Utilizing an Ensemble Model with Anomalous Label Smoothing to Detect Generated Scientific Papers

Yuan Zhao[1] and **Junruo Gao**[1] and **Junlin Wang**[1] and **Gang Luo**[1,*] and **Liang Tang**[1]
[1]China Telecom Cloud Technology Co., Ltd
{zhaoyuan1,gaojr1,wangjl52,luog6,tangl33}@chinatelecom.cn
* Corresponding author

Abstract

Generative AI, as it becomes increasingly integrated into our lives, has brought convenience, though some concerns have arisen regarding its potential impact on the rigor and authenticity of scientific research. To encourage the development of robust and reliable automatically-generated scientific text detection systems, the "DAGPap24: Detecting Automatically Generated Scientific Papers" competition was held and shared the same task with the 4th Workshop on Scholarly Document Processing (SDP 2024) to be held at ACL 2024. In the DAGPap24 competition, participants were tasked with constructing a generative text detection model that could accurately distinguish between the human written fragment, the synonym replacement fragment, the ChatGPT rewrite fragment, and the generated summary fragment of a paper. In this competition, we first conducted a comprehensive analysis of the training set to build a generative paper detection model. Then we tried various language models, including SciBERT, ALBERT, DeBERTa, RoBERTa, etc. After that, we introduced an Anomalous Label Smoothing (ALS) method and a majority voting method to improve the final results. Finally, we achieved 0.9948 and 0.9944 F1 scores during the development and testing phases respectively, and we achieved second place in the competition.

1 Introduction

With the rapid development of NLP technology, especially with the emergence of ChatGPT[1], there is an increasing amount of text generated by non-human entities. However, machine-generated text may sometimes contain errors that are not easily discernible by humans, leading to a decline in the rigor and credibility of scientific papers. Current NLP generation technology has nearly reached parity with human writing, which poses the challenge of distinguishing them (Zhang et al., 2023). In response to this issue, Mitchell et al. (2023) propose a detection method that does not require training a classifier or collecting additional datasets. They argue that model-generated text is more sensitive to minor rewriting perturbations than human-written text. To further distinguish the sentences with various lengths, researchers in (Tian et al., 2023) propose a length-sensitive Multiscale positive-unlabeled Loss that can improve the ability of detection of BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). To further understand the achievements, this survey (Tang et al., 2024) provides an overview of existing techniques for detecting LLM-generated text.

In this paper, we mainly detail our work during the shared task DAGPap2024. We are tasked with detecting automatically generated papers. We can abstract this task into a token-level multi-classification problem. There are three challenges in this task. 1) Difficulty in efficiently bridging the gap between word-level tokenization and subword-level tokenization more rationally. The dataset provided in the competition is tokenized in word-level, while most of the existing language models using the subword-level tokenization method. If the two tokenized methods are not accurately mapped, it will not only lead to a waste of useful information but may also lead to misalignment of the prediction results and the test set. 2) Difficulty in ensuring that the distribution of the generated prediction results is consistent with the label distribution of the training set. We found that only relying on the model's predictions will lead to inconsistency in the distribution between the training set and test set. For example, there will be several other labels predicted in the middle of a continuous text. It leads to inconsistency in the distribution between training set and development set. 3) Difficulty in determining which model performs best on token-level classification tasks. Different models, each

[1]https://chatgpt.com/

Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), pages 130–134

with unique training parameters, yield varying predictions for this task, resulting in different label prediction outcomes for the same token. The selection of the final result directly impacts the F1 score. To overcome the aforementioned challenges and enhance our performance in the competition, we propose the following solutions:

- Based on our analysis, we adopted two ways to divide the data into sub-sentences. One is to divide the data into sub-sentences according to the token length of 240, and the other is to divide the tokenized data into sub-sentences with a max length of 500. These two tokenization ways can ensure that the prediction results can correspond back to the original text without the loss of information.

- Then, we fine-tuned DeBERTaV3$_{\text{large}}$(He et al., 2023), DeBERTaV2$_{\text{xxlarge}}$(He et al., 2021) and ALBERTV2$_{\text{xxlarge}}$(Lan et al., 2020) using token-level classification. We proposed an Anomalous Label Smoothing method (ALS) to guarantee the predicted results and the label distribution of the training set. Specifically, by scanning under different window lengths, we smooth out predicted anomalous labels, ensuring that there are no anomalous labels within the window.

