Doccano JSONL to BIO Rules

The provided script, named ‘jsonl2bio’, is designed to process JSONL (JSON Lines) files containing annotations of text data with associated entity labels. The script extracts the labeled entities, converts them into a BIO (Beginning, Inside, Outside) tagging format, and then stores the resulting data in a CSV (Comma-Separated Values) file. This documentation provides a detailed explanation of the script’s functionality, components, and usage.

Multiple Labels for a Single Word

When processing the annotations and generating BIO labels, the script considers the order of labels as they appear in the JSONL file. If multiple labels exist for a word, only the label that appears later in the JSONL file is written to the resulting CSV file (`ner.csv`). This ensures that only one label is assigned to a word, even if multiple labels were annotated for it. Labels that are annotated multiple times with the same label are not affected by this issue and are handled correctly.

Partial Annotations

In cases of partial labeling, where only a part of a word is labeled, the script assigns a label to each partial segment of the word. This behavior is a result of the script’s design, which splits the word based on the provided entity offsets. Consequently, for a single word that is partially annotated, it may have both ‘O’ and ‘B-’ labels (or vice versa), representing the partial annotations.

Functions

`map_labels(label)`

This function maps multi-word labels to single labels using a predefined label mapping dictionary.

`extract_entities(text, entities)`

This function takes a text string and a list of entities as input. It extracts entities' start and end offsets from the input JSON data and maps the labels. It then generates BIO labels for the words in the text based on the entity information. It also assigns ‘Proof of Concept’, ‘Steps to Reproduce’ and ‘Impact’ O labels due to the length of these particular labels.

`convert_to_csv(jsonl_file, csv_file, invalid_json_file, original_csv_file))`

This function is the main component of the script. It reads a JSONL file containing annotations, processes the data, and outputs the extracted information in CSV format. The process involves the following steps:

1. Parsing JSONL Data:
    - Reads the input JSONL file line by line.
    - Attempts to parse each line as JSON data.
    - Stores valid annotations in the annotations list and invalid lines in the invalid_json_lines list.
2. Saving Invalid JSON Lines:
    - Writes the lines with invalid JSON format to a separate file specified by invalid_json_file.
3. Processing Annotations:
    - Generates BIO labels for the text using the extract_entities function.
    - Combines words and labels into a single string for each annotation.
    - Writes the combined data (text and labels) to the specified CSV file.

`create_multi_label_csv(jsonl_file, csv_file)`

Generates a CSV file containing pairs of different overlapping entity labels for further review.

Undersampling

The second script, named `undersampling_ner_dataset`, is an extension of the first script and introduces additional functionalities to create a balanced CSV file by undersampling O-labels while maintaining a similar distribution of B-labels. It also includes a function to count B-labels in the original CSV file.

Strategy

1. Count the total number of B-labels in the original CSV file.
2. Collect the indices of O-labels for each annotation.
3. Randomly sample O-label indices equal to the number of B-labels.
4. Create a dictionary of O-label indices to keep for each annotation.
5. Generate the final CSV file by forming text from B-label indices and selected O-label indices, and insert them into the CSV file.