Understanding the Effects of Human-written Paraphrases in LLM-generated Text Detection

Hiu Ting Lau, Arkaitz Zubiaga

School of Electronic Engineering and Computer Science, Queen Mary University of London London E1 4NS

ARTICLE INFO

Keywords:
LLM-generated text detection
human-written paraphrases
large language models

ABSTRACT

Natural Language Generation has been rapidly developing with the advent of large language models (LLMs). While their usage has sparked significant attention from the general public, it is important for readers to be aware when a piece of text is LLM-generated. This has brought about the need for building models that enable automated LLM-generated text detection, with the aim of mitigating potential negative outcomes of such content. Existing LLM-generated detectors show competitive performances in telling apart LLM-generated and human-written text, but this performance is likely to deteriorate when paraphrased texts are considered. In this study, we devise a new data collection strategy to collect Human & LLM Paraphrase Collection (HLPC), a first-of-its-kind dataset that incorporates human-written texts and paraphrases, as well as LLM-generated texts and paraphrases. With the aim of understanding the effects of human-written paraphrases on the performance of state-of-the-art LLM-generated text detectors OpenAI RoBERTa and watermark detectors, we perform classification experiments that incorporate human-written paraphrases, watermarked and non-watermarked LLM-generated documents from GPT and OPT, and LLM-generated paraphrases from DIPPER and BART. The results show that the inclusion of human-written paraphrases has a significant impact of LLM-generated detector performance, promoting TPR@1%FPR with a possible trade-off of AUROC and accuracy.

1. Introduction

Large language models (LLMs) have become essential in Natural Language Processing (NLP) thanks to their advanced capabilities for text processing and generation, which is achieved through analysis of patterns and relationships between words and sentences using transformer models [42]. Consequently, LLMs have had a significant impact on Natural Language Generation (NLG), as they have provided improved capacity for automatically generating high quality text [4].

While the advancement of LLMs in the context of NLG has aided tasks such as machine translation [38] and text summarization [40], it has also given rise to undesired social problems, including intentional malicious usage, ethical concerns and information inaccuracy. This has brought about the need for researching the development of methods for automated LLM-generated text detection which distinguishes if a text is human- or LLM-generated [12]. Currently, there are 2 major streams of LLM-generated text detectors: (i) zero-shot classifiers [25, 33], which identify LLM-generated text based on the pattern and characteristics of the input, and (ii) watermark detectors [17], which rely on detecting the presence of watermarks which are imprinted into the text during the generation process [16], and are effective in the cases where the watermarks have been added by the LLM. The detectors then examine the input, classifying it as LLM-generated if the level of watermarking exceeds a set threshold, or as human-generated otherwise.

Both kinds of detectors have demonstrated excellent performance in LLM-generated text detection. However, research testing these detectors has primarily focused on datasets involving texts which are exclusively generated by humans or by LLMs. There can be, however, more complicated cases, such as paraphrased texts, which have been seldom considered in previous research. Paraphrasing is defined as the rewriting of context in a simpler and shorter form [8]; an LLM-generated text which is then paraphrased by humans leads to modified texts where the statistical properties of watermarks in the LLM-generated text is no longer identifiable. Since the above detectors perform classification based on token patterns and watermarks, paraphrasing could effectively evade both zero-shot classifiers and watermark detectors while preserving semantic information from the original LLM-generated text. It is

ml21305@qmul.ac.uk (H.T. Lau); a.zubiaga@qmul.ac.uk (A. Zubiaga)
ORCID(s): 0009-0008-7085-1800 (H.T. Lau); 0000-0003-4583-3623 (A. Zubiaga)

important to identify that a text originated from an LLM, despite having been subsequently paraphrased, as this can still be leveraged for massive generation of texts for malicious purposes.

In this work, we are the first to comprehensively study the effectiveness of LLM-generated text detectors in the presence of human-paraphrased texts, in turn assessing the impact of these edited texts on the model performance. In this study, we aim to address this problem by using human-written paraphrases for classification, with the notion that human-written paraphrases and LLM-generated paraphrases might contain different characteristics, which potentially improve the classifiers' performances. We set forth the following research question: "What are the effects of including human-written paraphrases in LLM-generated text detection?" With this aim and research question in mind, we make the following contributions:

- We perform a review of the literature to investigate the existing NLG developments, the importance of LLM-generated text detection, and the performances of existing detectors.
- We describe a data collection process which enables us to build and release the Human & LLM Paraphrase Collection (HLPC) dataset with human-generated and LLM-generated documents, along with their paraphrases.
- We perform classification experiments using state-of-the-art AI paraphrasers and detectors, along with human-written paraphrases.
- Our study contributes to the domain of LLM-generated text detection by examining the effects of including human-written paraphrases in classification and providing insights on data inclusion in future detector building.

2. Related Work

We review existing research in NLG with a focus on LLMs, the importance of LLM-generated text detection and existing detection models, following with a discussion of the main research gaps that our study addresses.

2.1. Natural Language Generation (NLG)

NLG represents a substantial branch of research within NLP, where existing NLG tasks include question-answering [1], text summarization [11], and machine translation [23].

LLMs in NLG. Very recently, the development of LLMs has brought significant improvements to NLG, primarily due to their ability of learning linguistic patterns from very large-scale corpora. Before the adoption of LLMs, two techniques were used for NLG. The earliest NLG systems used templates and rules [26] and the later systems utilized conditional probability between words to account for context dependency [13]. Fundamentally, these systems lack flexibility and adaptability since text generation is restricted, resulting in unfavorable generation patterns, such as inaccuracy in question answering in the rule-based model and word repetitions in the probabilistic model.

