Relation Leakage in Elicited Natural Language Inference Datasets

Michael Saxon, Xinyi Wang, Wenda Xu, William Yang Wang
University of California, Santa Barbara
Department of Computer Science
saxon@ucsb.edu, xinyi_wang@ucsb.edu, wendaxu@ucsb.edu, william@cs.ucsb.edu

Abstract

Natural language inference (NLI) is an important task for producing useful models of human language. Unfortunately large-scale NLI dataset production relies on crowdworkers who are prone to introduce biases in the sentences they write, enabling models to predict sentence pair relations from a single sentence in the pair, better than chance. This elicited sentence relation leakage property is undesirable as it enables models to cheat rather than learn the desired reasoning capabilities, and hasn’t gone away since its 2018 discovery. We analyze this problem in 8 modern NLI datasets, using a combination of previously established and novel model-based techniques, so as to enable ameliorating this leakage in future NLI datasets.

1 Introduction

Natural language inference (NLI) is an important task for producing useful models of human language. The NLI task is to, given a pair of sentences, typically referred to as the premise $s_1$ and hypothesis $s_2$, infer $r \in \{\text{neutral, entail, contradict}\}$, the logical relationship between the two sentences. Ideally, examples in an NLI dataset are drawn from a distribution where $r$ is conditionally dependent on the pair of sentences, but independent from either sentence individually (Wang et al., 2021c).

Large-scale NLI datasets are typically produced as follows: first, a set of seed sentences is sampled from some existing collection. Then, for each seed three new sentences are elicited from crowdworkers, who are instructed to write a corresponding sentence that satisfies a selected logical relation $r$ (Bowman et al., 2015). Usually, all the seeds in a dataset are assigned either $s_1$ or $s_2$, meaning either the the premise or hypothesis distribution is balanced and inferring the relation from a seed sentence alone is impossible. However, the elicited sentences aren’t balanced in the same way. Often undesired side-information about the label is inadvertently present in the elicited sentences, making better-than-chance estimation of $r$ from a single sentence possible.

These “leakage features” encoded in the elicited sentences (Zhang et al., 2019) are visible to NLI models employing machine learning (Poliak et al., 2018), enabling them to “cheat” rather than learning the desired logical reasoning ability, calling into question the appropriateness of NLI datasets as benchmarks for language understanding (Bowman and Dahl, 2021). In this work we rigorously analyze this problem of elicited sentence relation leakage in both popular and recent NLI datasets.
using novel techniques to enable targeted interventions to create higher quality future resources.

1.1 Background

Poliak et al. (2018) first introduced the problem of “hypothesis-only baselines” achieving high accuracy on several datasets, including the “judged” dataset SICK (Marelli et al., 2014), and “elicited” datasets SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018). They showed that a baseline, consisting of the InferSent (Conneau et al., 2017) BiLSTM encoder and an MLP classifier could predict the correct relation label at rates considerably higher than chance (e.g., hypothesis-only baseline achieving >2x the majority baseline performance on SNLI).

It is widely thought that systematic, shared biases in the types of words, sentence structures, or ideas that crowdworkers tend to choose when prompted with a logical relation drive this relation leakage in the hypothesis sentence (Gururangan et al., 2018; Zhang et al., 2019). For example, crowdworkers might slightly tend toward choosing certain non-sequitur topics when asked to write a neutral sentence, or occasionally not-invert the premise when given contradict, while avoiding using the word “not” when given entail.

Simple heuristics can be used to craft very challenging test sets that models trained on existing NLI resources struggle at, for example by creating premise-hypothesis pairs with high lexical overlap, but contradictory meanings (McCoy et al., 2020).

This thinking has led to producing adversarial (Nie et al., 2020) and counterfactually augmented (Kaushik et al., 2020) NLI datasets that are intended to provide more challenging generalization scenarios as well as to eliminate undesirable relation leakage biases. However, Bowman and Dahl (2021) observe that while these techniques do ameliorate the effects of benchmark saturation, “collecting examples on which current models fail can incentivize developing models that are different rather than better.” Multi-task learning (such as by larger, more diverse NLI datasets) can improve model robustness to fitting to spurious features (Tu et al., 2020), by essentially overwhelming model capacity to fit to all the various bias patterns available. However, as long as language understanding benchmarks include a core reliance on NLI datasets for which high elicited sentence relation leakage is present, the degree to which good results on them reflect the desired reasoning capability is questionable (Poliak et al., 2018).

