Defending Against Paraphrasing Attacks with RoBERTa-Based Model

Gui Chi

The Chinese University of Hong Kong, Shenzhen

CSC3160
Spring 2023

TL; DR: An investigation of the vulnerability of ChatGPT Detector to paraphrasing attacks.

What is paraphrasing attacks?

Generating restated text without substantially changing the meaning, can significantly reduce the accuracy of detectors.

My Contributions

This research presents an investigation of the vulnerability of ChatGPT Detector to three paraphrasing attacks: Parrot, PEGASUS, ChatGPT-T5-based paraphraser.

- **Reproduced ChatGPT Detector and achieved a high accuracy on HC3 dataset.**
- **Generated paraphrases of HC3 dataset using three paraphraser.**
- **Evaluated accuracy of detector on paraphrases and analyze the reasons behind the decrease in the accuracy.**
- **Proposed practical RoBERTa-based detectors that can successfully defend against these three paraphrasing attacks.**

Motivation

ChatGPT Detector is not reliable

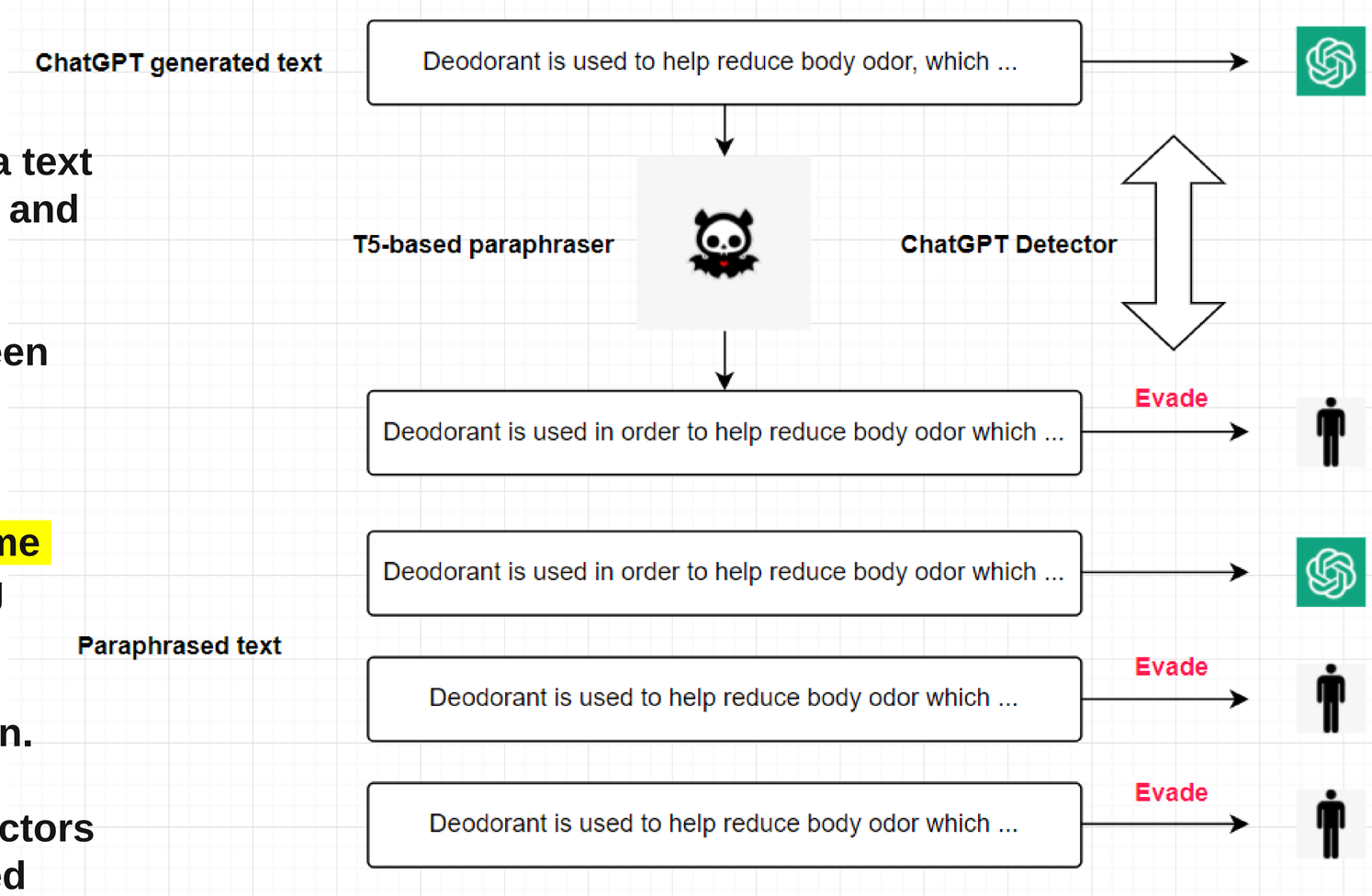
As ChatGPT is upgraded, detecting whether a text is generated by it becomes more challenging and significant.

