Project Title: NLP Named Entity Recognition (NER) Using BERT

Week 2: NER using BERT

- By Hareetima Sonkar

Project Overview

This week, I focused on laying the foundation for the project titled **"NLP Named Entity Recognition (NER) Using BERT (Bidirectional Encoder Representation from Transformers)"**. The goal was to familiarize myself with the basics of NER and BERT, set up the required environment, and explore sample datasets.

Objectives for Week 2

1. Explore the structure of a labeled NER dataset (CoNLL-2003).
2. Implement data cleaning and formatting to structure the dataset properly.
3. Tokenize text using BERT's tokenizer and align labels with subword tokens.
4. Save the processed dataset in CSV and JSON formats for training.

Tasks Completed:

1. Dataset Exploration

- Loaded the CoNLL-2003 dataset (eng.train, eng.testa, eng.testb).
- Extracted sentences and their corresponding Named Entity Recognition (NER) labels.
- Identified that the dataset uses the BIO tagging scheme (B-, I-, O).

Key Observations:

- The dataset contains tokens with corresponding entity labels like PER (Person), LOC (Location), ORG (Organization), MISC (Miscellaneous).
- Sentences are structured with tokens and labels in separate columns.
- The special -DOCSTART- marker is present to indicate document boundaries.

2. Data Cleaning and Formatting

- Converted dataset into a structured Pandas DataFrame for easy manipulation.
- Ensured all sentences and labels were correctly extracted and formatted.

Challenges & Solutions:

- Needed to ensure labels remain aligned with their corresponding words after processing.
- Used list structures to maintain sentence integrity before tokenization.

3. Tokenization Using BERT's Tokenizer

- Used WordPiece Tokenization with bert-base-cased.
- Mapped original words to subword tokens, ensuring label alignment.
- Added special tokens:
  - [CLS] → Marks the beginning of input.
  - [SEP] → Separates sentences.
  - [PAD] → Used for padding sequences to a uniform length.

Challenges & Solutions:

- BERT's tokenizer splits words into subwords, requiring careful alignment of labels.
- Used Word IDs mapping to correctly assign entity labels to subword tokens.
- Special tokens [CLS] and [SEP] were ignored (-100) to prevent label misalignment.

4. Saving the Processed Dataset

- Stored cleaned and tokenized data in both CSV and JSON formats:
  - CSV: Easy for visualization.
  - JSON: Structured format for direct model training.
- Verified the saved dataset by reloading it and ensuring data integrity.

Deliverables

✓ Explored and structured dataset

✓ Cleaned and formatted raw text data

✓ Tokenized text using BERT's tokenizer

✓ Aligned labels with subword tokens

✓ Saved dataset in CSV & JSON formats for training

Challenges Faced & Solutions

Challenge	Solution
BERT tokenizer splits words into subwords, causing misalignment of labels.	Used Word ID mapping to correctly assign labels to subword tokens.
Padding affects label alignment.	Assigned -100 to [CLS], [SEP], and [PAD] tokens to ignore them in training.
Ensuring dataset is saved in a structured format.	Saved in both CSV & JSON and verified integrity by reloading files.

Key Points:

- NER datasets require precise label alignment, especially when tokenizing text using subword tokenization.
- Padding and special tokens must be carefully handled to avoid errors in model training.
- BERT's tokenizer splits words into meaningful subwords, making it more effective than traditional word-level tokenization.

Completing Week 2 tasks has prepared a fully processed dataset, ready for fine-tuning BERT for NER tasks. The structured and tokenized dataset ensures that BERT can correctly learn entity relationships, improving model performance in the next phase.