Temporal Biases in Language Models When Performing QA Tasks

Anonymous Submission

Abstract

Temporal reasoning is a key component within the performance of language models on different question answering (QA) tasks. We analyze a model's performance on datasets with temporal vocabulary, such as "before", "after", and more to identify performance and potential weaknesses/bias in a language model's temporal reasoning. Through adversarial training, we demonstrate that exposing the model to temporal vocabulary can potentially improve a models performance with QA tasks that include temporal vocabulary and temporal reasoning. We also uncover potential biases that the model might have when it comes to predicting answers given a range of dates.

1 Introduction

Temporal reasoning is the ability for one to distinguish relationships between things, people, and events within a timeline. This can include chronological orders of events and durations of time. The main way we convey temporal significance through language is through vocabulary such as 'before', 'after', 'between', and many other examples. While this is simple for humans, it can be difficult for a language model to be able to reason temporally depending on the data that it is trained on.

This paper examines the performance of the ELECTRA-small model (Clark et al., 2020) in handling QA tasks using the Stanford Question Answering Dataset (SQuAD; Rajpurkar et al., 2016) as a baseline for training and evaluation. We look at baseline metrics given by the trained model, then we use an adversarial data set focused on using temporal vocabulary to identify weaknesses in temporal reasoning the model makes. Then we train the model with data using different kinds of temporal vocabulary to see how it affects the model in terms of the original dataset and examples containing temporal vocabulary and potential temporal reasoning problems.

2 Analysis

The model used for this paper is an ELECTRA-small model. The initial dataset used to train and evaluate the model was the SQuAD dataset, a QA dataset containing 100,000+ question-answer pairs from various contexts sourced from Wikipedia articles.

2.1 Baseline metrics

The model was trained over 3 epochs, and achieved a final accuracy score of 77.96 and a final f1 score of 85.81. a visual representation of loss can be seen in figure 1.

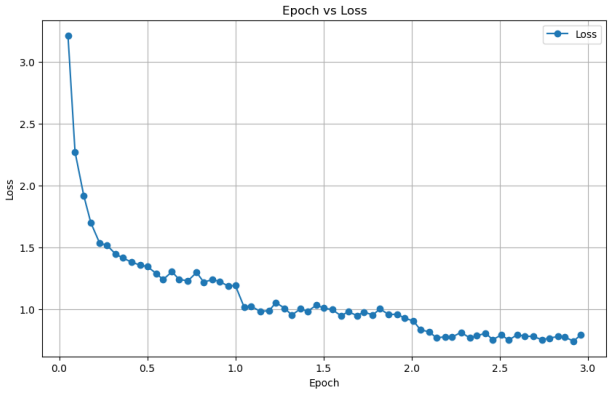

Figure 1: Loss over the 3 epochs.

Table 1 shows the proportion of correct/incorrect predictions based on the type of question.

The type of question that yields the highest accuracy are 'when' type questions. These kinds of questions can be largely split into two groups based on their corresponding answer; date-based

Type	Total questions	Percent correct
where	431	67.29
what	4763	71.3
who	1096	83.03
how	1090	74.59
when	695	87.19
why	151	50.99
other	2344	76.41

Table 1: Accuracy based on question type.

and text-based. Date-based answers use dates to answer a question. There isn't much room for interpretation, either the answer is correct or incorrect (ex. Q: When was Abraham Lincoln born? A: 1809). Text-based answers have a greater potential for error, as the model must answer the question with a string of text, usually taken from the context itself (ex. Q: When did World War 1 start? A: After the murder of Franz Ferdinand).

2.2 Temporal Vocabulary

Another kind of distinction to be made within 'where' QA pairs is the presence of temporal vocabulary. Temporal vocabulary refers to vocabulary that indicate an order of events relative to each other. Some words include 'before', 'after', 'during', etc. When looking at how temporal vocabulary affects the predictions made by the model, it appears that the model is less accurate when temporal vocabulary is present within questions, contexts, or answers. Table 2 shows the accuracy of the model when temporal vocabulary exists within a QA pair and/or context.

