Detecting Large Language Model-Generated Japanese Text:
Dataset Construction and Model-Agnostic Classification

2410181 Amane Morikawa

In recent years, the technology of Large Language Models (LLMs) has advanced rapidly, making it possible to generate natural and fluent text that is indistinguishable from human-written text. LLMs have the potential to support humans in a wide range of fields, such as education, law, and medicine. On the other hand, their advanced generative capabilities are giving rise to increasingly serious social issues, including the dissemination of fake news and the illicit use of these models in student reports. Against this background, the demand for methods to detect whether a text is LLM-generated text or human-written text is increasing. Most existing research targets English, and research on LLM-generated text detection for Japanese has not been paid much attention. Furthermore, it has been reported that in existing methods, the detection performance decreases when the domain or the LLM of the text differs between the training data and the test data. Under this background, this study aims to establish a general-purpose LLM-generated text detection method for Japanese.

First, we construct a large-scale dataset that does not depend on specific LLMs or domains. We targeted research paper abstracts as the primary data. We collected 9,343 research paper abstracts from 11 academic journals indexed in J-STAGE and regarded them as human-written text. Next, we used LLMs to generate abstracts for the same papers. In this process, we extracted text corresponding to the "Introduction" and "Conclusion" from the collected papers and provided them in prompts. In addition, we performed processes such as generating abstracts whose length is close to the actual abstracts and normalizing punctuation to generate text similar to human-written text. For the LLMs, in addition to the latest commercial models such as GPT-4o and Gemini 1.5 Flash, we used models specialized for Japanese: Llama-3-ELYZA-JP-8B (Llama3) and Llama-3.1-Swallow-8B (Swallow-8B). Furthermore, to verify the generalization performance of the classifier against different domains, we constructed datasets for three domains: Yahoo! Answers, Wikinews, and Wikipedia. Finally, we constructed a dataset consisting of approximately 37,000 pairs of human-written texts and LLM-generated texts for research paper abstracts, and approximately 10,000 pairs for the other domains. This dataset enables the establishment of a general-purpose LLM-generated text detection model and the comprehensive verification of its robustness against different domains.

Using the constructed dataset, we analyze the differences in linguistic features between human-written text and LLM-generated text. First, we in-

vestigate statistics such as the average number of characters and the diversity of the vocabulary used. As a result, it was confirmed that human-written text uses a more diverse vocabulary compared to LLM-generated text. Next, we calculated the perplexity (PPL) of each text using Llama3, Swallow-8B, and Japanese GPT-2, and compared their distributions. As a result, in many domains, LLM-generated text had a lower PPL compared to human-written text, and this tendency was particularly strong when evaluated by the model that generated the text itself. However, in the domain of research paper abstracts, the difference in PPL between them was small, and it was found that it is difficult to perform LLM-generated text detection using only PPL. Furthermore, we conducted analyses from various perspectives, such as part-of-speech distribution, dependency structure, sentiment polarity, and sentence embeddings using SBERT, but no feature that could clearly distinguish between human-written and LLM-generated texts was confirmed.

To detect LLM-generated text with high precision, we construct an LLM-generated text detection model using a Japanese pre-trained RoBERTa model. Specifically, we fine-tune RoBERTa using the constructed dataset to build a classifier. We also conduct experiments to evaluate the performance of the LLM-generated text detection model using our dataset. We used Accuracy, Recall, Precision, F1, and AUROC as evaluation metrics. Furthermore, we created the training data and test data with various combinations of domains and LLMs used for text generation to evaluate the robustness of the classifier.

As a result of the experiments, the classifier trained using all subsets achieved extremely high detection performance, with an Accuracy of approximately 0.96 and an AUROC of 0.99 or higher. On the other hand, the accuracy of LLM-generated text detection by humans was approximately 62%, which was close to random prediction. It implies that the machine learning model captures features that are imperceptible to humans and can detect with high precision. In the evaluation of the robustness of the classifiers, we confirmed that a classifier trained with text from a single domain and a single LLM as training data experienced a decrease in detection performance against test data of unknown domains or LLMs. For example, a classifier trained only on the "Yahoo! Answers" subset had an Accuracy of approximately 0.52 against other domains. In contrast, when one subset was excluded and all the remaining data was used as training data, an Accuracy of 0.91 or higher was maintained even for the test data of the excluded unknown subset. This result demonstrates that the robustness of LLM-generated text detection increases by training a classifier on a dataset that mixes texts from various domains. Furthermore, even against attacks that paraphrased part of the input text using a T5 model, the decrease in the detection accuracy of the classifier was small, confirming that it possesses robustness against

detection evasion.

Finally, to explore the factors behind why the classifier demonstrated high performance, we performed visualization of feature words using Integrated Gradients (IG), evaluation of the effectiveness of the layers of the RoBERTa model, and analysis using linear classifiers. In the visualization by IG, it was found that the model focuses on typical phrases in academic papers that have high frequency in LLM-generated texts. In the analysis of the effectiveness of each layer of RoBERTa, it was found that abstract representations capable of performing LLM-generated text detection with sufficiently high precision are obtained at the stage of the middle layers (6th and 7th layers), which are said to reflect syntactic information. As a result of training linear classifiers whose features are only surface features, syntactic features, or semantic features and evaluating their performance, an Accuracy comparable to the RoBERTa classifier (approximately 0.93) was obtained using only Bag-of-words (BoW) features. These results suggest that machine-learned classifiers do not necessarily detect LLM-generated text through advanced contextual understanding or semantic judgment, but rather use features such as "lexical bias" and "formulaic expression patterns", which are difficult for humans to recognize, as primary clues.