Comparative Analysis Report on NLP Project

12213021 陈泽南，12213023 何家阳

Experiment Overview

This report analyzes the OOD text classification performance of BERT-based models for both Chinese and English, and compares it with the zero-shot method. The analysis covers both in-domain and out-of-domain testing results.

1. Experimental Setup

1.1 Data Processing

- **Long Text Segmentation:** For texts exceeding the max_length (512), long documents are divided into multiple shorter segments, ensuring the model can process each chunk effectively without information loss.
- **Domain-Specific Preprocessing:** Across domains such as Reuters (news), WP (novels), and Essay (academic papers), text styles differ greatly. Preprocessing is tailored for each domain to handle these differences.
- **Data Sources:** The dataset covers a broad range including news (reuter), academic essays (essay), and web novels (wp), ensuring a diverse representation of language styles.
- **Noise Removal:** Preprocessing includes removing noise, such as special symbols and irrelevant content, to enhance data purity.
- **Text Cleaning:** Steps include removing redundant whitespaces and empty texts to further improve data quality.
- **Label Balance:** The dataset is inherently balanced. During dataloader construction, samples are randomly shuffled, and human/AI text counts are kept balanced to prevent model bias.
- **Dynamic Data Balancing:** Human and AI data are kept at a 1:1 ratio to avoid training bias.
- **Data Split Strategy:** An 8:1:1 split is used to allocate training, validation, and test data, ensuring accurate model evaluation.

1.2 Model Configuration

- English Model: bert-base-uncased
- Training Parameters:
    - Batch size: 16
    - Epochs: 5
    - Learning rate: 2e-5
    - Weight decay: 0.01
    - Warmup ratio: 0.1

    - Max sequence length: 512
- Chinese Model: hfl/rbtl3
- Training Parameters:
    - Batch size: 16
    - Epochs: 10
    - Learning rate: 2e-5
    - Weight decay: 0.01
    - Warmup ratio: 0.1
    - Max sequence length: 512

Note: For the zero-shot method, the fast-detect-gpt approach is used with gpt2-xl and Qwen2-0.5B models, respectively.

1.3 Evaluation Metrics

- Accuracy
- Precision
- Recall
- F1-score
- AUROC

2. Performance Comparison and Analysis

2.1 Supervised Learning Performance Comparison

2.1.1 English Domain

Source Domain: essay

Domain	Language	Accuracy	Precision	Recall	F1	AUROC
essay	English	0.9824	0.9949	0.9843	0.9896	0.9950
reuter	English	0.9793	0.9813	0.9948	0.9880	0.9931
wp	English	0.7563	0.8370	0.8854	0.8605	0.3251

Source Domain: reuter

Domain	Language	Accuracy	Precision	Recall	F1	AUROC
essay	English	0.8491	0.8497	0.9991	0.9184	0.8235
reuter	English	0.8237	0.8574	0.9527	0.9026	0.6690

wp	English	0.8594	0.9374	0.8941	0.9152	0.8839

Source Domain: wp

Domain	Language	Accuracy	Precision	Recall	F1	AUROC
essay	English	0.8397	0.8993	0.9137	0.9064	0.8114
reuter	English	0.7771	0.9694	0.7640	0.8546	0.8762
wp	English	0.9699	0.9872	0.9772	0.9822	0.9881

2.1.2 Chinese Domain

Source Domain: news

Domain	Language	Accuracy	Precision	Recall	F1	AUROC
xsum_webnovel	Chinese	0.5910	0.7220	0.2960	0.4199	0.5742
xsum-news	Chinese	0.5938	0.6094	0.5201	0.5612	0.6473
xsum-wiki	Chinese	0.6012	0.7011	0.3527	0.4693	0.7045

Source Domain: webnovel

Domain	Language	Accuracy	Precision	Recall	F1	AUROC
xsum_webnovel	Chinese	0.8440	0.9076	0.7660	0.8308	0.9142
xsum-news	Chinese	0.5206	0.5223	0.4699	0.4947	0.5251
xsum-wiki	Chinese	0.4719	0.4832	0.8096	0.6052	0.4163

Source Domain: wiki

Domain	Language	Accuracy	Precision	Recall	F1	AUROC
xsum_webnovel	Chinese	0.5020	0.5018	0.5660	0.5320	0.5080
xsum-news	Chinese	0.5155	0.5136	0.5663	0.5387	0.5431
xsum-wiki	Chinese	0.6573	0.7017	0.5471	0.6148	0.7621