Thus, producing NLI datasets with restricted relation leakage is important. The ideal NLI dataset would contain no observable relation leakage in either the premise or hypothesis distribution. While a dataset containing exactly one example of each relation for every sentence in both the premise and hypothesis populations would axiomatically achieve this, in practice building a non-trivial dataset that satisfies this condition at scale is challenging and has yet to be achieved.

We seek to address the following two needs:
1. An accounting of the relation leakage problem in datasets produced since (Poliak et al., 2018).
2. Concrete steps to enable future datasets that exhibit the relation leakage problem less.

1.2 Contributions

Our elicited sentence baselines generally achieve higher accuracy than Poliak et al. (2018)’s implementations on datasets where both were tested, demonstrating that the problem is even worse than previously suspected.

We show that the problem of relation leakage in NLI datasets still hasn’t been solved, by demonstrating hypothesis-only (or premise-only, when appropriate) baselines for 8 crowd-sourced NLI datasets including 5 released since 2018.

We compare state-of-the-art and replication model performance to the biased baselines and investigate how much they agree across the datasets.

Finally, we introduce a novel model-based metric and dataset analysis tool, the Progressive Evaluation of Cluster Outliers (PECO) score, that will enable future work producing NLI datasets that exhibit less elicited sentence relation leakage.

2 Datasets

We study eight datasets containing a total of 12 test or validation sets in which bias can be analyzed. We chose to focus on elicited datasets according to Poliak et al. (2018)’s typology (with the exception of SICK), where the data production process involves first sampling a set of sentences from a population of text and then asking annotators (usually crowd workers) to produce corresponding sentences to satisfy relations. A brief description of the datasets we analyze follows. Each dataset is introduced with its abbreviated name in bold. Table 1
### Table 1: Information on the NLI datasets we compare in this study. The ANLI dataset is further decomposed into partitions “A1,” “A2,” and “A3”. Vertical line denotes identical value in cell as above (sub-elements of same dataset)

| Dataset | Year | Elicited from | Lang | Elicited pop. | Train size | Dev size | Test size |
|---------|------|---------------|------|---------------|------------|----------|-----------|
| SICK    | 2014 | Image+Video Captions | En   | Hypothesis   | 4.4k       | 0.5k     | 4.9k      |
| SNLI    | 2015 | Image Captions + KB | En   | Hyp.          | 550k       | 10k      | 10k       |
| MNLI    | 2018 | Multiple Genre | En   | Hyp.          | 393k       | 20k      | 20k       |
| MNLI-m  |       |               |      |               | —          | 10k      | 10k       |
| MNLI-u  |       |               |      |               | —          | 10k      | 10k       |
| XNLI    | 2018 | MNLI          | 14 Langs. | Hyp. | —          | 70k      | 35k       |
| FEVER   | 2019 | Wikipedia     | En   | Premise       | 208k       | 20k      | 20k       |
| ANLI    | 2020 | Wikipedia+HotpotQA | En | Hyp. | 163k       | 3.2k     | 3.2k      |
| ANLI-A1 |       |               |      |               | 17k        | 1k       | 1k        |
| ANLI-A2 |       |               |      |               | 45k        | 1k       | 1k        |
| ANLI-A3 |       |               |      |               | 100k       | 1.2k     | 1.2k      |
| OCNLI   | 2020 | Multiple Genre | Zh   | Hyp.          | 47k        | 3k       | 3k        |
| CAugNLI | 2020 | IMDb+SNLI    | En   | Prem. & Hyp.  | 8.3k       | 1k       | 8.3k      |

Provides an overview of this information along with statistics such as train/dev/test set size, and which sentence is elicited (potentially biased).

**SICK** Sentences Involving Compositional Knowledge (Marelli et al., 2014) was produced by instructing annotators to label existing sourced pairs from 8K ImageFlickr data set (Young et al., 2014) and SemEval 2012 STS MSR-Video Description data set (Agirre et al., 2012). The dataset is in English. Each sentence pair was annotated for relatedness and entailment by means of crowdsourcing techniques.

