CS378: Final Project - Data Artifacts
https://github.com/ziqing26/cs378_fp

Hsin Ning Lee
hl32499

Ziqing Zhang
zz8366

Apr 2023

Abstract

In this work, we are trying to reproduce the analysis - Competency Problems: On Finding and Removing Artifacts in Language Data (Gardner et al., 2021). The competency problem represents a class of problems in which the correlation of input features and output labels is not uniform. We assumed that all simple feature correlations between the input data and predicted labels are spurious. We used the statistical test "competency problems" framework to identify the bias in the dataset. We found that the competency test requires a relatively balanced dataset to be effective. Moreover, the local edit method showed some improvement in reducing the bias in the WANLI dataset (Liu, Swayamdipta, Smith, & Choi, 2022). The distribution of rejected samples moved downward, and a greater amount of tokens fell within the confidence region after local editing. However, there are limitations to this method, such as unintended consequences of label editing on correlated features, which can dilute the effectiveness of this method on larger datasets. Additionally, the human annotation process is time-consuming and prone to bias and errors.

1 Introduction

While more datasets have been applied on a variety of NLP challenges, a growing corpus of research has shown the basic limits in existing datasets, including spurious correlations, bias samples, and dataset artifacts (Gururangan et al., 2018). In our project, the main goal is to use the "competency problems" framework to analyze WANLI dataset and provide some insights on removing data artifacts. Our dataset of choice, WANLI, which was generated collaboratively by humans and AI, is claimed to improve the model's performance by offering more diverse and challenging cases for training and evaluation (Liu et al., 2022). With the assumption: all simple feature correlations between the input data and predicted labels are spurious, we evaluate the WANLI dataset by addressing the issue of "competency problem" which is defined by our assumption. To further explain, the single feature itself should not give any information about the output label. Instead, it is the context of a whole sentence that affects the prediction.

By conducting the statistical test, we find that there are significant deviations from the competency assumption in a balanced dataset using conservative Bonferroni-corrected statistical test (Bonferroni, 1936). Figure 1 demonstrates the result of statistical test for deviation from competency problems. We notice that the distribution of labels significantly affects the accuracy of the competency analysis. Thus, we believe that collecting data that have balanced class label is crucial. In Figure 2, we have shown the improved version by pre-processing the data with balanced label.

To reduce the data artifacts, we suggest to use the theoretical treatment - local edits on a small sample from the WANLI dataset. The results are visualized and compared in Figure 3 and Figure 4. These proofs provide dataset builders with tools to monitor the data gathering process and ensure that the resulting datasets are as free of artifacts as possible. We hope that our analysis on WANLI dataset can provide some insights for future approaches to improve NLP data collecting, and minimizing dataset artifacts that could be introduced to improve model performance.

2 Task/Dataset/Model Description

We decided to impose a statistical test using a competency problem framework to find and reduce the artifacts in the chosen dataset. The com-

petency problem is defined as one in which, given any single feature, the marginal distribution over labels is uniform. In our analysis, x_i represents the unique words in the training sentences, and $y \in \{0, 1, 2\}$ represents the label corresponding to entailment, neutral, and contradiction in WANLI dataset (Liu et al., 2022). We define competency in the settings of NLI dataset to be $p(y|x_i) = 1/3$ for every $i \in N$ where N is the number of samples in the dataset. This is because a word should not be able to convey a sentiment label on its own, regardless of the context in which it appears (Gardner et al., 2021). Thus, the dimension i has artifacts if the observed probability $\hat{p}(y|x_i)$ does not equal to 1/3 in our case. To restate, the objective of our statistical test is to find the spurious correlations and to minimize the existing artifacts. The evaluation of model accuracy is secondary to our main aim of removing bias in the dataset.

Compared to standard PMI-based analysis that uses $\log \frac{\hat{p}(y|x_i)}{\hat{p}(y)}$, our competency analysis takes into account not only $\hat{p}(y|x_i)$ (y-axis in Figure 2), but also the number of times the feature is seen (x-axis) which determines the threshold for a deviation from $\hat{p}(y|x_i)$. So the result of our statistical test gives a more complete picture of data artifacts.

Our dataset of choice, WANLI (Liu et al., 2022), is a dataset created through collaboration between human crowd workers and GPT-3 language model. It was created to address the lack of linguistic diversity in NLP datasets by introducing machine-generated samples that are filtered and labeled by humans. The dataset consists of 107,885 NLI examples (including 5,000 test data) and has shown to have unique empirical strengths over the existing NLI dataset. In this analysis, we aim to determine whether there exists any improvement with respect to the bias in the training dataset.

