Named Entity Recognition using MultiCoNER II

Ananya Mantravadi, Sai Yaaminie Ganda, and Siddharth Saini, *IIIT Raichur*

Abstract

Named Entity Recognition aims to detect various entities such as person name, company name, location, etc. which are present in a piece of text. In this assignment, we were required to study and develop solutions for the named entity recognition problem defined in the SemEval 2023 MultiCoNER II Task. MultiCoNER II is a large multilingual dataset (12 languages) for fine-grained Named Entity Recognition. We present our experiments, observations, and evaluation of two methods. We chose the English language in the mono-lingual learning setup. The two methods we chose are Chain Conditional Random Fields, a statistical model (non-neural network-based sequence labeling model), and BiLSTM, a recurrent neural network.

I. Dataset & Analysis

MultiCoNER II - Multilingual Complex Named Entity Recognition is a multilingual dataset with data collected from 12 languages for the NER task [1]. The dataset represents challenges of complex named entities like imperative clauses, complex syntactic structures, and long-tail entity distributions. The sentences are taken from various domains, including news, social media, and Wikipedia, to provide a diverse range of text types and styles. The dataset consists of 6 broad range named entities - medical, location, creative works, group, product, and person; each further divided into finer groups. It is intended to be used for developing and evaluating machine learning models for NER in a multilingual and multi-domain context. For the purpose of this project, we intend to work in a mono-lingual setting of the English language. Data for the same can be found at https://registry.opendata.aws/multiconer/

Following is the tagset of MultiCoNER that belong to the 6 major classes of named entities:

- Location (LOC): Facility, OtherLOC, HumanSettlement, Station
- Creative Work (CW): VisualWork, MusicalWork, WrittenWork, ArtWork, Software
- Group (GRP): MusicalGRP, PublicCORP, PrivateCORP, AerospaceManufacturer, SportsGRP, CarManufacturer, ORG
- Person (PER): Scientist, Artist, Athlete, Politician, Cleric, SportsManager, OtherPER
- Product (PROD): Clothing, Vehicle, Food, Drink, OtherPROD
- Medical (MED): Medication/Vaccine, MedicalProcedure, AnatomicalStructure, Symptom, Disease

Figure 1 shows a histogram of the different sentence lengths found in the dataset. There are varying lengths in the sentences. While most of the sentences contain about 10 to 20 words, the maximum length observed is 68 words.

Figures 2 and 3 depict the distribution of NER tags in the training and validation datasets. Most of the tags in both datasets are 'O' as expected. The tag 'Artist' of the class PER has the highest frequency of 7544 in the training dataset and 429 in the validation dataset. The tags 'Drink' and 'Clothing' of the class PROD appear the least number of times in both datasets.

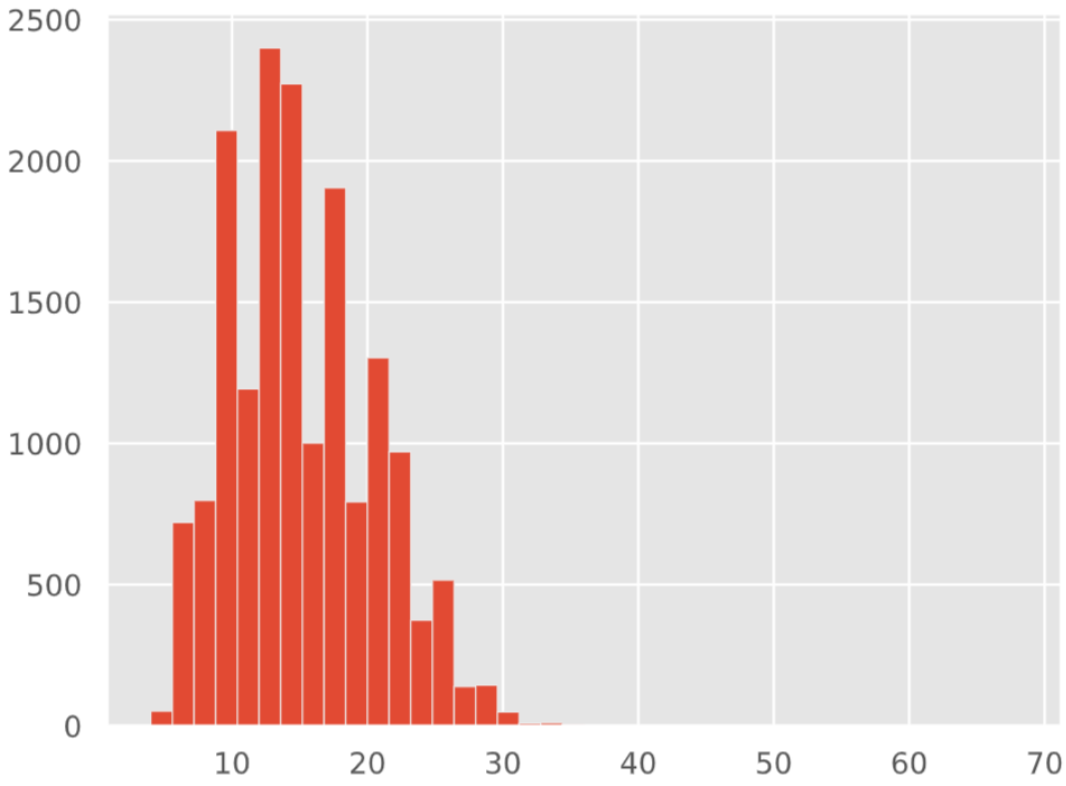


Fig. 1: Length of sentences

Tag	Frequency
O	198853
I-Artist	3837
B-Artist	3707
B-HumanSettlement	2616
I-VisualWork	2273
...	...
I-Symptom	106
I-CarManufacturer	102
I-Medication/Vaccine	89
I-Clothing	87
I-Drink	84

Fig. 2: NER tags in training dataset

Tag	Frequency
O	10570
I-Artist	217
B-Artist	212
I-VisualWork	122
I-OtherPER	122
...	...
I-Medication/Vaccine	6
I-AnatomicalStructure	4
I-Symptom	3
I-Drink	2
I-Clothing	2

Fig. 3: NER tags in validation dataset

II. Conditional Random Fields

Conditional Random Fields is a discriminative log-linear model used for sequence labeling tasks by using conditional probability $P(Y|X)$ of the output sequence given the input sequence, rather than modeling the joint probability of the input and output sequences as generative models like Hidden Markov Models do. It assigns a probability to an output sequence of labels Y, out of all possible sequences Y, given the input sequence data X. CRF is flexible compared to other generative models since they are able to look into the context of a word. Feature functions are a key component of CRF that one can decide to be able to capture arbitrary features, capitalization, or morphology. For example, linear chain CRF is a special case of CRF where the feature functions that we select are restricted to depend on only the current and previous labels, rather than arbitrary labels throughout the sentence. If there are K features from 1 to k, the general form of a feature function for an input of size n is $F_k(X, Y) = \sum_{i=1}^{n} f_k(y_{i-1}, y_i, X, i)$. Here, y_{i-1} and y_i are the previous and current tokens, X is the input sequence, and i is where we are in the sequence. The weights assigned to these features are learned during training so that more weightage is given to more important features. In the following sections, we briefly describe our methodology and different feature engineering iterations for improvising CRF.