Type	Correct	Incorrect	Accuracy
Baseline	606	89	87.19
'Before'	50	17	74.7
'After'	135	26	83.9
'During'	88	16	84.8

Table 2: Accuracy of the model when certain temporal vocabulary is present.

Temporal vocabulary appears within incorrect predictions at a higher rate than within correct predictions. For example, 'before' appears within the 89 incorrect predictions 27 times, while only 67 times within the correct predictions. This leads to an appearance rate in incorrect predictions that is almost 3 times greater. Inaccuracy caused by temporal vocabulary also stacks. For example, the more times the word 'before' appears in a context, the less likely the model is to correctly predict an answer.

2.3 Potential Bias

There seems to be a slight bias within the model when it comes to predicting dates, particularly within ranges. When faced with answering a question given a range of dates, the model consistently uses the latter range of the date as an answer.

(example - context: The last glacial ran from 74,000 BP to 11,600 BP. Q: When did the last glacial start? A: 11,600 BP).

The model also might have a bias to select a more recent date than an older date when it is unsure about the actual answer. In incorrect predictions where the predicted answer is a date, the predicted answer is a later, more recent date than the actual answer almost 60 percent of the time. This could signify a slight bias to select dates that are closer to the current date. It could also signify a slight bias to select dates that occur later in the context, as dates tend to increase chronologically as the context continues.

There isn't enough evidence to conclude that there is a bias within the model, but this observation was made.

2.4 Adversarial Analysis

In 2017 Jia and Liang published Adversarial examples for evaluating reading comprehension systems (Jia and Liang, 2017). In their paper, they explain how adversarial challenge datasets, or datasets that try to confuse or deceive the language model, can cause drops in accuracy, allowing one to further view shortcomings of the model and/or data being used.

We manually created an adversarial dataset using excerpts from Wikipedia as contexts. Each context within the data set used at least one word that could be categorized as temporal vocabulary. The main categories of temporal vocabulary each observation fell into were 'before', 'after', and 'other forms of temporal vocabulary'(such as 'during' and date ranges). There are 100 observations of each category, making 300 observations in total.

When evaluated against the adversarial dataset, the baseline model achieves an accuracy of 84.2 percent. While this is lower than the overall baseline accuracy, it does not seem to be bad compared to the accuracy the model achieved with original squad observations with temporal vocabulary.

The errors that the model makes are consistent with the mistakes it made with the squad data set. There are a number of mistaken dates, and when a context includes a date range, it chooses the later, more recent date. We wondered if training the model with adversarial examples that contain temporal vocabulary would help the model perform better when faced with such QA pairs and contexts. We also wondered whether training on more temporal vocabulary would address the bias the model has when picking more recent dates, given a date range.

3 Methods

We created training and evaluation data sets with each kind of temporal vocabulary, 'before' 'after', 'other', and all three combined. Each training set was made up of 150 QA pairs(450 total). After training the baseline model on one of the training sets or a combination of the three, they were then evaluated on the SQuAD dataset to compare the performance to the baseline model. By training and evaluating the model one by one, we can see how each kind of temporal vocabulary affects the model.

3.1 'Before'

The results of the evaluation of this model can be viewed in table 3. After training the baseline model on QA pairs containing the word 'before', we see that the overall accuracy on 'when' questions drops quite a bit, and it performs the worst of all models trained in this area. When analyzing the mistakes that this model makes, it seemed to generate many text answers in place of date answers. For example, the model could have predicted an answer like 'around the time of the invasion' rather than the actual answer of '1836'. This could indicate that the training data was structurally different enough from the SQuAD data to affect the accuracy this much.

However, this model performed the best of all models when it came to QA pairs containing temporal vocabulary. The model significantly increases accuracy when predicting answers across all three groups of temporal vocabulary that have been assigned. When viewing predictions the model makes, it seems that the model it seems that the model started to pick earlier dates when confronted with date ranges and multiple dates when trained with 'before' QA pairs, rather than later dates as the original model would pick. This could have been a coincidence, but it could be the case that training on the 'before' QA pairs improved the temporal reasoning of the model by addressing the supposed bias towards later dates that was described earlier.