Recent research proposed different LLMs that adopt deep learning and neural networks in NLP, contributing to a remarkable improvement in NLG [13]. With the use of a transformer architecture, the models can capture long-range text dependencies with positional encoding, allowing the models to understand both in-words and in-sentence relationships. The models can also be fine-tuned to cater to specific NLG needs. Coupled with large-scale datasets and parameters, LLMs can understand complex linguistic patterns and relationships. For example, GPT-3 is trained with 499 billion tokens and 175 billion parameters [6, 39]. This promotes models' learning of language representations, including syntactic, contextual and semantic information [13]. Text data from large corpora are used to train LLMs, such as CommonCrawl, WebText2, BookCorpus, etc [6, 39], which allow the models to learn knowledge from a broad range of disciplines. With the above advancements, LLMs can thus generate human-like outputs solely using user prompts as inputs [19]. As a result, the use of LLMs along with a simple chat interface has attracted massive usage and attention from the general public [34], with over 180 million users for OpenAI ChatGPT [32].

While LLMs provide a new direction for text generation in NLG, they are also being widely used to support evaluation of NLG outputs [21], which is however beyond the scope of this study.

Paraphrase Generation. One of the NLG tasks for which LLMs have brought a significant boost is paraphrase generation, with various AI paraphrasers built on top of existing LLMs. Paraphrasing is defined as the rewriting of context in a simpler and shorter form [8], and is used extensively to avoid content to be flagged as plagiarism. A paraphrase generator takes sentences or paragraphs as inputs, and creates rewritten outputs which preserve the semantics of the original text [13]. Currently, there are 2 types of AI paraphrasers: (i) systems that are inherently built for paraphrasing, specifically built to paraphrase text automatically and evade LLM-generated text detectors; an example of this is DIPPER [19], which could paraphrase long paragraphs and control output diversity, and (ii) AI chatbots that receive paraphrasing prompts to produces paraphrases; an example of this is T5-paraphraser Parrot [7].

To account for the quality of paraphrases, grammar accuracy and content semantic preservation are considered either with human experts or automatic model evaluation. In [31], human evaluation was conducted for the paraphrases generated from DIPPER and Llama-2-7B-Chat. Paraphrases from both models exhibit high ratings in terms of content preservation and grammar accuracy. [19] uses the P-SP model to compute semantic similarity scores which reflects the level of contextual relationship between the paraphrases and original text. DIPPER effectively paraphrases text with high semantic similarity. This shows that existing AI paraphrasers, coupled with the use of LLM, have shown great advancement in aiding automatic paraphrase generation.

2.2. Importance of LLM-generated Text Detection

Despite the convenience brought by LLMs for improved NLG, they can also cause various problems with a negative impact on society, which has motivated the need for investigating methods for LLM-generated text detection. Next, we discuss three key problems arising from the use of LLMs, especially when the generated texts is not flagged or labelled as being LLM-generated: intentional malicious usage, ethical concerns and information inaccuracy [41].

Intentional malicious usage. Malicious actors may exploit LLMs to produce fake content or to produce content in circumstances where LLM-generated content is unacceptable. Examples include the creation of fake news in political elections to denigrate competitors [15] and the creation of writing or code to then claim full credit for it in an academic environment [17]. As proposed in [33], malicious actors can be categorized into 3 levels: low-skilled, moderate-skilled and advanced-skilled based on their programming level and available resources. Moderate-skilled actors could already produce fake news or build spambots for social media, let alone the adverse effects that advanced-skilled actors could bring.

Ethical concerns. Ethical considerations concerning gender, race and ethnicity bias are raised attributing to the inherent sampling bias of the LLM, which potentially leads to social inequality and discrimination [2]. As presented in Section 2.1, LLMs are trained from very large corpora. However, the sampling of content from these corpora might itself be biased. For example, a study found that Reddit data used to train GPT-2 [29] is composed of context generated mainly by young males in the United States [28]. Since the training of models gives equal weighting for each document in the sample [15], the resulting model thus shows an under-representation of female groups and groups in other age ranges. These biases can be further extended to race, ethnicity and disability [14, 5]. As a result, LLMs might generate biased content that instills negative stereotypes and sentiments toward certain demographics [14], leading to exacerbated social inequality and discrimination.

Information inaccuracy. Information accuracy is not guaranteed by LLMs, which can end up misleading users who are inexperienced or who otherwise overly trust LLM outputs. LLMs can in fact be negatively impacted by inaccurate information that is inherent in the large corpora used for training them. For example, Reddit as a source of GPT-2 [29] has a risk of containing information that is not verified, leading to potential information inaccuracy and credibility issues. As a result, LLM-generated text might be inaccurate, even if the information is factual [25, 15, 22]. With LLM-generated content being perceived as a credible source by many users [18], it might mislead non-professional users with its outputs of varying levels of quality. Meanwhile, if the problem persists, it will lead to a degradation in LLM-generated text accuracy or cause complications in future training of LLMs.

With the massive creation of LLM-generated content on the web, there is a risk that future training of LLMs could include LLM-generated content without necessarily knowing, which can then amplify accuracy issues on the LLMs. Research has suggested that LLM-generated content should be excluded from training to avoid this problem [29] indicating that unnecessary overhead is caused due to the inaccuracy of LLMs. To conclude, considering the increasing usage and potential problems of automatic text generation, it is important to identify LLM-generated text to safeguard the quality of output and to mitigate the aforementioned problems with human oversight.

2.3. State-of-the-art LLM-generated Text Detection

Approaches to LLM-generated text detection. Currently, there are 2 major streams of LLM-generated text detectors: i) zero-shot classifiers [33, 25], and ii) watermark detectors [17]. Zero-shot classifiers aim to identify