SemEval-2024 Task 8: Final System Proposal

Anna Smirnova, Aziz Baran Kurtuluş, Vitalii Hirak
University of Groningen, the Netherlands
{a.smirnova.3, a.b.kurtulus, v.hirak}@student.rug.nl

1 Introduction

As generative language models develop, it is getting harder to distinguish machine generated texts from the ones written by humans. As a result, the task of distinguishing between such texts assumes great significance. In this paper we introduce our approach to the Task 8 of SemEval-2024 competition: Multigenerator, Multidomain, and Multilingual Black-box Machine-Generated Text Detection. We provide a comprehensive description of our system for both Subtask A (monolingual and multilingual) and Subtask B.

2 Subtask A: Machine-Generated VS Human-Written text classification

The Subtask A aims to solve the problem of classifying texts into ones created by generative models and ones written by humans. For the binary classification task, we want to focus on using pretrained Language Models (LMs). As was mentioned by Uchendu and colleagues (2020), transformer-based models are usually highly efficient in artificial text detection tasks. Moreover, the embeddings obtained from generative models could be used as an input to the classification model. Generative models themselves could be used for the classification task because they are closer in their nature to the models used for initial data generation.

Additionally, we want to experiment with the learning techniques such as ensembling models and freezing layers. The ensembling technique allows us to use several language models' predictions with additional decisive meta-model on top of them, so that we have more confidence about the final prediction (see for example: Maloyan et al., 2021). Freezing layers is, however, an ambiguous technique in terms of model's quality improvement. It could be useful when dealing with a small amount of data to prevent overfitting.

Below we present some transformer-based models, which we want to experiment with when dealing with mono- and multilingual parts of Subtasks A, however, we are not limited to this list.

2.1 Monolingual

As the monolingual task is about binary classification of English texts, we choose the models pretrained on English data only.

- **BERT**: An encoder-only model based on a pretraining consisting of two objectives: *Masked Language Modeling* (MLM) and *Next Sentence Prediction* (NSP) (Devlin et al., 2018).
- **RoBERTa**: A BERT based model achieved by dropping the pretraining scheme of NSP objective and using the byte-level BPE tokenizer while trained longer and larger batches (Liu et al., 2019).
- **GPT-2**: A generative model from the GPT family with *Causal Language Modeling* (CLM) pretraining objective (Radford et al., 2019).
- **GPT-J-6B**: A GPT model from the EleutherAI with 6B parameters developed by Wang and Komatsuzaki (2021).

2.2 Multilingual

When dealing with multilingual text classification, there are two ways to approach the issue: (1) to use the models pretrained on all target languages, i.e., multilingual models, or (2) to identify the language of text and then use the model pretrained on that corresponding language. Below we propose some models for both these methods.

2.2.1 Multilingual models method

- **mBERT**: The multilingual version of BERT (Devlin et al., 2018). It was trained on 104 languages using Wikipedia.
- **XLM-R**: Based on the XLM model by Lample and Conneau (2019), XLM-R is a multilingual model pretrained on Common Crawl corpus. It supports 100 languages and was pretrained on MLM.
- **mDeBERTaV3**: Based on the DeBERTaV3, which is an improvement of DeBERTa using *Replaced Token Detection* (RTD) pretraining (He et al., 2021). It was trained on the CC100 multilingual data.
- **XGLM**: XGLMs are multilingual generative models that have different versions in terms of parameters (Lin et al., 2021). They all are trained on 30 languages.

2.2.2 Language detection method

To detect the language of a text, we use the *langdetect*[1] Python library, which supports all the languages present in our dataset. After successfully detecting the language of the text, we pass the text to the model pretrained on that language. However, we did not find any monolingual pretrained LMs for Urdu, Indonesian, and Bulgarian languages – for them, multilingual models might be used.

- **AraBERT**: The BERT model pretrained on Arabic news articles, Arabic Corpus and OSIAN corpus (Antoun et al., 2020).
- **ruBERT**: The BERT model pretrained on the Russian part of Wikipedia and news data (Kuratov & Arkhipov, 2019).
- **BERT**: The BERT model pretrained on English data (Devlin et al., 2018). Also the BERT model pretrained on Chinese data is available (bert-base-chinese).

3 Subtask B: Mixed Machine-generated vs. Human-written Text Classification

Similarly to Subtask A, Subtask B is focused on investigating the origin of the texts, i.e., whether they were generated by models or authored by humans. The difference is that, if the text is artificially synthesized, the model that generated it needs to be determined as well. Because this task implies the execution of Subtask A in it, we want to use the best model obtained in Subtask A to classify whether text is model-generated. If it is the case, the text will subsequently undergo another classification process to determine the particular generative language model responsible for its creation.

Because of its similarity to Subtask A as a classification task, we aim to use the same language models, such as encoder-only model like BERT and RoBERTa, as well as generative models like GPT-2 and GPT-J-6b. We are particularly interested in the generative models, since Subtask B is monolingual only, and the generative models can may be especially useful for predicting which model the text was generates by. The learning techniques such as freezing layers and ensembles of models are also to be experimented with under this subtask.

[1] https://github.com/Mimino666/langdetect

References

Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based Model for Arabic Language Understanding. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, pages 9–15, Marseille, France. European Language Resource Association.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv (Cornell University)*. https://arxiv.org/pdf/1911.02116v2

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv (Cornell University)*. https://arxiv.org/pdf/1810.04805v2

He, P., Gao, J., & Chen, W. (2021). DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled

Embedding Sharing. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2111.09543

Kuratov, Y. & Arkhipov, M. (2019). Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language.

Lample, G., & Conneau, A. (2019). Cross-lingual language model pre-training. *arXiv (Cornell University)*. http://export.arxiv.org/pdf/1901.07291

Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., Pasunuru, R., Shleifer, S., Koura, P. S., Chaudhary, V., O'Horo, B., Wang, J., Zettlemoyer, L., Kozareva, Z., Diab, M., . . . Li, X. (2021). Few-shot Learning with Multilingual Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2112.10668

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). ROBERTA: A robustly optimized BERT pre-training approach. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1907.11692

Maloyan, N., Nutfullin, B., & Ilyushin, E. (2022). DIALOG-22 RuATD Generated Text Detection. https://doi.org/10.48550/arXiv.2206.08029.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Wang, B., & Komatsuzaki, A. (2021). GPT-J-6B: A 6 Billion parameters autoregressive language model.

Uchendu, A., Le, T., Shu, K. & Lee, D., 2020, January. Authorship attribution for neural text generation. In Conf. on Empirical Methods in Natural Language Processing (EMNLP).