A. Basic Features

In the first iteration, the basic feature functions include the following - bias, if the word is at sentence beginning or end, if the word is alphanumeric, numeric, or digit, uppercase, lowercase, word length, hash length, and byte length of the word. While training, the CRF parameters c1 and c2 which are the coefficients for L_1 and L_2 regularization have been set at 0.1 each. The F1-score achieved was 83.75% (Figure 4). We also observed the most likely/unlikely transitions and state features with large positive/weights. Among the likely transitions were B-OtherPER $\rightarrow$ I-OtherPER, B-VisualWork $\rightarrow$ I-VisualWork, B-Athlete $\rightarrow$ I-Athlete which shows how the model is able to correctly find the relationships between the tags since the B tags are followed by I tags in the same class which we generally expect. A few of the top state features were bias, lowercase, and hash length for the label 'O'.

B. Basic Features + Neighbours

In the next iteration, we included features of the words around the target word - one previous word and one after. This resulted in an F1-score of 85.31%, which is an increase of 1.56% after including neighbours' features (Figure 5).

C. Basic Features + POS Tags

Part Of Speech (POS) tagging can be used in NER since it can help in identifying named entities. The intuition here is that parts of speech like proper nouns are more likely to be named entities rather than verbs, interjections, and conjunctions. With the idea that POS tagging can be a useful feature, we assigned POS tags to tokens using Natural Language Toolkit (NLTK). In this version, we observed an F1-score of 84.21% (Figure 6). This shows that compared to II-A where we did not include the POS Tags, the F1-score increases from 83.75% to 84.21%. But compared to II-B where we included the neighbours, the F1-score decreases from 85.31% to 84.21%. Therefore we observe that we should include both POS tags and Neighbours in order to get the best results.

D. Basic Features + POS Tags + Neighbours

After combining features including those of neighbours' along with POS tags, the F1-score improved to 85.61% (Figure 7) which is better than all the previous iterations.

	precision	recall	f1-score	support
B-AerospaceManufacturer	0.83	1.00	0.91	10
B-AnatomicalStructure	0.73	0.47	0.57	17
B-ArtWork	0.33	0.08	0.12	13
B-Artist	0.52	0.42	0.47	212
B-Athlete	0.32	0.35	0.34	79
B-CarManufacturer	0.67	0.62	0.64	13
B-Cleric	0.67	0.27	0.38	15
B-Clothing	0.33	0.20	0.25	10
B-Disease	0.67	0.44	0.53	18
B-Drink	1.00	0.64	0.78	11
B-Facility	0.69	0.42	0.52	52
B-Food	0.67	0.11	0.18	19
B-HumanSettlement	0.70	0.57	0.63	109
B-MedicalProcedure	0.57	0.31	0.40	13
B-Medication/Vaccine	0.40	0.11	0.17	18
B-MusicalGRP	0.59	0.35	0.44	37
B-MusicalWork	0.28	0.11	0.16	61
B-ORG	0.65	0.51	0.57	78
B-OtherLOC	0.67	0.25	0.36	16
B-OtherPER	0.42	0.27	0.33	91
B-OtherPROD	0.80	0.24	0.38	49
B-Politician	0.45	0.26	0.33	53
B-PrivateCorp	0.75	0.55	0.63	11
B-PublicCorp	0.78	0.25	0.38	28
B-Scientist	0.11	0.07	0.08	15
B-Software	0.64	0.27	0.38	26
B-SportsGRP	0.79	0.54	0.64	41
B-SportsManager	0.50	0.06	0.11	16
B-Station	0.50	0.30	0.37	20
B-Symptom	0.67	0.60	0.63	10
B-Vehicle	1.00	0.20	0.33	20
B-VisualWork	0.08	0.03	0.05	61
B-WrittenWork	0.70	0.35	0.47	54
I-AerospaceManufacturer	0.50	1.00	0.67	7
I-AnatomicalStructure	0.00	0.00	0.00	4
I-ArtWork	0.38	0.14	0.20	22
I-Artist	0.52	0.43	0.47	217
I-Athlete	0.32	0.36	0.34	78
I-CarManufacturer	0.67	0.33	0.44	6
I-Cleric	0.57	0.25	0.35	16
I-Clothing	0.33	0.50	0.40	2
I-Disease	0.69	0.41	0.51	22
I-Drink	1.00	1.00	1.00	2
I-Facility	0.61	0.58	0.59	64
I-Food	0.67	0.22	0.33	9
I-HumanSettlement	0.84	0.68	0.75	72
I-MedicalProcedure	0.75	0.67	0.71	9
I-Medication/Vaccine	0.00	0.00	0.00	6
I-MusicalGRP	0.40	0.41	0.40	46
I-MusicalWork	0.43	0.37	0.40	108
I-ORG	0.51	0.59	0.55	108
I-OtherLOC	0.57	0.25	0.35	32
I-OtherPER	0.46	0.30	0.37	122
I-OtherPROD	0.69	0.26	0.38	42
I-Politician	0.31	0.30	0.31	60
I-PrivateCorp	0.67	0.40	0.50	10
I-PublicCorp	0.56	0.40	0.47	25
I-Scientist	0.08	0.06	0.07	16
I-Software	0.04	0.04	0.04	27
I-SportsGRP	0.69	0.63	0.66	46
I-SportsManager	0.50	0.06	0.11	17
I-Station	0.93	0.52	0.67	27
I-Symptom	0.50	1.00	0.67	3
I-Vehicle	0.36	0.20	0.26	20
I-VisualWork	0.22	0.11	0.15	122
I-WrittenWork	0.79	0.66	0.72	90
O	0.91	0.98	0.95	10570
accuracy			0.85	13323
macro avg	0.55	0.38	0.42	13323
weighted avg	0.83	0.85	0.84	13323

