Nota AI at GenAI Detection Task 1: Unseen Language-Aware Detection System for Multilingual Machine-Generated Text

Hancheol Park **Jaeyeon Kim** **Geonmin Kim** **Tae-Ho Kim**
Nota Inc.
{hancheol.park,jaeyeon.kim,geonmin.kim,thkim}@nota.ai

Abstract

Recently, large language models (LLMs) have demonstrated unprecedented capabilities in language generation, yet they still often produce incorrect information. Therefore, determining whether a text was generated by an LLM has become one of the factors that must be considered when evaluating its reliability. In this paper, we discuss methods to determine whether texts written in various languages were authored by humans or generated by LLMs. We have discovered that the classification accuracy significantly decreases for texts written in languages not observed during the training process, and we aim to address this issue. We propose a method to improve performance for unseen languages by using token-level predictive distributions extracted from various LLMs and text embeddings from a multilingual pre-trained langauge model. With the proposed method, we achieved third place out of 25 teams in Subtask B (binary multilingual machine-generated text detection) of Shared Task 1, with an F1 macro score of 0.7532.[1]

1 Introduction

Recently proposed large language models (LLMs) have demonstrated the ability to generate natural language with a level of fluency akin to that of humans, but they can still produce content that includes incorrect information (Azaria and Mitchell, 2023; Ji et al., 2023). Due to this fluency, people may not realize that the generated text contains inaccuracies, making it easier for false information to spread as if it were true. This can lead to various negative consequences. As a result, detecting text generated by LLMs has become increasingly important. In particular, with numerous language models now supporting multilingual text generation, identifying LLM-generated text across different languages has also become a significant research topic.

In this paper, we discuss methods to determine whether texts written in various languages are authored by humans or generated by LLMs. More specifically, we describe the system we developed for Subtask B of Shared Task 1 (i.e., binary multilingual machine-generated text (MGT) detection) (Wang et al., 2025) at the COLING 2025 Workshop on Detecting AI Generated Content (DAIGenC). The goal of this task is to develop a high-performance binary classification system. To create the dataset used in this task, a variety of LLMs, ranging from closed models such as GPT-4 to open-source models such as the LLaMA series (Dubey et al., 2024), were utilized. One of the main challenges in this shared task is that a significant number of samples in the evaluation set are written in languages that the models did not observe during the training phase.

It is known that when pretrained language models (PLMs)[2], which have been pre-trained on multilingual raw corpora, are fine-tuned with large-scale natural language understanding task datasets in specific languages, these models can effectively perform those tasks even on samples written in languages not observed during training (Gaim et al., 2023). This is referred to as zero-shot cross-lingual transfer learning (Artetxe et al., 2020). In this work, we conducted a preliminary study to examine the effectiveness of zero-shot cross-lingual transfer learning in multilingual MGT detection. We trained the *multilingual E5-large* (Wang et al., 2024) model on two different datasets. One dataset consisted of the entire training data provided for Subtask B of Shared Task 1, and the other was composed only

[1]Our code is available at https://github.com/nota-github/NotaAI_Multilingual_MGT_Detection.

[2]Since LLMs are also pre-trained language models, there could be confusion regarding the terminology. In this paper, we will refer to encoder-based language models, such as *BERT* (Devlin et al., 2019) and *RoBERTa* (Liu et al., 2019), which have been used for natural language understanding tasks, as PLMs.

Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect), pages 191–196
January 19, 2025.

of samples written in English from that dataset. Evaluation was conducted on 27,045 samples from the development set, excluding English samples. The model trained on the entire dataset showed a very high F1 score (0.9806), as there were no samples written in unseen languages. However, when the model was trained only on English samples, although there was some evidence of zero-shot cross-lingual transfer learning, a significant drop in performance was observed (F1 score: 0.7965).

In this work, samples written in seen languages are inferred using a multilingual PLM trained through a standard supervised fine-tuning approach. To more accurately distinguish the source of texts written in seen languages, we investigate various multilingual PLMs. However, for texts written in unseen languages, inference is performed differently from the traditional approach. To determine whether samples written in unseen languages are MGTs, we explore features that can be commonly used for the multilingual MGT detection task, regardless of the language.