**SNLI** The Stanford NLI dataset was produced using (Bowman et al., 2015). The corpus contains content from the Flickr 30k Corpus (Young et al., 2014), VisualGenome corpus (Krishna et al., 2017) and Gururangan et al. (2018). The corpus is in English. The dataset is collected through human-written English sentence pairs.

**MNLI** The Multi-genre NLI Corpus (Williams et al., 2018) is modeled on the SNLI corpus (Bowman et al., 2015) but it differs in the range of genres of spoken and written English text supporting cross-genre evaluation.

**XNLI** The Cross-Lingual NLI Corpus (Conneau et al., 2018) consists of manually-translated dev and test samples from MNLI in 14 languages: French, Spanish, German, Greek, Belgian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili, and Urdu. It is interesting for analysis because on a high level the semantics of the data follow MNLI. The corpus is made to evaluate the inference in any language when only English data is presented at training time.

**FEVER** NLI-style FEVER (Nie et al., 2019) is an NLI reformulation of the FEVER claim verification dataset (Thorne et al., 2018). The original dataset was collected by eliciting annotators to write fact sentences that are supported, refuted, or unverifiable relative source passages drawn from Wikipedia. This is converted into an NLI task by treating the elicited sentences as premises and the source passages as NLI pairs with relations entail, contradict, or neutral respectively. This dataset is unique in that the premises were elicited from seed hypotheses, meaning it has a balanced hyp. distribution but potentially biased prem. distribution.

**ANLI** The adversarial NLI corpus (Nie et al., 2020) is collected through crowdworkers and the purpose of this dataset creation is to make the state-of-art results fail in this dataset. The sentences are selected from the Wikipedia and manually curated HotpotQA training set (Yang et al., 2018). The language is in English. It contains three partitions of increasing complexity and size, which we refer to hereafter as A1, A2, and A3. Detailed data statistics are in Table 1.

**OCNLI** The Original Chinese NLI corpus was collected following MNLI-procedures but with strategies intended to produce challenging inference pairs (Hu et al., 2020). No translation was employed in producing this data; the source premise sentences and elicited hypotheses are original.

**CAugNLI** Kaushik et al. (2020) produced counterfactually augmented datasets for NLI and sentiment analysis using human annotators, instructing them to make minimal changes to the sentences beyond those necessary to change the label. It extends the work of Maas et al. (2011) and Bowman...
et al. (2015). They find that a BiLSTM classifier achieves negligible performance over chance when trained on hypothesis only. However, since their dataset includes elicited modified sentences in both the premise and hypothesis populations, there are opportunities for bias on both.

3 Methods

In this section, we explain how all numbers in the results table were produced. In particular this involved training models in the “normal” replication NLI condition and the “biased” (elicited population sentences only) condition, sourcing SOTA models from publicly available leaderboards, and deriving dataset-level metrics about the agreement between normal and biased condition models (to assess their agreement), and model-based latent space cluster analysis (to identify samples exhibiting maximum bias) for PECO.

3.1 SOTA Models

We searched publicly available leaderboards and Google Scholar citations to attempt to find the highest-test accuracy-performing models yet produced for each dataset. They are as follows:

| Dataset | Model                  |
|---------|------------------------|
| SICK    | NeuralLog (Chen et al., 2021) |
| SNLI    | EFL (Wang et al., 2021b)  |
| MNLI    | T5 (Raffel et al., 2019)  |
| ANLI    | InfoBERT (Wang et al., 2021a) |
| XNLI    | ByT5 (Xue et al., 2021)   |
| FEVER   | KILT (Petroini et al., 2021) |
| OCNLI   | RoBERTa-xwm-ext-1 (Xu et al., 2020) |
| CAugNLI | as of March 2022 we were unable to find SOTA results for this model surpassing our RoBERTa-large fine-tuned implementation. |

3.2 Model Training

We utilize a simple, yet relatively powerful replication technique that we apply equally to all datasets. Starting from pretrained transformer checkpoints (Wolf et al., 2019) we fine-tune roberta-large (Liu et al., 2019), xlm-roberta-base (Conneau et al., 2019), and bert-base-chinese (Devlin et al., 2018) using Pytorch Lightning. Table 2 lists relevant hyperparameters. All models were trained on NVIDIA A-100 GPUs. All models are the HuggingFace xForSequenceClassification with num_classes=3 and no other modifications. All models are trained using the Adam optimizer with cross entropy loss. We train using hyperparameters as specified in Table 2.

