

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**  
**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**  
**KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN**

---



**MÔN XỬ LÝ NGÔN NGỮ TỰ NHIÊN CHO KHOA HỌC DỮ LIỆU**  
**BÀI TẬP THỰC HÀNH TUẦN 2**

**GVHD:** ThS. Huỳnh Văn Tín

Sinh viên thực hiện:

Lê Ngọc Thiên Phúc – 22521117

๑๑๑ Tp. Hồ Chí Minh, 09/2024 ๑๑๑

**Câu 1:** Sử dụng mô hình LSTM để giải quyết bài toán NER cho tiếng Anh.

- Đánh giá mô hình theo độ đo F1-score (micro và macro):

F1-score micro test set: 80.53%  
F1-score macro test set: 54.79%

- Kết quả của F1-score và Accuracy cho từng nhãn thực thể:

	precision	recall	f1-score	support
B-art	0.00	0.00	0.00	140
B-eve	1.00	0.02	0.03	113
B-geo	0.85	0.89	0.87	11090
B-gpe	0.95	0.94	0.95	4755
B-nat	1.00	0.06	0.12	62
B-org	0.81	0.66	0.73	5952
B-per	0.85	0.78	0.82	5052
B-tim	0.92	0.86	0.89	6118
I-art	0.00	0.00	0.00	128
I-eve	0.17	0.03	0.05	101
I-geo	0.80	0.78	0.79	2194
I-gpe	0.88	0.38	0.53	60
I-nat	0.00	0.00	0.00	17
I-org	0.80	0.75	0.78	4917
I-per	0.86	0.83	0.85	5093
I-tim	0.85	0.70	0.77	1950
O	0.99	1.00	1.00	671658
accuracy			0.99	719400
macro avg	0.69	0.51	0.54	719400
weighted avg	0.98	0.99	0.98	719400

Accuracy test = 98.54%		
	Label	Accuracy
0	B-art	0.000000
1	B-eve	0.017699
2	B-geo	0.892606
3	B-gpe	0.936698
4	B-nat	0.064516
5	B-org	0.664987
6	B-per	0.783056
7	B-tim	0.858450
8	I-art	0.000000
9	I-eve	0.029703
10	I-geo	0.775752
11	I-gpe	0.383333
12	I-nat	0.000000
13	I-org	0.749644
14	I-per	0.833890
15	I-tim	0.702564
16	O	0.998067

**Câu 2:** Áp dụng mô hình trên để giải quyết bài toán NER cho tiếng Việt.

- Đánh giá mô hình theo độ đo F1-score (micro và macro):

```
F1-score micro test set: 85.88%
F1-score macro test set: 79.95%
```

- Kết quả của F1-score và Accuracy cho từng nhãn thực thể:

	precision	recall	f1-score	support
B-AGE	0.91	0.93	0.92	568
B-DATE	0.98	0.96	0.97	1640
B-GENDER	0.88	0.94	0.91	447
B-JOB	0.62	0.55	0.58	172
B-LOCATION	0.92	0.90	0.91	4425
B-NAME	0.91	0.55	0.69	317
B-ORGANIZATION	0.88	0.78	0.83	769
B-PATIENT_ID	0.96	0.91	0.93	1978
B-SYMPTOM_AND_DISEASE	0.91	0.82	0.86	1134
B-TRANSPORTATION	0.95	0.82	0.88	193
I-AGE	0.00	0.00	0.00	6
I-DATE	0.96	0.99	0.98	1726
I-JOB	0.43	0.17	0.24	114
I-LOCATION	0.92	0.87	0.89	4892
I-NAME	0.00	0.00	0.00	13
I-ORGANIZATION	0.90	0.80	0.84	2007
I-PATIENT_ID	0.00	0.00	0.00	27
I-SYMPTOM_AND_DISEASE	0.91	0.77	0.83	1479
I-TRANSPORTATION	0.79	0.59	0.68	69
0	0.99	1.00	0.99	214584
accuracy			0.98	236560
macro avg	0.74	0.67	0.70	236560
weighted avg	0.98	0.98	0.98	236560