	Correct	Incorrect	Accuracy
Overall	554	141	79.7
'Before'	54	12	81.8
'After'	148	13	91.9
'Other'	92	12	88.4

Table 3: Accuracy of the model when trained on QA pairs focused around the word 'before'.

3.2 'After'

The results of the evaluation of this model can be viewed in table 4. After training the baseline model on QA pairs containing the word 'after', we can see that the overall accuracy on the data set drops slightly, while accuracy on temporal examples improves slightly. The model does very well when faced with 'before' examples, but it is lackluster when faced with other temporal examples, not improving much from the baseline model. It seems odd that the models trained with 'before' and 'after' QA pairs perform better when evaluated by the other grouping, but it could be the case that the models would need more training data to draw any convincing conclusions.

	Correct	Incorrect	Accuracy
Overall	590	105	84.9
'Before'	55	11	83.3
'After'	139	22	86.3
'Other'	87	17	83.6

Table 4: Accuracy of the model when trained on QA pairs focused on the word 'after'.

3.3 'Other'

The results of the evaluation of this model can be viewed in table 5. After training the baseline model on QA pairs containing other forms of temporal vocabulary, it seems to have struck the best balance between accuracy on the SQuAD dataset as a whole and examples using temporal vocabulary. This model achieves a comparable accuracy compared to the baseline model while performing

quite well with QA pairs using temporal vocabulary.

When looking through predictions the model makes, it seems to strike more of a balance between choosing earlier dates vs. later dates, with the frequency each is chosen hovered around 50 percent. It is possible that introducing more date ranges to the data set forces the model to think about events more in how they relate to each other along a timeline, which improves its temporal reasoning while still performing well with the original data set.

	Correct	Incorrect	Accuracy
Overall	600	95	86.3
'Before'	55	11	83.3
'After'	144	17	89.4
'Other'	92	12	88.4

Table 5: Accuracy of the model when trained on QA pairs focused on words such as 'during', date ranges, and other types of temporal vocabulary.

3.4 All types

The results of the evaluation of this model can be viewed in table 6. Unfortunately, training the model on all three types of data resulted in poor accuracy metrics. I suspect that this could be that there is a subtle difference in the training QA pairs from the QA pairs in the SQuAD data set that causes the model to predict answers in ways that are not included within gold labels given by the data set.

	Correct	Incorrect	Accuracy
Overall	575	120	82.7
'Before'	51	15	77.2
'After'	132	29	81.9
'Other'	83	21	79.8

Table 6: Accuracy of the model when trained on QA pairs of all kinds.

4 Conclusion

Training and evaluating the baseline ELECTRA small model on data sets focused around different kinds of temporal vocabulary dropped its overall accuracy when evaluated against the SQuAD data set, but accuracy could be improved based on the types of temporal vocabulary the model is trained on. This suggests that targeted training can address weaknesses within these models such as biases toward selecting earlier or later dates in date ranges.

While training the model on QA pairs containing temporal vocabulary improved accuracy on temporal tasks for the purposes of this paper, possible future improvements upon this research include training on a larger data set of a wider variety of topics, and a deeper look into the nuances of how SQuAD data is structured, to ensure that the model is equipped with consistent data to be able to give the best possible predicted answer. Training the model with adversarial data and temporal vocabulary during training could improve the robustness of a model, improving its performance across QA tasks including and without temporal vocabulary.

5 References

[Clark et al.2020] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Man- ning. 2020. ELECTRA: Pretraining Text En- coders as Discriminators Rather Than Generators. In Proceedings of the International Conference on Learning Representations (ICLR)

[Jia and Liang2017] Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2021–2031, Copenhagen, Denmark, September. Association for Computational Linguistics.

[Rajpurkar et al.2016] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392, Austin, Texas, November. Association for Computational Linguistics.