Assignment – 2 Report

In this assignment, task 1, the two comparisons are done between RNN models with different number of layers, and different number of nodes.

Comparison based on Different Number of Layers:

In this comparison, we experimented with RNN models having different numbers of LSTM layers. The goal is to assess the impact of model depth on its ability to learn hierarchical representations and capture long-term dependencies in the data. The models have been trained with one layer (model m2 in our program), and two layers (model m1 in our program) to observe how adding layers influences the model's performance. This comparison helps determine the optimal depth for the dataset.

Comparison based on Different Number of Nodes:

In this comparison, we experimented with RNN models having different numbers of nodes (neurons) in the LSTM layers. The goal of this comparison is to observe how the performance of the models changes with varying neuron sizes. Different node sizes affect the capacity of the model to capture patterns in the data. The models have been trained with 10 neurons (model m2 in the program), 20 neurons (model m3 in the program) to analyse the impact on precision, recall, and F-measure metrics. This comparison helps understand the accommodation between model complexity and performance.

Results in term of precision, recall and f-measure for model m1 in program (eval1.txt):

Entity type	Precision	Recall	F-measure
geo	74.96	81.55	78.12
gpe	82.04	87.34	84.61
per	60.23	52.12	55.88
org	49.75	37.91	43.03
tim	68.34	60.59	64.24
art	Cannot be computed	0.0	Cannot be computed
nat	Cannot be computed	0.0	Cannot be computed
eve	Cannot be computed	0.0	Cannot be computed

Entity type	Precision	Recall	F-measure
Overall	69.31	65.59	67.4

Results in term of precision, recall and f-measure for model m2 in program (eval2.txt):

Entity type	Precision	Recall	F-measure
geo	82.82	84.03	83.42
gpe	94.66	92.64	93.64
per	72.41	71.84	72.12
org	69.23	59.41	63.94
tim	86.48	74.03	79.77
art	Cannot be computed	0.0	Cannot be computed
nat	Cannot be computed	0.0	Cannot be computed
eve	Cannot be computed	0.0	Cannot be computed

Entity type	Precision	Recall	F-measure
Overall	81.36	76.5	78.85

Results in term of precision, recall and f-measure for model m3 in program (eval3.txt):

Entity type	Precision	Recall	F-measure
geo	81.67	85.5	83.54
gpe	94.71	93.58	94.14
per	69.74	72.02	70.72
org	73.99	58.47	65.32
tim	79.3	76.95	78.11
art	0.0	0.0	Cannot be computed
nat	Cannot be computed	0.0	Cannot be computed
eve	Cannot be computed	0.0	Cannot be computed

Entity type	Precision	Recall	F-measure
Overall	80.09	77.58	78.81

In task 1, let's analyse the results for each model:

Model m1:

- **geo:** Moderate precision and recall, resulting in a reasonable F-measure.

- **gpe:** Good precision and recall, leading to a high F-measure.

- **per:** Lower precision but decent recall, contributing to a moderate F-measure.

- **org:** Low precision and recall, resulting in a low F-measure.

- **tim:** Decent precision and recall, leading to a reasonable F-measure.

- **art, nat, eve:** These categories couldn't be computed, likely due to no true positives (0 recall).

Overall:

- The overall F-measure is moderate, reflecting a balanced performance across entity types.

Model m2:

- **geo:** High precision and recall, resulting in a high F-measure.

- **gpe:** Very high precision and recall, leading to an excellent F-measure.

- **per:** Good precision and recall, contributing to a high F-measure.

- **org:** Moderate precision and recall, resulting in a reasonable F-measure.

- **tim:** Very high precision and decent recall, leading to a high F-measure.

- **art, nat, eve:** These categories couldn't be computed, likely due to no true positives (0 recall).

Overall:

- The overall F-measure is high, indicating strong performance across entity types.

Model m3:

- **geo:** High precision and recall, resulting in a high F-measure.

- **gpe:** Very high precision and recall, leading to an excellent F-measure.

- **per:** Moderate precision and recall, contributing to a moderate F-measure.

- **org:** Moderate precision and low recall, resulting in a moderate F-measure.

- **tim:** Good precision and recall, leading to a reasonable F-measure.

- **art, nat, eve:** These categories couldn't be computed, likely due to no true positives (0 recall).

Overall:

- The overall F-measure is moderate, indicating balanced performance across entity types.