Accuracy test = 98.42%

	Label	Accuracy
0	B-AGE	0.926056
1	B-DATE	0.964634
2	B-GENDER	0.939597
3	B-JOB	0.546512
4	B-LOCATION	0.899435
5	B-NAME	0.552050
6	B-ORGANIZATION	0.782835
7	B-PATIENT_ID	0.908999
8	B-SYMPTOM_AND_DISEASE	0.821869
9	B-TRANSPORTATION	0.818653
10	I-AGE	0.000000
11	I-DATE	0.991889
12	I-JOB	0.166667
13	I-LOCATION	0.867539
14	I-NAME	0.000000
15	I-ORGANIZATION	0.795715
16	I-PATIENT_ID	0.000000
17	I-SYMPTOM_AND_DISEASE	0.768087
18	I-TRANSPORTATION	0.594203
19	O	0.996407

**Câu 3:** Cải thiện kết quả của câu 1, 2 bằng cách cài đặt thêm các mô hình khác.

❖ **Cải thiện câu 1 bằng GRU:**

- **Đánh giá mô hình theo độ đo F1-score (micro và macro):**

F1-score micro test set: 81.06%  
F1-score macro test set: 56.13%

- **Kết quả của F1-score và Accuracy cho từng nhãn thực thể:**

	precision	recall	f1-score	support
B-art	0.00	0.00	0.00	140
B-eve	0.83	0.26	0.39	113
B-geo	0.87	0.88	0.87	11090
B-gpe	0.94	0.94	0.94	4755
B-nat	1.00	0.03	0.06	62
B-org	0.80	0.69	0.74	5952
B-per	0.85	0.79	0.82	5052
B-tim	0.93	0.85	0.89	6118
I-art	0.00	0.00	0.00	128
I-eve	0.47	0.07	0.12	101
I-geo	0.79	0.78	0.79	2194
I-gpe	0.90	0.45	0.60	60
I-nat	0.00	0.00	0.00	17
I-org	0.78	0.77	0.78	4917
I-per	0.85	0.87	0.86	5093
I-tim	0.85	0.71	0.78	1950
O	0.99	1.00	1.00	671658
accuracy			0.99	719400
macro avg	0.70	0.53	0.57	719400
weighted avg	0.98	0.99	0.99	719400

Accuracy test = 98.58%

	Label	Accuracy
0	B-art	0.000000
1	B-eve	0.256637
2	B-geo	0.878990
3	B-gpe	0.941115
4	B-nat	0.032258
5	B-org	0.686996
6	B-per	0.792953
7	B-tim	0.850768
8	I-art	0.000000
9	I-eve	0.069307
10	I-geo	0.780766
11	I-gpe	0.450000
12	I-nat	0.000000
13	I-org	0.766728
14	I-per	0.868840
15	I-tim	0.714872
16	O	0.997945

❖ Cải thiện câu 2 bằng GRU:

- Đánh giá mô hình theo độ đo F1-score (micro và macro):

F1-score micro test set: 85.36%  
F1-score macro test set: 78.76%

- Kết quả của F1-score và Accuracy cho từng nhãn thực thể:

	f1-score	support	Accuracy test = 98.38%	
			Label	Accuracy
B-AGE	0.93	568	0 B-AGE	0.964789
B-DATE	0.97	1640	1 B-DATE	0.966463
B-GENDER	0.93	447	2 B-GENDER	0.950783
B-JOB	0.55	172	3 B-JOB	0.488372
B-LOCATION	0.91	4425	4 B-LOCATION	0.886554
B-NAME	0.69	317	5 B-NAME	0.558360
B-ORGANIZATION	0.85	769	6 B-ORGANIZATION	0.837451
B-PATIENT_ID	0.93	1978	7 B-PATIENT_ID	0.911527
B-SYMPTOM_AND_DISEASE	0.87	1134	8 B-SYMPTOM_AND_DISEASE	0.828924
B-TRANSPORTATION	0.83	193	9 B-TRANSPORTATION	0.756477
I-AGE	0.00	6	10 I-AGE	0.000000
I-DATE	0.97	1726	11 I-DATE	0.990151
I-JOB	0.10	114	12 I-JOB	0.052632
I-LOCATION	0.89	4892	13 I-LOCATION	0.830335
I-NAME	0.00	13	14 I-NAME	0.000000
I-ORGANIZATION	0.85	2007	15 I-ORGANIZATION	0.824614
I-PATIENT_ID	0.00	27	16 I-PATIENT_ID	0.000000
I-SYMPTOM_AND_DISEASE	0.83	1479	17 I-SYMPTOM_AND_DISEASE	0.759973
I-TRANSPORTATION	0.49	69	18 I-TRANSPORTATION	0.347826
O	0.99	214584	19 O	0.996705
accuracy	0.98	236560		
macro avg	0.68	236560		
weighted avg	0.98	236560		