Named Entity Recognition using MultiCoNER II

CS5803: Natural Language Processing

Ananya Mantravadi (CS19B1004)
Sai Yaaminie Ganda (CS19B1022)
Siddharth Saini (CS19B1024)
IIIT Raichur

Contents

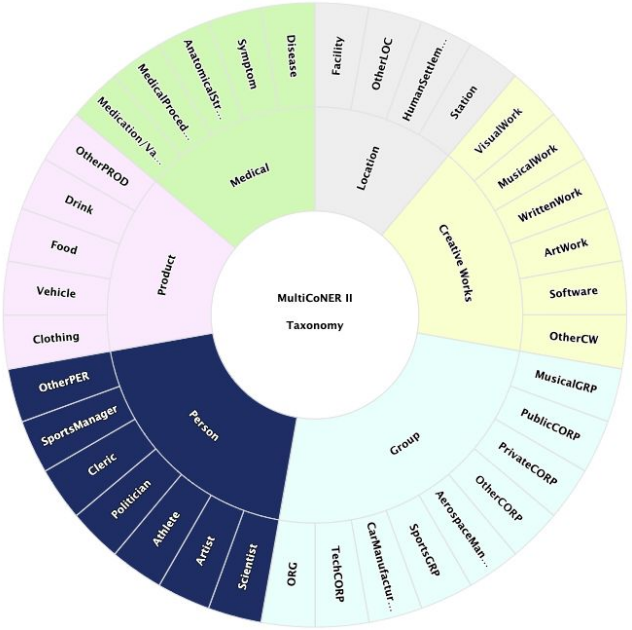
Problem Statement

Study and develop solutions for the Named Entity Recognition problem defined in the SemEval 2023 MultiCoNER 2 Task:

- Task 1 - Apply one statistical model (non-neural network based sequence labelling model) such as HMM, MEMM, or CRF to perform NER defined in the MultiCoNER 2 Task.
- Task 2 - Apply one deep learning model (RNNs, transformers etc.) to perform NER defined in the MultiCoNER 2 Task.

MultiCoNER II Multilingual Complex Named Entity Recognition Dataset

- Represents challenges of complex named entities like imperative clauses, complex syntactic structures, and long-tail entity distributions.
- Dataset sources - Various domains, including news, social media, and Wikipedia, providing 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.



Source: https://multiconer.github.io/

Dataset Preparation and Loading

- The dataset is hosted on the AWS cloud. First, install the AWS CLI (AWS command line interface). Then load the multiconer2023 dataset from the multiconer bucket. (Amazon S3 is an object storage service that stores data as objects within buckets.)
- Choice of language = English (EN) (monolingual task).
- The CoNLL file has 4 parts :
    - Id (in comments) - one id for one sentence
    - Domain = EN
    - Words/Tokens - words in English vocabulary
    - Corresponding NER tag

Statistical model - CRF

- 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
- 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.
- 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.

general form of a feature function for an input of size *n* -

$$F_k(X, Y) = \sum_{i=1}^{n} f_k(y_{i-1}, y_i, X, i)$$

- y_{i-1} - previous tokens
- y_i - current tokens
- X - input sequence
- i - position in the sequence

CRF:
Methodology and feature engineering iterations

1. Basic Features

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 L1 and L2 regularization have been set to 0.1 each.

The F1-score - 83.75%

Observations - Most likely/unlikely transitions; state features with large positive/weights.
Likely transitions: B-OtherPER -> I-OtherPER, B-VisualWork -> I-VisualWork, B-Athlete -> I-Athlete.
Top state features: bias, lowercase, and hash length for the label 'O'.

Conclusion - 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.

	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

2. Basic Features + Neighbours

Included features of the words around the target word - one previous word and one after.

The F1-score - 85.31%

Observations - Increment in F1 score by 1.56%

Conclusion - Including neighbours' features helps model to predict the target tag better.

	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

3. Basic Features + POS tags

Part Of Speech (POS) tagging can be used in NER since it can help in identifying named entities.

The F1-score - 84.21%

Observations - Increment in F1 score by 0.46% compared to A), where we consider only basic features and decrement of 1.1% compared to B), including neighbouring features.