To replicate on the normal condition, we tokenize the premise and hypothesis sequences, with starting [CLS] token or model equivalent, and separator token </s> or equivalent depending on the model.

To train each dataset’s corresponding bias condition(s) model, we use the same setup as the replication model but follow Poliak et al. (2018)’s formulation of fine-tuning the chosen classification model on only the bias condition sentence, premise only or hypothesis only.

All datasets except FEVER and CAugNLI exhibit their bias in exclusively the hypothesis distribution, while FEVER exhibits bias in the premise distribution, and CAugNLI exhibits bias in both (Table 1).

3.3 Comparing Results

For each dataset we investigate three derived metrics from the model performance. They are:

- **Improvement over Chance:** We subtract the biased condition test accuracy from the accuracy achieved by a “guess majority label” strategy following (Poliak et al., 2018):
  \[ \Delta_{maj} = \text{Acc}(f_{bias}(X_{test}), Y_{test}) - \text{Acc}(Y_{maj}, Y_{test}) \]  

- **SOTA Accuracy Recovered:** Accuracy achieved by biased condition model over SOTA accuracy:
  \[ \%R_S = \frac{\text{Acc}(f_{bias}(X_{test}), Y_{test})}{\text{Acc}(f_{SOTA}(X_{test}), Y_{test})} \]

Table 2: Training information for the models used on each dataset. “R” denotes roberta-large, “XR” denotes xlm-roberta, and “BCh” denotes bert-base-chinese. Hyperparameters other than batch size kept equal between biased condition and normal condition training runs. “P.Epoch” denotes the fine-tuning epoch after which maximum validation accuracy was achieved.
Replication Accuracy Recovered: Accuracy achieved by biased model over replication:

\[
\%R = \frac{\text{Acc}(f_{\text{bias}}(X_{\text{test}}), Y_{\text{test}})}{\text{Acc}(f_{\text{Repl.}}(X_{\text{test}}), Y_{\text{test}})}
\] (3)

3.4 Comparing Normal and Biased Models

We are particularly interested in understanding the degree to which models trained in the normal and biased conditions “reason” similarly. To enable this analysis we define a simple agreement function \(A_g(Y_1, ..., Y_N)\) as the total number of entries that are identical for all \(Y_i\). This function is equivalent to accuracy for a vector of model decisions and ground truth labels (\(\text{Acc}(f(X), Y) = A_g(f(X), Y)\)).

Normal-Bias Agreement: The total number of samples for which the biased and normal models agree over the total number of samples in the set:

\[
NBA = \frac{A_g(f_{\text{bias}}(X_{\text{test}}), f_{\text{Repl.}}(X_{\text{test}}))}{|X_{\text{test}}|}
\] (4) \[\text{Normal-Bias Recovery: The total number of samples for which the biased and normal models agree, and both classify correctly over the total number of samples they agree on:}

\[
NBR = \frac{A_g(f_{\text{bias}}(X_{\text{test}}), f_{\text{Repl.}}(X_{\text{test}}), Y_{\text{test}})}{A_g(f_{\text{bias}}(X_{\text{test}}), f_{\text{Repl.}}(X_{\text{test}}))}
\] (5)

Token Relevance Agreement: To compare “reasoning” similarity between the two models, we produce normalized local explanation maps \(m(f(X)) = \text{norm}(\nabla_X(f(X)))\) using the gradient-based technique from Sundararajan et al. (2017). We then compute the cosine similarity:

\[
\text{TRA} = \frac{m(f_{\text{Repl.}}(X_{\text{test}})) \cdot m(f_{\text{Repl.}}(X_{\text{test}}))}{\|m(f_{\text{Repl.}}(X_{\text{test}}))\| \|m(f_{\text{Repl.}}(X_{\text{test}}))\|}
\] (6)

3.5 Cluster-based Bias Evaluation

We are interested in investigating how the biased distributions of the elicited sentences in NLI datasets are captured in the learned representation spaces of models trained on them. In particular, we are interested in answering this question: is elicited sentence label leakage captured semantically in regions of latent space?