Fig. 4: Classification report - CRF (II-A)

	precision	recall	f1-score	support
B-AerospaceManufacturer	1.00	0.80	0.89	10
B-AnatomicalStructure	0.70	0.41	0.52	17
B-ArtWork	0.00	0.00	0.00	13
B-Artist	0.61	0.56	0.58	212
B-Athlete	0.41	0.44	0.42	79
B-CarManufacturer	0.67	0.46	0.55	13
B-Cleric	0.67	0.27	0.38	15
B-Clothing	1.00	0.10	0.18	10
B-Disease	0.64	0.39	0.48	18
B-Drink	0.83	0.45	0.59	11
B-Facility	0.60	0.35	0.44	52
B-Food	0.00	0.00	0.00	19
B-HumanSettlement	0.73	0.61	0.66	109
B-MedicalProcedure	0.57	0.31	0.40	13
B-Medication/Vaccine	1.00	0.06	0.11	18
B-MusicalGRP	0.57	0.32	0.41	37
B-MusicalWork	0.56	0.30	0.39	61
B-ORG	0.61	0.44	0.51	78
B-OtherLOC	0.86	0.38	0.52	16
B-OtherPER	0.47	0.29	0.36	91
B-OtherPROD	0.67	0.16	0.26	49
B-Politician	0.50	0.38	0.43	53
B-PrivateCorp	1.00	0.55	0.71	11
B-PublicCorp	0.88	0.25	0.39	28
B-Scientist	0.00	0.00	0.00	15
B-Software	0.67	0.31	0.42	26
B-SportsGRP	0.91	0.51	0.66	41
B-SportsManager	0.50	0.19	0.27	16
B-Station	0.77	0.50	0.61	20
B-Symptom	0.71	0.50	0.59	10
B-Vehicle	0.88	0.35	0.50	20
B-VisualWork	0.39	0.31	0.35	61
B-WrittenWork	0.71	0.37	0.49	54
I-AerospaceManufacturer	1.00	1.00	1.00	7
I-AnatomicalStructure	0.50	0.25	0.33	4
I-ArtWork	0.00	0.00	0.00	22
I-Artist	0.58	0.55	0.57	217
I-Athlete	0.41	0.46	0.43	78
I-CarManufacturer	1.00	0.33	0.50	6
I-Cleric	0.71	0.31	0.43	16
I-Clothing	0.00	0.00	0.00	2
I-Disease	0.75	0.41	0.53	22
I-Drink	1.00	1.00	1.00	2
I-Facility	0.57	0.53	0.55	64
I-Food	0.00	0.00	0.00	9
I-HumanSettlement	0.84	0.74	0.79	72
I-MedicalProcedure	0.67	0.44	0.53	9
I-Medication/Vaccine	0.00	0.00	0.00	6
I-MusicalGRP	0.46	0.35	0.40	46
I-MusicalWork	0.61	0.43	0.50	108
I-ORG	0.56	0.52	0.54	108
I-OtherLOC	0.94	0.47	0.62	32
I-OtherPER	0.52	0.34	0.41	122
I-OtherPROD	0.67	0.19	0.30	42
I-Politician	0.41	0.40	0.40	60
I-PrivateCorp	1.00	0.70	0.82	10
I-PublicCorp	0.75	0.36	0.49	25
I-Scientist	0.00	0.00	0.00	16
I-Software	0.30	0.22	0.26	27
I-SportsGRP	0.91	0.63	0.74	46
I-SportsManager	0.50	0.18	0.26	17
I-Station	0.95	0.67	0.78	27
I-Symptom	0.25	0.33	0.29	3
I-Vehicle	0.80	0.20	0.32	20
I-VisualWork	0.46	0.40	0.43	122
I-WrittenWork	0.74	0.62	0.67	90
O	0.92	0.98	0.95	10570
accuracy			0.87	13323
macro avg	0.61	0.38	0.45	13323
weighted avg	0.85	0.87	0.85	13323

Fig. 5: Classification report - CRF (II-B)

E. Basic Features + POS Tags + Neighbours + Suffix Symbols + URL + Emotion Symbols + Word Forms

Further feature additions were done to help make CRF more robust in correctly identifying named entities. Features that help in identifying the context through suffix symbols like the presence of an exclamation mark or a question mark were added. To identify URLs, we also added a feature to check if the word starts with ”https://” or ”http://”. The MultiCoNER dataset has three sources for its data - Wikipedia, MS-MARCO QnA corpus [2], and Bing user queries from the ORCAS dataset [3]. These may contain some words that are generally used on social media that indicate positive or negative emotions like ”XD”, ”yay!”, ”:D” or ”:’(”, ”:(”, ”-.-”. We added a list of possible words like this that can be added as features. Also, features for words with specific suffixes like those ending with ”ing”, ”es”, ”ent”, ”ly”, ”ery” were added since they can help establishing the named entity recognition along with the corresponding POS tags. The F1-score obtained in this version was 85.62% (Figure 8), which is a negligible increase. It can probably be inferred that too many of these handcrafted features did not improve the model’s ability.

F. Hyperparameter Tuning

In order to select the best hyperparameters, we performed a Randomized Search which gave us:

```
best params: {'c1': 0.14913613658393327, 'c2': 0.021561693769247266}
best CV score: 0.7837934189380391
```

	precision	recall	f1-score	support
B-AerospaceManufacturer	0.91	1.00	0.95	10
B-AnatomicalStructure	0.75	0.53	0.62	17
B-ArtWork	0.50	0.08	0.13	13
B-Artist	0.49	0.47	0.48	212
B-Athlete	0.33	0.35	0.34	79
B-CarManufacturer	0.64	0.54	0.58	13
B-Cleric	0.67	0.27	0.38	15
B-Clothing	0.38	0.30	0.33	10
B-Disease	0.54	0.39	0.45	18
B-Drink	1.00	0.55	0.71	11
B-Facility	0.60	0.40	0.48	52
B-Food	0.67	0.11	0.18	19
B-HumanSettlement	0.74	0.61	0.67	109
B-MedicalProcedure	0.62	0.38	0.48	13
B-Medication/Vaccine	0.20	0.06	0.09	18
B-MusicalGRP	0.59	0.35	0.44	37
B-MusicalWork	0.33	0.13	0.19	61
B-ORG	0.63	0.51	0.57	78
B-OtherLOC	0.50	0.31	0.38	16
B-OtherPER	0.39	0.26	0.31	91
B-OtherPROD	0.85	0.22	0.35	49
B-Politician	0.50	0.26	0.35	53
B-PrivateCorp	0.86	0.55	0.67	11
B-PublicCorp	0.75	0.21	0.33	28
B-Scientist	0.12	0.07	0.09	15
B-Software	0.62	0.31	0.41	26
B-SportsGRP	0.84	0.51	0.64	41
B-SportsManager	0.25	0.06	0.10	16
B-Station	0.64	0.45	0.53	20
B-Symptom	0.70	0.70	0.70	10
B-Vehicle	1.00	0.35	0.52	20
B-VisualWork	0.15	0.05	0.07	61
B-WrittenWork	0.73	0.35	0.48	54
I-AerospaceManufacturer	0.54	1.00	0.70	7
I-AnatomicalStructure	0.50	0.25	0.33	4
I-ArtWork	0.20	0.09	0.13	22
I-Artist	0.50	0.49	0.49	217
I-Athlete	0.31	0.36	0.34	78
I-CarManufacturer	0.67	0.33	0.44	6
I-Cleric	0.57	0.25	0.35	16
I-Clothing	0.20	0.50	0.29	2
I-Disease	0.60	0.41	0.49	22
I-Drink	1.00	1.00	1.00	2
I-Facility	0.55	0.58	0.56	64
I-Food	0.40	0.22	0.29	9
I-HumanSettlement	0.80	0.71	0.75	72
I-MedicalProcedure	0.75	0.67	0.71	9
I-Medication/Vaccine	0.00	0.00	0.00	6
I-MusicalGRP	0.45	0.41	0.43	46
I-MusicalWork	0.51	0.38	0.43	108
I-ORG	0.50	0.59	0.54	108
I-OtherLOC	0.47	0.44	0.45	32
I-OtherPER	0.41	0.30	0.35	122
I-OtherPROD	0.69	0.26	0.38	42
I-Politician	0.35	0.30	0.32	60
I-PrivateCorp	0.67	0.40	0.50	10
I-PublicCorp	0.50	0.40	0.44	25
I-Scientist	0.09	0.06	0.07	16
I-Software	0.07	0.07	0.07	27
I-SportsGRP	0.83	0.63	0.72	46
I-SportsManager	0.25	0.06	0.10	17
I-Station	0.85	0.63	0.72	27
I-Symptom	0.43	1.00	0.60	3
I-Vehicle	0.59	0.50	0.54	20
I-VisualWork	0.34	0.13	0.19	122
I-WrittenWork	0.82	0.66	0.73	90
O	0.92	0.98	0.95	10570
accuracy			0.86	13323
macro avg	0.55	0.40	0.44	13323
weighted avg	0.84	0.86	0.84	13323