Conclusion - Need to include both POS tags and Neighbours in order to get the best results.

	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

4. Basic Features + POS tags + Neighbours

Combining features including those of neighbours' along with POS tags.

The F1-score - 85.61%, improvement over all previous combinations.

Observations - Highest F1 score among all previous iterations.

Conclusion - Neighborhood features with POS tags results in improvement in model performance.

	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

5. Basic Features + POS tags + Neighbours + Suffix Symbols + URL + Emotion Symbols + Word Forms

Add additional features to make crf more robust for NE prediction task.

- Suffix symbols: ‘!’, ‘?’
- URL identification: Word starts with "https://" or "http://"
- Positive/negative emotion words: XD, yay!, :D or :'(, :(, -.-
- Word Forms: Words ending with "ing", "es", "ent", "ly", "ery", etc.

The F1-score - 85.62%, but negligible increase over 85.61%.

Observations - Highest F1 score among all the iterations.

Conclusion - It can be inferred that too many of these handcrafted features does not improve the model's ability.

	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

6. Hyperparameter Tuning

In order to select the best hyperparameters, we performed a Randomized Search which gave us: 'c1': 0.149, 'c2': 0.0215

Accuracy - 87.09%
Precision - 85.66%
Recall - 87.09%
F1-score - 85.79%, improvement over all previous combinations.

Observations - Classes with 0 F1-scores: B-ArtWork, B-Clothing, I-AnatomicalStructure, and I-Clothing (very few samples, I-Clothing: 87). Even second-majority samples class I-Artist has only 58% F1-score

Conclusion - 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.

	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

Qualitative Analysis

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)

Recurrent Neural Network: BiLSTM

1. BiLSTM

- BiLSTM is a deep learning model, that specialize in sequential data that has temporal characteristics, or time dependencies.
- Preprocessing data before sending as input to BiLSTM
    - Assign id to each word in train and validation set (vocabulary)
    - Assign id to each ‘ner_tag’ in the dataset.
    - Pad the sentences to create dataset with sentences of uniform length (max_length).
- Model is trained on the train set and validated on dev set.
- Hyperparameters - learning_rate = 0.0001
- Optimizer - Adam

BiLSTM Model Summary & Results

```
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
_________________________________________________________________
```

	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

Observations

- Accuracy is 95.35% and the weighted F1-score is 93.08%. But the the macro average F1-score is 1.46%.
- 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.

2. BiLSTM with GloVe Embeddings

- Glove stands for Global Vectors, a word embedding technique.
- We have used pre-trained Glove with 6B tokens and 50-dimensional representation for each vector.
- It creates real-valued vector representations for words based on the co-occurrence statistics of the words in a corpus.

Source (pre-trained glove) - http://nlp.stanford.edu/data/glove.6B.zip

BiLSTM with GloVe Model Summary & Results

```
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
_________________________________________________________________
```

Since we used pre-trained word embeddings, the non-trainable parameters have increased from 0 to 1,707,000.

	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

Observations

- The macro F1 score for the BiLSTM model with both assign integer value to words and tags and using GloVe embeddings is very less (~ 1.5%).
- **Reason** - 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.

Qualitative Analysis

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)

Conclusion

- We have explored two different approaches to named entity recognition (NER) - Conditional Random Fields (CRF) and Bi-directional Long Short-Term Memory (BiLSTM) models.
- Tried different feature sets for CRF and different preprocessing/word embedding techniques for BiLSTM.
- Presented Qualitative and Quantitative results for both Task 1 and Task 2. Through various observations, we concluded that CRF model has better qualitative and quantitative results on the given dataset.
- 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 and can better handle syntactic ambiguity through handcrafted features based on linguistic properties.
- Processing complex and ambiguous NEs is challenging; therefore, it is important to use models that are appropriate for the task.

References

- https://multiconer.github.io/
- Bajaj, Payal, et al. "Ms marco: A human generated machine reading comprehension dataset." arXiv preprint arXiv:1611.09268 (2016).
- 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.
- 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.

Thank you