To answer this we produce dimensionality-reduced elicited sentence embeddings for the test set, using the normal condition replication models, then fit a high-\(N\) kNN clustering to this collection of embeddings. This will allow us to analyze how the local distribution of labels varies over the elicited sentence embedding space. By comparing the KL-divergence of the label distribution within each cluster and the global label distribution, we can compute the Progressive Evaluation of Cluster Outliers (PECO) score (Figure 2).

Elicited Sentence Embeddings: To embed the elicited sentences as they’re learned by a model in the standard condition, we feed the elicited sentences \(s_e\) through the normal replication fine-tuned NLI model encoder. We extract the latent codes produced at the output very last fully connected layer of the model before the linear classifier to collect latent codes for every \(s_e\) in the test set. We then embed these codes into their 30 principal components to produce the embeddings (Figure 2 (a)).

Clustering: We fit a high-\(k\) (in this case, \(k = 50\)) \(k\)-means clustering over the distribution of elicited sentence embeddings to provide a set of local bins for analysis. For each cluster, we count the relation labels its samples contain, to produce a set of 50 cluster-label distributions (Figure 2 (b)).

Computing Cluster Divergences: For each cluster distribution \(p_i\), we assess the Kullback-Leibler divergence between it and the global label distribution \(p_G\) to produce divergence scores \(s_i\):

\[
s_i = D_{\text{KL}}(p_i||p_G) = \sum_{j=1}^{3} \frac{p_i \log \left( \frac{p_i}{p_G} \right)}{}
\] (7)

This step is depicted in Figure 2 (c).

Progressive Evaluation: Finally, we compute the PECO score for this collection of cluster divergences as the area under the curve produced by
4 Results and Discussion

Table 3 show the full results for the experiments. For all numbers, lower is better, but to differing degrees. For SOTA and Repl. accuracy, lower numbers denote “harder” datasets that are less saturated benchmarks. However, Bowman and Dahl (2021) point out that a less saturated benchmark isn’t necessarily a “better” benchmark, as it might reward difference rather than improvement.

For the remaining scores and metrics, lower is unambiguously better. Lower Bias Cond. and Maj. denote datasets that contains less observable elicited sentence relation leakage to sophisticated and very simple models, respectively. Lower Result Comparisons metrics are all directly reduced as a consequence of the aforementioned. PECO gets lower when the clusters, on average, are closer to the global distribution. Lower Normal-Biased Agreement numbers additionally indicate a measure of different “reasoning” between the biased condition model and normal model, an expected trait of a task where both sentences are consistently present), while the underlying meanings encoded by those words remain.

We find that the simple replication procedure produces models with quite close test accuracy to the SOTA, with a maximum relative accuracy difference of 8%, and a 92% Pearson’s correlation coefficient (PCC) between SOTA and Repl. accuracy across the datasets. We are confident in the suitability of our normal condition and bias models for comparing the NLI datasets.

4.1 Relation Leakage Remains a Problem

Elicited sentence relation leakage remains a problem for all evaluated NLI datasets. In particular, the XNLI, FEVER, and OCNLI datasets, for which no documented debiasing or adversarial data preparation efforts were made, exhibit high absolute bias condition performance over majority ($\Delta_{maj} > 20$).

Although ANLI (Nie et al., 2020) and CAugNLI (Kaushik et al., 2020) are improvements over the others in terms of $\Delta_{maj}$, with CAugNLI shining particularly in this regard, none eliminate the relation leakage problem entirely.

4.2 Model Capacity and ANLI

While ANLI “improves” on the other benchmarks by serving as a harder task with lower SOTA accuracy, it exhibits concerning bias patterns. In particular, biased condition models achieve $\%R_R > 95%$! Shockingly, RoBERTa-large trained in the hypothesis only condition is able to almost reach the SOTA accuracy for ANLI R2 and R3.

Interestingly, the bias condition and normal condition models only agree $\approx 50%$ of the time. This means that 25% of the samples in ANLI R3 test are only correctly classified by the normal model, 25% are correctly classified by both, and 25% are only correctly classified in the biased condition. This might suggest that the ANLI production process has indeed produced a pattern in its samples complex enough that fitting to both the bias pattern in the hypotheses and the true logical relationships between the premises and hypotheses is beyond the capacity of RoBERTa-large or the training techniques that have been attempted thus far.

