# Understanding and Fixing Bias in a Natural Language Inference Model to Increase Overall Accuracy on the Standford Natural Language Inference 'validation' Split

**Jason Schwarz, PT, DPT, MSDS**
**jasonschwarz@utexas.edu**

## Abstract

Dataset artifacts existing in natural language datasets may teach natural language inference (NLI) models bias. Bias in inference models will allow those models to see the hypotheses only, and yet predict more accurately than guessing the most common label because of the artifacts in the hypotheses. Understanding this bias will help researchers train models with less, or best-case scenario, no bias. This paper has two objectives. First, I reconstruct bias exploration and understanding. Second, I train a new model on the residuals of a biased model to fix the bias while improving overall accuracy on Standford Natural Language Inference's (SNLI) 'validation' split.

## 1 Introduction

Natural language processing (NLP) attempts to provide a digital transformation of abstract ideas (eg. words) to binary representations. This requires enormously large amounts of compute power, data, and parameters. Researchers are currently exploring a myriad of language tasks to see how NLP can appropriately and efficiently make the transformations. Moving beyond the digital transformations themselves, the value of their applications far exceeds the value of the transformations themselves. For example, word processing programs use some level of NLP to predict grammar and spelling suggestions to make language clearer.

A subset of NLP is natural language inference (NLI). As the name suggests, NLI attempts to infer an idea from a piece of text or pieces of text. For instance, NLI can compare two pieces of text and predict if the statements have similar meanings ("entailment"), cannot be determined ("neutral"), or contradictory meanings ("contradiction") (Dagan et al., 2006, 2013). Researchers refer to the first sentence as the premise and to the second sentence as the hypothesis (Dagan et al., 2006, 2013).

Training an NLI model to make such predictions requires vast amounts of examples. Each example needs a premise, hypothesis, and label. Labels are typically represented as classes "0", "1", and "2" for entailment, neutral, and contradiction, respectively. Generating these examples is labor intensive when done by humans.

The SNLI dataset contains over 500,000 human generated examples with the above format. Humans looked at a picture, then wrote four sentences about it, one for each class of inference and one premise. Researchers may now use this dataset to train a specific NLI model.

Many times, a pretrained model will be imported then trained on a specific NLI task. This pretrained model is trained on a large corpus of data. This trains the word embeddings. Embeddings are digital representations of single words in many dimensions. It is not uncommon to have 300 dimensions to represent each word as in GloVe vectors (Pennington et al., 2014).

## 2 Motivation

A valuable application of a trained NLI model is to compare any two sentences generated from any source and infer their entailment status. For instance, in the medical field, having a computer recognize the phrase "a broken bone" is entailed with "a fractured femur" will help clinicians sort through troves of medical records in less time. This will allow them to perform other duties not done well by computers.

Ideally, a trained NLI model will use only the premise and hypothesis to make predictions.

However, when a model works well seeing only the hypothesis, something else beyond "comparing the hypothesis to the premise" is learned during training. This is referred to as model bias. The specifics of the dataset that engender the bias are referred to as dataset artifacts. Working well is defined as making more correct predictions than if the model were to simply predict the most common label for every hypothesis.

The hypotheses in the SNLI dataset have been shown to contain dataset artifacts by Rudinger et al. (2017). When training an NLI model with the hypotheses-only from the SNLI dataset and the ELECTRA-small pre-trained model (Clark et al., 2020), the final model can then predict using the hypotheses-only in evaluation mode with an accuracy of 69.2% (Poliak et al., 2018). The most common label in this dataset is 0 (entailment); it represents 33.8% percent of the examples. Thus, the final model has an accuracy that is an absolute increase of 35.4%, or a relative increase of 104.7% more than if it guessed the most common label. These numbers indicate the model can predict significantly more examples correctly just based on what it sees in the hypotheses. The model learned bias from the dataset artifacts found in the hypotheses.

When a model does learn bias, understanding what the bias is will help create more accurate, less biased models in the future. The bias can cause models to perform poorly when making predictions on contrast sets (Gardener et al., 2020), which is not desirable. This paper explores various contributions to bias and ways to reduce it.