In monolingual MGT detection, LLMs are known to assign high probability value to each generated token of the MGT (Sarvazyan et al., 2024). However, it is unclear whether this will be useful in multilingual MGT detection. This is because many unseen languages are likely low-resource languages with insufficient training data, meaning LLMs may not have learned many tokens for these languages. Therefore, it is uncertain whether LLMs will assign high probabilities to all tokens in texts written in such unseen languages. In this study, we examine the effectiveness of a model that uses token-level predictive distributions extracted from various LLMs as features for multilingual MGT detection. The previous study (Sarvazyan et al., 2024) used only the LLaMA-2 models, but we utilized various models to reflect the characteristics of different LLMs. To address the issue of differing tokenization results across models, we also propose a novel network architecture. Additionally, we found that using meaning representations extracted from a multilingual PLM that had not been fine-tuned further improved performance. Previously, such meaning representations were not utilized together.

The experimental results showed that token-level predictive distributions extracted from various LLMs and embeddings from a multilingual PLM are useful in multilingual MGT detection. The system we proposed achieved third place out of 25 teams in the Shared Task 1, with an F1 macro score of 0.7532.

2 System Overview

Our MGT detection system first identifies the language in which the given text is written. We use LangID[3] as the language identification tool. If the given text is written in a language observed during training, it is inferred using a model fine-tuned with supervised learning on a multilingual PLM (§2.1). Otherwise, it is inferred using a model that utilizes token-level predictive distributions extracted from various LLMs as features, along with a meaning representation from a multilingual PLM (§2.2).

2.1 Fine-Tuning a Multilingual PLM on a Labeled MGT Detection Dataset

If the given text is in a language present in the training data, inference is performed using a supervised fine-tuned multilingual PLM. The PLM is trained to solve the binary classification problem by minimizing the cross-entropy loss. After fine-tuning, we evaluated the performance of various multilingual PLMs using a development set composed solely of samples written in seen languages. We use the PLM that performed the best among them.

2.2 Multilingual MGT Detector for Unseen Languages

As features that can be commonly used in the multilingual MGT detection task regardless of the language, we utilize information from the predictive distributions of each token when a text is fed into LLMs. We extract three features, which are known to be useful in monolingual MGT detection (Sarvazyan et al., 2024), from the predictive distributions of each token.

- **Log probabilty of the predicted token (F1):** This feature represents the log probability of the next token predicted with the highest probability for a given token input from an LLM.
- **Log probabilty of the generated token (F2):** This feature is the log probability of the token actually generated for a given token input in the LLM.
- **Entropy of the predictive distribution (F3):** It represents the entropy value of the probability distribution of the predicted next tokens.

[3] https://github.com/saffsd/langid.py

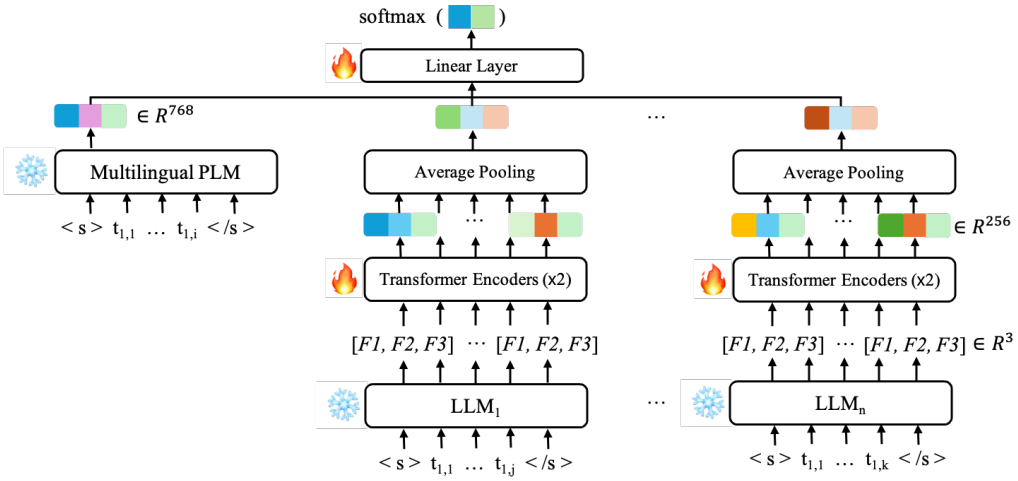

Figure 1: A description of the proposed model for classifying whether a text written in an unseen language is MGT or not. The ice symbol represents a module whose parameters are not updated during training, while the flame symbol indicates a module whose parameters are updated.

