Report on NER Model Training and Evaluation Process Using spaCy

1. Data Preparation
Data Loading: The code begins by defining a function load_data_from_csv to load data from a CSV file into a pandas DataFrame. This function takes the file path as an input and returns the loaded DataFrame.

Entity Extraction from Tags: The function add_entities_from_tags parses a specific column in the DataFrame to extract entity information (start, end, label) and adds it to a new 'entities' column.

Data Splitting: The function split_data_into_sets splits the data into training, testing, and validation sets. It uses train_test_split from sklearn.model_selection to achieve this, with configurable test and validation sizes.

2. Data Conversion to spaCy Format
Processing for spaCy: The process_data function takes the prepared data and converts it into a format compatible with spaCy. It uses DocBin to create binary spaCy training data.

Normalization and Adjustment of Spans: Functions like normalize_label and adjust_span are used to standardize entity labels and adjust span indices to align with token boundaries.

3. Model Training
Configuration Initialization: The spaCy command spacy init fill-config is used to initialize a configuration file for training.

Model Training: The spaCy train command is used to train the NER models on the prepared data. The process includes specifying paths for training and validation data.

Saving the Tok2Vec Layer: The save_tok2vec_layer function saves the Tok2Vec layer of the trained model, which is crucial for subsequent fine-tuning.

4. Model Fine-Tuning
Combining Data for Fine-Tuning: The code combines new data with a sample of previous training data to create a dataset for fine-tuning.

Adjustments in Config File: Modifications are made to the config file to use the saved Tok2Vec layer (init_tok2vec) from the previously trained model.

Further Training: The model is further trained (fine-tuned) on the new combined dataset using the updated configuration.

5. Evaluation

Test Data Preparation: Similar processing steps are applied to the test data for each training iteration (T1, T2, T3).

Evaluation Function: The evaluate_model function loads the trained model and evaluates it on the test data. It calculates precision, recall, F1 scores for each category, global scores, and weighted average F1 score.

Knowledge retention The model is not prone to catastrophic forgetting.

Forward transfer The model learns a new task while reusing knowledge acquired from previous tasks.
Backward transfer The model achieves improved performance on previous tasks after learning a new task.

Fixed model capacity Memory size is constant regardless of the number of tasks and the length of a data stream.

The above parameters can be judged by evaluating tasks across different test sets:
Here are the results for the same , indicating that the model has

T1 ON G1
F1 Scores for Each Category:
Treatment: F1 Score = 0.9886760323541932
Chronic Disease: F1 Score = 0.9791275307868921
Cancer: F1 Score = 0.9712292938099389
Allergy: F1 Score = 0.9614035087719298
Other: F1 Score = 0

T2 ON G2
F1 Scores for Each Category:
Treatment: F1 Score = 0.9838623108025906
Chronic Disease: F1 Score = 0.9861878453038674
Cancer: F1 Score = 0.9783046828689982
Allergy: F1 Score = 0.9450392576730907
Other: F1 Score = 0

T3 ON G3
F1 Scores for Each Category:
Treatment: F1 Score = 0.9827564631979377
Chronic Disease: F1 Score = 0.9855567373901359
Cancer: F1 Score = 0.9793418796439946
Allergy: F1 Score = 0.9580674567000911
Other: F1 Score = 0

T2 ON G1
F1 Scores for Each Category:
Treatment: F1 Score = 0.9879559532002754
Chronic Disease: F1 Score = 0.9887751083103584
Cancer: F1 Score = 0.9824665676077267
Allergy: F1 Score = 0.945054945054945
Other: F1 Score = 0

T3 ON G2
F1 Scores for Each Category:
Treatment: F1 Score = 0.985743816380355
Chronic Disease: F1 Score = 0.9878285691630654
Cancer: F1 Score = 0.9800824175824175
Allergy: F1 Score = 0.9666666666666667
Other: F1 Score = 0

T3 ON G1
F1 Scores for Each Category:
Treatment: F1 Score = 0.9849237622066129
Chronic Disease: F1 Score = 0.9879470460383324
Cancer: F1 Score = 0.9799081515499425
Allergy: F1 Score = 0.9535714285714287
Other: F1 Score = 0

From these scores it can be observed that the fine-tuned models through continual learning are either similar or better across different categories. Hence, we can see that the model has knowledge retention ,Backward transfer and fixed model size(1.1gb)