The model m2 generally outperforms m1 and m3 across most entity types. Both m2 and m3 show better overall performance compared to m1. All models struggle with certain entity types (art, nat, eve), where no instances were correctly identified. The choice of the best model depends on the trade-offs between precision, recall, F-measure and accuracy of the models.

Error analysis from Task 2:

Error analysis of “eval1.txt”:

False Negative Error Analysis:

1.Missed entities:

The model failed to recognize the person entity 'Schoch' in the sentence.

The possible reason for this might be that the complexity of the sentence structure or the presence of uncommon names might have posed a challenge for the model.

```
Test example: 11
Sentence: ['Schoch', 'leads', 'the', 'World', 'Cup', 'standings', 'with', '1,700',
'points', '.']
Target: ['B-per', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']
Predicted: ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']

At 0 ('per', 'Schoch') Missed.
```

False Positive Error Analysis:

2.Incorrectly extracted entities:

The model failed to recognize the entities such as organization, time, and geopolitical entities.

The possible reason for this might be that the model might struggle with disambiguating between different entity types, especially when they co-occur in similar contexts.

```
Test example: 36
Sentence: ['The', 'High', 'Court', 'dropped', 'rape', 'charges', 'against', 'Besigye',
'last', 'month', ',', 'and', 'President', 'Museveni', 'said', 'Besigye', 'would', 'not',
'be', 'tried', 'by', 'a', 'military', 'court', 'that', 'had', 'charged', 'him', 'with',
'terrorism', 'and', 'illegal', 'arms', 'possession', '.']
Target: ['O', 'B-org', 'I-org', 'O', 'O', 'O', 'O', 'B-per', 'O', 'O', 'O', 'O', 'B-per',
'I-per', 'O', 'B-per', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O',
'O', 'O', 'O', 'O', 'O', 'O']
Predicted: ['O', 'B-org', 'B-tim', 'O', 'O', 'O', 'O', 'B-gpe', 'O', 'O', 'O', 'O', 'B-
per', 'I-per', 'O', 'B-gpe', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O',
'O', 'O', 'O', 'O', 'O', 'O', 'O']

At 1 ('org', 'High Court') Missed.
At 7 ('per', 'Besigye') Missed.
At 12 ('per', 'President Museveni') Extracted.
At 15 ('per', 'Besigye') Missed.
At 1 ('org', 'High') Incorrectly extracted.
At 2 ('tim', 'Court') Incorrectly extracted.
At 7 ('gpe', 'Besigye') Incorrectly extracted.
At 15 ('gpe', 'Besigye') Incorrectly extracted.
```

Error analysis of “eval1.txt” and “eval2.txt”:

3.Model variation analysis:

Analysis: The three models (m1, m2, and m3) have different architectures, including variations in the number of layers, and nodes.

Implication: These architectural differences can lead to variations in how the models capture and learn patterns, affecting their overall performance.

```
Test example: 48
Sentence: ['Another', 'Jordanian', 'man', ',', 'Shadi', 'Abdalla', 'was', 'convicted',
'separately', 'in', '2003', '.']
Target: ['O', 'O', 'O', 'O', 'B-per', 'I-per', 'O', 'O', 'O', 'O', 'B-tim', 'O']
Predicted: ['O', 'O', 'O', 'O', 'B-org', 'I-org', 'O', 'O', 'O', 'O', 'B-tim', 'O']

At 4 ('per', 'Shadi Abdalla') Missed.
At 10 ('tim', '2003') Extracted.
At 4 ('org', 'Shadi Abdalla') Incorrectly extracted.
```

```
Test example: 48
Sentence: ['Another', 'Jordanian', 'man', ',', 'Shadi', 'Abdalla', 'was', 'convicted',
'separately', 'in', '2003', '.']
Target: ['O', 'O', 'O', 'O', 'B-per', 'I-per', 'O', 'O', 'O', 'O', 'B-tim', 'O']
Predicted: ['O', 'O', 'O', 'O', 'B-per', 'I-per', 'O', 'O', 'O', 'O', 'B-tim', 'O']

At 4 ('per', 'Shadi Abdalla') Extracted.
At 10 ('tim', '2003') Extracted.
```

Error analysis of “eval3.txt”:

4.Contextual error:

The model might be misinterpreting the term based on superficial patterns, or it may have encountered inconsistencies in the training data regarding the entity type of 'Japanese.'

```
Test example: 23
Sentence: ['Gollnisch', ',', 'a', 'member', 'of', 'the', 'European', 'parliament', ',',
'received', 'a', 'five-year', 'suspension', 'from', 'his', 'post', 'as', 'a', 'professor',
'of', 'Japanese', 'at', 'Lyon', 'University', 'after', 'he', 'made', 'the', 'comments',
'at', 'a', 'press', 'conference', 'in', '2004', '.']
Target: ['B-per', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O',
'O', 'O', 'O', 'O', 'O', 'B-art', 'O', 'B-org', 'I-org', 'O', 'O', 'O', 'O', 'O', 'O', 'O',
'O', 'O', 'O', 'B-tim', 'O']
Predicted: ['B-per', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-tim', 'O', 'O',
'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-gpe', 'O', 'B-org', 'I-org', 'O', 'O', 'O', 'O', 'O', 'O',
'O', 'O', 'O', 'O', 'O', 'O']

At 0 ('per', 'Gollnisch') Extracted.
At 20 ('art', 'Japanese') Missed.
At 22 ('org', 'Lyon University') Extracted.
At 34 ('tim', '2004') Missed.
At 11 ('tim', 'five-year') Incorrectly extracted.
At 20 ('gpe', 'Japanese') Incorrectly extracted.
```

5. Multi-Word Entity Errors:

The model might struggle with recognizing entities that span multiple words, particularly when the context or semantic cues for the entire phrase are crucial. It may also be sensitive to the order of words in the training data. In this analysis it fails to recognise the “per” entity.

```
Test example: 26
Sentence: ['Pakistan', "'s", 'Information', 'Minister', 'Sheikh', 'Rashid', 'Ahmed',
'made', 'the', 'announcement', 'Tuesday', ',', 'saying', 'Pakistan', 'needs', 'his',
'leadership', '.']
Target: ['B-org', 'O', 'B-per', 'I-per', 'I-per', 'I-per', 'I-per', 'O', 'O', 'O', 'B-tim',
'O', 'O', 'B-geo', 'O', 'O', 'O', 'O']
Predicted: ['B-geo', 'O', 'B-org', 'I-per', 'I-per', 'I-per', 'I-per', 'O', 'O', 'O', 'B-
tim', 'O', 'O', 'B-geo', 'O', 'O', 'O', 'O']

At 0 ('org', 'Pakistan') Missed.
At 2 ('per', 'Information Minister Sheikh Rashid Ahmed') Missed.
At 10 ('tim', 'Tuesday') Extracted.
At 13 ('geo', 'Pakistan') Extracted.
At 0 ('geo', 'Pakistan') Incorrectly extracted.
At 2 ('org', 'Information') Incorrectly extracted.
```

Results and comments on Task 3:

In task 3 the model selection to evaluate from step 1 is done on the basis of the accuracy of the respective models.

Model 1 accuracy = 94.97%

```
390/400 [============================>.] - ETA: 1s - loss: 0.1753 - accuracy
391/400 [============================>.] - ETA: 1s - loss: 0.1753 - accuracy
392/400 [============================>.] - ETA: 1s - loss: 0.1752 - accuracy
393/400 [============================>.] - ETA: 1s - loss: 0.1750 - accuracy
394/400 [============================>.] - ETA: 1s - loss: 0.1749 - accuracy
395/400 [============================>.] - ETA: 0s - loss: 0.1747 - accuracy
396/400 [============================>.] - ETA: 0s - loss: 0.1746 - accuracy
397/400 [============================>.] - ETA: 0s - loss: 0.1745 - accuracy
398/400 [============================>.] - ETA: 0s - loss: 0.1744 - accuracy
399/400 [============================>.] - ETA: 0s - loss: 0.1743 - accuracy
400/400 [==============================] - ETA: 0s - loss: 0.1742 - accuracy
400/400 [==============================] - 67s 168ms/step - loss: 0.1742 - a
ccuracy: 0.9497
```