## 3 Methodology

One way to measure bias involves testing a model by allowing it to see the hypotheses only. This clearly speaks to the evaluation of the model. Significantly, both the training and evaluation methods play an important role.

(Dis)similarity of training and evaluation methodologies describe whether models are evaluated similarly to how they were trained. For instance, a model trained using premises, hypotheses, and labels, and then evaluated on premises and hypotheses is said to have similar training and evaluation methodologies. On the other hand, a model trained using premises, hypotheses, and labels, and then evaluated on hypotheses only is said to have *dis*similar training and evaluation methodologies.

Models tend to do best when they are evaluated in a similar fashion as to how they were trained. Intuitively, changing the training and evaluation methodologies will decrease predictive power of a model because it is being asked to do something differently than how it was trained.

When attempting to measure bias, one must consider any change in training/evaluation methodology. Changing the methodology could be a confounding variable when measuring bias. Therefore, to measure bias in a model using the hypothesis only, a model is trained in this report using only the hypothesis and label. Of note, the hypothesis-only model may have limited real-world applicability, but it serves as a proxy for detecting bias in the premise and hypothesis model.

Because of the differences in bias and bias with a confounding variable, I will make a distinction between "bias" and "confounding bias".

In Part 1 of this experiment, I train two models over three epochs on the SNLI dataset ("train" split) starting with the ELECTRA-small pre-trained model. The first model is trained using the premise, hypothesis, and label and is called the PHL model henceforth. The second model is trained using only the hypothesis and label and is called the HL model. The results are compared to the findings by Poliak et al. (2018). "Part 1: Understanding the Bias" is performed as a reconstruction attempt. (Their repository is not used in this attempt).

Both PHL and HL models are evaluated using premise-hypothesis and hypothesis-only methodologies using the "validation" split of the SNLI dataset.

The main indicators of bias will be how well a model predicts over the most common label occurrence of 33.8%.

I investigate potential sources of bias including sentence length, unique words to each class, and words with high percentage use in each class.

In "Part 2: Removing the Bias", I attempt to fix the bias while increasing overall accuracy on the 'validation' split. Ideally, the newly trained model executes the above by learning less bias while relying more on abstract meanings between the premises and hypotheses.

I train this new model on the residuals of the biased HL model along with the premises, hypotheses, and labels and call it the RPHL model. Here I attempt to recreate the format used by He et

| | Premise and hypothesis | | Hypothesis only | |
|---|---|---|---|---|
| | **Accuracy (%)** | **Mean Probability** | **Accuracy (%)** | **Mean Probability** |
| PHL | 89.5 | 0.975 | (48.1) | (0.941) |
| HL | (57.5) | (0.738) | 69.5 | 0.811 |
| RPHL | 90.7 | 0.963 | (51.8) | (0.906) |
| RHL | (77.2) | (0.759) | 77.2 | (0.759) |

Table 1. Accuracies of models when evaluated on premise and/or hypothesis. Mean probabilities are calculated from examples that the model predicted correctly. Parenthetical values indicate when then model was evaluated differently than how it was trained. For instance, "(57.5)" indicates the HL model's accuracy when evaluated on "premise and hypothesis" when it was trained on only the "hypothesis".

al. (2019). (Their repository is not used in this attempt). The idea behind using the residuals is the following four paragraphs.

When training a model on hypotheses only, the model learns bias in the hypotheses as it has no other information on which to train.

An optimum model can be described as the model that predicts as accurately as a human using both bias and abstract meanings. It makes its final prediction requiring abstract meanings of both premise and hypothesis without using bias. Training on the SNLI dataset is believed to learn bias *and* abstract meanings. To tease out the learning of abstract meanings, one can first learn the bias. Then, one can train on the residuals of the bias. This is argued to learn the abstract meanings in a separate model. (He et al., 2019).

Let $m^*$ be the optimum model function. Let $m_b$ be the biased model function. Let $m_{am}$ be the abstract-meanings model function. Let $x$ be the dataset. Let $x_{ho}$ be a hypothesis only version of the dataset. Let $m(x)$ be the logits that model function $m$ produces when provided input $x$.