Fig. 6: Classification report - CRF (II-C)

	precision	recall	f1-score	support
B-AerospaceManufacturer	1.00	0.70	0.82	10
B-AnatomicalStructure	0.71	0.29	0.42	17
B-ArtWork	0.00	0.00	0.00	13
B-Artist	0.59	0.58	0.59	212
B-Athlete	0.45	0.49	0.47	79
B-CarManufacturer	0.62	0.38	0.48	13
B-Cleric	0.67	0.27	0.38	15
B-Clothing	1.00	0.10	0.18	10
B-Disease	0.55	0.33	0.41	18
B-Drink	0.83	0.45	0.59	11
B-Facility	0.66	0.40	0.50	52
B-Food	1.00	0.05	0.10	19
B-HumanSettlement	0.71	0.61	0.65	109
B-MedicalProcedure	0.67	0.31	0.42	13
B-Medication/Vaccine	0.50	0.06	0.10	18
B-MusicalGRP	0.63	0.32	0.43	37
B-MusicalWork	0.53	0.28	0.37	61
B-ORG	0.62	0.46	0.53	78
B-OtherLOC	0.86	0.38	0.52	16
B-OtherPER	0.46	0.27	0.34	91
B-OtherPROD	0.67	0.20	0.31	49
B-Politician	0.49	0.36	0.41	53
B-PrivateCorp	1.00	0.55	0.71	11
B-PublicCorp	0.86	0.21	0.34	28
B-Scientist	0.20	0.07	0.10	15
B-Software	0.47	0.27	0.34	26
B-SportsGRP	0.92	0.59	0.72	41
B-SportsManager	0.60	0.19	0.29	16
B-Station	0.71	0.50	0.59	20
B-Symptom	0.75	0.60	0.67	10
B-Vehicle	0.86	0.30	0.44	20
B-VisualWork	0.41	0.30	0.34	61
B-WrittenWork	0.73	0.41	0.52	54
I-AerospaceManufacturer	1.00	0.86	0.92	7
I-AnatomicalStructure	0.00	0.00	0.00	4
I-ArtWork	0.00	0.00	0.00	22
I-Artist	0.57	0.59	0.58	217
I-Athlete	0.44	0.50	0.47	78
I-CarManufacturer	1.00	0.33	0.50	6
I-Cleric	0.71	0.31	0.43	16
I-Clothing	0.00	0.00	0.00	2
I-Disease	0.69	0.41	0.51	22
I-Drink	1.00	1.00	1.00	2
I-Facility	0.62	0.56	0.59	64
I-Food	1.00	0.22	0.36	9
I-HumanSettlement	0.86	0.75	0.80	72
I-MedicalProcedure	0.80	0.44	0.57	9
I-Medication/Vaccine	0.00	0.00	0.00	6
I-MusicalGRP	0.59	0.35	0.44	46
I-MusicalWork	0.58	0.44	0.50	108
I-ORG	0.56	0.55	0.55	108
I-OtherLOC	0.94	0.47	0.62	32
I-OtherPER	0.48	0.34	0.40	122
I-OtherPROD	0.60	0.21	0.32	42
I-Politician	0.39	0.38	0.39	60
I-PrivateCorp	1.00	0.70	0.82	10
I-PublicCorp	0.73	0.32	0.44	25
I-Scientist	0.20	0.06	0.10	16
I-Software	0.11	0.15	0.12	27
I-SportsGRP	0.89	0.70	0.78	46
I-SportsManager	0.60	0.18	0.27	17
I-Station	0.95	0.67	0.78	27
I-Symptom	0.25	0.33	0.29	3
I-Vehicle	0.62	0.25	0.36	20
I-VisualWork	0.40	0.33	0.36	122
I-WrittenWork	0.76	0.68	0.72	90
O	0.92	0.98	0.95	10570
accuracy			0.87	13323
macro avg	0.63	0.38	0.45	13323
weighted avg	0.86	0.87	0.86	13323

Fig. 7: Classification report - CRF (II-D)

With these settings, we got our final results of accuracy of 87.09%, the precision of 85.66%, recall of 87.09%, and F1-score of 85.79% (Figure 9). The model size is 14.14M. This model has good predictive capacity given that most of the classes have non-zero scores even when the total number of classes is very large, i.e., 67. A few classes that have not been identified correctly include B-ArtWork, B-Clothing, I-AnatomicalStructure, and I-Clothing. Given these classes contain very few samples, for example, 87 for I-Clothing vs 3837 for I-Artist, they could have not been learned during training and predicted correctly by the model. Interestingly, even the highest sample class I-Artist has only 58% F1-score which suggests that just the number of samples is not a determining factor, but the semantic, syntactic, and contextual meaning of the word also matter in this task of NER.