There are many detection algorithms have been proposed and achieve great performance.

However, people often do not use the text generated by ChatGPT directly, but make some artificial changes, such as deleting, changing words, changing the word order, and so on.

Paraphrasing is similar to human modification.

Recent studies have shown that various detectors are not reliable for detection of LLM-generated texts after paraphrasing attacks.



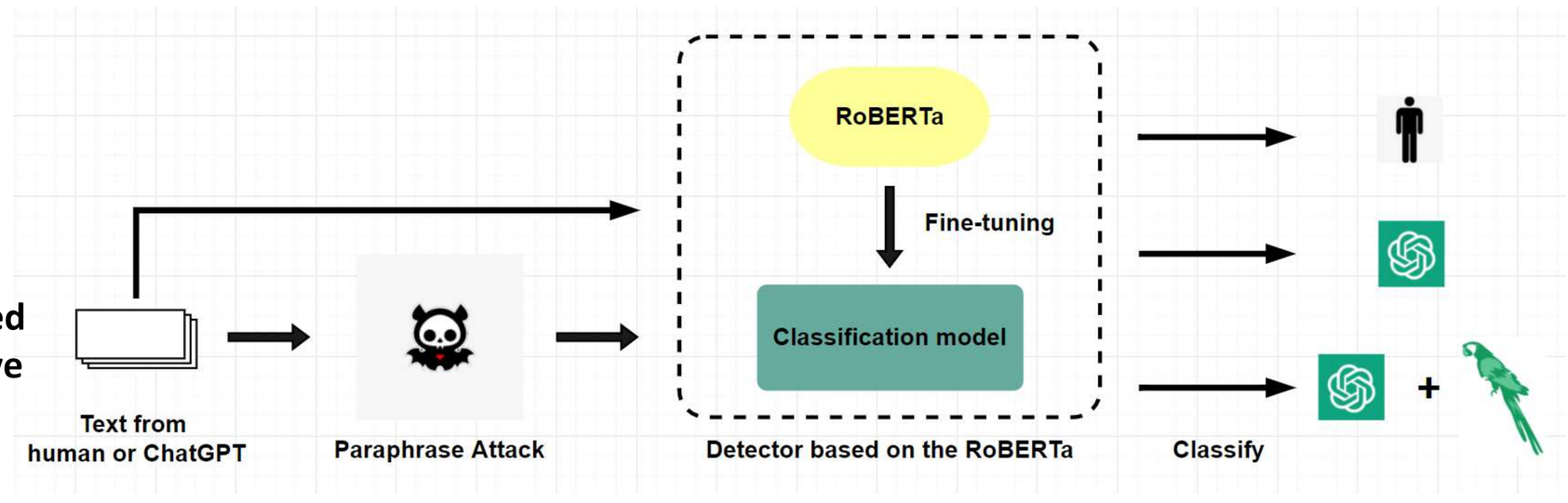
AI generated texts may evade detection after paraphrasing attacks

Model

Model Architecture

I opted to modify the original binary classification detector into a three-way classification detector.

I anticipate that training our RoBERTa-based detector with paraphrased data will improve its accuracy and robustness in detecting paraphrased data.



Different Detectors

I fine-tuned the Roberta with a Roberta-base checkpoint for all detectors.

I trained one binary classification detector and five three-way classification detectors. The later use same model but different training data.

Detectors	Training Dataset				Classification Type
	HC3-English	HC3-Parrot	HC3-PEGASUS	HC3-ChatGPT	
ChatGPT Detector	✔				Binary
Parrot Detector	✔	✔			Three-way
PEGASUS Detector	✔		✔		Three-way
ChatGPT-T5 Detector	✔			✔	Three-way
Large Detector	✔	✔	✔	✔	Three-way
Large* Detector	✔	✔	✔	✔	Three-way

Contact: chigui@link.cuhk.edu.cn

Visit GitHub for more information.

