CS310 - Natural Language Processing
Assignment 1 Report

Name: Tan Hao Yang SID: 12212027

Introduction

This assignment focuses on training a neural network-based text classification model to detect humor in Chinese text using the Chinese Language Humor Detection (CLEVA) dataset, available at https://github.com/LaVi-Lab/CLEVA.

Data Processing

The data processing stage effectively prepares the Chinese humor detection dataset for training by leveraging PyTorch's DataLoader. It begins with loading the JSON Lines dataset, followed by implementing two tokenizers: a basic one that extracts individual Chinese characters while discarding non-Chinese tokens, and an improved version that captures special patterns like digits, English words, and punctuations. The vocabulary is built to map tokens to numerical IDs, ensuring proper text representation. The batching process uses a custom collate function, efficiently handling variable-length sequences and labels, making the data ready for neural network training.

Build the Model

The model-building phase successfully constructs a neural network for Chinese humor detection using the torch.nn module, adhering to the specified requirements. The implementation leverages the bag-of-words approach with nn.EmbeddingBag to efficiently process variable-length text sequences, enhancing computational efficiency. The fully-connected component includes at least two hidden layers, implemented conveniently with torch.nn.Sequential, ensuring sufficient depth for feature learning. The model is initialized with appropriate parameters, tested with a sample batch, and paired with a training setup featuring a loss function, optimizer, and learning rate scheduler, setting a solid foundation for effective training.

Train and Evaluate

The training and evaluation phase successfully implemented a robust pipeline for the Chinese humor detection model, utilizing a sufficient number of epochs to achieve optimal performance. The process involved training the model on a split dataset, evaluating on validation data, and reporting comprehensive metrics on the test set. The basic tokenizer yielded a test accuracy, precision, recall, and F1-score, reflecting a balanced performance with room for improvement in distinguishing nuanced humor. In contrast, the improved tokenizer, which captures special patterns like digits and punctuations, showed slightly lower metrics, possibly due to increased noise from additional tokens, highlighting the trade-off between complexity and precision in tokenization strategies.

	Accuracy	Precision	Recall	F1-Score
Basic Tokenizer	0.737	0.684	0.737	0.671
Improved Tokenizer	0.722	0.663	0.722	0.669

Explore Word Segmentation

The exploration of word segmentation as an advanced tokenizer significantly enhanced the data processing phase for the Chinese humor detection model. By installing the jieba package (https://github.com/fxsjy/jieba) and utilizing its cut_for_search method instead of the standard cut, the segmentation process was optimized for more precise and context-aware word grouping, which is particularly effective for search-oriented tasks and can better capture semantic units in humor detection. The segmented data, with a reduced vocabulary size due to word-level grouping, was processed through the train and evaluate pipeline. Compared to the original character-based approach, the segmented data showed improved performance, likely due to better representation of meaningful phrases, though the F1-score suggests a potential trade-off in balancing precision and recall, possibly reflecting challenges in aligning segmented units with humor-specific features.

	Accuracy	Precision	Recall	F1-Score
Basic Tokenizer	0.737	0.684	0.737	0.671
Improved Tokenizer	0.722	0.663	0.722	0.669
Jieba Tokenizer	0.753	0.776	0.753	0.664

Conclusion

This project successfully developed and evaluated a neural network-based text classification model for Chinese humor detection, utilizing the CLEVA dataset and PyTorch framework. The data processing phase effectively implemented basic and improved tokenizers, meeting the requirements for character-level and pattern-recognized tokenization, while the exploration of jieba-based word segmentation with `cut_for_search` provided valuable insights into advanced text representation. The model, built with nn.EmbeddingBag and multiple hidden layers, demonstrated robust training and evaluation capabilities, with performance metrics reflecting the impact of different tokenization strategies. The results suggest that word segmentation enhances accuracy by capturing semantic units, though further tuning could optimize F1-scores. Future work could explore transformer-based models, incorporate cultural context features, or expand the dataset to refine humor detection, offering exciting opportunities to advance cross-lingual NLP research.