As shown in Figure 1, tokenized sample texts are fed into N different LLMs. Since each LLM has a different tokenization method, the given input text may be split into a different number of tokens. The three features extracted from the predictive distributions for each token are then used as inputs to the transformer encoder, producing 256 dimensional hidden states. These features are combined into a single feature vector for all input tokens through average pooling. The feature vectors extracted from the LLMs are concatenated and serve as input to the classifier (i.e., linear layer). In this study, we use three different LLMs.

We also utilize the meaning representation extracted from a multilingual PLM. As a preliminary study, we translated 20 random English texts into all the languages used in the training and devtest sets provided for this shared task, and then extracted text embeddings from *XLM-RoBERTa-base* (Conneau et al., 2020). We considered the hidden state of the <s> token in the last layer as the text embedding, and visualized these embeddings in 2D using t-SNE. As a result, texts sharing similar meanings were positioned close to each other, regardless of the language, while texts with different meanings were positioned farther apart. This phenomenon has been widely discussed in previous research (Ding et al., 2022). Although this feature is language-agnostic and its relation to multilingual MGT is still uncertain, empirically, we observed that this feature improves the MGT detection performance. Before applying this feature, the F1 score on the development set was 0.7114, but it improved significantly to 0.7370. As a result, our system utilizes this feature as well.

One plausible reason why the meaning representation could be useful is the following: LLMs often generate texts that deviate from common sense. Unless multilingual PLMs are intentionally trained to learn noise, these texts are likely to differ significantly from the common-sense knowledge learned by the PLMs. In other words, texts containing such incorrect information may be out-of-distribution samples and could be represented far from samples containing accurate knowledge in the embedding space. We will examine this hypothesis further in future work.

We trained this detector using the features described so far, optimizing it to minimize the cross-entropy loss. The specific language models we used are described in more detail in Section 3.2.

3 Implementation Details

3.1 Datasets

The datasets provided in Shared Task 1 is as follows: For model training, 674,083 training samples and 288,894 samples from the development set are used. Both of these datasets consist of samples written in the same nine languages. For leaderboard evaluation, the devtest dataset and the test set contain 74,081 and 151,425 samples, respectively, with samples written in 11 and 16 different languages.

	F1
XLM-R$_{Base}$	0.9426
XLM-R$_{Large}$	0.9648
mE5$_{Base}$	0.9653
mE5$_{Large}$	**0.9728**

Table 1: Performance of fine-tuned multilingual PLMs on the development set

Rank	Team	F1 Macro	F1 Micro
1	Grape	0.7916	0.7962
2	rockstart	0.7557	0.7564
3	**Nota AI (Ours)**	**0.7532**	**0.7591**
4	LuxVeri	0.7513	0.7527
5	TechExperts(IPN)	0.7463	0.7474
6	azlearning	0.7436	0.7449
7	nampfiev1995	0.7427	0.7440
	Baseline	0.7416	0.7426

Table 2: Top 7 leaderboard for Shared Task1

3.2 Models

The models used in our proposed system are as follows: For samples written in seen languages, we used the fine-tuned *multilingual e5-large* model because this model showed the best performance on the development set among various mutlingual PLMs (see Table 1). For samples written in unseen languages, we used *XLM-RoBERTa-base* as the multilingual PLM. Additionally, since our method uses various LLMs, we aimed to reduce computational costs for inference by employing the following smaller LLMs (sLLMs): *Llama-3.2-1B-Instruct* (Dubey et al., 2024), *Qwen2.5-1.5B-Instruct* (Team, 2024), and *Phi-3-mini-128k-instruct* (Abdin et al., 2024). We have confirmed that these three sLLMs have already learned all the languages used in the training, development, and devtest sets.

3.3 Hyperparameters

All PLMs used in this experiment were trained with the same hyperparameters. The learning rate was set to 5e-5 with a linear decay. We trained for 3 epochs, with a warmup ratio of 0.01, and selected the models that showed the best performance on the development set. The AdamW optimizer (Loshchilov and Hutter, 2019) was used for parameter updates, and the weight decay was set to 0.1. The model that uses predictive distribution information was trained with the following hyperparameters: it was trained for 50 epochs with a learning rate of 2e-4 and linear decay. Both the weight decay and warmup ratio were set to 0.01.