SNLI dataset was previously used in competency problem exploration (Gardner et al., 2021). Compared to SNLI which is sourced from a specific corpus of texts, WANLI contains a wider range of language styles, including informal language, dialects, and writing styles that may not be present in SNLI. Moreover, testing the competency assumption on the less focused and standardized WANLI dataset have a broader application compared to using SNLI datasets.

To perform our tasks, we use the ELECTRA-small model(K. Clark, Luong, Le, & Manning, 2020) which has an improved training method compared to ELECTRA. The name "ELECTRA" stands for "Efficiently Learning an Encoder that Classifies Token Replacements Accurately". This smaller architecture of the model requires fewer computational resources but still obtains high accuracy on a range of linguistic tasks. After being pre-trained on a large corpus of text data, the model can be fine-tuned for particular tasks, such as sentiment analysis or machine translation.

3 Performance Analysis

To evaluate WANLI dataset under the competency settings, we set up a hypothesis test to identify the bias in dataset. We apply a one-sided binomial proportion hypothesis test, as the rejection sampling can only result in binomial proportions of $p(y|x_i)$ higher than 1/3. We represent bias as rejection sampling from the desire competency distribution based on single feature value, which is a single word.

$$H_0 : p(y|x_i) = 1/3 = p_0$$
$$H_1 : p(y|x_i) >= 1/3$$

To compute z-statistic using the standard formula:

$$z^* = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0/n}} \Rightarrow \hat{p} = \frac{z^*}{2\sqrt{n}} + 1/2$$

,where n is the number of samples and z^* is the confidence level where $z^* = 1 - \alpha$ and $\alpha = 0.05$ in this experiment.

To perform the statistical test for competency problem, we first uses 18,000 data from WANLI dataset. The result is shown in Figure 1. In the figure, we can see that examples tends to be biased towards neutral and contradiction label compared to contradiction label. We found that the distribution of labels significantly affects the accuracy of competency analysis. This is because WANLI dataset is unbalanced with a ratio of 4:5:1 (entailment:neutral:contradiction) (Liu et al., 2022).

To eliminate the effect of unbalanced dataset, we then balance out the label representations by randomly selecting 18,000 samples with even number of each label. This ensures the artifacts come from single word biases instead of skewed distribution of classes. The statistical test result from the balanced representation is shown in Figure 2. We find a significant amount of deviations from the competency assumption. There are more bias towards prediction of contradiction class, which is shown by the larger proportion of words above the statistical test line (Figure 2).

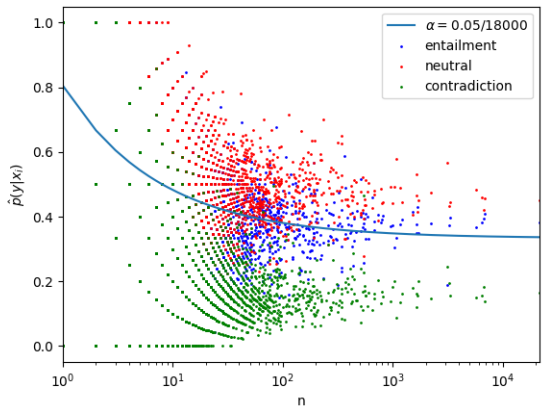

Figure 1: A statistical test for deviation from a competency problem on 18,000 samples from WANLI dataset without balancing the dataset. Contradiction classes has signicantly fewer number of samples, resulting in a distribution of bias against it. The plot shows the number of occurrences of each word n against the conditional probability $\hat{p}$ of the label y given the presence of the word x. All features above the blue line have detectable correlation with class labels, using a Bonferroni-corrected statistical test. The result is compounded by the imbalanced number of samples.

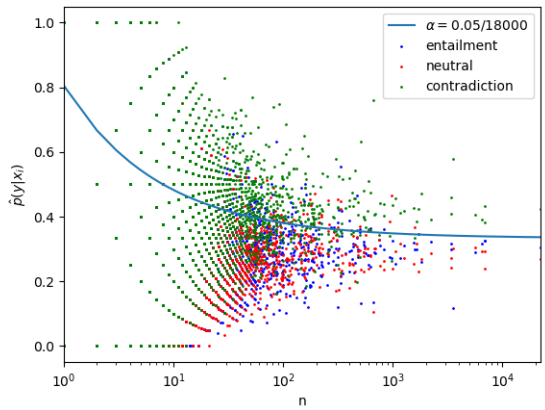

Figure 2: The same statistical test as Figure 1 using 18,000 samples from WANLI but with balanced number of samples for each classes.