4.3 XNLI and Semantic vs Lexical

A topic of debate surrounding cheating features in NLI (and NLP more broadly) is the extent to which annotator’s biases toward certain surface-level lexical features or deeper semantic preferences are driving the relation leakage.

While previous work has focused on finding words, phrases, patterns, and heuristics in the surface form of the data, our biased condition study of XNLI provides an interesting opportunity to investigate the potential for the influence of deeper semantics.

For testing XNLI we use the multilingual xlm-roberta model which is initially pretrained on a massive multilingual corpus in order to have a latent space that captures crosslingual semantic representations.

However, XNLI does not come exclusively with val and test sets, and no training set. We are forced to train on MNLI alone. However, the test and val sentences in XNLI are produced by having human annotators translate the meanings of MNLI test and val samples. This produces a natural experiment wherein surface form biases present in the training data are completely eradicated in the test set (as only the 14 non-English languages Table 1 are present), while the underlying meanings encoded by those words remain.

In Table 3 we indeed find that XNLI and MNLI
| Dataset   | SOTA   | Repl. | Bias Cond. | Maj. | $\Delta_{\text{maj}}$ | $\%_S$ | $\%_R$ | PECO | Labels | Recovery | Tok. Rel. |
|-----------|--------|-------|------------|------|---------------------|-------|-------|------|--------|-----------|-----------|
| SICK      | 90.3   | 87.8  | 60.0       | 56.0 | 4.0                 | 66.4  | 68.3  | 0.407| 48.8   | 49.7      | 32.5      |
| SNLI      | 93.1   | 90.6  | 71.6       | 33.8 | 37.8                | 76.9  | 79.0  | 0.400| 70.0   | 71.8      | 20.3      |
| MNLI-m    | 92.0   | 88.7  | 59.8       | 35.6 | 24.2                | 65.0  | 67.4  | 0.357| 32.5   | 33.4      | 7.8       |
| MNLI-u    | 91.7   | 88.3  | 60.9       | 36.5 | 24.4                | 66.4  | 69.0  | 0.371| 47.5   | 48.6      | 7.8       |
| XNLI      | 83.7   | 73.7  | 55.0       | 33.3 | 21.7                | 65.7  | 74.6  | 0.366| 52.3   | 54.0      | 20.1      |
| FEVER     | 86.3   | 74.7  | 63.5       | 33.3 | 30.2                | 73.6  | 85.0  | —    | 39.4   | 38.0      | 50.7      |
| ANLI      | 58.3   | 54.0  | 48.2       | 33.5 | 14.7                | 82.7  | 89.3  | 0.363| 37.6   | 53.4      | 8.4       |
| CAug-s1   | 75.5   | 61.9  | 51.0       | 33.4 | 17.6                | 67.5  | 82.4  | 0.550| 53.8   | 57.9      | 12.2      |
| CAug-s2   | 58.6   | 50.0  | 48.1       | 33.4 | 14.7                | 82.1  | 96.2  | 0.432| 53.9   | 60.3      | 11.1      |
| OCNLI     | 78.2   | 71.8  | 61.5       | 36.8 | 24.7                | 78.6  | 85.7  | 0.457| 69.0   | 74.1      | 18.7      |
| CAug-s1   | 84.7   | 84.7  | 39.0       | 33.9 | 5.1                 | 46.0  | 46.0  | —    | 43.4   | 42.5      | 18.2      |
| CAug-s2   | 84.7   | 84.7  | 41.9       | 33.9 | 8.0                 | 49.5  | 49.5  | 0.351| 32.9   | 32.7      | 39.5      |

Table 3: Results table for test (or val, if no labeled test set is provided), and derived metrics. SOTA denotes state-of-the-art model test accuracy. Repl. denotes replication model accuracy. Bias Cond. denotes the test accuracy of a model that exclusively uses the elicited sentence (either premise or hypothesis, see Table 1). Maj. denotes accuracy of a “guess the majority label” strategy. The Result Comparisons metrics are described in subsection 3.3. PECO, the Progressive Evaluation of Cluster Outliers score, is described in subsection 3.5. The Normal-Biased Agreement metrics are described in subsection 3.4. For all numbers, lower is better.

exhibit very similar result comparisons. The models on both datasets have a $\Delta_{\text{max}} \approx 20\%$ and $\%_S \approx 65$. These leakage feature results, being robust to manual translation into 14 different languages, seem to indicate that there is a strong fundamental semantic component to the human biases driving the elicited sentence relation leakage.