G. Qualitative Results

A few sentences were picked at random from the validation set which is shown in Figure 10. We formatted them in the following manner to analyse the model performance in each sentence - 'token (ground-truthTag predictedTag)'. In sentence 670, all 'O' tags are predicted correctly. 'eldritch' was also predicted correctly as 'B-MusicalGRP'. However, 'my sharona' was wrongly classified as 'VisualWork' instead of 'MusicalWork', although they both belong to the major class Creative Works. Sentence 39 shows a case where 'OtherPER' was classified as 'Artist', probably due to bias towards the 'Artist' class with a large number of samples. In sentence 399, all the tags are predicted correctly, 'west japan railway company' as 'PublicCorp'.

	precision	recall	f1-score	support
B-AerospaceManufacturer	1.00	0.80	0.89	10
B-AnatomicalStructure	0.71	0.29	0.42	17
B-ArtWork	0.00	0.00	0.00	13
B-Artist	0.60	0.61	0.60	212
B-Athlete	0.40	0.49	0.44	79
B-CarManufacturer	0.67	0.46	0.55	13
B-Cleric	0.80	0.27	0.40	15
B-Clothing	0.00	0.00	0.00	10
B-Disease	0.55	0.33	0.41	18
B-Drink	0.83	0.45	0.59	11
B-Facility	0.63	0.37	0.46	52
B-Food	0.00	0.00	0.00	19
B-HumanSettlement	0.69	0.64	0.66	109
B-MedicalProcedure	0.60	0.23	0.33	13
B-Medication/Vaccine	0.75	0.17	0.27	18
B-MusicalGRP	0.61	0.38	0.47	37
B-MusicalWork	0.53	0.30	0.38	61
B-ORG	0.64	0.46	0.54	78
B-OtherLOC	0.75	0.38	0.50	16
B-OtherPER	0.41	0.26	0.32	91
B-OtherPROD	0.67	0.20	0.31	49
B-Politician	0.44	0.36	0.40	53
B-PrivateCorp	0.83	0.45	0.59	11
B-PublicCorp	0.86	0.21	0.34	28
B-Scientist	0.00	0.00	0.00	15
B-Software	0.47	0.27	0.34	26
B-SportsGRP	0.93	0.61	0.74	41
B-SportsManager	0.50	0.19	0.27	16
B-Station	0.69	0.55	0.61	20
B-Symptom	0.75	0.60	0.67	10
B-Vehicle	0.90	0.45	0.60	20
B-VisualWork	0.42	0.30	0.35	61
B-WrittenWork	0.72	0.39	0.51	54
I-AerospaceManufacturer	1.00	1.00	1.00	7
I-AnatomicalStructure	0.00	0.00	0.00	4
I-ArtWork	0.00	0.00	0.00	22
I-Artist	0.58	0.60	0.59	217
I-Athlete	0.38	0.50	0.43	78
I-CarManufacturer	1.00	0.33	0.50	6
I-Cleric	0.83	0.31	0.45	16
I-Clothing	0.00	0.00	0.00	2
I-Disease	0.58	0.32	0.41	22
I-Drink	1.00	1.00	1.00	2
I-Facility	0.60	0.48	0.53	64
I-Food	0.00	0.00	0.00	9
I-HumanSettlement	0.77	0.74	0.75	72
I-MedicalProcedure	0.75	0.33	0.46	9
I-Medication/Vaccine	0.00	0.00	0.00	6
I-MusicalGRP	0.55	0.35	0.43	46
I-MusicalWork	0.58	0.46	0.52	108
I-ORG	0.55	0.54	0.54	108
I-OtherLOC	0.62	0.47	0.54	32
I-OtherPER	0.41	0.32	0.36	122
I-OtherPROD	0.60	0.21	0.32	42
I-Politician	0.34	0.38	0.36	60
I-PrivateCorp	0.75	0.60	0.67	10
I-PublicCorp	0.60	0.36	0.45	25
I-Scientist	0.00	0.00	0.00	16
I-Software	0.11	0.15	0.12	27
I-SportsGRP	0.81	0.76	0.79	46
I-SportsManager	0.50	0.18	0.26	17
I-Station	0.78	0.67	0.72	27
I-Symptom	0.25	0.33	0.29	3
I-Vehicle	0.70	0.35	0.47	20
I-VisualWork	0.46	0.36	0.40	122
I-WrittenWork	0.75	0.64	0.69	90
O	0.93	0.98	0.95	10570
accuracy			0.87	13323
macro avg	0.55	0.38	0.43	13323
weighted avg	0.85	0.87	0.86	13323

Fig. 8: Classification report - CRF (II-E)

	precision	recall	f1-score	support
B-AerospaceManufacturer	1.000000	0.900000	0.947368	10
B-AnatomicalStructure	0.750000	0.352941	0.480000	17
B-ArtWork	0.000000	0.000000	0.000000	13
B-Artist	0.585586	0.613208	0.599078	212
B-Athlete	0.377778	0.430380	0.402367	79
B-CarManufacturer	0.666667	0.461538	0.545455	13
B-Cleric	0.800000	0.266667	0.400000	15
B-Clothing	0.666667	0.200000	0.307692	10
B-Disease	0.545455	0.333333	0.413793	18
B-Drink	0.833333	0.454545	0.588235	11
B-Facility	0.689655	0.384615	0.493827	52
B-Food	1.000000	0.052632	0.100000	19
B-HumanSettlement	0.693069	0.642202	0.666667	109
B-MedicalProcedure	0.500000	0.230769	0.315789	13
B-Medication/Vaccine	0.666667	0.111111	0.190476	18
B-MusicalGRP	0.545455	0.324324	0.406780	37
B-MusicalWork	0.500000	0.278689	0.357895	61
B-ORG	0.644068	0.487179	0.554745	78
B-OtherLOC	0.857143	0.375000	0.521739	16
B-OtherPER	0.413793	0.263736	0.322148	91
B-OtherPROD	0.647059	0.224490	0.333333	49
B-Politician	0.463415	0.358491	0.404255	53
B-PrivateCorp	0.875000	0.636364	0.736842	11
B-PublicCorp	0.750000	0.214286	0.333333	28
B-Scientist	0.000000	0.000000	0.000000	15
B-Software	0.571429	0.307692	0.400000	26
B-SportsGRP	0.925926	0.609756	0.735294	41
B-SportsManager	0.500000	0.187500	0.272727	16
B-Station	0.687500	0.550000	0.611111	20
B-Symptom	0.666667	0.600000	0.631579	10
B-Vehicle	0.875000	0.350000	0.500000	20
B-VisualWork	0.409091	0.295082	0.342857	61
B-WrittenWork	0.766667	0.425926	0.547619	54
I-AerospaceManufacturer	1.000000	1.000000	1.000000	7
I-AnatomicalStructure	0.000000	0.000000	0.000000	4
I-ArtWork	0.500000	0.272727	0.352941	22
I-Artist	0.568966	0.608295	0.587973	217
I-Athlete	0.369565	0.435897	0.400000	78
I-CarManufacturer	1.000000	0.333333	0.500000	6
I-Cleric	0.833333	0.312500	0.454545	16
I-Clothing	0.000000	0.000000	0.000000	2
I-Disease	0.692308	0.409091	0.514286	22
I-Drink	1.000000	0.500000	0.666667	2
I-Facility	0.586957	0.421875	0.490909	64
I-Food	1.000000	0.222222	0.363636	9
I-HumanSettlement	0.770270	0.791667	0.780822	72
I-MedicalProcedure	0.600000	0.333333	0.428571	9
I-Medication/Vaccine	0.000000	0.000000	0.000000	6
I-MusicalGRP	0.500000	0.347826	0.410256	46
I-MusicalWork	0.544444	0.453704	0.494949	108
I-ORG	0.600000	0.583333	0.591549	108
I-OtherLOC	0.937500	0.468750	0.625000	32
I-OtherPER	0.431818	0.311475	0.361905	122
I-OtherPROD	0.588235	0.238095	0.338983	42
I-Politician	0.348485	0.383333	0.365079	60
I-PrivateCorp	0.777778	0.700000	0.736842	10
I-PublicCorp	0.727273	0.320000	0.444444	25
I-Scientist	0.000000	0.000000	0.000000	16
I-Software	0.214286	0.222222	0.218182	27
I-SportsGRP	0.857143	0.782609	0.818182	46
I-SportsManager	0.500000	0.176471	0.260870	17
I-Station	0.620690	0.666667	0.642857	27
I-Symptom	0.250000	0.333333	0.285714	3
I-Vehicle	0.800000	0.200000	0.320000	20
I-VisualWork	0.475248	0.393443	0.430493	122
I-WrittenWork	0.772152	0.677778	0.721893	90
O	0.927824	0.981457	0.953887	10570
accuracy			0.870975	13323
macro avg	0.606976	0.385133	0.448126	13323
weighted avg	0.856652	0.870975	0.857924	13323

