Does Riddle Sense Make Sense?

Yuval Omer
yuval.omer2@mail.huji.ac.il

Matan Cohen
Matan.cohen11@mail.huji.ac.il

Liron Naccache
liron.naccache@mail.huji.ac.il

1 Introduction

The research objective of this paper is to analyze the RiddleSense dataset (Lin et al., 2021), specifically aiming to identify any biases and spurious correlations. Moreover, we intend to employ Data Maps (Swayamdipta et al., 2020) a data visualization tool to enhance our comprehension of the RiddleSense's distribution patterns.

2 Data

RiddleSense is a multiple-choice question-answering dataset designed to test reasoning and deduction abilities in language models. It consists of 5,700 riddles. Each riddle is uniquely represented as a multiple-choice task with five possible answers. The riddles span a diverse range of themes and abstract concepts which requires the capability to perform complex commonsense reasoning, counterfactual thinking, and figurative interpretation. While humans reached 91% accuracy on Riddle-sense SOTA models have yet to surpass the 70% mark [2].

3 Experiments and results

In this project, we have conducted numerous experiments to explore several facets of the RiddleSense dataset, among this experiments, were: Baseline Replication, Spurious correlations exploration and Dataset Cartography.

Baseline Replication

In order to create a baseline we can compare results with on RiddleSense, we first tried to replicate the result of the original paper. We fine-tuned the multiple-choice version of the BERT uncased model (Devlin et al., 2018) on the RiddleSense dataset. Predicting on a sample was done by the model giving a score for each answer independently and choosing the highest score as the answer. After training over three epochs, using the default HuggingFace configurations, our model achieved an accuracy of 53.18%, slightly below the original paper's 54.16% using a similar model.

Spurious correlations exploration

As solving riddles is a task that requires a high level of language understanding, we wished to check whether the model's performance can be in part attributed to some bias in the data.

To do this, we fine-tuned the model on a "Answers dataset", in which each sample contained only the answers to the riddle, without the riddle itself. We got an accuracy of 40% over the validation set using this method. This performance is double the expected performance of a trivial model given five answer options, thus we can infer that there is a spurious correlation between question answers and the probability of them being the correct answer.

Consequently, we focused our attention on a specific statistic: the frequency of occurrence for each word as a potential answer to a riddle, calculated as the ratio of times the word/phrase was the correct answer to the total number of times the word/phrase appeared as a possible answer. Our analysis revealed that 27% of the words consistently served as the correct answer across all related questions, while 70% of the words were consistently incorrect whenever they appeared. Merely 3% of the words exhibited variability, appearing as both correct and incorrect answers in different samples. These findings confirms the presence of statistical patterns within the data that the model can over-fit to instead of generalizing.

In order to further examine the findings shown above, we created "Unique answers dataset". Which is made of only answers, similarly to "Answers dataset", but we made sure that each word/phase appears as a possible answer at most one time in a possible answer. After the actions above, we were left with 657 unique samples.

Then, we fine-tuned our model on this dataset over the same number of training steps as the model trained on "Answers dataset". We got an accuracy of 31% [2] over the validation set which is 9% decrease compared to "Answers dataset". We can infer that 9% out of the 20% of the bias is originating from word repetition, however, there is still 11% bias that is unexplained by our theory.

Dataset Cartography

Another aspect we sought to examine in RiddleSense is the quality of the samples. To do this we turned to Data Maps, a data visualization method introduced in (Swayamdipta et al., 2020) with the intent of mapping data into categories to improve model performance and ease data exploration.

Following (Swayamdipta et al., 2020), we calculated two objectives:

- Confidence - the model's confidence in assigning the correct answer to the riddle's golden label answer.
- Variability - the model's variability of its confidence across different training epochs

In (Swayamdipta et al., 2020) and (Sar-Shalom and Schwartz, 2023) two methods of implementing Data Maps were introduced, Swayamdipta fine-tuned a model over several epochs and calculated the confidence and variability from the training process. While Sar-Shalom suggests that even after seeing a sample once, the model already develops a bias to that sample. To combat this issue he suggests training several models each for one epoch and using the mean of their observations for the Data Map. In both methods the dataset is seperated into three distinct regions:

- Ambiguous region - samples where the variability is high
- Easy to Learn region - samples where confidence is high and variability is low.
- Hard to Learn region - samples where both confidence and variability are low.

