Competency Problems: On Finding and Removing Artifacts in Language Data

Presenter: Yujie Qiao

Motivation

- Problems:
  - Popular datasets in NLP are prone to shortcuts, dataset artifacts, bias, and spurious correlations between input features and output labels
  - Bias in data collection is pervasive and not easily addressed with current learning techniques
- Question:
  - What exactly makes a correlation “spurious”, instead of a feature that is legitimately predictive of some target label, i.e. how to tell which features have “spurious” instead of legitimate correlations?

↓

A theoretical framework

Competency Problems

- The marginal distribution over labels given any single feature is uniform
    - Key assumption:
        - in a language understanding problem, no single feature on its own should contain information about the class label **=>** all simple correlations between input features and output labels are spurious: $p(y|xi)$, for any feature xi, should be uniform over the class label
        - Assume an input vector x and an output value y, where $x \in \{0, 1\}$ and $y \in \{0, 1\}$. In this setting, competency means $p(y|xi) = 0.5$ for all i **=>** the information mapping x to y is found in complex feature interactions, not in individual features

Competency Problems - Example

- Sentiment analysis on movie reviews
  - A single feature might be the presence of the word “**amazing**”, which could be legitimately correlated with **positive** sentiment in some randomly-sampled collection of actual movie reviews.
    - the word “amazing” on its own should **NOT** give information about a sentiment label independent of the context in which it appears, which could include negation, metaphor, sarcasm, etc

Core Claims

- If a model picks up on **individual** feature correlations in a dataset, it has learned something extra-linguistic, such as information about human biases, not about how words come together to form meaning, which is the heart of natural language understanding
- To push machines towards linguistic competence, we must **control** for all sources of extra-linguistic information, ensuring that no simple features contain information about class labels

Biased Sampling

- Humans suffer from blind spots, social bias, priming, and other psychological effects that make collecting data for competency problems challenging.
  - E.g:
    - instructions in a crowdsourcing task that prime workers to use particular language
    - the “amazing” example previously
    - racial bias in face recognition
    - abusive language detection datasets
- A plausible model for accounting for the bias

⇩

Rejection sampling from the target competency distribution based on single feature values

Rejection Sampling

- Not a psychological model of dataset construction, but a reasonable first-order approximation of the outcome of human bias on data creation!
- Procedure:
  - A person samples an instance from an unbiased distribution $p_u(x, y)$ where the competency assumption holds.
  - The person examines this instance, and if feature $x_i = 1$ appears with label $y = 0$, the person rejects the instance and samples a new one, with probability r_i
  - With no bias ($r_i = 0$)

Rejection Sampling (cont'd)

- Let:
  - Pu(y|xi) => conditional probability of y = 1 given xi = 1 under the **unbiased** distribution
  - Pb(y|xi) => conditional probability of y = 1 given xi = 1 under the **biased** distribution
  - Pˆ(y|xi) => empirical probability within a biased dataset of n samples
  - fi => marginal probability Pu(xi)
- Pu(y|xi) is 0.5 by assumption
- Artifact present => if the empirical probability pˆ(y|xi) statistically differs from 0.5.
- **With no bias (ri = 0), this probability is 0.5, as expected, and it rises to 1 as ri increases to 1**

$$p_b(y, x_i) = \frac{1}{2} f_i + \frac{1}{2} f_i r_i p_b(y, x_i)$$

$$\therefore p_b(y, x_i) = \frac{f_i}{2 - f_i r_i}$$

$$p_b(x_i) = \frac{1}{2} f_i + \frac{1}{2} f_i (1 - r_i) + \frac{1}{2} f_i r_i p_b(x_i)$$

$$\therefore p_b(x_i) = \frac{2 f_i - f_i r_i}{2 - f_i r_i}$$

$$\therefore p_b(y \mid x_i) = \frac{p_b(y, x_i)}{p_b(x_i)} = \frac{1}{2 - r_i}$$

Rejection Sampling (cont'd)

- By the central limit theorem (CLT)
  - pˆ(y|xi) ≈ μpˆ
  - As the rejection probability ri increases, the center of this distribution tends from 0.5 to 1
- Increasing the sample size n concentrates the distribution inversely proportional to $n^{1/2}$ but the expected value is unchanged
- So… *simply sampling more data from the same biased procedure will NOT omit artifacts created by rejection sampling—the empirical probability will still be biased by ri even if n increases arbitrarily*

$$\mu_{\hat{p}} = p_b(y \mid x_i) = \frac{1}{2 - r_i}$$

$$\sigma_{\hat{p}}^2 = \left(\frac{1 - r_i}{(2 - r_i)^2}\right)^2 \cdot \frac{1}{n}.$$

