MINI1

SHIVANGI MAHTO
UT AUSTIN, TEXAS
shivangi@cs.utexas.edu

Abstract

In this paper, a Named-entity recognition (NER) scheme has been presented for localizing persons' name in a given text. We studied various kind of feature extractions for logistic regression and could achieve up to 91% F1 on development set with 18 kinds of token features. Additionally, 88% F1 could be achieved with just context dependent and character n-gram features.

1 Experiments

1.1 Data

As provided, binary labeled data has been used for experimental purpose.

1.2 Feature Extraction

We used two kind of features for representing each token in addition to the baseline 'current word' feature:

- Context dependent features to include context information of the word in the sentence : 'previous word', 'previous to previous word', 'next word', 'next to next word'
- Word features to include the specific nature of the word itself: 'character 3-gram', 'length(word)', 'is word a title', 'is word all upper case', 'is word all lower case', 'is word a digit', 'is it all alphabet', and so on.

Bag of words (BOW) was created using 'indexer' class with all possible kinds of tokens present in the training data. Each feature was represented by weights in the space of BOW.

1.3 Optimizer

We tried SGD, Regularized AgaGrad and Unregularized AdaGrad. AdaGrad was pretty slow on our working machine, so we continued with SGD. The learning rate for SGD was kept at 0.10. 50 epochs were found optimum for training the classifier.

1.4 Classifier

For classification purpose, two classifiers: logistic regression classifier and random forest classifier were explored. On training random forest with basic word features, we could achieve better F1 than baseline, however, it is pretty slower as compared to logistic regression. Thus we continued our experiments with logistic regression.

1.5 Results

For the combination of all kinds of features, we could observe F1 of 91% for the development set. Data reading and training took almost 596 seconds.

Features	Prec.	Recall	F1
Current word (baseline)	0.94	0.64	0.76
+ Context dep. feats.	0.95	0.75	0.83
+ Character 3-gram	0.95	0.83	0.88
+ Other word feat	0.92	0.91	0.91

Table 1: Performance of the logistic regression classifier on development set for different kinds of features.

1.6 Conclusion

Context dependency and character n-gram are helpful features towards entity chunk recognition. We only tried 3-gram but other n-grams would be worth to try.

Acknowledgments

The author has no collaborators but would like to thanks the course T.A. and a student 'Nidhi Kadkol' for helping in solving some crucial doubts.