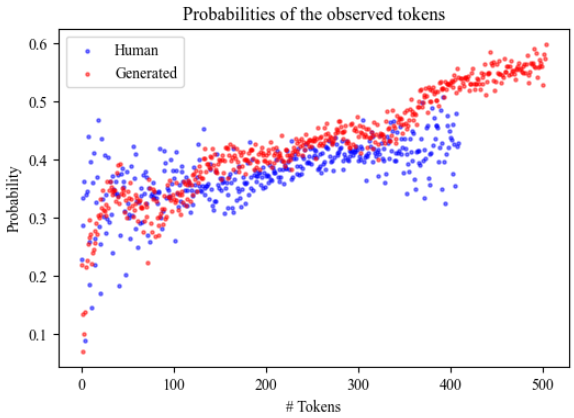

Figure 2: The average probability assigned by the LLM to the generated tokens, based on the token length of the text. The measurements were taken using samples of unseen languages from the devtest set, and the probabilities of the generated tokens were calculated using *Llama-3.2-1B-Instruct*.

4 Results and Discussion

4.1 Leaderboard

As described in Table 2, we achieved third place out of 25 teams in Subtask B (binary multilingual MGT detection) of Shared Task 1, with an F1 macro score of 0.7532.

4.2 Discussion

As mentioned in Section 2.2, it is uncertain whether token-level predictive distributions are helpful for samples written in unseen languages in multilingual MGT detection. To investigate this, we examined the average probabilities of generated tokens for samples from unseen languages in the devtest set, categorized by token length. As shown in Figure 2, we observed that higher probabilities are assigned to generated text compared to human-written text. In other words, this can be considered a discriminative feature for determining whether the text is MGT.

Furthermore, we investigated whether the proposed method is actually effective for samples in unseen languages. We trained the model on English data only and then evaluated the MGT detection performance on samples written in unseen languages from the devtest set. When compared to the *multilingual E5-large* model (F1: 0.9030), which performed zero-shot cross-lingual transfer

learning, our method showed better performance with an F1 score of 0.9175.

5 Conclusion

In this work, we proposed a system for determining whether samples written in unseen languages are MGTs or not, and our approach achieved third place in Subtask B of Shared Task 1. However, while we obtained a relatively high F1 score compared to the baseline, it was not significantly higher. For future work, we should focus more on investigating features that can better distinguish samples in unseen languages.

Limitations

We achieved a high rank of third place in this shared task, but there are some limitations in our methods. First, our approach relies on distinguishing between languages, which means that misidentifying the language type increases the likelihood of making an incorrect classification for that sample. Additionally, while we utilize sLLMs, extracting token-level predictive distributions involves significant computational costs.

Acknowledgments

This work was supported by Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT (MSIT, Korea) & Gwangju Metropolitan City.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kunbo Ding, Weijie Liu, Yuejian Fang, Weiquan Mao, Zhe Zhao, Tao Zhu, Haoyan Liu, Rong Tian, and Yiren Chen. 2022. A simple and effective method to improve zero-shot cross-lingual transfer learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4372–4380, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Fitsum Gaim, Wonsuk Yang, Hancheol Park, and Jong Park. 2023. Question-answering in a low-resourced language: Benchmark dataset and models for Tigrinya. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11857–11870, Toronto, Canada. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Adrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55:1–38.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Areg Mikael Sarvazyan, José Ángel González, and Marc Franco-salvador. 2024. Genaios at SemEval-2024 task 8: Detecting machine-generated text by mixing language model probabilistic features. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 101–107, Mexico City, Mexico. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2402.05672*.

Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, Jinyan Su, Minh Ngoc Ta, Mervat Abassy, Kareem Elozeiri, Saad El Dine Ahmed, Maiya Goloburda, Tarek Mahmoud, Raj Vardhan Tomar, Alexander Aziz, Nurkhan Laiyk, Osama Mohammed Afzal, Ryuto Koike, Masahiro Kaneko, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2025. GenAI content detection task 1: English and multilingual machine-generated text detection: AI vs. human. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.