Hypothesis Test

- Test if there is enough evidence to reject the null hypothesis (ri = 0, i.e., that the data is unbiased)
  - A one-sided binomial proportion hypothesis test, as the rejection sampling can only lead to binomial proportions for Pb (y | xi) that are greater than ½
  - Null hypothesis:
    - Pb(y | xi) = 0.5 = p0, or equivalently, that ri = 0
  - Alternative hypothesis:
    - Pb (y | xi) ≥ 0.5
  - Z-statistic (The use of a z-statistic depends on the normal approximation to a binomial distribution, which holds for large n)
    - if our observed proportion pˆ is far from p0 = 0.5, we will have enough evidence to reject the null hypothesis that ri = 0

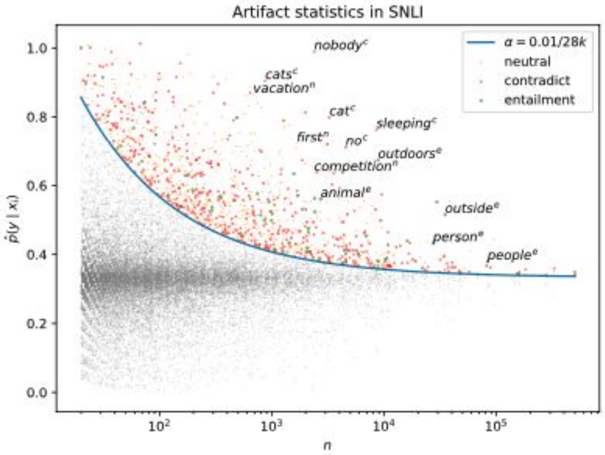
$$z^* = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Experiments

- Evaluation Data:
  - SNLI
  - Universal Dependencies English Web Treebank
- Other details:
  - **SNLI:**
    - Each feature xi represents the presence of a word in a given example
    - p0 = 1/3 , as SNLI has three labels
  - **UD English Web Treebank:**
    - Prepositional phrase (PP) attachment problem - determining whether a PP attaches to a verb (e.g., We ate spaghetti with forks) or a noun (e.g., We ate spaghetti with meatballs).
    - Extracted (verb, noun, prepositional phrase) constructions with ambiguous attachment from the UD English Web Treebank (EWT) training data
    - Treat (verb, preposition) tuples as features and attachment types (noun or verb) as labels

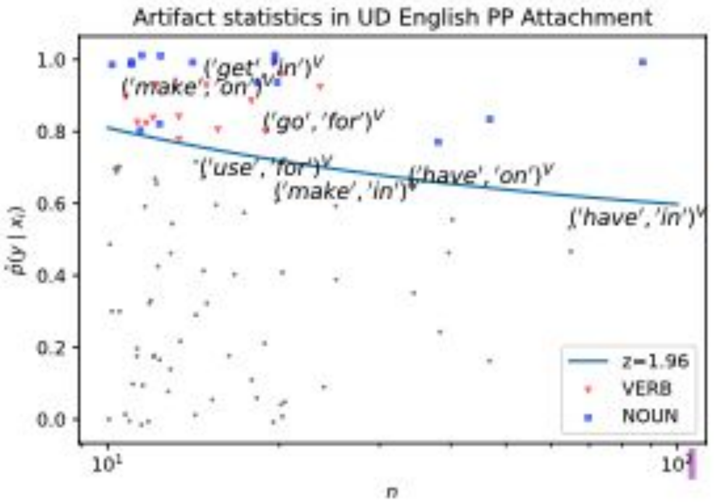
Revelation - individual word artifacts in the SNLI dataset

- z-statistic for each token vs. n (the number of times the token appears in the data)
- Blue curve: the value of the z-statistic at which the null hypothesis (that ri = 0) should be rejected
  - significance level of α = 0.01 & a conservative Bonferroni correction



Revelation - artifacts in the UD English Web Treebank

- z-statistic for each tuple that appears 10 or more times in the data



Do NLP models learn to bias their predictions based on artifacts?