Fig. 9: Classification report - CRF (II-F)

Sentence 670

reverse (O O) is (O O) the (O O) fourth (O O) album (O O) of (O O) the (O O) progressive (O O) metal (O O) band (O O) eldritch (B-MusicalGRP B-MusicalGRP) containing (O O) a (O O) cover (O O) of (O O) my (B-MusicalWork B-VisualWork) sharona (I-MusicalWork I-VisualWork) . (O O)

Sentence 39

da (B-OtherPER B-Artist) yanlin (I-OtherPER I-Artist) a (O O) distant (O O) relative (O O) of (O O) the (O O) defunct (O O) balhae (B-HumanSettlement O) regime (O O) rebels (O O) ; (O O) he (O O) is (O O) defeated (O O)

Sentence 399

it (O O) is (O O) partially (O O) owned (O O) by (O O) the (O O) west (B-PublicCorp B-PublicCorp) japan (I-PublicCorp I-PublicCorp) railway (I-PublicCorp I-PublicCorp) company (I-PublicCorp I-PublicCorp) . (O O)

Fig. 10: Token, (Ground-truth Tag, Predicted Tag) of Few Sentences in Validation Set

III. Deep learning model - BiLSTM

BiLSTM is a variant of Recurrent neural networks (RNNs), that specialize in sequential data that has temporal characteristics, or time dependencies. RNNs are often used in fields like natural language processing and speech recognition which generally deal with linguistic properties where words or phrases also depend on those prior to them, or on the context around them.

A. BiLSTM

To prepare input for neural network architecture, token-to-id, and tag-to-id bidirectional mapping dictionaries were created. As the BiLSTM requires all the input sentences to be of the same length, we used 'post' padding for each sentence. Since the maximum length of the sentence is 68, we pad all of the sentences to match the length. We trained the model with an early stopping setting if the validation accuracy continues to go down. The model architecture is as shown below:

```
model = Sequential()
model.add(Embedding(input_dim=vocab_size, output_dim=output_dim, input_length=input_length))
model.add(Bidirectional(LSTM(units=25, return_sequences=True, dropout=0.2), merge_mode = 'concat'))
model.add(TimeDistributed(Dense(num_labels, activation='softmax')))
```

```
Model: "sequential_3"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_3 (Embedding)     (None, 68, 50)            1707000

 bidirectional_1 (Bidirectio  (None, 68, 50)           15200
 nal)

 time_distributed_3 (TimeDis  (None, 68, 67)           3417
 tributed)

=================================================================
Total params: 1,725,617
Trainable params: 1,725,617
Non-trainable params: 0
_________________________________________________________________
```

Fig. 11: BiLSTM Model Summary with Parameters

Embedding layer: This layer assigns a dense vector of a fixed size specified in 'output_dim' to each integer input. By passing 'vocab_size' as 'input_dim', we create dense vectors for each word in the vocabulary of the dataset.
Bidirectional(LSTM) layer: We pass the LSTM layer as input to the Keras Bidirectional layer which concatenates the output of forward and backward LSTM.
TimeDistributed(Dense) layer: This layer allows a fully connected neural network, or a dense layer, to be applied to each time step of the input sequence - the same dense layer is applied independently to each time step of the sequence, and a separate set of weights is learned for each time step. The output of each dense layer at each time step is then concatenated to produce a final output.

From the results depicted in Figure 13, we see that the accuracy is 95.35% and the weighted F1-score is 93.08%. But the predictive power of the model when looking at the macro average F1-score of 1.46% tells us another story. The accuracy and weighted F1-score suggest that the model is predicting only the majority class 'O' as we can clearly observe, possibly due to imbalanced class distribution. However, the very low macro average F1 score indicates that the model is performing very poorly on most of the other classes. Therefore, while the overall performance of the model may seem good based on the weighted

metrics, it is important to examine the macro average metrics to identify which classes the model is struggling with and to address any issues with class imbalance or model representational power for those classes.

B. BiLSTM + Pre-trained Word Embeddings

In the next iteration, we wanted to see if the model would give better results if pre-trained word embeddings were used. We chose GloVe (Global Vectors) for Word Representation here [4]. It is a word embedding technique that creates real-valued vector representations for words based on the co-occurrence statistics of the words in a corpus. The resulting word embeddings are dense vectors of fixed size with a meaningful substructure, ranging from 50 to 300 dimensions, that capture semantic and syntactic similarities between words. We have used the glove model pre-trained on the combination Gigaword5 + Wikipedia2014, which has 6 billion tokens and 50-dimensional vectors.

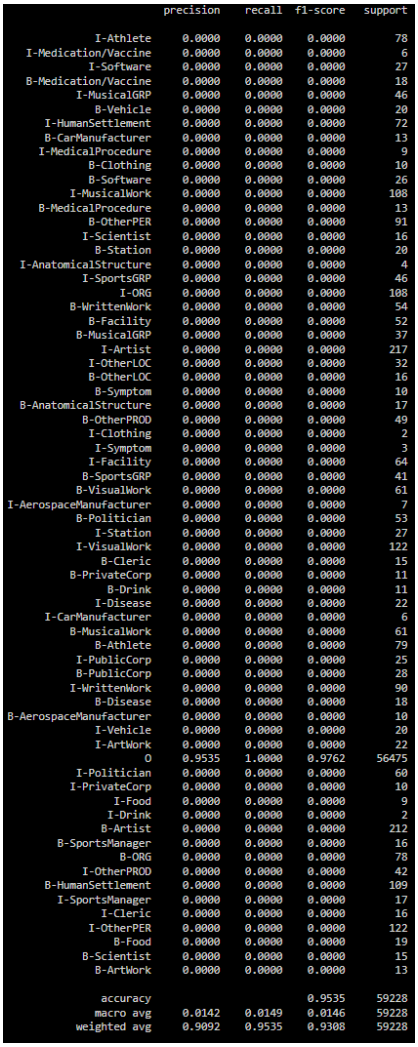
```
Model: "sequential_5"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_4 (Embedding)     (None, 68, 50)            1707000

 bidirectional_3 (Bidirectio  (None, 68, 50)           15200
 nal)

 time_distributed_5 (TimeDis  (None, 68, 67)           3417
 tributed)

=================================================================
Total params: 1,725,617
Trainable params: 18,617
Non-trainable params: 1,707,000
_________________________________________________________________
```