Then, mathematically, one can define:

$$m^*(x) = m_b(x_{ho}) + m_{am}(x) \qquad (1)$$

The above equation and ideas are adapted from He et al., (2019).

A modified training loss is used to train the abstract-meanings model (aka. RPHL model) on the residuals of a biased model (aka. HL model). For each example provided during training, the biased model makes a prediction based on the hypothesis only. The abstract-meanings model makes a prediction based on the premise and hypothesis. Each model outputs logits. If the biased model's logits result in at least a 0.5 probability for the correct class, then those logits are added to the logits of the abstract-meanings model. Then, loss is calculated using cross entropy loss. Ground truth labels are used as the optimal model.

Because there are three classes, initial probabilities in an untrained model are roughly 33-33.7% for each class. Thus, choosing a threshold of 0.5 would indicate some level of bias. Moreover, I chose this threshold of 0.5 to capture weaker forms of bias in hopes to eliminate it before it strengthens.

During evaluation time, only the abstract-meanings model is used to make predictions. Since bias is already learned in HL, RPHL is thought to learn non-bias features (He et al., 2019). These non-bias features are thought to be the abstract meanings of the sentences.

Furthermore, I train a similar model to RPHL but on hypotheses only and call it RHL for residuals, hypotheses, and labels. This is used to compare to the HL model as they have similar training methodologies. Although this model is suspicious, it provides interesting analysis and is kept in this report.

## 4 Results

### 4.1 Part 1: Understanding the Bias

#### 4.1.1 Accuracies of Models with Similar Training and Evaluation Methodologies

As shown in Table 1, the PHL model predicts 89.5% labels correctly when provided premises

| | Entailment | Neutral | Contra-diction |
|---|---|---|---|
| Means | 6.8 | 8.3 | 7.4 |
| Standard deviations | 2.9 | 3.4 | 2.9 |

Table 2. Means and standard deviations of hypothesis sentence lengths (in words) in each class in the "evaluation" split of the SNLI dataset.

| Entailment | | Neutral | | Contradiction | |
|---|---|---|---|---|---|
| **Word** | **Count** | **Word** | **Count** | **Word** | **Count** |
| opposing | 5 | vacation | 13 | Nobody | 52 |
| physical | 4 | celebrating | 10 | sleep | 12 |
| liquid | 4 | happily | 10 | nothing | 11 |
| photographing | 3 | tour | 7 | anything | 8 |
| Theres | 3 | Halloween | 7 | napping | 7 |
| motion | 3 | last | 7 | sofa | 6 |
| motorcycle | 3 | meet | 6 | spaceship | 5 |
| accessories | 3 | discuss | 5 | porch | 5 |
| speaks | 3 | upcoming | 5 | ignoring | 5 |
| hills | 3 | likes | 5 | cow | 5 |

Table 3. Top 10 words unique to each class and their occurrence counts in the “evaluation” split of the SNLI dataset.

| Entailment | | | Neutral | | | Contradiction | | |
|---|---|---|---|---|---|---|---|---|
| **Word** | **Percent** | **Count** | **Word** | **Percent** | **Count** | **Word** | **Percent** | **Count** |
| instrument | 90.0 | 20 | tall | 93.2 | 44 | sleeping | 88.0 | 108 |
| touching | 83.3 | 12 | competition | 87.5 | 24 | driving | 81.1 | 53 |
| least | 90.0 | 10 | because | 82.6 | 23 | Nobody | 100 | 52 |
| Humans | 87.5 | 8 | birthday | 85.0 | 20 | alone | 90.0 | 50 |
| activity | 85.7 | 7 | got | 81.3 | 16 | cat | 83.7 | 49 |
| arts | 85.7 | 7 | win | 87.5 | 16 | asleep | 90.7 | 43 |
| screen | 85.7 | 7 | trip | 93.3 | 15 | no | 83.9 | 31 |
| speaking | 85.7 | 7 | married | 86.7 | 15 | empty | 92.9 | 28 |
| transportation | 85.7 | 7 | tries | 86.7 | 15 | eats | 83.3 | 24 |
| opposing | 100 | 5 | favorite | 86.7 | 15 | sleeps | 95.0 | 20 |

