CDEC midterm report

Jingyuan Shi

February 26, 2025

1 Introduction

Cross-Document Event Coreference (CDEC) Resolution determines whether two event mentions in different sentences refer to the same real-world event. This project fine-tunes ModernBERT and Qwen2.5-0.5B-Instruct for this task.

ModernBERT improves upon BERT with GeGLU layers for better activation, RoPE for enhanced positional encoding, and training on a larger dataset for improved efficiency and performance.

Qwen2.5-0.5B-Instruct is a decoder-only language model designed for instruction-following tasks. I selected the instruct variant because it is explicitly optimized for providing direct answers, making it a strong baseline for binary classification tasks like event coreference. Its ability to understand natural language instructions should help in distinguishing whether two event mentions refer to the same event.

By fine-tuning both models, this project aims to enhance event coreference resolution and compare their performance.

2 Methods

The dataset consists of approximately 227,000 samples in the training set. However, the data distribution is highly imbalanced, with a 10:1 ratio between non-coreferent and coreferent event pairs.

Most of the samples lack location, time, and participant information. Since retaining these features does not impact the classification results, they were dropped, leaving only the original sentences and their respective event trigger words.

2.1 Data Preprocessing

For ModernBERT, data was formatted as follows:

```
[CLS] First sentence: {sentence1}
Event trigger: {trigger1} [SEP]
Second sentence: {sentence2}
Event trigger: {trigger2}
```

The [CLS] token representation was extracted and passed into a classification layer for binary classification, using the HuggingFace default ModernBertForSequenceClassification model.

For Qwen2.5-0.5B-Instruct, a prompt-based approach was used with the following template:

```
Task: Determine if two event words refer to the same event.
First sentence: {sentence1}
Event word in first sentence: {trigger1}
Second sentence: {sentence2}
Event word in second sentence: {trigger2}
Question: Do the event words {trigger1} and {trigger2} refer to the same event?
Answer only with Yes or No.
Answer:
```

Applying Qwen's chat template results in the following structured prompt:

```
<|im_start|>system
You are Qwen, created by Alibaba Cloud. You are a helpful assistant.
<|im_end|>
<|im_start|>user
Task: Determine if two event words refer to the same event.
First sentence: {sentence1}.
Event word in first sentence: {trigger1}
Second sentence: {sentence2}.
Event word in second sentence: {trigger2}
Question: Do the event words {trigger1} and {trigger2} refer to the same event?
Answer only with Yes or No.
Answer:
<|im_end|>
<|im_start|>assistant
Yes
<|im_end|>
```

This template ensures that the instruct model outputs only "Yes" or "No" responses, making it well-suited for classification.

2.2 Fine tuning

To improve event coreference resolution, both models were fine-tuned using different strategies, considering their architectures and computational constraints.

ModernBERT is a relatively lightweight model with 149M parameters, allowing for full fine-tuning of all parameters. The model was optimized for binary classification using cross-entropy loss. The preprocessed input format was passed into the ModernBertForSequenceClassification model from Hugging Face, and the '[CLS]' token representation was used for classification.

Due to its significantly larger size, full fine-tuning of Qwen2.5-0.5B-Instruct was infeasible given time and memory constraints. Instead, LoRA (Low-Rank Adaptation) was applied with a rank of 16 to reduce memory usage while maintaining efficiency.

Furthermore, to balance the highly imbalanced dataset, oversampling was applied to the minority (coreferent) class (oversample 3x). However, due to computational constraints, Qwen was trained on only part of the oversampled dataset.

This approach allowed efficient adaptation of Qwen for binary event classification while leveraging its instruction-tuned capabilities.

3 Results and Analysis

3.1 ModernBERT Results

ModernBERT was trained for 3 epochs with a learning rate of 1×10^{-5}, using 500 warmup steps and a cosine learning rate schedule. The batch size was set to 64, and the best model was selected at 6000 steps, based on the highest F1-score on the evaluation set.

The final performance metrics on the test set are presented in Table **??**.