Fig. 12: BiLSTM + GloVe Model Summary with Parameters

Since we used pre-trained word embeddings, in the model summary, although the total parameters are the same as the first version, the non-trainable parameters have increased from 0 to 1,707,000. Coming to the results, there is no difference at all compared to using BiLSTM without GloVe embeddings. We think the following reasons can explain why deep learning models are performing significantly worse while processing complex and ambiguous named entities due to the challenges posed by these types of entities. Complex NEs can take the form of any linguistic constituent and may not look like traditional NEs. This syntactic ambiguity can make it challenging for BiLSTM models to recognize them based on context alone. DL-based models generally are effective in automatically learning useful representations and underlying factors from raw data, but the RNN BiLSTM, in this case, fails to perform well. These models perform significantly worse on complex/unseen entities.

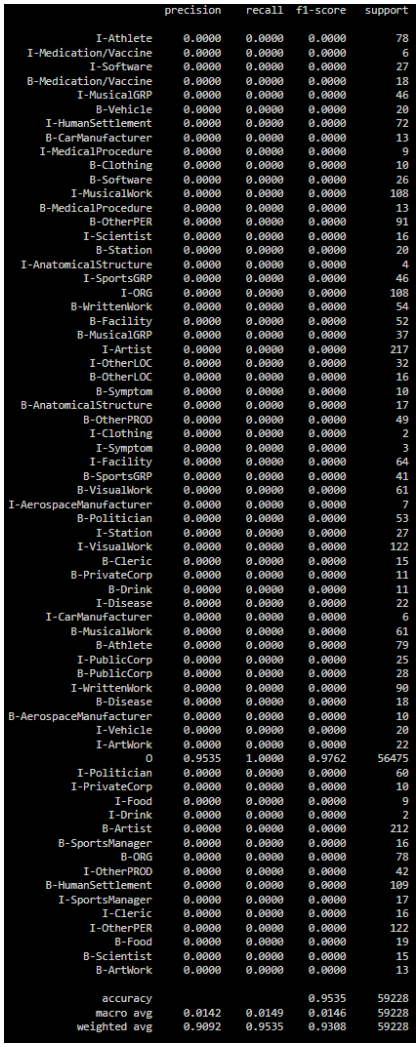
C. Qualitative Results

For showing the qualitative results, a few sentences were picked at random from the validation set present in Figure 15 that give the tokens with their ground-truth and predicted tag. In all three sentences, all the words including named entities like 'OtherPER' were predicted as 'O'.



	precision	recall	f1-score	support
I-Athlete	0.0000	0.0000	0.0000	78
I-Medication/Vaccine	0.0000	0.0000	0.0000	6
I-Software	0.0000	0.0000	0.0000	27
B-Medication/Vaccine	0.0000	0.0000	0.0000	18
I-MusicalGRP	0.0000	0.0000	0.0000	46
B-Vehicle	0.0000	0.0000	0.0000	20
I-HumanSettlement	0.0000	0.0000	0.0000	72
B-CarManufacturer	0.0000	0.0000	0.0000	13
I-MedicalProcedure	0.0000	0.0000	0.0000	9
B-Clothing	0.0000	0.0000	0.0000	10
B-Software	0.0000	0.0000	0.0000	26
I-MusicalWork	0.0000	0.0000	0.0000	108
B-MedicalProcedure	0.0000	0.0000	0.0000	13
B-OtherPER	0.0000	0.0000	0.0000	91
I-Scientist	0.0000	0.0000	0.0000	16
B-Station	0.0000	0.0000	0.0000	20
I-AnatomicalStructure	0.0000	0.0000	0.0000	4
I-SportsGRP	0.0000	0.0000	0.0000	46
I-ORG	0.0000	0.0000	0.0000	108
B-WrittenWork	0.0000	0.0000	0.0000	54
B-Facility	0.0000	0.0000	0.0000	52
B-MusicalGRP	0.0000	0.0000	0.0000	37
I-Artist	0.0000	0.0000	0.0000	217
I-OtherLOC	0.0000	0.0000	0.0000	32
B-OtherLOC	0.0000	0.0000	0.0000	16
B-Symptom	0.0000	0.0000	0.0000	10
B-AnatomicalStructure	0.0000	0.0000	0.0000	17
B-OtherPROD	0.0000	0.0000	0.0000	49
I-Clothing	0.0000	0.0000	0.0000	2
I-Symptom	0.0000	0.0000	0.0000	3
I-Facility	0.0000	0.0000	0.0000	64
B-SportsGRP	0.0000	0.0000	0.0000	41
B-VisualWork	0.0000	0.0000	0.0000	61
I-AerospaceManufacturer	0.0000	0.0000	0.0000	7
B-Politician	0.0000	0.0000	0.0000	53
I-Station	0.0000	0.0000	0.0000	27
I-VisualWork	0.0000	0.0000	0.0000	122
B-Cleric	0.0000	0.0000	0.0000	15
B-PrivateCorp	0.0000	0.0000	0.0000	11
B-Drink	0.0000	0.0000	0.0000	11
I-Disease	0.0000	0.0000	0.0000	22
I-CarManufacturer	0.0000	0.0000	0.0000	6
B-MusicalWork	0.0000	0.0000	0.0000	61
B-Athlete	0.0000	0.0000	0.0000	79
I-PublicCorp	0.0000	0.0000	0.0000	25
B-PublicCorp	0.0000	0.0000	0.0000	28
I-WrittenWork	0.0000	0.0000	0.0000	90
B-Disease	0.0000	0.0000	0.0000	18
B-AerospaceManufacturer	0.0000	0.0000	0.0000	10
I-Vehicle	0.0000	0.0000	0.0000	20
I-ArtWork	0.0000	0.0000	0.0000	22
O	0.9535	1.0000	0.9762	56475
I-Politician	0.0000	0.0000	0.0000	60
I-PrivateCorp	0.0000	0.0000	0.0000	10
I-Food	0.0000	0.0000	0.0000	9
I-Drink	0.0000	0.0000	0.0000	2
B-Artist	0.0000	0.0000	0.0000	212
B-SportsManager	0.0000	0.0000	0.0000	16
B-ORG	0.0000	0.0000	0.0000	78
I-OtherPROD	0.0000	0.0000	0.0000	42
B-HumanSettlement	0.0000	0.0000	0.0000	109
I-SportsManager	0.0000	0.0000	0.0000	17
I-Cleric	0.0000	0.0000	0.0000	16
I-OtherPER	0.0000	0.0000	0.0000	122
B-Food	0.0000	0.0000	0.0000	19
B-Scientist	0.0000	0.0000	0.0000	15
B-ArtWork	0.0000	0.0000	0.0000	13
accuracy			0.9535	59228
macro avg	0.0142	0.0149	0.0146	59228
weighted avg	0.9092	0.9535	0.9308	59228

