CS59000: Machine Learning for Natural Language Processing

Homework 2

November 22, 2017

1 Task

A Named Entity Recognition Task (NER) is performed in this homework. The sentences are labelled with beginning, inside and outside. Moreover, there are 127 different labels. This is a sequential prediction problem. An MEMM model is implemented to predict the next state given the previous state, current word. The problem is modelled as a supervised label prediction problem.

1.1 Features

Each of the words and labels are assigned a random embeddings of different length.

1. Size of word embeddings: 100
2. Size of label embeddings: 20
3. Input features consisted of embeddings of previous label, current word, previous word, and the next word

1.2 MLP Configuration

1. λ (Regularization parameter) = 0.03
2. α (learning rate) = 0.1
3. Stopping criteria : max_iterations = 1000 ; min_thresh b/w weights = $1e-2$
4. No. of features = $20 + 3 * 100 = 320$
5. No. of instances in each batch = 1000
6. Momentum = 0.9
7. No. of hidden layers = 1 ; No. of nodes in the hidden layer = 200

2 Results

The results obtained on the test data are the following:

1. Precision: 0.57
2. Recall: 0.61
3. F1 Score: 0.59

3 Comments

1. Important features are the word embeddings corresponding to not only the current word but also the previous and next words in the sequence. This required having two dummy embeddings for start and end of the sequence for both in the case of labels and the words.
2. The idea of assigning embeddings randomly is counter intuitive in the sense that the words that have similar meaning or the words that correspond to the nouns do not lie closer in the ND space. This might have drastically effected the F1 scores of performance.
3. Utilizing more informative word embeddings either obtained from Glove or word2vec can significantly improve the performance.