CE807 – Text Analytics

Assignment 1: Using distant learning for Named Entity Recognition

Task: 4

Siddharth Shyamsunder: ss18812

Prior Analysis:

From our past analysis in task 1, we looked into the needs of a more refined approach of training data through distant learning, to remove the cumbersome latency and labor costs in manually annotating the NER information on all the data. From earlier, we discussed the need of evaluation measures, which can identify the efficiency of our system. Some of the critical claims for which we found some new results include making use of fewer grams in terms of n grams while training data, and the staleness nature of the fact that increasing the dataset size would lead to better performance.

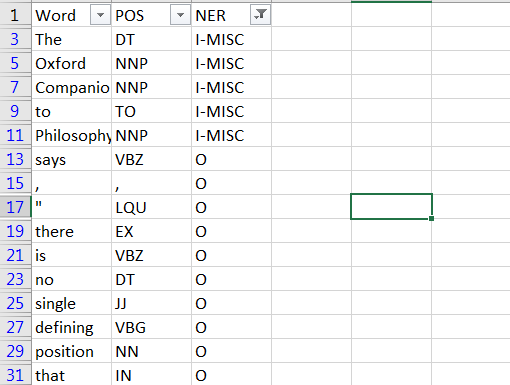
Experiments:

We have made use of the wikiner datasetm the standard dataset consisting of values taken from Wikipedia websites. The aim of our experiment is to train the information from the Wikiner website and once we have got an accuracy value, we can make use of the training to test the CoNLL dataset, and determine the accuracy in predictions.

In Task 2, we looked into training the model. For the beginning, we had a thorough analysis of the entire training dataset. The training dataset resulted in a memory loss due to the huge dimensionality of the dataset. Hence we made use of a part of the dataset to test our evaluation.

In our first step, we tokenized the information provided in the dataset, and identified the dataset to hold 3 values, the word, itd NER Tagger, and its POS Tagger. We then created a matrix of recordsx3 to get the feature matrix. In our case, the Word column plays an important role. The NER would be the target labels.

As we have decided based on our analysis to reduce the number of grams on the n gram model, we have dropped off the POS tag values, as the POS tag values only find meaning in the setence structuring not word structuring. In doing so we had a structure as such:



In our next step, we perform the same few steps on the test set. Although our unique test will not be used until the model has been trained, we need the words for encoding the output feature list.

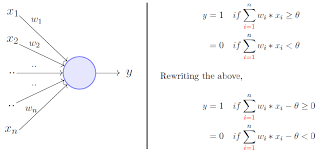
Now as the machine learning techniques that we use belong to scikit learn and scikit learn does not support categorical data like words, we made use of one hot encoding to create a feature matrix of all words, where should the text appear in a row, the word corresponding to that row and the word column would become 1, similarly, we trained the labels to be converted to numerical value using label encoder.

Distant Learning Algorithm used:

The Machine Learning technique that we used in our algorithm was the perceptron model. Yje biggest reason that made us choose this algorithm over many other algorithms was this classifier algorithm’s sheer size to handle hugr datasets.

Another advantage of using the perceptron model especially in the case of distant supervision is this model is a oassively aggressive model, in other words, this model focuses not on learning from what it predicts correctly, but rather, this document learns on what part it predicts wrong. In a more technical term, the Perceptron learns from its false positives, false negatives and true negatives than from true positives.

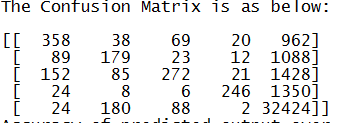
The perceptron model works on the concept of treshhold frequency. Wherein this model takes in the input and decides the output based on a particular treshholdfrequency range.

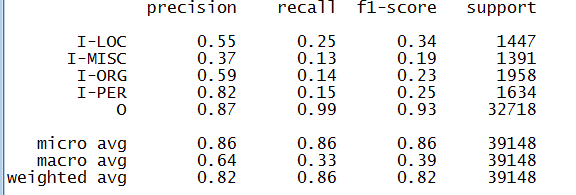


Once we use the perceptron model, we get an accuracy of 90.81% on 30000, which is clearly good in terms of a cross validated accuracy.

Then we move on to task 3, where we use our perceptron modelled classifier to work on the CoNLL dataset. And once we predicted the output we compared the generated output with that of the previously manually annotated output and we received an 85.5% accuracy in the prediction.

We then generate the classification report and confusion matrix as follows.





Evaluation:

One factor that corrected our assumption of Mikheev’s statement that the dataset size does not depend on performance can be seen in this evaluation. When, we initially tested the evaluation on very few records, we had a high accuracy but a low percentage of precision and recall, furthermore, when we tested between 250 to 500 records, there was a sudden increase of 80% in the total accuracy between the true output and tested output. This shows the idea of localization of data, and hence if a dataset is highly localised or rather a variety if data is assigned in the 1st few records, then there can be an efficient output even despite of taking small number of records.

As can be seen from the classification report the weighted average of the precision, recall and F measure go above 80, which means, using distant learning techniques do perform a higher f measure increase as compared to standard manually annotated ethods.

Since we made use of unigrams, we avoided the need to include embeddings like Word-Vec on the data, which would have caused many redundant features.

Similsrily, using this method, NER can be calculated for other languages as well.

Conclusion:

From our experiments, we can see that using distant learning, a corpus of 39000 records can be annotated within minutes, in comparison to earlier manual labor.

Also we discovered the idea of localization, wherein most features would only be associated with part of the data.

Future scope can be included in building a distant learning model to associate words for languages aside from English.