Detection of Hallucinations in Textual Outputs: SHROOM Shared Task

lakshya kuma raikwar

April 2024

1 Abstract

This report details our approach for the SHROOM shared task, specifically the model-agnostic subtask aimed at detecting hallucinations in text. We employ a fusion model combining the outputs of fine-tuned DistilBert and RoBERTa models to predict hallucinatory content effectively.

2 Introduction

The task of detecting hallucinations in machine-generated text is crucial for ensuring the reliability of language models in real-world applications. This report describes our participation in the SHROOM shared task, focusing on detecting hallucinations without access to the model that generated the text.

3 Methodology

3.1 Data Preprocessing

We preprocess the data by tokenizing text entries using the DistilBert tokenizer, preparing it for model input, which includes hypothesis, target, source, and reference texts, along with a binary hallucination label.

3.2 Model Architecture

Our methodology incorporates a two-stage modeling approach:

- First, we fine-tune a DistilBert model on the task-specific data to predict hallucinations directly.
- We then utilize a pre-trained RoBERTa model, combining it with the fine-tuned DistilBert model in a fusion architecture. This fusion model leverages the strengths of both individual models to enhance the detection capabilities.

3.3 Training and Validation

Both models are trained using a cross-entropy loss, optimized for binary classification. Validation performance is closely monitored to adjust hyperparameters and prevent overfitting.

4 Results

4.1 Evaluation Metrics

The models are evaluated using two primary metrics:

- **Accuracy:** Measures the proportion of correct predictions against the total predictions made.
- **Spearman Correlation:** Assesses how well the model's predicted probabilities of hallucination align with the annotators' judgments.

4.2 Performance of Fine-tuned DistilBert Model

The fine-tuned DistilBert model achieved the following results on the validation set:

- **Accuracy:** 62.1% — indicating the model's ability to correctly identify a significant number of instances.
- **Spearman Correlation:** 33.5 — reflecting its alignment with human judgment on the validation set.

4.3 Performance of Fusion Model

Our fusion model achieved the following results on the validation set:

- **Accuracy:** 66.4% — indicating a high rate of correctly identified instances.
- **Spearman Correlation:** 37.2 — showing strong alignment with human judgment.

These results demonstrate the effectiveness of our model in detecting hallucinations, performing robustly across diverse datasets.

5 Discussion

Integrating outputs from both DistilBert and RoBERTa allows our model to capture diverse linguistic indicators of hallucinations, contributing to robust performance across different types of text.

6 Conclusion

The fusion model demonstrates effective detection of hallucinations, suggesting potential for further development towards more reliable automated text generation systems. Future work will explore advanced ensemble methods and deeper integration of model outputs.