Table 4. Words with high (>= 80%) percentage use in one class and occur at least 5 times in the "evaluation" split of the SNLI dataset.

and hypotheses. The PHL model is thought to make predictions based on bias and abstract meanings. The HL model predicts 69.5% labels correctly when provided hypotheses only. The HL model is thought to use bias only to make predictions that accurately because it does not have access to the premises. Said differently, the increase from the most common label occurrence, which is 33.8%, to 69.5% in the HL model represents the bias learned. These percentages are similar to what Poliak et al., (2018) found. Their model predicted 69.17% of validation examples correctly. This slight difference in accuracies is likely due to preprocessing differences. I did not preprocess the data. Fortunately, these accuracies are close enough to proceed with this analysis.

#### 4.1.2 Accuracies of Models with Dissimilar Training and Evaluation Methodologies

As expected, the PHL model performs worse when provided hypotheses only. It drops from 89.5% to 48.1% accuracy. Similarly, the HL performs worse when provided premises and hypotheses. It drops from 69.5% to 57.5% accuracy. Most of the decrease in performance is thought to be due to changing the methodologies. Despite the confounding variable of a methodological change between training and evaluation time, bias still clearly presents itself.

| Word: "Nobody" | | | | | | |
|---|---|---|---|---|---|---|
| | | Correct prediction counts by class | | | | |
| | Training & Evaluation method | Entailment | Neutral | Contra-diction | Total occurrences | Mean Probability |
| Labels* | -- | 0 | 0 | 52 | 52 | -- |
| PHL | Premise & Hypothesis | 0 | 0 | 52 | 52 | 0.9996 |
| RPHL | Premise & Hypothesis | 0 | 0 | 52 | 52 | 0.9999 |
| HL | Hypothesis only | 0 | 0 | 52 | 52 | 0.9878 |
| RHL | Hypothesis only | 0 | 0 | 52 | 52 | 0.9876 |

Tabel 5. Label predictions of various models for sentences with "Nobody". Mean probability is weighted average across all correct predictions for that model. *Values in the "Labels" row are ground truth values, not predictions.

The PHL model displays bias, as it was able to accurately predict 48.1% examples. This is an absolute increase of 14.3% accuracy over the most common label occurrence of 33.8%. This amount of confounding bias should be used with caution but may yet provide a useful benchmark.

#### 4.1.3 Hypothesis Sentence Lengths

I initially analyze sentence lengths of the hypotheses. Table 2 shows the means and standard deviations between all three classes. Importantly, all the means fall within one standard deviation of the other class's means. This indicates the hypothesis-sentence-length distributions of each class have significant overlap with each other. Because of this obvious overlap, this is extremely unlikely to be causing bias. Therefore, I do not pursue this further.

#### 4.1.4 Words Unique to each Class

I next analyze words that are unique to each class. Table 3 shows the top ten in each class. The entailment class clearly has the lowest occurrence of unique words. The neutral and contradiction classes have roughly similar occurrences of unique words except for "Nobody" in the contradiction class. The total counts of unique words in each class are 540, 1356, and 1031 for entailment, neutral, and contradiction, respectively. Roughly 80-90% of the unique words are used only 1 or 2 times. The low prevalence of most of these words likely contribute only mildly to model bias.

Considering the words in the contradiction class, five of them have similar meanings. "Nobody", "sleep", "nothing", "napping", and "ignoring" all indicate a negation. Moreover, in each of the 8 occurrences of "anything", it was used as a negation. For instance, one actual sentence is, "A boy is not holding anything." Thus, a total of six words in the contradiction list have similar connotations.

#### 4.1.5 Words with High Percentage Use in each Class

Perhaps the most interesting source of dataset artifacts are words that are highly correlated to each class and exist at or above some occurrence threshold. I chose an occurrence threshold of 5. Words occurring fewer times than the occurrence threshold are thought to contribute only mildly to model bias. High correlation is set at 0.8. Table 4 shows the top ten words in each class with their percentage seen in its highest occurrence class and total occurrences of each word. This is similar to what Poliak et al., (2018) found in their analysis.

