarial	Modified Adversarial
EM	75.97	55.6	61.2
F1-Score	84.81	65.267	69.99

Table 2: Metrics – After debiasing.

Overall, the debiased model showed 2.4% decrease in EM score and 1.36% decrease in F1-score in the SQuAD development dataset. According to (Utama et al., 2020) this can be attributed to the fact that the shallow model f_b is likely to capture multiple types of biases leading to many examples being down-weighted in the debiased training objective. As a result, the effective training data size is reduced, and it leads to a drop in in-distribution performance in the debiasing methods. As a solution to this, they propose an annealing mechanism where the weight is gradually decreased throughout training using a linear schedule. (The annealing mechanism was not used in my experiment since their results showed better improvements on Question Answering without the annealing mechanism). Similarly, there is a 0.8% decrease in the EM score and 0.07% decrease in the F1-score in the debiased randomly sampled ADDONESENT dataset. Although, the debiased model is starting to show more examples being picked from the context rather than the adversary, it seems to pick wrong amounts of context in its prediction (Figure 6). Sometimes, it picks more context than present in the list of valid predictions. Other times, it picks less context than necessary to be considered a valid prediction (more on this pattern is discussed later).

Article: Victoria_(Australia)
Paragraph: "On 1 July 1851, writs were issued for the election of the first Victorian Legislative Council, and the absolute independence of Victoria from New South Wales was established proclaiming a new Colony of Victoria. Days later, still in 1851 gold was discovered near Ballarat, and subsequently at Bendigo. Later discoveries occurred at many sites across Victoria. This triggered one of the largest gold rushes the world has ever seen. The colony grew rapidly in both population and economic power. In ten years the population of Victoria increased sevenfold from 76,000 to 540,000. All sorts of gold records were produced including the \"richest shallow alluvial goldfield in the world\" and the largest gold nugget. Victoria produced in the decade 1851\u20131860 20 million ounces of gold, one third of the world's output[citation needed]. The population of Adelaide increased Thomas Coke in the eleven years after the discovery of gold."

Question: "How much did the population of Victoria increase in ten years after the discovery of gold?”
Valid Predictions: “sevenfold” , “ 76,000 to 54,000”
Original Prediction: sevenfold

Prediction under adversary (baseline): sevenfold
Prediction under modified adversary (baseline): sevenfold
Prediction under modified adversary (debiased): sevenfold from 76,000 to 540,000

Figure 6: A sample prediction under the debiased model that provides the correct answer but also picks context that makes the answer incorrect amongst the list of valid predictions.

Between the modified adversarial examples in the baseline model and the debiased model, there is no change in the EM score and a 0.66% increase in the F1-score. An analysis of the predictions between the two show that 391 of the 500 examples showed no change in the final prediction results from the modified ADDONESENT adversarial dataset in the baseline model to the modified ADDONESENT adversarial dataset in the debiased model. However, it was observed that for most cases the debiased model is more likely to pick shorter answers than the baseline model except for *‘what’* and *‘how many/much’* type questions where it is likely to pick more context than the gold predictions.

In the 109 examples where the prediction was influenced by the debiasing framework, 28% of these examples were incorrect in the sampled 500 examples from the ADDONESENT dataset and were predicted correctly in the modified adversarial dataset and then continued to be predicted correctly in the debiased model as well. As discussed originally, these were mostly *'how many'*, *'when'*, *'where'* or simpler *'what'* questions (like what date) that no longer contained an appropriate value in the adversary sentence for the model to predict. The debiased model could have continued to rely on its lexical bias to answer these questions.

The more interesting results exist in the other 72% of the 109 examples. 26% of these were incorrectly predicted in the baseline modified adversary and were predicted correctly in the debiased model. Out of this 26%, 64% were the more complex *'what'* type questions that the model failed to reason through appropriately in the baseline models and 21% of them were *'who'* type questions in a context containing more than one person or entity name/ For these questions, the baseline model had chosen to pick a random person or group from the context as its prediction but the debiased model predicted the correct person or entity.

6.5% of the 109 examples were *'what'* type questions that were picked incorrectly from the modified adversary in the baseline model and came close to the correct prediction in the debiased model but picked more words from the context than the gold prediction values. For example, predicting *meritocratic bureaucracy* instead of *meritocratic* in the gold predictions.

Roughly 16% of the 109 examples were predicted completely correctly in the modified adversary of the baseline model but only came close to predicting the correct answer in the debiased model. A general problem was that the debiased model picked varied amounts of context along with the correct prediction values (Figure 6) lowering the EM score. Regardless, the debiased model generally performed better than baseline on *'what'* type questions which as discussed earlier were more likely to be predicted incorrectly in the baseline adversarial dataset using its lexical biases, given that they require a deeper comprehension of the question in comparison to other question types.

5 Conclusion

Overall, it is fair to say that pre-trained models tend to rely on biases present in the dataset. For Question Answering, the model based most of its predictions for *'what'* type questions on a lexical overlap between the question and the sentences in the context. For other question types, such as *'how many'*. *'who'*, *'where'*, the question type for a given example is important given that the model seems to rely on it to predict the correct answer type instead of comprehending the question itself. Similar ideas are suggested through recent studies on passage only datasets (Kaushik et al., 2018). The self-debiasing framework by (Utama et al., 2020) yielded roughly a 1-point increase in the F1-score on the modified adversarial dataset. It has advantages over ensemble based debiasing models that require a knowledge of the bias *a-priori*. Since, the biased model is of great importance in this self-debiasing framework, there is room to tune the hyperparameters of the biased/shallow model which may help achieve more impactful results for future systems.

Acknowledgments

I would like to thank Dr. Durrett for all the informative lectures and the TAs for their help throughout the course.

References

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. 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)*, pages 4069–4082, Hong Kong, China. Association for Computational Linguistics.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. 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.

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.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Towards Debiasing NLU Models from Unknown Biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610, Online. Association for Computational Linguistics.