Experiments and Results

Paraphrasing Attacks evade ChatGPT Detect

I evaluated the ChatGPT Detector' robustness against distinct paraphrasing attacks.

I paraphrased the ChatGPT responses in HC3-English and created three paraphrased datasets, HC3-Parrot, HC3-PEGASUS, and HC3-ChatGPT, using the Parrot, PEGASUS, and ChatGPT-T5 based paraphrasers, respectively.

I then tested the paraphrased datasets with the original detectors and examined the reasons for the results obtained.

Paraphrase Method	Accuracy on different test set			
	HC3-Parrot	HC3-PEGASUS	HC3-ChatGPT	Average
Basic Paraphrase	**0.7770**	0.9763	**0.7174**	0.8235
Separator Paraphrase	0.8121	0.9890	0.8293	0.8768
Add "." to the end	0.4058	0.1722	0.1547	**0.2442**

Common differences between Paraphrases and their original text

- Paraphraser can't generate separators, such as "\n"
- Paraphrases have punctuation error
- Original texts contain some fixed statements

Roberta-based Detectors Defense

I anticipate that training our RoBERTa-based detector with paraphrased data will improve its accuracy and robustness in detecting paraphrased data.

I trained six different detectors based on different training datasets and did the evaluations.

Model	Accuracy on different test datasets				
	HC3-English	HC3-Parrot	HC3-PEGASUS	HC3-ChatGPT	Average
ChatGPT Detector	0.9992	0.7892	0.9709	0.7856	0.8862
Parrot Detector	0.9972	0.9987	-	-	-
PEGASUS Detector	0.9819	-	**0.8778**	-	-
ChatGPT-T5 Detector	0.9963	-	-	0.9950	-
Large Detector	0.9821	0.9990	**0.8918**	0.9890	0.9655
Large* Detector	0.9964	0.9970	0.9910	0.9850	**0.9923**

Our results

- All three detectors demonstrated excellent performance on the HC3-English dataset.
- Parrot Detector and ChatGPT-T5 Detector outperformed ChatGPT Detector, while PEGASUS Detector performed worse.
- Large* Detector can successfully defend three paraphrasing attacks while Large Detector has similar problems with PEGASUS Detector.

Case Study & Further Analysis

Take a close look

Common differences in Parrot and ChatGPT-T5 based paraphrases

1. Change the normal word order.
2. Minor errors, such as an extra word, make grammar incorrect.
3. The statement becomes a question.
4. Omit the comma so that the main and subordinate clauses are not separated.
5. Write proper nouns in smaller capitals.
6. It cannot generate punctuation except for the ending. Such as "[", "(", "/"

Similarities between PEGASUS paraphrases and original texts

1. Correct grammar and standard punctuation.
2. Proper nouns are capitalized.
3. Keep part of the content of the original text and remove some subordinate clauses.

Insights

Firstly, AI-generated texts after undergoing a specific paraphrasing attack share certain common features that can be learned during training.
Secondly, the three-way classification approach has proven effective as paraphrases have distinct features that are different from both human texts and origin AI-generated texts.

It is suspected that the high level of similarity between ChatGPT-generated texts and PEGASUS paraphrases may be responsible for the poor performance of the PEGASUS Detector and Large Detector.
I change the type of HC3-PEGASUS dataset to ChatGPT during training and testing in Large* Detector, resulting in high accuracies for all datasets.

Conclusions

Conclusions

- Our experiments revealed the vulnerability of the ChatGPT Detector to various paraphrasing attacks and identified factors contributing to decreased accuracy.
- Based on these reasons, we present three way classification detectors which can successfully defend these attacks.
- Our results support our hypothesis that fine-tuning a RoBERTa-based detector using paraphrased data leads to improved accuracy and robustness in detecting paraphrases.

Limitations

However, Large* Detector may be vulnerable to paraphrasing attacks that do not share common features with the three paraphrasers mentioned above. Our future work will be on making the detector able to resist other attacks and improve its robustness.

GitHub Repo

WeChat

Contact: chigui@link.cuhk.edu.cn

Visit GitHub for more information.