- Finally, to take full advantage of different models, we used a majority voting method to ensemble the predictions of multiple models. By ensembling the results from the models, we attained F1 scores of 0.9948 and 0.9944 in the development and testing phases, respectively.

2 Task Overview

To enforce the reliability of scientific papers, we focus on detecting automatically generated scientific papers. We should distinguish human-written fragments, synonym-replaced fragments, ChatGPT-generated fragments, and summarized fragments, which can be abstracted as a token-level classification task. Fig. 1 provides a visualization of the fragments extracted from the training set.

2.1 Data

The original dataset is selected from scientific papers, consisting of 5,000 training samples, 5,000 development samples, and 20,000 test samples. Each

Figure 1: Visualization result of different types of fragments in a training sample.

sample may consist of fragments from the four categories mentioned above, appearing in different sequences and frequencies. The summarized fragments are obtained via a deep learning model, the synonym replacement fragments are obtained by substituting the original words with the synonyms from NLTK, and the ChatGPT fragments are obtained by rewriting the original fragments using ChatGPT. As Fig. 1 shows, the number of labels in this fragment is three. We statistically analyzed the number of the sequence of continuously labeled identical fragments and found that they were identically distributed in the training set, validation set and test set.

We processed the training set according to subword-level tokenization to avoid the loss of useful information (e.g., the information involved in the truncated sentences) and ensure the consistency of prediction results with the test set. We conducted a statistical analysis on the length of tokenized fragments for each category in the training set. The statistical results are shown in Table 1. The column "1%" denotes that, upon arranging all fragment lengths into ascending order, 1% of the fragments have lengths equal to or under the specified value. Columns like "50%" and "75%" are similarly conveyed in the same manner.

Category	0.5%	50%	75%	Mean
human	35.00	1204.00	2234.75	1640.87
synonym	29.00	587.00	672.50	612.20
ChatGPT	21.00	382.00	476.00	397.32
summary	22.84	359.00	413.00	381.25

Table 1: Statistics on the length of the fragments of each category in the training set.

3 Our Work

In this section, we will provide a detailed description of our work, which includes data analysis, data processing, models, and the anomalous label smoothing method. Our work is illustrated in the structural flowchart as shown in Fig. 2.

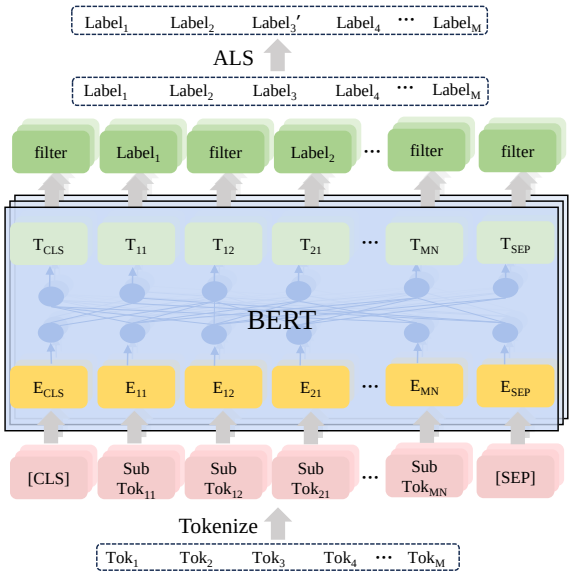

Figure 2: Overall architecture of our work.

3.1 Data Analysis and Data Processing

Initially, through analyzing the raw data, we discovered several key issues. The first was that sentences split according to the length of the word sequence may exceed the maximum length of the BERT sequence after word segmentation. This results in some training data being wasted since the information involved in the truncated sentences may be lost. It is difficult for the predicted label to correctly correspond to the test text. The second was that the 99.5% of continuous label lengths exceed 20.

During data pre-processing, we followed the method suggested by the given example and divided the raw train data into two parts, including the training set and the validation set. Based on the data analysis mentioned above, two data preprocessing methods were adopted. One method named TokV1 was to divide the sentences in the training set and Validation set into sub-sentences with a token list of length 240. The other one named TokV2 was to perform word tokenization before segmentation and then ensure that each sub-sentence after word tokenization was a tokenized text with a max of 500 tokens. This ensured that tokenized sentences would hardly exceed the maximum length of the pre-trained model, allowing the model to achieve the best training effect.

3.2 Models

For the token-level classification task, it is necessary to perceive the contextual relationships of the tokens. The models with a bidirectional attention mechanism might be a good choice, thus, we used models with an encoder structure, including DeBERTaV2$_{\text{xxlarge}}$, DeBERTaV3$_{\text{large}}$, and ALBERTV2$_{\text{xxlarge}}$ and conducted a series of experiments on these models. The details of these models are presented in Table 2.