Class	Precision	Recall	F1-score	Support
Not Coreferent	0.967	0.989	0.978	39111
Coreferent	0.858	0.654	0.742	3842
Accuracy	0.959			
Macro Avg	0.912	0.822	0.860	42953
Weighted Avg	0.957	0.959	0.957	42953

Table 1: ModernBERT classification report on the test set

The model achieves an overall accuracy of 95.9%, with a high F1-score (0.978) for the "Not Coreferent" class. However, due to the extreme class imbalance in the dataset, the performance on the "Coreferent" class is notably lower (F1-score = 0.742).

Observations:

- The precision for the Coreferent class (0.858) is relatively high, indicating that when the model predicts an event pair as coreferent, it is usually correct.
- The recall for the Coreferent class (0.654) is significantly lower, meaning that many actual coreferent pairs are not being identified by the model.
- The F1 score for the Coreferent class is 0.742, which is a fair numebr.
- The imbalanced dataset likely causes the model to favor predicting "Not Coreferent," leading to a high overall accuracy but lower recall on the minority class.

Potential Improvements:

- Exploring loss function modifications like focal loss to penalize incorrect predictions on the minority class more heavily.
- Experimenting with threshold tuning to adjust the decision boundary and improve recall on the Coreferent class.

3.2 Qwen2.5-0.5B-Instruct Results

Qwen2.5-0.5B-Instruct was fine-tuned using LoRA with rank 16. It was trained for 1 epochs with a learning rate of 1e-5, using 500 warmup steps and a cosine learning rate schedule. The batch size was set to 4 with 2 gradient accumulation steps.

I trained 2 models, one over 10% of oversampled dataset, and the other one over 50% of oversampled dataset. The result shows the one with 10% is the better one. It suggests that there is an overfitting problem in this training, a better way to solve this problem is to run over the evaluation set and monitor the performance.

The final evaluation results are presented in Table **??**.

Class	Precision	Recall	F1-score	Support
Not Coreferent	0.949	0.990	0.970	39111
Coreferent	0.828	0.464	0.594	3842
Accuracy	0.943			
Macro Avg	0.890	0.727	0.782	42953
Weighted Avg	0.939	0.943	0.936	42953

Table 2: Qwen2.5-0.5B-Instruct classification report on the test set

Qwen2.5-0.5B-Instruct achieves a high overall accuracy, but its performance on the "Coreferent" class is significantly weaker compared to the "Not Coreferent" class.

Observations:

- The recall for the Coreferent class is only 0.464, meaning that the model struggles to identify actual coreferent events, missing over two-thirds of them.
- The precision for the Coreferent class (0.828) is higher, indicating that when Qwen predicts an event pair as coreferent, it is often correct. However, it rarely makes such predictions.

Potential Improvements:

- Increasing the quality of fine-tuning data and training for additional steps may help the model learn more nuanced event coreference relationships.
- Exploring different prompting strategies, such as providing additional context or fine-grained explanations, to improve Qwen's ability to detect coreferent pairs.
- Using temperature tuning or decision threshold adjustment to encourage more balanced predictions across both classes.

4 Discussion and Comparison

Both ModernBERT and Qwen2.5-0.5B-Instruct were fine-tuned for the cross-document event coreference resolution task, but they exhibit distinct strengths and weaknesses. A comparative analysis of their performance is provided in Table **??**.

Metric	ModernBERT	Qwen2.5-0.5B-Instruct
Accuracy	95.9%	94.3%
F1-score (Coreferent)	0.742	0.594
Precision (Coreferent)	0.858	0.828
Recall (Coreferent)	0.654	0.463

Table 3: Comparison of ModernBERT and Qwen2.5-0.5B-Instruct performance

4.1 Performance on Coreference Detection

ModernBERT outperforms Qwen significantly in identifying coreferent events, with a higher F1-score and recall rate. This suggests that ModernBERT is better at detecting actual event coreference relationships, whereas Qwen tends to miss many coreferent pairs.

One possible reason for this difference is that ModernBERT is encoder based model and trained on a lot of language understanding tasks, meaning it can fully leverage contextual information from both sentences. In contrast, Qwen is a decoder-only model, which might lead it to focus more on typical language patterns rather than deeply understanding event semantics.