- Evaluation Data:
  - SNLI
  - RTE data from SuperGlue
- Experiments:
  - Focuses on words with high z-statistics, which are often words that show up very frequently with slight deviations from Pu(y|xi)
  - Models: RoBERTa-base fine-tuned on RTE, and ALBERT-base fine-tuned on SNLI

Experiments & Results

- Procedure:
  - Create two synthetic input examples:
    - the premise containing only the single token with an empty hypothesis
    - an empty premise and hypothesis containing the single token
  - Run a forward pass with each input and average the target class probabilities as an estimate of $\tilde{p}(y|x_i)$

Dataset	Class	$\Delta\hat{p}_y$
RTE	entailment	+2.2 %
SNLI	entailment	+14.7 %
SNLI	neutral	+7.9 %
SNLI	contradiction	+12.5 %

Mitigation - Local Edits

- Sensitive edit model
  - Sensitivity = how often a change to inputs results in the label changing
  - Edit sensitivity si = the probability that y changes during editing given the occurrence of a particular feature in the edited data $s_i = p_b(y' = \neg y \mid x'_i).$
  -
  - ei = the probability that dimension i gets flipped when going from x to x′

Mitigation - Local Edits

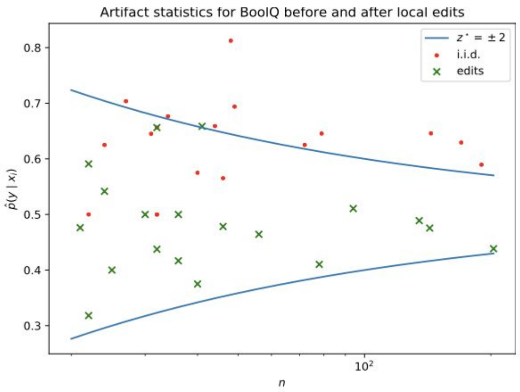
- 3 ways to achieve unbiased data from a local edit procedure that edits dimensions independently:
    - (1) start with unbiased data
    - (2) always flip every feature
    - (3) flip the label half the time for each feature
- **Proposition 1** (Proof in §B). *Assume x_i, x_j, e_i, e_j, s_i, and s_j are independent for all i, j. Then $p_e(y' \mid x'_i) = \frac{1}{2}$ if and only if $r_i = 0$ or $s_i = \frac{1}{2}$ or $e_i = 1$.*

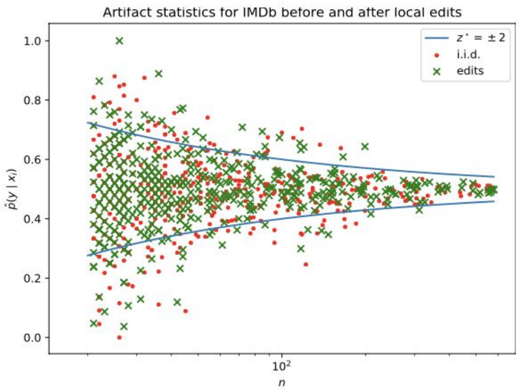
Investigate the effectiveness of local edits

- Evaluation Data:
  - the Boolean Questions dataset
  - IMDb
- Other details:
  - Define each feature xi as the occurrence of a particular word within q for BoolQ, and within the text of the review for IMDb
  - Make local edits to the question or review text and recording the updated binary label.

Investigate the effectiveness of local edits

- For BoolQ, many tokens in the original data exhibit artifacts in the positive (> 0.5) direction
  - within the edited data, almost all tokens fall within the confidence region.
- In contrast, there is no apparent distributional difference between artifact statistics for the original vs. edited texts on IMDb





Other Mitigation Strategies

1. Increase the number of annotators
    - Alleviate substantial **person-specific** correlations between features and labels

- Intuition:
    - more annotators **washes out correlations** & makes the data **less biased**
- Procedure:
    - Recall: a single possible rejection probability, where an instance is rejected with probability r_i if $x_i = 1$ and $y = 0$. What if we introduce additional rejection probabilities?
    - Split a dataset into k different slices that have their own **bias vectors r**
        - Uncorrelated r vectors: <u>as k increases, the probability that $\hat{p}(y|x_i)$ deviates from $p_u(y|x_i)$ tends towards zero</u>
        - Correlated r vectors: <u>increasing the number of annotators will not produce data reflecting the competency assumption</u>