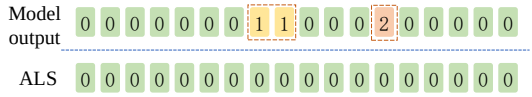

Figure 3: The process of Anomalous Label Smoothing to correct labels.

Model	Parameter	Hidden Size
SciBERT	110M	768
RoBERTa	355M	1024
ALBERTV2$_{\text{xxlarge}}$	223M	4096
DeBERTaV2$_{\text{xxlarge}}$	1.5B	1536
DeBERTaV3$_{\text{large}}$	304M	1024

Table 2: The configuration of the chosen models.

3.3 Anomalous Label Smoothing

We conducted data analysis on the models' outputs and discovered that within a sequence of continuously labeled identical data, there were discrete labels of other categories intermixed, which is different from our analysis results mentioned above. Thus, we designed a post-processing method named Anomalous Label Smoothing(ALS), which is similar to a Conditional Random Field (CRF) (Lafferty et al., 2001) in filtering out unreasonable labels. ALS method corrects label fragments smaller than the window size to the label of its left or right fragment by setting a window size and iterating through the prediction results. Initially select the label on the right, and for the rest, select the label of the fragment on the left. The size of the window is determined in specific experiments, based on the second key point from our previous data analysis. As shown in Fig. 3, through the ALS method, we have corrected the labels of other categories interspersed within a sequence of continuously identical labels.

Model	Score
SciBERT	0.8637
RoBERTa	0.8814
DeBERTaV3$_{\text{large}}$	0.8971
DeBERTaV2$_{\text{large}}$+TokV2	0.9892
ALBERTV2$_{\text{xxlarge}}$+TokV1+ALS$_{10,20}$	0.9885
DeBERTaV2$_{\text{xxlarge}}$+TokV1+ALS$_{10,20}$	0.9887
DeBERTaV3$_{\text{large}}$+TokV1+ALS$_{10,20}$	0.9844
ALBERTV2$_{\text{xxlarge}}$+TokV2+ALS$_{10,20,30}$	0.9897
DeBERTaV2$_{\text{xxlarge}}$+TokV2+ALS$_{10,20,30}$	0.9910
DeBERTaV3$_{\text{large}}$+TokV2+ALS$_{10,20,30}$	0.9919

Table 3: The F1 scores on the development set.

Ensembled	Base Model	Dev	Test
V1	DeBERTaV3$_{\text{large}}$ +TokV1+ALS$_{10,20}$ DeBERTaV2$_{\text{xxlarge}}$ +TokV1+ALS$_{10,20}$ ALBERTV2$_{\text{xxlarge}}$ +TokV1+ALS$_{10,20}$	0.9908	N/A
V2	EnsembledV1 DeBERTaV2$_{\text{large}}$ +TokV2 ALBERTV2$_{\text{xxlarge}}$ +TokV2+ALS$_{10,20,30}$ DeBERTaV2$_{\text{xxlarge}}$ +TokV2+ALS$_{10,20,30}$ DeBERTaV3$_{\text{large}}$ +TokV2+ALS$_{10,20,30}$	0.9948	0.9943
V2$_{\text{ALS}_5}$	EnsembledV1 DeBERTaV2$_{\text{large}}$ +TokV2 ALBERTV2$_{\text{xxlarge}}$ +TokV2+ALS$_{10,20,30}$ DeBERTaV2$_{\text{xxlarge}}$ +TokV2+ALS$_{10,20,30}$ DeBERTaV3$_{\text{large}}$ +TokV2+ALS$_{10,20,30}$	N/A	**0.9944**

Table 4: The F1 scores of the ensemble models.