4.4 SICK, CAugNLI and Small Data

It’s noteworthy that the two lowest-bias datasets in the sample we’ve tested, SICK and CAugNLI, are both the smallest. It’s probably easier for smaller datasets to get a clean sample of sentence pairs where there are few enough semantic overlaps such that bias does not emerge.

4.5 PECO for Understanding and Improving

Figure 1 shows how the PECO-derived cluster-bias T-SNE plots can be used directly as a visualization tool for analyzing and “debugging” biased datasets. In the plot, SNLI clearly has considerably more high-biased clusters taking up a considerable portion of the latent space as compared to CAugNLI.

For example, an intervention could be performed on identified bias regions in the distribution by having human annotators create new premise sentences from the given hypotheses, thereby forcing the PECO-based bias metrics to reduce. This idea is further backed up by the PCC of 35.8 that we find between PECO and $\%_R$, suggesting that producing datasets of lower PECO score will naturally lead to lower recovered performance in the biased condition, and thus less elicited sentence relation leakage.

4.6 Other Impressions

We were disappointed by the quality of the local explanation map agreements for predicting any of the other values in the table. It exhibited no statistically significant correlations for any. In future work we will try more sophisticated local explanation methods to assess if this lack of a relationship really reflects a fundamental difference in how the models “reason” over the sentences, or just too noisy of a local explanation technique.

5 Conclusion

We have introduced useful tools and techniques for analyzing elicited sentence relation leakage bias in NLI datasets, and applied them to a large, representative set of popular current datasets.

In the four years since Poliak et al. (2018)’s the state of affairs of single sentence bias in NLI hasn’t improved very much. Although efforts toward adversarial NLI dataset production have led to “better benchmarks” that haven’t saturated as quickly, they still continue to exhibit cheating features, and worse, sometimes lead to cheating feature performance approaching current state-of-the-art accuracy (such as with ANLI R3).
6 Acknowledgements

This work was supported in part by the National Science Foundation Graduate Research Fellowship under Grant No. 1650114. We would also like to thank the Robert N. Noyce Trust for their generous gift to University of California via the Noyce Initiative.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), pages 385–393, Montréal, Canada. Association for Computational Linguistics.

Samuel Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4843–4855.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. The snli corpus.

Zeming Chen, Qi Yue Gao, and Lawrence S. Moss. 2021. NeuralLog: Natural language inference with joint neural and logical reasoning. In Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics, pages 78–88. Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. CoRR, abs/1911.02116.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language data. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 670–680.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kuebler, and Larry Moss. 2020. Ocnli: Original chinese natural language inference. In Findings of EMNLP.

Divyansh Kaushik, Amrith Setlur, Eduard Hovy, and Zachary C Lipton. 2020. Explaining the efficacy of counterfactually augmented data. International Conference on Learning Representations (ICLR).

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In Int. J. Comput. Vision, 123(1):32–73.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In International Conference on Learning Representations.

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

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

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

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2020. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019, pages 3428–3448. Association for Computational Linguistics (ACL).
Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In Association for the Advancement of Artificial Intelligence (AAAI).

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. arXiv:1910.14599 [cs].

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.

Liu 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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Fun-towicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. 2020. Clue: A chinese language understanding evaluation benchmark. In Proceedings of the 28th International Conference on Computational Linguistics, pages 4762–4772.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2021. Byt5: Towards a token-free future with pre-trained byte-to-byte models. CoRR, abs/2105.13626.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2:67–78.

Guanhua Zhang, Bing Bai, Jian Liang, Kun Bai, Shiyu Chang, Mo Yu, Conghui Zhu, and Tiejun Zhao. 2019. Selection Bias Explorations and Debias Methods for Natural Language Sentence Matching Datasets. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4418–4429, Florence, Italy. Association for Computational Linguistics.