Lab3 Report

fmonteroperez1 - aca15fm

For this lab our task was to learn a named entity recogniser using the structured perceptron algorithms. Our perceptron given some words ans some weights will predict the label (Person, Location, Organisation, Misc) for each word

1 Structured Perceptron Implementation

Similar to the binary perceptron, we had to implement train, test and prediction functions as well as functions for feature extraction. Feature extraction functions come in pairs, one function analyzes the corpus an extracts the feature counts from the corpus, the other one (Phi) given a sentence and a set of labels, returns the counts of keys in the dictionary provided by the previous function.

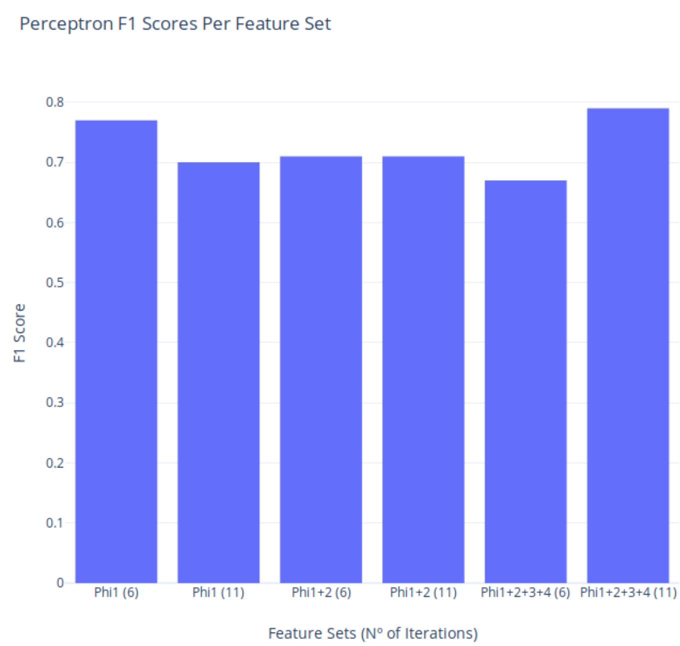

Figure 1

2 Feature Selection And Extraction

Two feature sets were required per the assignment brief, this is current_word - current_label (Phi 1) and previous_label - current_label (Phi 2). On top of this I have implemented functions to extract the following features: current_suffix - current_label (Phi 3) and previous_label - current_suffix (Phi 4).

I chose this features because I believed it would be interesting to see the effect of the suffix on the labels given our training corpus.

Once the perceptron was implemented, it can be executed in three different modes, one which only uses Phi1 for training and testing, one that uses Phi1+Phi2 and finally one that users all the Phis 1 to 4.

3 Results

With the perceptron setup to loop over the training data and to average the weights as iterates, the following results were obtained when doing 1, 6 and 11 iterations.

Features Selected	**Setup**	**F1-Score**
Phi1	1 iteration	0.75
Phi1	6 iterations averaged	0.77
Phi1	11 iterations averaged	0.70
Phi1+2	1 iteration	0.51
Phi1+2	6 iterations averaged	0.71
Phi1+2	11 iterations averaged	0.71
Phi1+2+3+4	1 iteration	0.68
Phi1+2+3+4	6 iterations averaged	0.67
Phi1+2+3+4	11 iterations averaged	0.79

Initially I was surprised about how good Phi1 was on its own, this could be because our training data and test data are both from the same domain so many of the words would be repeated thus our perceptron would classify it as the label its associated to that word. We will be using this as the base for comparing other feature sets.

For this feature, the highest weighted features are numbers with O tags and a name tagged as PER, this could be expected as our data is sports related and contains many scores and dates.

The second step was to analyse the perceptron using Phi1 in combination with Phi2. The initial scores for this feature set was pretty bad, but after iterating over the data set multiple times we managed to improve the F1 score although it was still lower than that for Phi1. So this feature sets are not as good as the previous one.

This feature sets top weighted features include names of various locations tagged as LOC and and some scores without recognisable tags.

Finally our last feature set which is based around words and suffixes whilst being worst than Phi1 only at the beginning, after 11 iterations we manage to obtain a F1 Score superior to that of the first feature we selected so we can say that this model is better than the other ones.

I believe the features I suggested improved the F1 Score of the perceptron because many times the suffix of a word can be changed based on what word comes before so this feature adds more data to the perceptron without adding too much noise (unlike the previous feature set) thus the F1 Score is higher.

For the top ranked features, we can see multiple tags from the last two Phi's which further supports that this feature sets add value to the classifier without adding too much noise.