The contradiction class seems to contribute to model bias the most based on total number of occurrences. Next highest contributor to model bias is the neutral class.

Looking through the top ten words in the contradiction class, patterns can be found. For example, the idea of sleeping shows up twice, and the connotations of "Nobody", "alone", "no", are "empty" are similar (Poliak et al., 2018). It is much harder to find similar patterns in the entailment and neutral classes.

| Word: “sleeping” | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Training & Eval method | Correct prediction counts by class | | | Mean probability of correct predictions by class | | | Total correct | Total occurrences |
| | | Entail | Neutral | Contra | Entail | Neutral | Contra | | |
| Labels* | -- | 7 | 6 | 95 | -- | -- | -- | -- | 108 |
| PHL | Prem & Hyp | 5 | 5 | 65 | 0.9895 | 0.9697 | 0.9996 | 75 | 108 |
| RPHL | Prem & Hyp | 5 | 5 | 65 | 0.9409 | 0.8979 | 0.9998 | 70 | 108 |
| HL | Hyp | 3 | 1 | 66 | 0.6338 | 0.5539 | 0.9588 | 70 | 108 |
| RHL | Hyp | 5 | 1 | 66 | 0.5216 | 0.5488 | 0.9582 | 72 | 108 |

Table 6. Label predictions of various models for sentences with "sleeping". Mean probability split up by each class. *Values in the “Labels” row are ground truth values, not predictions.

#### 4.1.6 Specific Examples

This section focuses on two words, a few of their examples, and how the various models classified their sentences. The words are taken from the contradiction list of words in Table 4 as they are thought to contribute heavily to bias due to number of occurrences. “sleeping” is the most common occurring word with a high percentage use in the contradiction class. “Nobody” is the third most common. “Nobody” is analyzed here instead of “driving” (the second most common) because it is also unique to the contradiction class and may show something different.

However, Table 5 shows no significant difference between models’ predictions of hypotheses with the word “Nobody”. Each of the 4 models predict all 52 hypotheses correctly with high probability. An example of a “Nobody” sentence (contradiction class only) is, “Nobody has umbrellas.”

One “sleeping” sentence example in the dataset contradiction class is, “A girl is sleeping”; one example in the entailment class is, “A man holds two sleeping children.” Notice in the contradiction class, “sleeping” describes the subject of the sentence, whereas in the entailment class, “sleeping” describes the direct object of the sentence.

Table 6 shows correct label predictions by class for various models for sentences with “sleeping” in them. Although the number of correct predictions is exactly the same between models PHL and RPHL for all classes, the confidence of the RPHL predictions (indicated by gray fill) are slightly lower for “entailment” and “neutral” classes. This may indicate decreased bias in the RPHL model, which is one of the objectives of this report.

### 4.2 Part 2: Fixing the Bias

#### 4.2.1 Training a Model to learn more Abstract Meanings (and less Bias)

As seen in Table 1, the RPHL model resulted in an increase of overall accuracy on the ‘validation’ split of SNLI when provided the premises and hypotheses. Overall accuracy improved from 89.5% in the PHL model to 90.7% in the RPHL model. This absolute increase of 1.2% (which translates to 120 examples) is considered an important step. It demonstrates one of the strengths of training on residuals of a biased model. Moreover, it is important to understand why it makes this significant jump in overall accuracy.

#### 4.2.2 RPHL Model, Confounding Bias

Recall the PHL model accurately classified 48.1% of examples when provided the hypothesis only. Again, this is considered confounding bias in this report due to the change in training and evaluation methodologies. The average probability of correct labels in this scenario is 0.941.

Interestingly, the RPHL model accurately classified 51.8% of examples when provided the hypothesis only. The average probability of correct labels is 0.906.

Thus, the RPHL model’s accuracy is an absolute increase of 3.6% over the PHL model’s accuracy

with hypothesis only. However, the average probability of correct labels decreased by an absolute value of 3.5%. While the number of correctly predicted labels increased, the confidence of these predictions decreased. It is difficult to say whether this indicates more or less confounding bias between the models.