3.4 Majority Voting

We performed majority voting on the prediction results from multiple models to enhance the robustness and accuracy of the overall outcome. We assume that the label corresponding to the token at position is $l_1, l_2 ... l_i ... l_n$, $l_i \in \mathbb{L} = \{0, 1, 2, 3\}$, where n represents the number of models. The frequency for each label is denoted as $\text{feq}(l_i)$. Thus, the voting rule is as Eq. 1.

$$f(l) = \text{argmax}(\text{feq}(l_i)), i \in \mathbb{L} \quad (1)$$

At the same time, we observed that when determining the final labels through majority voting, if there is a disagreement among the opinions of each model participating in the vote, i.e., the frequency of any label is same, then the final result cannot be determined through voting. Therefore, based on Eq. 1, we optimized the voting rules. We selected the relatively better model output l_T to serve as a final result for such inconclusive situation, as depicted in Eq. 2.

$$g(l) = \begin{cases} l_T & \text{feq}(l_1) = \text{feq}(l_2) ... = \text{feq}(l_n) \\ f(l) & others \end{cases} \quad (2)$$

4 Experiments

During the modeling process, we tried different state-of-the-art models, including SciBERT, RoBERTa, DeBERTaV2$_{\text{xxlarge}}$, DeBERTaV3$_{\text{large}}$, ALBERTV2$_{\text{xxlarge}}$. ALS$_{10,20,30}$ was used to denote the length of Anomalous Label Smoothing, i.e., the smoothing windows were 10, 20, 30 respectively. We conducted numerous parameter tuning experiments and applied various training techniques, such as early stopping (Prechelt, 2002), n-fold validation (Raschka, 2018), and so on. The learning rate for DeBERTaV3$_{\text{large}}$, ALBERTV2$_{\text{xxlarge}}$, and DeBERTaV2$_{\text{xxlarge}}$ were set to 5e-5, 2e-5 and 5e-6 respectively. Due to space constraints, we list representative experimental results in Table 3.

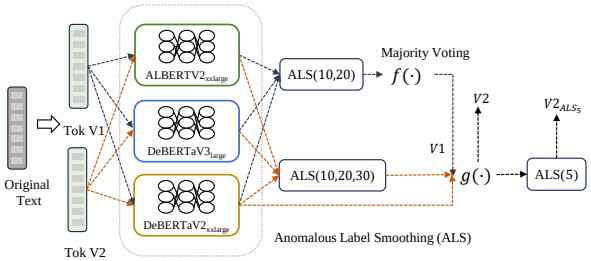

Figure 4: The process of our experiments.

Firstly, we finetuned some base models following the given settings, among which the DeBERTaV3$_{\text{large}}$ model performed best. The reason might be that the DeBERTaV3$_{\text{large}}$ model introduces the absolute word position embeddings, which can contribute to distinguishing between vocabulary usage in human language and machine-generated text. As analyzed in 3.1, differences in text tokenization methods, or inconsistencies of label distributions, would both decrease the F1 scores. Thus, we introduced two types of tokenization methods, which are named $TokV1$ and $TokV2$ to ensure that the length of tokenized input texts does not exceed the maximum length of the pre-trained model. Furthermore, we developed an Anomalous Label Smoothing method, referred to as ALS, to refine the predicted results, aiming to align the final results as closely as possible with the label distribution of the training set. We ultimately acquired several fine-tuned models base on DeBERTaV2$_{\text{large}}$, DeBERTaV2$_{\text{xxlarge}}$ and DeBERTaV3$_{\text{large}}$ models, all of which achieved F1 scores exceeding 0.98 on the development set. Finally, we employed model fusion to integrate the results of the models in Table 3. In that case, we can further improve the model's performance. Table 4 shows the results of the model fusion experiments.

5 Conclusion

In this paper, we mainly introduce the automatically generated papers detection and detail our solution for the DAGPap2024 competition. Firstly, based on the analysis of the data, we adopted two different tokenization ways to ensure that the pre-

dicted results can accurately correspond to the original text set. Then, we introduced an Anomalous Label Smoothing method to ensure that the distribution of predicted results is consistent with the label distribution of the training data set without the loss of information. Finally, we used model fusion to maximize the performance of different models. The above efforts ensured that we achieved a high F1 result in this competition, which was 0.9948 and 0.9944 on F1 score during the development and testing phases, and we achieved second place in the competition.

References

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR.

Lutz Prechelt. 2002. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer.

Sebastian Raschka. 2018. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2024. The science of detecting llm-generated text. *Communications of the ACM*, 67(4):50–59.

Yuchuan Tian, Hanting Chen, Xutao Wang, Zheyuan Bai, Qinghua Zhang, Ruifeng Li, Chao Xu, and Yunhe Wang. 2023. Multiscale positive-unlabeled detection of ai-generated texts. *arXiv preprint arXiv:2305.18149*.

Zheng Zhang, Jie Gao, Ranjodh Singh Dhaliwal, and Toby Jia-Jun Li. 2023. Visar: A human-ai argumentative writing assistant with visual programming and rapid draft prototyping. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–30.