In our experiminations we have seen that using Swayamdipta's method of creating a Data Map nearly all samples in RiddleSense were categorised as Easy to learn, while when using Sar-Shalom's method all samples where in the hard to learn region. We suggest a new method where we combine the strengths of both methods. Instead of training several models for one epoch each, we suggest training several models with incremental number of training epochs and collecting information at the end of each epoch. As such we believe our method will retain information on the learnablity of a sample over time while using different initializations and collecting more information from earlier stages of training to deal with bias, results shown in Fig. 1. To explore the regions we visually inspected samples from each region as seen in Table 1. In the Ambiguous category we saw that some samples share the same answer, thus we believe the riddles by themselves are hard for the model but the mere repetition of the answer might have caused a bias in the model that allowed it to improve on this samples. As for the Hard-to-Learn category, we identified several reasons for the classification:

- Some answers contain spelling errors
- In some samples the golden label is mislabeled, meaning it is not a correct answer to the riddle.
- In some samples there is ambiguity in the answers, meaning there is more than one possible correct answer in the choices presented.

We then fine-tuned our baseline model on the ambiguous samples to test the quality of samples in each region in the Data Map, and got the results detailed in Table 2. From this results we can learn that the samples in the hard-to-learn region seems to confuse the model and lower it's performance below trivial performance of 20%, this strengthens the findings shown above about the samples in this region. Following (Sar-Shalom and Schwartz, 2023) , we created a dataset containing: Three copies of each sample that is among the 25% of samples with the highest variability and one copy of any other sample. After fine-tuning, we got an accuracy of 54.46 on the validation set, an increase of 0.98% above the baseline. [2]

4 Conclusions

In this paper we have shown that RiddleSense has bias in the labels, many samples in the training set contains typos, mislabeling or ambiguity. But we haven't uncovered all biases present in the dataset. Additionaly we have developed and presented a new method to compute Data Maps which paired with the training method presented by Sar-Shalom improved model performance [2].

Type of Riddle	Riddle	Choices
Easy to learn	You can't hear me but I'm not dead. Who am I?? What is black and white and read all over?	spiritualist, silence+ , hearing, sound, look newspaper+ , print, piece, skim, reply
Ambiguous	What gets smaller as it gets fuller? What is weightless and colorless, but when put into a barrel, the barrel weighs less?	bit, put, hole+ , rice, unit hole+ , measuring, heft, color, gallon
Hard to learn	a little mountain cannot get WHAT HAS ONE EYE BUT COULD NEVER SEE? One day I sat on a mat, I looked up and saw a big bat, As it swooped down from the air it got caught in my hair, Finally it just took my rat. What kind of riddle is it?	hump, rockies, national park, tae+? , snowbell wink, glass eye , pin+ , tool, anatomy linerick+ , nature reserve, twister, have, bobby pin

Table 1: Riddles according to their type. A pink marker and a plus sign indicate the golden label. A question mark symbol means there's a typo in the answer. A green marker symbols that an answer which is not the golden label but can also be a correct answer.

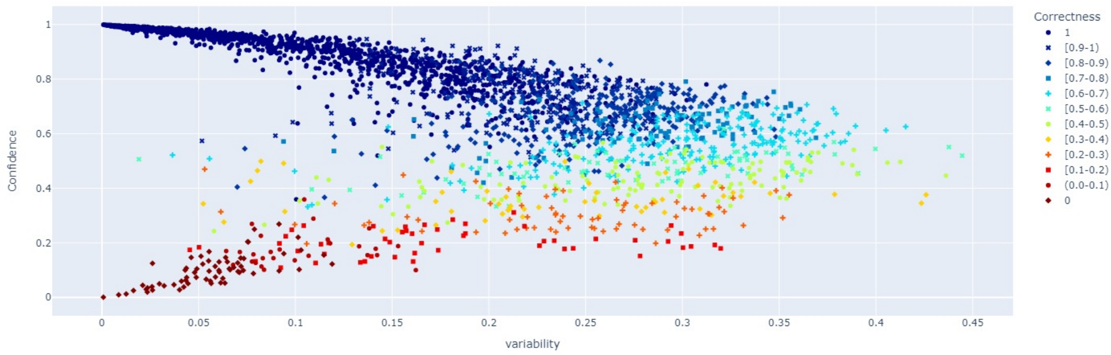

Figure 1: Data map on the train set

Model	Performance on validation set (accuracy)
Human Performance SOTA - UnifiedQA (T5-3B) Random Guess	91.33% 68.80% 20.0%
Our Results using BERT-BASE	
Baseline Training only on easy to learn Training only on hard to learn Swayamdipta's ambigious only training method Sar-Shalom's training method	53.18% 44.4% 17% 43.39% 54.46%
Answer Bias Unique samples Answer Bias	40.16% 31.21%

Table 2: Results on validation set

References

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*.

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. 2021. Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. *arXiv preprint arXiv:2101.00376*.

Aviad Sar-Shalom and Roy Schwartz. 2023. Curating datasets for better performance with example training dynamics. In *In Findings of ACL*.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. *arXiv preprint arXiv:2009.10795*.