#### 4.2.3 RHL Model, Bias

The final model to examine is the RHL model. Its results are suspicious because it makes the exact same predictions (and loss values) whether it is provided the premises and hypotheses or hypotheses only from the SNLI 'validation' split. Thus, these results should be interpreted with caution. These results were evaluated three times each to eliminate doubt in its implementation.

The RHL model accurately predicts 77.2% of examples with an average probability of 0.759. Compare these values of the HL model; the HL predicts 69.5% of examples accurately with an average probability of 0.811.

Again, it appears that the residual based model increased accuracy but decreased confidence in its predictions.

## 5 Discussion

Bias in the above models is not inherently wrong. If I wanted to deploy a model to accurately predict the entailment class for a human-written sentence about a photo, then the bias would likely be welcomed. However, this represents a narrow scope for this model and limits its real-world applicability. Therefore, training a model to learn less bias and more abstract meanings of sentences holds high value.

As shown, the widely used SNLI dataset contains dataset artifacts. The source of the artifacts likely come from how the dataset was created.

Recall, humans looked at a picture and then wrote four sentences. One was the premise sentence describing what was in the picture. The other three were hypothesis sentences, one for each class. The entailment sentence arguably required less creativity and imagination than the neutral and contradictory sentences. The variation in cerebral energy required to write sentences for various classes may explain part of the differences found in Tables 3 and 4 (unique words in each class and words with high percentage use in one class).

Understanding the bias plays a crucial role in fixing it. One way not explored here is creating a dataset without artifacts. This report shows an alternate method. I trained a model on the residuals of a biased model in hopes to remove bias and make the model learn more of the abstract meanings of sentences.

Training a model on the residuals of a biased model theoretically makes sense. In this report, the accuracy of the RHPL did increase. How it increases is a significant factor to understand.

This analysis shows a residual-hypothesis-only model (RHL) that increased its accuracy (compared to a HL model) when given just the hypotheses. This increase causes concern as this appears to be "more bias" learned by the model. However, the decrease in confidence of the model provides some relief. One might argue the bias did in fact become weaker but is more widespread. Whether this is a move in the right direction is yet to be seen.

Furthermore, a case can be made stating the RPHL model learned different and weaker bias than the PHL model while making it more uniform across different classes. Therefore, the improvements of the RPHL model over the PHL model should be used with caution.

Perhaps the HL model learns the easy bias. Then, as the RPHL model trains on the HL model's residuals, the RPHL model is forced to pay more attention to harder examples while still allowing it to learn the easy examples. This may be the reason it performed better overall *and* better when provided just the hypotheses.

In another light, there are no changes in predictions of sentences with "Nobody" or "sleeping". It is possible these words have too strong of a correlation to be easily swayed. This seems to indicate that the RPHL model does not make progress on stronger correlated words, but on weaker correlated words.

Lastly, perhaps hyperparameters of the RPHL model, specifically the confidence threshold of 0.5, may need tuning to make even better progress.

## 6 Conclusion

Dataset artifacts exist in datasets such as the SNLI dataset and influence models to learn bias. The bias causes models to perform poorly when exposed to contrast sets. Therefore, eliminating bias is desirable.

This report attempts to remove bias by training a model to learn the residuals of a biased model. This model, the RPHL model, appears to increase overall accuracy on the SNLI 'validation' split while decreasing the strength of the bias but allowing it to be more widespread.

Further investigations will do well to understand bias learned by RPHL models and how to remove it. Before going too far down the path, it will also be wise to carefully consider dataset creation, model training, and model applicability. Ideally, all three will be as similar as possible. This provides the most benefit to humanity.

## References


Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pretraining Text Encoders as Discriminators Rather Than Generators. In Proceedings of the International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.2003.10555

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment, pages 177– 190. Springer.

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.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, et al.. 2020. Evaluating Models' Local Decision Boundaries via Contrast Sets. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1307–1323, Online. Association for Computational Linguistics.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn Dataset Bias in Natural Language Inference by Fitting the Residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pages 1532– 1543.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in Natural Language Inference. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social Bias in Elicited Natural Language Inferences. In The 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL): Workshop on Ethics in NLP.