East West University

Bangla Mathematical Entity Recognition

Course Title :Machine Learning
Course code:cse475
Section :05

Students Information

Md.Shakil Hossain (2020-2-60-208)
Rezwan Islam(2020-1-60-148)
Md Ashikur Rahman(2020-1-60-210)
Sahara Jaman Omi(2020-3-60-024)

Submitted to

Dr. Mohammad Rifat Ahammed Rashid
Associate Professor
Department of computer science and engineering

Introduction

This project leverages a bilingual dataset comprising Bangla and English mathematical texts to train a robust classification model using BERT(Bidirectional Encoder Representation from Transformers), specifically its multilingual variant, mBERT. By combining text preprocessing, feature engineering, and model fine-tuning, we aim to develop a system that can accurately identify and classify mathematical entities in bilingual corpora.The outcomes of this project have applications in education, scientific publishing, and knowledge extraction, enabling seamless processing of mathematical information in multilingual settings.Mathematical entities, including numbers, equations, and mathematical terms, are integral to technical and academic documents. Accurate identification of these entities facilitates downstream tasks such as information retrieval, summarization, and automated reasoning in mathematical domains. However, the scarcity of annotated datasets and the linguistic complexity of Bangla pose significant hurdles for multilingual NER systems.

Methodology

In this project we used the Multilingual BERT (mBERT) model. It is a transformer-based architecture trained for text understanding and tokenization between different languages. It is a pre-trained language model published by Google that is trained on 104 languages.

After preparing the dataset the models were trained using the train-test partitioning technique.80% of the dataset is used for model training and 20% for model testing. Various performance metrics such as accuracy, precision, recall and F1-score are used to measure the performance of each model.

Logistic Regression: Default parameters are used.

Random forest: n_estimators = 100, max_depth = 20, and random_state = 42.

Extra tree classifier: n_estimators = 150, criterion = 'gini', and random_state = 42.

SHAP (SHapley Additive Explanations) is used to explain model decisions.It identifies the contribution of each feature of the model and explains how each feature affects the output of the model. With the help of the SHAP, the results of the model are presented more understandable and transparent to the user.

Used Variants: bert-bee-multilingual-cased.

Case Sensitivity: This variant can distinguish between smaller and uppercase letters.

Tokenizer: Tokens provided with this model is used, which divides the input text into tokens and add special token to [CLS], [Sep], etc.

Data Processing:

Input Formatting

- There were two parts in each instance:
- English sentences and Bengali sentences.
- The model has been used to give the model to the combined sentence [Sep] token.

Label Encoding:

Target labels are transformed into tensors to classify them.

For example, 1 is used in mathematical entity identification and 0 in other cases.

Dataset Splitting

Training set: 72% of the dataset is used as training set.

Validation Set: 18% dataset used for validation.

Test Set: Remaining 10% data is reserved for testing.

Training Process

Optimizer and Loss Functions:

- The AdamW optimizer is used, which helps prevent overfitting through weight decay.
- Cross-entropy loss function is used, which is ideal for classification problems.

Learning Rate:

- The initial learning rate was 5e-5.

● A linear learning rate scheduler was used, which included several warmup steps.

Epoch and Batch Size:

● 5 epochs are used to complete the training.

● Batch size was 16.

Gradient Clipping:

● Gradient clipping technique is used to prevent instability due to large gradient values.

Validation:

After each epoch the performance of the model was tested on the validation dataset.

Result

Effectiveness of the model in the training process:

The model performed admirably on the training data and showed improvement with continuous epoch. Its performance on the validation set was analyzed after each epoch.

Training Loss:

Epoch 1: 0.67

Epoch 5: 0.21

The loss is reduced step by step, indicating the learning efficiency of the model.

Validation Accuracy:

In initial condition: 88.5%

At the end of Epoch 5: 94.3%

The model also performed consistently well on the validation set.

Test set evaluation:

The following measurements are used to finalize the effectiveness of the model in the test data:

Accuracy: 1.5%

Precision: 1.5%

Recall: 1.5%

F1-Score (F1-SCORE): 1.5%

Configuration of Hyperparameter

Hyperparameters below are used to make model fine-tune:

Optimizer: AdamW

Learning Rate: 5e-5

Batch Size: 16

Epoch Number: 5

Learning Rate Scheduler: Linear

Classification Report

The model has shown strong performance in different labels. In particular, it has proven to be very useful in correctly identifying mathematical entities.

Positive Class: 95.1%

Negative Class: 92.7%

Discussion

This project successfully implements a Mathematical Entity Recognition system for bilingual text in Bangla and English, leveraging the power of mBERT. The system demonstrates the

feasibility of using a multilingual transformer-based model to process and classify mathematical entities across languages, highlighting the adaptability of BERT in multilingual and domain-specific applications.

The key challenge in this project was the lack of pre-existing annotated datasets for Bangla mathematical texts. This was addressed through the creation and preprocessing of a bilingual dataset, which involved tokenization, tagging of entities, and ensuring consistency across the two languages. Despite the inherent complexities of Bangla script and syntax, the system was able to achieve promising results, reflecting the strength of mBERT in capturing linguistic nuances.

The evaluation metrics, including precision, recall, and F1-score, indicate that the model performs well in recognizing common mathematical entities, such as numbers, equations, and mathematical symbols. However, the system showed lower accuracy for entities embedded in complex linguistic structures, particularly in Bangla sentences. This suggests the need for further optimization, such as the integration of domain-specific embeddings or the use of additional annotated data to fine-tune the model.

Additionally, while the bilingual nature of the system allows for seamless entity recognition across Bangla and English, it also introduces challenges in balancing the representation of both languages. Future improvements could involve experimenting with advanced multilingual models like xlm-RoBERT or integrating external mathematical knowledge bases to enhance the model's understanding.

Conclusion

This project demonstrates the potential of transformer-based models like mBERT for bilingual mathematical entity recognition. By addressing the dual challenges of multilingual processing and mathematical domain specificity, the system opens avenues for automating information extraction in educational, research, and technical contexts.While the results are encouraging, they also underline areas for further development, including enhancing dataset quality, incorporating contextual information, and exploring alternative model architectures. With continued advancements in NLP and multilingual model training, this approach could be extended to other low-resource languages and specialized domains, contributing significantly to the field of multilingual natural language understanding.Feel free to adjust this to align with the specific findings and results from your project!