2.2 Zero-Shot Detection Performance Comparison

Domain	Language	#Samples	Accuracy	Precision	Recall	F1	Sampling/Scoring Model
essay	English	6994	0.8450	0.9547	0.8602	0.9050	gpt2-xl/gpt2-xl
reuter	English	7000	0.7971	0.9726	0.7855	0.8691	gpt2-xl/gpt2-xl
wp	English	8000	0.8247	0.9317	0.8630	0.8960	gpt2-xl/gpt2-xl
xsum_webnovel	Chinese	4997	0.8664	0.8607	0.8743	0.8675	Qwen2-

							0.5B/Qwen2-0.5B
xsum-news	Chinese	4530	0.8693	0.8974	0.8340	0.8645	Qwen2-0.5B/Qwen2-0.5B
xsum-wiki	Chinese	3607	0.7040	0.9121	0.4516	0.6041	Qwen2-0.5B/Qwen2-0.5B

3. Main Findings

3.1 Supervised Learning Models

3.1.1 In-Domain Performance Differences

- The English model significantly outperforms the Chinese model overall.
- The performance gap is largest in the news domain, with English model accuracy 38-47 percentage points higher.
- The gap is smaller in the online novel domain, but the English model still leads by 7-21 percentage points.

3.1.2 Cross-Domain Generalization

- The English model shows much better cross-domain generalization.
- Its out-of-domain performance drops only slightly, remaining between 75-90%.
- The Chinese model's cross-domain performance drops sharply, mostly to the 50-60% range.

3.1.3 Model Stability

- The English model's metrics are more balanced, especially the difference between Precision and Recall.
- The Chinese model's metrics fluctuate more, especially in cross-domain scenarios.

3.1.4 AUROC Performance

- The English model's AUROC is generally high (80-99%), indicating more reliable classification.
- The Chinese model's AUROC is relatively low (52-91%), especially cross-domain.

3.2 Zero-Shot Detection Models

3.2.1 Overall Performance Comparison

- English domain (gpt2-xl):
    - All three domains show balanced metrics; accuracy is between 0.79~0.84, precision is extremely high (0.93~0.97), recall is also high (0.75~0.84), and F1 is above 0.869.

    - The model is stable, with precision clearly higher than recall, suggesting it is conservative in AI text identification (low false positives, but more false negatives), possibly due to more AI data in the English datasets.

- Chinese domain (Qwen2-0.5B):
    - Both xsum_webnovel and xsum-news domains outperform English, with accuracy around 0.87 and high precision, recall, and F1, reflecting strong discrimination.
    - The xsum-wiki domain drops sharply, with recall only 0.45 and F1 only 0.60, indicating the fast-detect-gpt approach performs poorly in this domain, with more missed AI text.

3.2.2 Model Capability Analysis

- gpt2-xl (English):
    - Performs best in essay and wp domains, slightly lower in reuter, with room to improve recall.
    - Precision is higher than recall, showing more robust discrimination of negative (human) texts, but some AI texts are not detected.
- Qwen2-0.5B (Chinese):
    - Performs very well in webnovel and news, with balanced recall and precision, reflecting strong discrimination for both classes.
    - Recall is much lower in the wiki domain, indicating higher miss rates for AI text and poor generalization.

3.2.3 Comparison of Chinese and English Models

- F1 scores for the English model are generally higher than for the Chinese model.
- The English gpt2-xl model is more stable, with extremely high precision and slightly lower recall, resulting in strong robustness.

3.2.4 Metric Comparison

- Precision: All models have high precision, especially in English, indicating strict discrimination of AI text.
- Recall: Chinese wiki domain recall is much lower than other domains, highlighting the need for better generalization.
- F1: As a comprehensive measure, English models generally outperform Chinese models.

4. Future Work

1. Data Quality Improvement:
    - Increase quality and quantity of news domain training data
    - Ensure diversity and representativeness of training data
    - Optimize data preprocessing pipeline

2. Model Optimization:
    - Improve training strategies for better cross-domain generalization
    - Optimize model architectures for better handling of Chinese features
    - Consider implementing domain adaptation layers

5. Conclusion

This study shows that while BERT models perform well on both Chinese and English text classification tasks in supervised learning, there remain significant differences in cross-domain generalization. The English model demonstrates more robust performance and generalization capability, whereas the Chinese model's performance drops sharply out-of-domain. Zero-shot detection results are also more stable and reliable for English. Improving Chinese model generalization and stability—especially in more challenging domains—remains an important direction for future work.