Fig. 13: BiLSTM Results (III-A)



	precision	recall	f1-score	support
I-Athlete	0.0000	0.0000	0.0000	78
I-Medication/Vaccine	0.0000	0.0000	0.0000	6
I-Software	0.0000	0.0000	0.0000	27
B-Medication/Vaccine	0.0000	0.0000	0.0000	18
I-MusicalGRP	0.0000	0.0000	0.0000	46
B-Vehicle	0.0000	0.0000	0.0000	20
I-HumanSettlement	0.0000	0.0000	0.0000	72
B-CarManufacturer	0.0000	0.0000	0.0000	13
I-MedicalProcedure	0.0000	0.0000	0.0000	9
B-Clothing	0.0000	0.0000	0.0000	10
B-Software	0.0000	0.0000	0.0000	26
I-MusicalWork	0.0000	0.0000	0.0000	108
B-MedicalProcedure	0.0000	0.0000	0.0000	13
B-OtherPER	0.0000	0.0000	0.0000	91
I-Scientist	0.0000	0.0000	0.0000	16
B-Station	0.0000	0.0000	0.0000	20
I-AnatomicalStructure	0.0000	0.0000	0.0000	4
I-SportsGRP	0.0000	0.0000	0.0000	46
I-ORG	0.0000	0.0000	0.0000	108
B-WrittenWork	0.0000	0.0000	0.0000	54
B-Facility	0.0000	0.0000	0.0000	52
B-MusicalGRP	0.0000	0.0000	0.0000	37
I-Artist	0.0000	0.0000	0.0000	217
I-OtherLOC	0.0000	0.0000	0.0000	32
B-OtherLOC	0.0000	0.0000	0.0000	16
B-Symptom	0.0000	0.0000	0.0000	10
B-AnatomicalStructure	0.0000	0.0000	0.0000	17
B-OtherPROD	0.0000	0.0000	0.0000	49
I-Clothing	0.0000	0.0000	0.0000	2
I-Symptom	0.0000	0.0000	0.0000	3
I-Facility	0.0000	0.0000	0.0000	64
B-SportsGRP	0.0000	0.0000	0.0000	41
B-VisualWork	0.0000	0.0000	0.0000	61
I-AerospaceManufacturer	0.0000	0.0000	0.0000	7
B-Politician	0.0000	0.0000	0.0000	53
I-Station	0.0000	0.0000	0.0000	27
I-VisualWork	0.0000	0.0000	0.0000	122
B-Cleric	0.0000	0.0000	0.0000	15
B-PrivateCorp	0.0000	0.0000	0.0000	11
B-Drink	0.0000	0.0000	0.0000	11
I-Disease	0.0000	0.0000	0.0000	22
I-CarManufacturer	0.0000	0.0000	0.0000	6
B-MusicalWork	0.0000	0.0000	0.0000	61
B-Athlete	0.0000	0.0000	0.0000	79
I-PublicCorp	0.0000	0.0000	0.0000	25
B-PublicCorp	0.0000	0.0000	0.0000	28
I-WrittenWork	0.0000	0.0000	0.0000	90
B-Disease	0.0000	0.0000	0.0000	18
B-AerospaceManufacturer	0.0000	0.0000	0.0000	10
I-Vehicle	0.0000	0.0000	0.0000	20
I-ArtWork	0.0000	0.0000	0.0000	22
O	0.9535	1.0000	0.9762	56475
I-Politician	0.0000	0.0000	0.0000	60
I-PrivateCorp	0.0000	0.0000	0.0000	10
I-Food	0.0000	0.0000	0.0000	9
I-Drink	0.0000	0.0000	0.0000	2
B-Artist	0.0000	0.0000	0.0000	212
B-SportsManager	0.0000	0.0000	0.0000	16
B-ORG	0.0000	0.0000	0.0000	78
I-OtherPROD	0.0000	0.0000	0.0000	42
B-HumanSettlement	0.0000	0.0000	0.0000	109
I-SportsManager	0.0000	0.0000	0.0000	17
I-Cleric	0.0000	0.0000	0.0000	16
I-OtherPER	0.0000	0.0000	0.0000	122
B-Food	0.0000	0.0000	0.0000	19
B-Scientist	0.0000	0.0000	0.0000	15
B-ArtWork	0.0000	0.0000	0.0000	13
accuracy			0.9535	59228
macro avg	0.0142	0.0149	0.0146	59228
weighted avg	0.9092	0.9535	0.9308	59228

Fig. 14: BiLSTM + GloVe Results (III-B)

Sentence 44

the (O O) cup (O O) is (O O) named (O O) after (O O) joe (B-OtherPER O) mcdonagh (OtherPER O) . (O O)

Sentence 19

it (O O) was (O O) described (O O) by (O O) edward (B-OtherPER O) meyrick (I-OtherPER O) in (O O) 1928 (O O) . (O O)

Sentence 24

akira (B-OtherPER O) nishiguchi (I-OtherPER O) : (O O) killed (O O) five (O O) people (O O) and (O O) engaged (O O) in (O O) fraud (O O) ; (O O) executed (O O) in (O O) 1970 (O O) . (O O)

Fig. 15: Token, (Ground-truth Tag, Predicted Tag) of Few Sentences in Validation Set

IV. Conclusion

In this project, our work focused on exploring two different approaches to named entity recognition (NER) - Conditional Random Fields (CRF) and Bi-directional Long Short-Term Memory (BiLSTM) models. We began by experimenting with different combinations of basic features, POS tags, and neighboring words to train CRF models. We then expanded our feature set to include additional suffix symbols, URLs, emotion symbols, and word forms. Through hyperparameter tuning, we were able to achieve high accuracy and F1 scores for our CRF models. We also explored the effectiveness of BiLSTM models for NER and experimented with leveraging pre-trained word embeddings. While our BiLSTM models achieved high weighted accuracy and F1 scores by predicting only one single majority class, we found that they struggled with recognizing complex and ambiguous named entities.

Our findings suggest that CRF models can be better equipped to handle complex and ambiguous NEs because they take into account the dependencies between adjacent labels in the sequence. This can help ensure that the predicted labels form a coherent sequence and can better handle syntactic ambiguity through handcrafted features based on linguistic properties. There are challenges in processing complex and ambiguous NEs in open-domain settings and it is important to use models that are appropriate for the task at hand. Experimenting with other deep learning models like transformers, or a combination of statistical and deep learning models such as BiLSTM+CRF might produce better results as they compensate for their individual limitations and can capture complementary features or patterns in the dataset.

References

[1] https://multiconer.github.io/dataset
[2] Bajaj, Payal, et al. "Ms marco: A human generated machine reading comprehension dataset." arXiv preprint arXiv:1611.09268 (2016).
[3] Craswell, Nick, et al. "ORCAS: 20 million clicked query-document pairs for analyzing search." Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020.
[4] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.