CS563-NLP

Assignment 2: NER

Group Name: **1801cs31_1801cs33**
Students:

Names	**Roll No.**	**Batch**
M **Maheeth Reddy**	1801CS31	B.Tech.
Nischal A	1801CS33	B.Tech.

Solution:

1. Without Context Emission Probabilities
   - ➔ Bigram - *bigram_tagger_no_context.py*
   - ➔ Trigram - *python trigram_tagger_no_context.py*
2. Without Context Emission Probabilities
   - ➔ Bigram - *python bigram_tagger_with_context.py*
   - ➔ Trigram - *python trigram_tagger_with_context.py*

Each file has two options to run: -

Choice 1 – BIO Tags

Choice 2 – Fine-Grained NER Tags

NOTE: For dealing with unknown words during the test time, we first encode all words of the train document which occur less than 5 times using "$RARE$" symbol. We learn the "$RARE$" transition and emission probabilities. We replace every unknown word of the test set by the "$RARE$" token.

The results obtained are discussed from the next page. The classification matrix which includes the class-wise and overall precision, recall and F1 score along with the accuracy is included for each run.

Without Using Context Results

N-Gram Used	Type	Macro Average F1 Score	Accuracy (%)
Bigram	BIO Tags	0.36	91
	Fine-Grained NER Tags	0.11	89
Trigram	BIO Tags	0.38	90
	Fine-Grained NER Tags	0.11	90

Classification Matrices

BIGRAM

	precision	recall	f1-score	support
B	0.67	0.03	0.07	3473
I	0.49	0.03	0.06	2482
O	0.91	1.00	0.95	55941
accuracy			0.91	61896
macro avg	0.69	0.36	0.36	61896
weighted avg	0.88	0.91	0.87	61896

	precision	recall	f1-score	support
O	0.91	0.98	0.94	55941
company	0.78	0.04	0.08	886
facility	0.00	0.00	0.00	619
loc	0.77	0.02	0.04	1101
movie	0.00	0.00	0.00	82
musicartist	0.00	0.00	0.00	331
other	0.17	0.09	0.11	1140
person	0.11	0.02	0.03	782
product	0.00	0.00	0.00	746
sportsteam	0.00	0.00	0.00	195
tvshow	0.00	0.00	0.00	73
accuracy			0.89	61896
macro avg	0.25	0.10	0.11	61896
weighted avg	0.85	0.89	0.86	61896

TRIGRAM

	precision	recall	f1-score	support
B	0.69	0.04	0.08	3473
I	0.31	0.06	0.10	2482
O	0.91	0.99	0.95	55941
accuracy			0.90	61896
macro avg	0.64	0.36	0.38	61896
weighted avg	0.87	0.90	0.87	61896

	precision	recall	f1-score	support
O	0.91	0.99	0.95	55941
company	0.78	0.04	0.08	886
facility	0.00	0.00	0.00	619
loc	0.79	0.02	0.04	1101
movie	0.00	0.00	0.00	82
musicartist	0.00	0.00	0.00	331
other	0.13	0.07	0.09	1140
person	0.21	0.02	0.04	782
product	0.00	0.00	0.00	746
sportsteam	0.00	0.00	0.00	195
tvshow	0.00	0.00	0.00	73
accuracy			0.90	61896
macro avg	0.26	0.10	0.11	61896
weighted avg	0.85	0.90	0.86	61896

Comparing Bigram and Trigram: We can see that some classes which are under-represented have very low F1 scores as the HMM is not able to learn them efficiently. For this, we believe that we must compare the Macro Average F1 scores of the bigram and the trigram models. We observe an increase in the class-wise F1 scores of the under-represented classes as well as an increase in the macro-average F1 score from bigram to trigram for both BIO tags as well as Fine-Grained tagging. Hence, we can say that in this case, trigram model is better than the bigram one.

With Using Context Results

N-Gram Used	Type	Macro Average F1 Score	Accuracy (%)
Bigram	BIO Tags	0.42	90
	Fine-Grained NER Tags	0.13	87
Trigram	BIO Tags	0.44	89
	Fine-Grained NER Tags	0.14	88

Classification Matrices

BIGRAM

	precision	recall	f1-score	support
B	0.47	0.07	0.12	3473
I	0.40	0.11	0.18	2482
O	0.91	0.99	0.95	55941
accuracy			0.90	61896
macro avg	0.59	0.39	0.42	61896
weighted avg	0.87	0.90	0.87	61896

	precision	recall	f1-score	support
O	0.91	0.96	0.94	55941
company	0.49	0.04	0.07	886
facility	0.01	0.03	0.02	619
loc	0.74	0.04	0.07	1101
movie	0.01	0.02	0.02	82
musicartist	0.00	0.00	0.00	331
other	0.18	0.15	0.16	1140
person	0.15	0.04	0.06	782
product	0.05	0.01	0.02	746
sportsteam	0.19	0.04	0.06	195
tvshow	0.00	0.00	0.00	73
accuracy			0.87	61896
macro avg	0.25	0.12	0.13	61896
weighted avg	0.85	0.87	0.85	61896

TRIGRAM

	precision	recall	f1-score	support
B	0.42	0.10	0.16	3473
I	0.29	0.16	0.21	2482
O	0.91	0.98	0.94	55941
accuracy			0.89	61896
macro avg	0.54	0.41	0.44	61896
weighted avg	0.86	0.89	0.87	61896

	precision	recall	f1-score	support
O	0.91	0.97	0.94	55941
company	0.42	0.04	0.07	886
facility	0.12	0.04	0.06	619
loc	0.36	0.04	0.07	1101
movie	0.02	0.02	0.02	82
musicartist	0.00	0.00	0.00	331
other	0.22	0.14	0.17	1140
person	0.19	0.04	0.06	782
product	0.03	0.03	0.03	746
sportsteam	0.33	0.04	0.07	195
tvshow	0.00	0.00	0.00	73
accuracy			0.88	61896
macro avg	0.24	0.12	0.14	61896
weighted avg	0.85	0.88	0.86	61896

Comparing Bigram and Trigram: We observe an increase in the class-wise F1 scores of the under-represented classes as well as an increase in the macro-average F1 score from bigram to trigram for both BIO tags as well as Fine-Grained tagging. Hence, we can say that in this case, trigram model is better than the bigram one.

Comparing Without and With Context: We see a consistent increase in the Macro-Average F1 score using context as compared to without context. The F1 scores of the under-represented classes have also increased. We see 6%, 2%, 6% and 3% improvement respectively on Macro-Average F1 score for Bigram BIO, Bigram Fine Grained, Trigram BIO, Trigram Fine-grained respectively. Hence, using context for emission probability performs better.