澳門大學

協同創新研究院

協同

人文學院

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

基於語法錯誤差異識別大型語言模型生成的文本：一種黑盒的零樣本方法

吳俊潮, 學生編號: MC256533

理學碩士學位（數據科學 – 計算語言學）課程

項目報告

項目導師

袁毓林教授

2024 年 5 月 5 日

内 容 摘 要

大型語言模型生成文本的檢測器的有效性在很大程度上取決於是否具備一定規模的訓練數據。傳統的白盒的零樣本檢測器擺脫了龐大的訓練資源限制，但在很大程度上仍受限於文本生成模型的可訪問性。本文提出了一種簡單而有效的黑盒的零樣本檢測方法，該方法出於人類寫作更容易犯錯，即人類寫作文本存在更多的語法錯誤的直覺。具體來說，我們發現語法糾錯後的大型語言模型生成的文本和原始文本的相似度往往大於糾錯後的人類寫作的文本和原始文本之間的相似度。利用這一觀察結果，我們提出了一個新的檢測方法，通過引入一個外部的語法糾錯模型來計算輸入文本的語法錯誤差異分數 (GECScore)，分數更高的文本越傾向於是大型語言模型生成的文本。實驗表明，本文提出的零樣本技術優於目前最新的零樣本的和有監督的檢測方法，獲得平均 98.7% 的 AUROC 效能，同時在面對文本釋義和對抗性擾動等潛在的攻擊時具備強大的魯棒性。

關鍵字: 生成文本檢測　語法錯誤　零樣本技術

ABSTRACT

The efficacy of an large language model (LLM) generated text detector depends substantially on the availability of sizable training data. White-box zero-shot detectors, which require no such data, are nonetheless limited by the accessibility of the source model of the LLM-generated text. In this paper, we propose an simple but effective black-box zero-shot detection approach, predicated on the observation that human-written texts typically contain more grammatical errors than LLM-generated texts. This approach entails computing the **G**rammar **E**rror **C**orrection **Score** (GECScore) for the given text to distinguish between human-written and LLM-generated text. Extensive experimental results show that our method outperforms current state-of-the-art (SOTA) zero-shot and supervised methods, achieving an average AUROC of 98.7% and showing strong robustness against paraphrase and adversarial perturbation attacks.

KEY WORDS: generated text detection grammar error zero-shot detection