Examining closely the biased features, we found that words such as "economy", are biased towards prediction of contradiction; words such as "make" and "some" are biased towards prediction of entailment; words such as "claim" are biased towards netural label. Since single features (in this case the words) should not imply a certain class, it is thus pertinent to try to reduce bias based on single features.

4 Describing the Fix

Previous research (Sennrich, 2017; Zhao, Wang, Yatskar, Ordonez, & Chang, 2018) has tried to make small edits to datasets to minimize artifacts. Data augmentation can be effective with an appropriately sensitive edit model (Gardner et al., 2021), where sensitivity is defined to as the the frequency of a change to inputs results in the label changing.

The objective of performing local edit is to balance each label's occurrence for word x_i. Here, we made changes only to the hypothesis of each selected sample which leads to changes in label y.

Using local edit from the rejection sampling, we want to effectively reduce the artifacts in dataset and the prediction bias.

In this task, we perform local edits on a small subset of dataset D, which has $n = 300$ samples from WANLI, to form a new dataset D', which consists of samples x_0, y'. Here are the procedure to make D':

1. Randomly sample an instance x from D of n instances created under the statistical test under D. The chosen instances are rejected by null hypothesis ($\hat{p}(y|x_i)$).

2. Make some changes to the hypothesis of x to get x'.

3. Manually change label from y to y' and add $\langle x', y' \rangle$ to D'.

5 Evaluating the Fix

In the above discussion, we have identified the biases in WANLI dataset, and derived the method of local edit. The fix is performed on a smaller settings (300 samples) to simplify the process and give a overall understanding on the improvement. Figure 3 visualizes the balanced-class data without any local editing. In Figure 4, the distribution of rejected samplings have moved downward compare to the non-processed data. Before local editing, many samples in the original data exhibit artifacts in the positive direction, where $\hat{p}(y|x_i)$ is rejected by hypothesis test. While, using local edit, the distribution of tokens is more even across

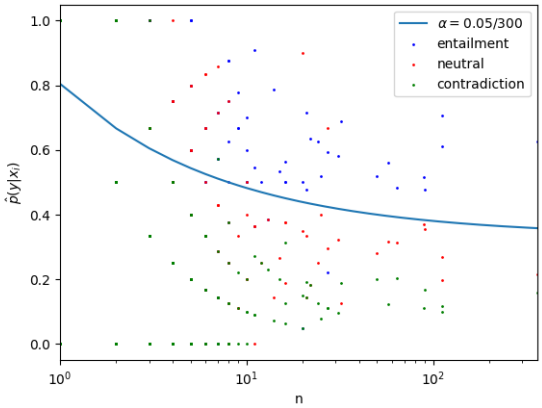

Figure 3: Data artifact using 300 examples from WANLI without local edits.

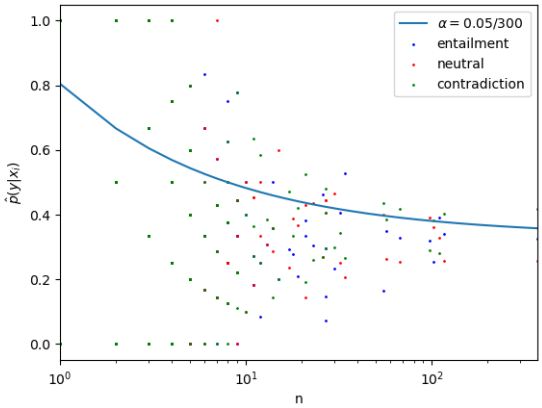

Figure 4: Data artifact using 300 examples from WANLI with local edits to examples involving the 30 most deviated feature. Less deviation from the hypothesis is observed compared to Figure 3.

all classes and a greater amount of tokens are fall within the confidence region.

Though the local edit did make some improvements to the WANLI dataset, we noticed that editing the label of a sentence may have a larger effect on the performance of competency problem, because single features (i.e. words) are correlated, which arise from their co-occurence within a sentence. Performing local edits on a sentence in the hope of correcting the bias for a single feature x_a will results in unintended increase in $\hat{p}(y|x_i)$ for another feature x_b. Such unintended consequence will be diluted with larger datasets. However, larger dataset means more annotation would be needed. Since human annotation is needed in making these changes, bias in this process can lead to the introduction of new artifacts. In addition, performing human annotation can be time-consuming, and the annotation error cannot be neglected. Thus, performing edits on large training datasets is diffcult to be completely free from artifacts (Shwartz, Rudinger, & Tafjord, 2020).

