CS 6320.001: Natural Language Processing
Fall 2019
Project 2 Evaluation sheet

Name: Rishita Bansal
NetID: rxb180044

Task no	Points	Comments
1		
2		
3		
4		
5		
6		
7		
Report		
TOTAL		

Name Entity Recognition

Rishita Bansal
The University of Texas at Dallas,
Richardson, TX- 75080 USA
rishita.bansal@utdallas.edu

Abstract
Whenever we are given some sentence and we want to determine the meaning of the sentence, it can sometimes be difficult due to ambiguities in the meanings of some words. We can make this easy if we can find the context of the sentence. We can find context if we have labels telling us the category a word belongs to. We can also have different parameters as features to predict the category of the words.

Introduction
We have a dataset where every word is labelled by the category it belongs to. We are also given POS tags. We can add more features by getting hypernyms, meronyms etc for all the words. We can then train the data using Logistic Regression and test it on test data and get the accuracy scores.

TASK 1&2

Pre-processing of data
In preprocessing we rename the column names, handle nan values and create columns for our new features and initialize them. We also convert the the columns into object type so that inserting into the columns gets easy.

Adding Features
To get better accuracy of prediction, we can add extra features to the data. Here I have added Synonyms, Antonyms, Hypernyms, Hyponyms, Holonyms and Meronyms.

Training data

We vectorize the name tags by giving integer values to the tags. We then pass the training data into the Logistic Regression classifier.

Obeservations from the Models

Tags	precision	recall	f1-score	support
B-LOC	0.91	0.88	0.89	2075
B-MISC	0.91	0.79	0.85	934
B-ORG	0.89	0.59	0.71	1211
B-PER	0.84	0.81	0.82	1754
I-LOC	0.84	0.71	0.77	276
I-MISC	0.82	0.61	0.70	323
I-ORG	0.88	0.62	0.73	814
I-PER	0.67	0.97	0.79	1309
O	0.99	1.00	0.99	41304
micro avg	0.96	0.96	0.96	50000
macro avg	0.86	0.78	0.81	50000
weighted avg	0.96	0.96	0.96	50000

Accuracy Score: 0.95956

TASK 3

Pre-processing of data:

Here we first rename the headers and then drop the unwanted columns as we need only the words and the name entity tags.

Getting Vocabulary:

We get the list of unique words and unique tags to get a vocabulary of words and tags present in our dataset.

Making Dictionary:

We need to assign a number to all the unique words and tags and make a dictionary which will be used for encoding.

Encoding data:

Based on the dictionary obtained in the previous step we need to encode our data by replacing words and tags with their corresponding numbers.

Vectorizing Data:

Based on the pre-trained word-vec model given to us, we need to get tensors for all the words in our data.

Training Data:

We load all the data and create a data loader. Once our data loader is ready, we will start training our neural network.

Conclusion

We trained model to predict name entities for given words. We trained our model using 2 methods and found out the accuracy on test set.

References

[1] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Proceedings of the Seventh Conference on Natural Language Learning at HLTNAACL 2003, pages 142{147, 2003.

[2] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity. In Proceedings of ACL, pages 271–278, 2004.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.