Model 2 accuracy = 97.63%

```
390/400 [============================>.] - ETA: 1s - loss: 0.0823 - accuracy
391/400 [============================>.] - ETA: 1s - loss: 0.0823 - accuracy
392/400 [============================>.] - ETA: 1s - loss: 0.0823 - accuracy
393/400 [============================>.] - ETA: 1s - loss: 0.0823 - accuracy
394/400 [============================>.] - ETA: 1s - loss: 0.0822 - accuracy
395/400 [============================>.] - ETA: 0s - loss: 0.0822 - accuracy
396/400 [============================>.] - ETA: 0s - loss: 0.0822 - accuracy
397/400 [============================>.] - ETA: 0s - loss: 0.0822 - accuracy
398/400 [============================>.] - ETA: 0s - loss: 0.0822 - accuracy
399/400 [============================>.] - ETA: 0s - loss: 0.0822 - accuracy
400/400 [==============================] - ETA: 0s - loss: 0.0822 - accuracy
400/400 [==============================] - 68s 169ms/step - loss: 0.0822 - a
ccuracy: 0.9763
```

Model 3 accuracy = 97.58%

```
390/400 [============================>.] - ETA: 1s - loss: 0.0823 - accuracy
391/400 [============================>.] - ETA: 1s - loss: 0.0823 - accuracy
392/400 [============================>.] - ETA: 1s - loss: 0.0823 - accuracy
393/400 [============================>.] - ETA: 1s - loss: 0.0823 - accuracy
394/400 [============================>.] - ETA: 1s - loss: 0.0823 - accuracy
395/400 [============================>.] - ETA: 0s - loss: 0.0823 - accuracy
396/400 [============================>.] - ETA: 0s - loss: 0.0823 - accuracy
397/400 [============================>.] - ETA: 0s - loss: 0.0823 - accuracy
398/400 [============================>.] - ETA: 0s - loss: 0.0823 - accuracy
399/400 [============================>.] - ETA: 0s - loss: 0.0823 - accuracy
400/400 [==============================] - ETA: 0s - loss: 0.0823 - accuracy
400/400 [==============================] - 69s 171ms/step - loss: 0.0823 - a
ccuracy: 0.9758

(base) C:\Users\Neeraja Vudhanthi\AppData\Roaming\Microsoft\Windows\Start Me
nu\Programs\Python 3.10>
```

Since the model m2 has a higher accuracy we select m2 to evaluate on our own sentences.

Note: In our program, for better efficiency of the code during the run time, we create model m4 which is similar to m2.

After evaluation of the model m2 on our own sentences, the accuracy is **76.67%** .

```
Epoch 1/2
1/5 [=====>.....................] - ETA: 16s - loss: 2.5679 - accuracy: 0
4/5 [=======================>......] - ETA: 0s - loss: 2.5627 - accuracy: 0.
5/5 [==============================] - 4s 16ms/step - loss: 2.5606 - accurac
y: 0.2533
Epoch 2/2
1/5 [=====>.....................] - ETA: 0s - loss: 2.5495 - accuracy: 0.
5/5 [==============================] - ETA: 0s - loss: 2.5328 - accuracy: 0.
5/5 [==============================] - 0s 16ms/step - loss: 2.5328 - accurac
y: 0.7667

(base) C:\Users\Neeraja Vudhanthi\AppData\Roaming\Microsoft\Windows\Start Me
nu\Programs\Python 3.10>
```

The below are the results in the "eval4.txt" after task 3.

Test example: 0

Sentence: ['Tokyo', 'hosted', 'the', 'Summer', 'Olympics', 'in', '2020', '.']

Target: ['B-loc', 'O', 'O', 'B-eve', 'I-eve', 'O', 'B-tim', 'O']

Predicted: ['B-tim', 'B-tim', 'B-tim', 'B-tim', 'B-tim', 'O', 'O', 'O']

At 0 ('loc', 'Tokyo') Missed.

At 3 ('eve', 'Summer Olympics') Missed.

At 6 ('tim', '2020') Missed.

At 0 ('tim', 'Tokyo') Incorrectly extracted.

At 1 ('tim', 'hosted') Incorrectly extracted.

At 2 ('tim', 'the') Incorrectly extracted.

At 3 ('tim', 'Summer') Incorrectly extracted.

At 4 ('tim', 'Olympics') Incorrectly extracted.

Test example: 1

Sentence: ['The', 'Eiffel', 'Tower', 'is', 'located', 'in', 'Paris', '.']

