Assignment-3 Report

Anshuman Mourya
Computer Science and Automation
Indian Institute of Science , Bangalore
`anshumanm@iisc.ac.in`

Abstract

The goal of the assignment is to build an NER system for diseases and treatments. The input of the code will be a set of tokenized sentences and the output will be a label for each token in the sentence. Labels can be D, T or O signifying disease, treatment or other.

1 Dataset Description

Model is trained on given dataset. The format of each line in the training dataset is token label. There is one token per line followed by a space and its label. Blank lines indicate the end of a sentence. It has a total of 3655 sentences.

2 Literature Review

To take advantage of the surrounding context when labelling tokens in a sequence, a commonly used method is conditional random field (CRF). It is a type of probabilistic graphical model that can be used to model sequential data, such as labels of words in a sentence.In a CRF, each feature function is a function that takes in as input: a sentence s ,the position i of a word in the sentence ,the label

$$l_i l_{i-1}$$

of the current word ,the label

$$l_i l_{i-1}$$

of the previous word and outputs a real-valued number (though the numbers are often just either 0 or 1). In CRF, we design a set of feature functions to extract features for each word in a sentence. During model training, CRF will try to determine the weights of different feature functions that will maximise the likelihood of the labels in the training data. Assign each feature function fjfj a weight jj . Given a sentence s, we can now score a labeling l of s by adding up the weighted features over all words in the sentence:

$$score(l|s) = \sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j(s, i, l_i, l_{i-1})$$

Finally, we can transform these scores into probabilities p(l—s)p(l—s) between 0 and 1 by exponentiating and normalizing:

$$p(l|s) = \frac{\exp[score(l|s)]}{\sum_L \exp[score(L|s)]}$$

3 Implementation

3.1 Model Used

A Python binding to CRFSuite , pycrfsuite is used. It uses the CRF implementation to perform sequence tagging. Most of the functions used in code are inbuilt. Self defined features are added.

3.2 Faetures Used

Features are added in incremental way ,to observe the change in precision, recall, f1 score and accuracy. Features used are as follows :

- The word w itself (converted to lowercase for normalisation)

- The prefix/suffix of w (e.g. -ion)

- The words surrounding w, such as the previous and the next word

- Whether w is in uppercase or lowercase

- Whether w is a number, or contains digits

- The POS tag of w, and those of the surrounding words

- Whether w is or contains a special character (e.g. hypen, dollar sign)

Table 1: CRFSuite Implementation

Without Using POS based features				
Label	Precison	Recall	F1	Support
T	0.73	0.66	0.70	872
O	0.95	0.97	0.96	11006
D	0.73	0.53	0.62	702
Avg/Total	0.91	0.92	0.91	12580
Accuracy	0.915			
Using POS based features				
Label	Precision	Recall	F1	Support
T	0.75	0.68	0.72	872
O	0.95	0.97	0.96	11006
D	0.74	0.55	0.63	702
Avg/Total	0.93	0.93	0.93	12580
Accuracy	0.929			

Table 2: One Test data example

Word	True Label	Predicted Label
of	(O)	O
three	O	O
different	O	O
doses	O	O
of	O	O
epidural	T	T
neostigmine	T	T
coadministered	T	T
with	T	O
lidocaine	T	T
for	O	O
postoperative	D	D
analgesia	D	D

3.3 Parameters Used

pycrfsuite Trainer function takes in model parameters. I have passed following parameter values.

- Coefficient of L1 penalty : fine tuned using dev set and finally set as 0.1

- Coefficient of L2 penalty : also fine tuned and set as 0.01

- Number of iterations : 200

4 Result

Table 1 shows the results obtained using CRFSuite Implementation. It contains two types of observations : with and without POS based features . A little improvement has been observed in case when POS based features are added as compared to their without counterparts . Most of the words (around 87 %) are labeled as 'O' . Due to this skewness , scores and accuracy obtained is high. Other used features doesn't have any significant impact on the metrics. Table 2 shows output obtained from one of the sentence in test data .

5 Accuracy/Measures

Precision, Recall, F1 measure and accuracy are used as metrics for evaluation.

6 Link to github code

https://github.com/
anshumanmourya/NLU-Assignment3.
git