4.2 Computational Efficiency

ModernBERT, with only 149M parameters, is much smaller and more efficient to train. Full fine-tuning was feasible, leading to strong performance improvements.

Qwen2.5-0.5B-Instruct, being a decoder-only model with significantly more parameters, required LoRA-based adaptation. Additionally, it was only trained on 10% of the over-sampled dataset, which may have impacted its ability to generalize to coreferent events.

4.3 Suitability for the Task

For this particular task, ModernBERT is the better choice, as it provides higher recall and better balance between the two classes. While Qwen has advantages as a generative model optimized for instruction following, its low recall makes it unreliable for detecting coreferent events.

Future improvements for Qwen could include:

- Increasing fine-tuning data and training iterations.
- Using a larger model.
- Experimenting with better prompting strategies to encourage balanced predictions.
- Applying class-weighted loss or adjusting the decision threshold to favor recall.

Overall, while both models demonstrate strong performance in non-coreferent event detection, ModernBERT provides a more reliable and practical solution for event coreference resolution.

5 Challenges and Solutions

5.1 ModernBERT Training Efficiency

One of the key optimizations for ModernBERT was the use of FlashAttention 2, which significantly speeded up training. This allowed for faster convergence without requiring substantial changes to the training setup.

5.2 Challenges with Qwen2.5-0.5B-Instruct

Most of the challenges encountered were related to fine-tuning Qwen2.5-0.5B-Instruct, primarily due to the large dataset and higher memory requirements.

- Initially, Unsloth was considered for a more efficient SFT (Supervised Fine-Tuning) process. However, attempts to train using this library resulted in frequent out-of-memory (OOM) errors, preventing successful training.
- Despite Qwen's total FLOPs (6.1e16) being half that of ModernBERT's training (1.3e17), fine-tuning Qwen took approximately 3 hours for just half of the dataset, which was unexpectedly slow.
- Increasing the batch size to improve throughput resulted in OOM errors, suggesting that memory constraints were a major bottleneck.

To address these inefficiencies, the following strategies could be explored:

- Implementing packing to merge multiple short sequences into a single input batch, improving GPU utilization.
- Utilizing FlashAttention to reduce memory overhead and accelerate training, as was done successfully with ModernBERT.

These improvements could help reduce training time and enable more effective fine-tuning of large-scale models like Qwen in future experiments.

5.3 Dataset Imbalance and Redundancy

Although the dataset is large, it suffers from two significant issues:

- Severe class imbalance, with a 10:1 ratio between non-coreferent and coreferent events.
- High redundancy, where many sentences are repeated multiple times in the non-coreferent samples.

A key insight from training Qwen on 10% versus 50% of the dataset was that even with a smaller subset, the model was able to learn useful information, suggesting that excessive redundant data harms performance rather than helping it.

To address these inefficiencies, the following strategies could be explored:

- Applying data engineering techniques to remove redundant samples, ensuring that training data is both diverse and informative.
- Doing experiments with other data balancing strategies, such as under-sampling non-coreferent ones, to improve recall on the minority class.

Addressing these challenges would lead to a more efficient and effective fine-tuning process, particularly for large models like Qwen2.5-0.5B-Instruct.

6 Conclusion

In this project, two models—ModernBERT and Qwen2.5-0.5B-Instruct—were fine-tuned for cross-document event coreference resolution. ModernBERT, with full fine-tuning, achieved higher recall and better overall balance, making it the superior choice for this task. Qwen, despite being instruction-tuned, struggled with low recall, likely due to its decoder-only architecture and prompt-based approach.

Key challenges included training inefficiencies, particularly with Qwen, where memory constraints and slow fine-tuning led to significant delays. Future improvements could include packing, FlashAttention, and better prompt engineering to enhance Qwen's performance.

Overall, this study highlights the strengths of encoder based models like ModernBERT in structured classification tasks and the challenges of adapting decoder-only models for event coreference resolution. Future work could explore hybrid approaches that leverage both architectures for improved performance.