SCUNLP-2 at the NTCIR-18 FigArg-2 Task:

Apply Repeat-Error-Correction Learning on Text Classification

Tong-Ru Wu, Jheng-Long Wu

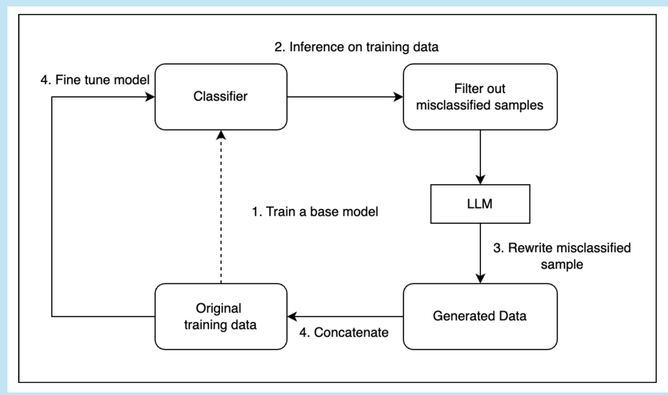
Department of Data Science, Soochow University, Taiwan

Introduction

Large Language Models (LLMs) like GPT-4 have significantly advanced NLP capabilities, but their resource intensity makes deployment challenging for all use cases. Smaller models, such as BERT and its variants, are more cost-effective to run but typically need a large amount of high-quality training data to perform well. Obtaining large-scale human-labeled datasets is often difficult, especially in specialized areas. This situation has increased interest in data-centric methods that use LLMs to improve the data available for training smaller models. Our paper proposes a strategy called **Repeat-Error-Correction Learning** to address this, combining **LLM-based data augmentation** and **error correction**. This framework was developed for the NTCIR-18 FinArg-2 shared task, specifically for detecting Argument Temporal References in Earnings Conference Calls. The goal is to improve the performance of text classification when large labeled datasets are limited.

Method

The Repeat-Error-Correction Learning process starts with training an initial base classification model, such as BERT or DistilBERT, on the available labeled training data without augmentation or rewriting initially. The core of the method is an iterative cycle. The trained base model is used to infer labels on the training dataset, and **misclassified samples** are identified where the predicted label differs from the ground truth. These misclassified samples reveal the model's weaknesses and form a challenging subset. A state-of-the-art Large Language Model, specifically GPT-4o-mini, is then used as a data generator to **rewrite these error examples**. The key principle for rewriting is to preserve the original meaning and label but modify the surface form through rephrasing, restructuring, or stylistic adjustments. These newly synthesized text–label pairs, which transform misclassified samples into new examples while keeping their original labels, are merged with the original training set to create an expanded dataset. The base classifier is then fine-tuned again on this augmented corpus. This cycle of error detection, text rewriting, and model fine-tuning is repeated to progressively improve the model's ability to generalize. Safeguards like quality filtering of generated samples and appropriate regularization are included to maintain class balance and prevent error reinforcement. The dataset used is the FinArg-2 Earnings Conference Call (ECC) dataset, containing fields like claim text, premise texts, year, quarter, and a label indicating the temporal reference (no time reference, long past, or short past). The label distribution shows three classes: label 0 (No time reference) at 50.0%, label 1 (long past) at 29.1%, and label 2 (short past) at 20.9% across the full dataset. The texts contain fewer than 512 tokens, which aligns with BERT's limitations.



Method Overview: Repeat-Error-Correction Learning

1.Initial Training: Train a base classifier (e.g., BERT, DistilBERT) on the original labeled dataset.

2.Error Detection: Use the trained model to infer labels on the training data and identify misclassified samples.

3.Text Rewriting: Utilize GPT-4o-mini to rewrite misclassified samples by rephrasing or restructuring them **without altering the meaning or label**.

4.Augmentation & Fine-tuning: Combine the rewritten samples with the original data to form an expanded training set, then fine-tune the classifier.

5.Iterative Cycle: Repeat the above steps to continuously enhance performance on challenging samples.

Result

The experiments structured around model selection, dataset processing, and the Repeat-Error-Correction Learning process were conducted, with three sets of prediction results submitted for the NTCIR-18 FinArg-2 task. Base models included "bert-base-uncased" and "distilbert-base-uncased". OpenAI's GPT-4o-mini was utilized to generate new expressions from misclassified samples. The training dataset was expanded by combining this newly generated data with the original data. The Repeat-Error-Correction Learning involved iterative cycles of detecting misclassified samples, rewriting them, and fine-tuning the base model. Model performance was evaluated using Micro-F1 and Macro-F1 scores according to FinArg-2 ECC Task guidelines. For base model selection on the validation set, "distilbert-base-uncased" (Base Model 2) achieved a higher Validation Micro-F1 of 75.33% than Base Model 1. After applying repeat-error-correction cycles, Model 4 recorded the highest **Validation Macro-F1 of 74.85%**, while Model 4 (SCUNLP_ECC_1 submission) also achieved the highest **Validation Micro-F1 of 77.33%**. These results highlight a **trade-off between optimizing for overall accuracy (Micro-F1) and ensuring robust performance across all classes (Macro-F1)**. Three final models were submitted based on superior Macro-F1 scores on the validation set. On the test dataset, SCUNLP_ECC_1, despite its high Validation Micro-F1, had a lower Test Macro-F1 (59.54%) and rank (19) compared to SCUNLP_ECC_2 and SCUNLP_ECC_3, which had lower Validation scores but higher Test Macro-F1 scores (63.41% and 63.37% respectively) and better ranks (13 and 14). This outcome **underscores the importance of evaluating both Micro-F1 and Macro-F1 metrics** as they provide complementary perspectives essential for understanding model robustness and generalization. While the approach enhanced validation performance, the **results on the test dataset suggest challenges in generalization**.

Model	Pre-trained model	Learning rate	Epochs	Validation Micro-F1	Validation Macro-F1
Base model 1	bert-base-uncased	2E-5	5	72.67%	69.74%
Base model 2	distilbert-base-uncased	2E-5	5	75.33%	72.36%
Model 3	Base model 2	2E-5	5	75.33%	72.33%
Model 4	Model 3	2E-5	5	77.33%	74.85%

The performance of models on validation dataset

Submission name	Model	Validation Micro-F1	Validation Macro-F1	Test Micro-F1	Test Macro-F1	Rank
SCUNLP-2_ECC_1	Model 4	77.33%	74.85%	63.10%	59.54%	19
SCUNLP-2_ECC_2	Model 3	75.33%	72.33%	66.67%	63.41%	13
SCUNLP-2_ECC_3	Model 2	75.33%	72.36%	66.67%	63.37%	14

Submission results for FinArg-2 ECC Task

Conclusion

The study introduced a **repeat-error-correction learning framework** that enables models to learn from newly generated training samples created by rewriting misclassified instances. The approach successfully used LLMs to enhance performance on the validation dataset by addressing model weaknesses identified through misclassified samples. However, despite improvements on the validation set, the approach **did not generalize as well to the test dataset**. A potential reason suggested is that the model might have forgotten crucial information from earlier training stages when incorporating the newly generated samples. Future work will focus on refining the framework by exploring strategies to **retain essential knowledge** while integrating augmented data and investigating additional methods to **mitigate overfitting** and improve generalization to unseen data.