Other Mitigation Strategies

2. Data filtering

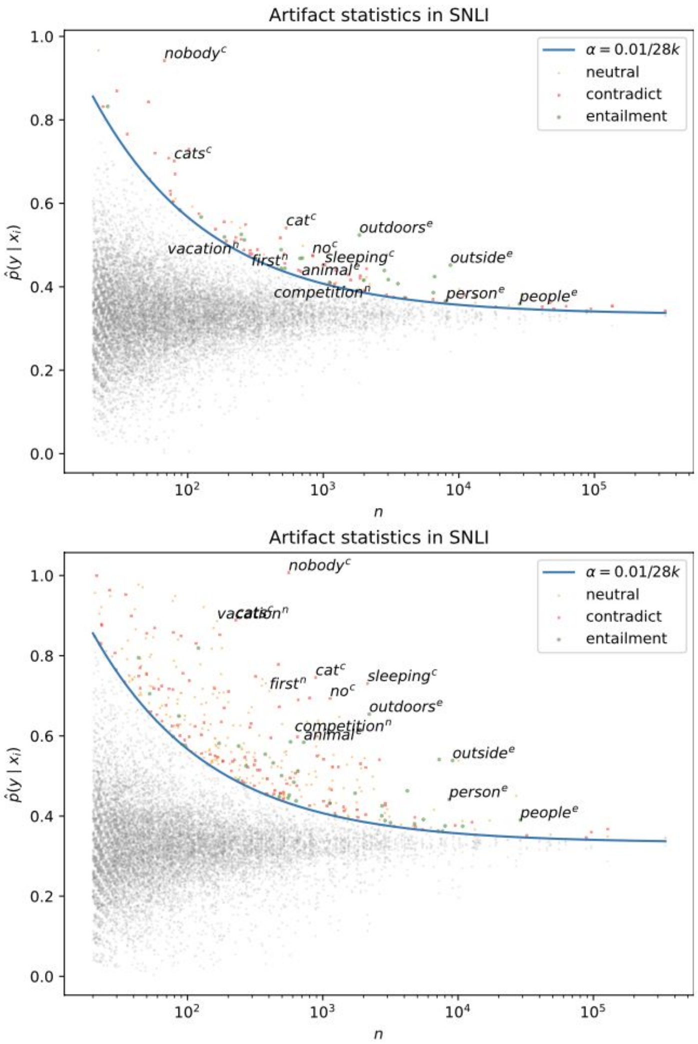
    - Remove data from a training set that is biased in some way in order to get a model that generalizes better

- Pros:
    - **In the extreme case** where ri ≈ 1, such as with “nobody” in SNLI, this process could **effectively** remove xi from the observed feature space.
- Cons:
    - **Undesirable to remove entire instances** because of bias in a single feature
- Procedure:
    - “Ambiguous” training data vs. original training set
        - Ambiguous” instances: data classified as “ambiguous” according to Dataset Cartography
        - Original training set**:** a random (same-size) sample from the SNLI training set

Other Mitigation Strategies

2. Data filtering

- Results: the “ambiguous” instances have many fewer deviations from the competency assumption, across the entire range of the hypothesis test!



Other Mitigation Strategies

1. Increase the number of annotators
    - Alleviate substantial **person-specific** correlations between features and labels
- Intuition:
    - more annotators **washes out correlations** & makes the data **less biased**
- Procedure:
    - Recall: a single possible rejection probability, where an instance is rejected with probability r_i if $x_i = 1$ and $y = 0$. What if we introduce additional rejection probabilities?
    - Split a dataset into k different slices that have their own **bias vectors r**
        - Uncorrelated r vectors: <u>as k increases, the probability that $\hat{p}(y|x_i)$ deviates from $p_u(y|x_i)$ tends towards zero</u>
        - Correlated r vectors: <u>increasing the number of annotators will not produce data reflecting the competency assumption</u>

Other Related Work

- What's different?
    - Here, they introduced a competency assumption and discussed its implications
- Can we discourage relying on individual features?
    - ensemble weak models together with strong models during training
    - ensembles of models with unaligned gradients

Conclusion

- Examined existing datasets for evidence of statistically-significant feature bias, and then explore the extent to which this bias impacts models supervised with this data
- Theoretically analyzed data collection under this biased sampling process, showing that any amount of bias will result in increasing probability of statistically-significant spurious feature correlations as dataset size increases
- This framework allowed us to examine the theoretical impact of proposed techniques to mitigate bias, including performing local edits after data collection and filtering collected data

Discussion

- This paper set up initially in binomial random variable settings - can it be generalized to multiple labels/variables?
- How to effectively/empirically measure the ri in the rejection sampling procedure?
- Any particular reason to use significance level of $\alpha = 0.01$ instead of the conventional level of 0.05?