Target: ['O', 'B-loc', 'I-loc', 'O', 'O', 'O', 'B-loc', 'O']

Predicted: ['B-tim', 'B-tim', 'O', 'O', 'O', 'O', 'O', 'O']

At 1 ('loc', 'Eiffel Tower') Missed.

At 6 ('loc', 'Paris') Missed.

At 0 ('tim', 'The') Incorrectly extracted.

At 1 ('tim', 'Eiffel') Incorrectly extracted.

Test example: 2

Sentence: ['Jane', 'works', 'for', 'Microsoft', 'in', 'Seattle', '.']

Target: ['B-per', 'O', 'O', 'B-org', 'O', 'B-loc', 'O']

Predicted: ['B-tim', 'B-tim', 'B-tim', 'B-tim', 'B-tim', 'B-loc', '_PAD_']

At 0 ('per', 'Jane') Missed.

At 3 ('org', 'Microsoft') Missed.

At 5 ('loc', 'Seattle') Extracted.

At 0 ('tim', 'Jane') Incorrectly extracted.

At 1 ('tim', 'works') Incorrectly extracted.

At 2 ('tim', 'for') Incorrectly extracted.

At 3 ('tim', 'Microsoft') Incorrectly extracted.

At 4 ('tim', 'in') Incorrectly extracted.

Test example: 3

Sentence: ['Mozart', 'composed', 'many', 'classics', '.']

Target: ['B-per', 'O', 'O', 'O', 'O']

Predicted: ['B-tim', 'B-tim', 'B-tim', 'B-tim', 'O']

At 0 ('per', 'Mozart') Missed.

At 0 ('tim', 'Mozart') Incorrectly extracted.

At 1 ('tim', 'composed') Incorrectly extracted.

At 2 ('tim', 'many') Incorrectly extracted.

At 3 ('tim', 'classics') Incorrectly extracted.

Test example: 4

Sentence: ['The', 'Mona', 'Lisa', 'is', 'displayed', 'at', 'the', 'Louvre', 'Museum', 'in', 'Paris', '.']

Target: ['O', 'B-per', 'I-per', 'O', 'O', 'O', 'O', 'B-loc', 'I-loc', 'O', 'B-loc', 'O']

Predicted: ['B-tim', 'B-tim', 'B-tim', 'B-tim', 'B-tim', 'O', 'O', 'B-tim', 'O', 'O', 'O', 'O']

At 1 ('per', 'Mona Lisa') Missed.

At 7 ('loc', 'Louvre Museum') Missed.

At 10 ('loc', 'Paris') Missed.

At 0 ('tim', 'The') Incorrectly extracted.

At 1 ('tim', 'Mona') Incorrectly extracted.

At 2 ('tim', 'Lisa') Incorrectly extracted.

At 3 ('tim', 'is') Incorrectly extracted.

At 4 ('tim', 'displayed') Incorrectly extracted.

At 7 ('tim', 'Louvre') Incorrectly extracted.

Results in term of precision, recall and f-measure task 3 (eval4.txt):

Entity type	**Precision**	**Recall**	**F-measure**
loc	100.0	16.67	28.57
org	Cannot be computed	0.0	Cannot be computed
eve	Cannot be computed	0.0	Cannot be computed
per	Cannot be computed	0.0	Cannot be computed
tim	0.0	0.0	Cannot be computed

Entity type	**Precision**	**Recall**	**F-measure**
Overall	4.35	8.33	5.71

From the above the below summarizations can be made:

1. **Imbalance in Precision and Recall:** The high precision for the 'loc' entity type suggests that when the model predicts a location, it tends to be accurate. However, the low recall indicates that it misses many actual locations, resulting in an imbalance between precision and recall.
2. **Challenges Across Multiple Entity Types:** The inability to compute metrics for 'org,' 'eve,' 'per,' and 'tim' indicates serious challenges in the model's understanding and classification of these entity types.
3. **Room for Improvement:** The low overall performance metrics underscore the need for further model refinement. Strategies such as additional training data, fine-tuning, or exploring more advanced architectures could help address these issues.
4. **Consideration of Entity Type Importance:** Depending on the specific application, the importance of different entity types may vary. Understanding the context and criticality of each entity type is crucial for refining the model effectively.

In summary, the results highlight specific areas where the model struggles, and further optimization is required to enhance its capability to accurately identify and classify entities, especially for less represented types ('org,' 'eve,' 'per,' 'tim').