To mitigate the effect of high correlation of features, further work can focus on increasing the number of annotators (Geva, Goldberg, & Berant, 2019). Since we highly depend on the human annotated data when reducing the dataset artifacts, there can be new bias and correlations between features and labels introduced in the process. With more annotators, the associations are lessen in aggregate, making the data less biased overall.

6 Related Work

There are recent works on theoretical treatments of bias (Shah, Schwartz, & Hovy, 2020). Our work is different by having experimental comparison of data artifact treatment using the competency assumption.

Meanwhile, competency problem framework (Gardner et al., 2021) was used to find spurious n-gram correlations with answers. They set up a hypothesis test for a binary classficiation problem to determine whether there is sufficient evidence to reject null hypothesis: $\hat{p}(y|x_i) = 1/2$. To mitigate the artifacts, Gardner et al.'s research empirically analyzes the effectiveness of local edits for single feature artifact reduction using two datasets: Boolean Questions dataset (C. Clark et al., 2019) and IMDb. They argued that removing dataset artifacts can potentially improve model performance and reduce overfitting. Our work is different in focusing on empirical data using WANLI which has three-class classification and more worker and AI collaboration in data annotation.

Local editing is carried out by creating a new dataset, which contains the new samples that being annotated manually. This type of data augmentation can be effective when combined with a properly sensitive edited model, where sensitivity refers to how frequently a change in inputs causes the label to change. This technique is similar to modify model's local decision boundaries

using contrast sets (Gardner et al., 2020). By making local edits to the text while monitoring the sensitivity of those edits, the correlations between train features and test labels can be reduced.

7 Conclusion

To analyze the effectiveness of competency problem assumption on correcting bias in data artifacts, we performed a Bonferroni-corrected statistical test on 18,000 WANLI examples using both unbalanced and balanced examples for each label. We explored the effectiveness of local edits in reducing data artifacts and more features lie in the confidence interval after local edits.

We found that the local edit method showed some improvement in reducing the bias in the WANLI dataset. The distribution of rejected samples moved downward, and a greater amount of tokens fell within the confidence region after local editing. However, there are limitations to this method, such as unintended consequences of label editing on correlated features, which can dilute the effectiveness of this method on larger datasets. Additionally, the human annotation process is time-consuming and prone to bias and errors.

To further improve the effectiveness of bias reduction methods, future work can focus on increasing the number of annotators to reduce the correlations between features and labels in aggregate. This can help make the data less biased overall. It is important to carefully evaluate and assess the effectiveness and limitations of bias reduction methods in natural language processing to ensure the development of fair and accurate models.

References

Bonferroni, C. (1936). *Teoria statistica delle classi e calcolo delle probabilità*. Seeber. Retrieved from `https://books.google.com/books?id=3CY-HQAACAAJ`

Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., & Toutanova, K. (2019). *Boolq: Exploring the surprising difficulty of natural yes/no questions.*

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). *Electra: Pre-training text encoders as discriminators rather than generators.*

Gardner, M., Artzi, Y., Basmov, V., Berant, J., Bogin, B., Chen, S., ... Zhou, B. (2020, November). Evaluating models' local decision boundaries via contrast sets. In *Findings of the association for computational linguistics: Emnlp 2020* (pp. 1307–1323). Online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2020.findings-emnlp.117` doi: 10.18653/v1/2020.findings-emnlp.117

Gardner, M., Merrill, W., Dodge, J., Peters, M. E., Ross, A., Singh, S., & Smith, N. A. (2021). *Competency problems: On finding and removing artifacts in language data.*

Geva, M., Goldberg, Y., & Berant, J. (2019, November). Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 1161–1166). Hong Kong, China: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/D19-1107` doi: 10.18653/v1/D19-1107

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. (2018). *Annotation artifacts in natural language inference data.*

Liu, A., Swayamdipta, S., Smith, N. A., & Choi, Y. (2022). *Wanli: Worker and ai collaboration for natural language inference dataset creation.*

Sennrich, R. (2017, April). How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: Volume 2, short papers* (pp. 376–382). Valencia, Spain: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/E17-2060`

Shah, D. S., Schwartz, H. A., & Hovy, D. (2020, July). 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* (pp. 5248–5264). Online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2020.acl-main.468` doi: 10.18653/v1/

2020.acl-main.468

Shwartz, V., Rudinger, R., & Tafjord, O. (2020, November). "you are grounded!": Latent name artifacts in pre-trained language models. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 6850–6861). Online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2020.emnlp-main.556` doi: 10.18653/v1/2020.emnlp-main.556

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018, June). Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 2 (short papers)* (pp. 15–20). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/N18-2003` doi: 10.18653/v1/N18-2003