E1-246 Assignment 3

G P Shrivatsa Bhargav
SR 14865
bhargavs@iisc.ac.in

1 Tasks

To build a Named Entity Recognition system for the given corpus. The possible tags are disease(D) , treatment(T) and other(O).

2 Dataset

The given dataset has 3655 sentences(64k words). All the words are annotated. The annotation has some inaccuracies (ex: Sarcoidosis is labeled "T").
The dataset is shuffled sentence wise and split into training, validation and test set in the ratio 0.7,0.1,0.2 . The testing data has 10978 "O", 947 "D" and 720 "T" tags.

3 Metrics

As the class label distribution is unbalanced, percentage accuracy based on correct and incorrect classification is not a good metric. Predicting everything as "O" will give 86 percent accuracy and it doesn't tell us how good the classification actually is.
Therefore, to measure the performance, the average F1 score of the three labels is used.

4 Model

CRF from the open source tool Mallet is used. It was found that the default parameters for the order(2) and number of iterations(500) works best.

5 Selected features

This section will cover the selected features and the rationale behind their usage. Other features like word shape, stemming,MESH categories were tried but there were no gains. These will be discussed in the later sections.

- The word itself.
- Lemma. Reducing a word to its lemma has the effect of "grouping" multiple words to its base form. The presence of different forms of the same word will not have a big role to play in degrading performance.
- POS tag. The words belonging to D and T are most likely NN,JJ,NNP and NNS. These tags will distinguish D and T from O.
- The above features for the next and the previous word. Knowing features of the next and previous words matter because the probability of the current word being tagged D or T increases if the previous or next words are marked D or T
- 50 dimensional GloVe embedding.These embeddings are trained to capture word similarities so all the words of the same tag are likely to be closer.
- Prefixes and suffixes . There are some common prefixes and suffixes that disease and treatment words have. For example, many disease words end with 'itis'(gastroenteritis) or 'sis' (Scoliosis). Many treatment words end with 'tomy' (Appendectomy). So 1 to 4 length prefixes and suffixes are used.

6 Features vs performance

Table 1 shows the performance gained by adding each feature.

7 Other features

Multiple other features were tried. Some of them are:

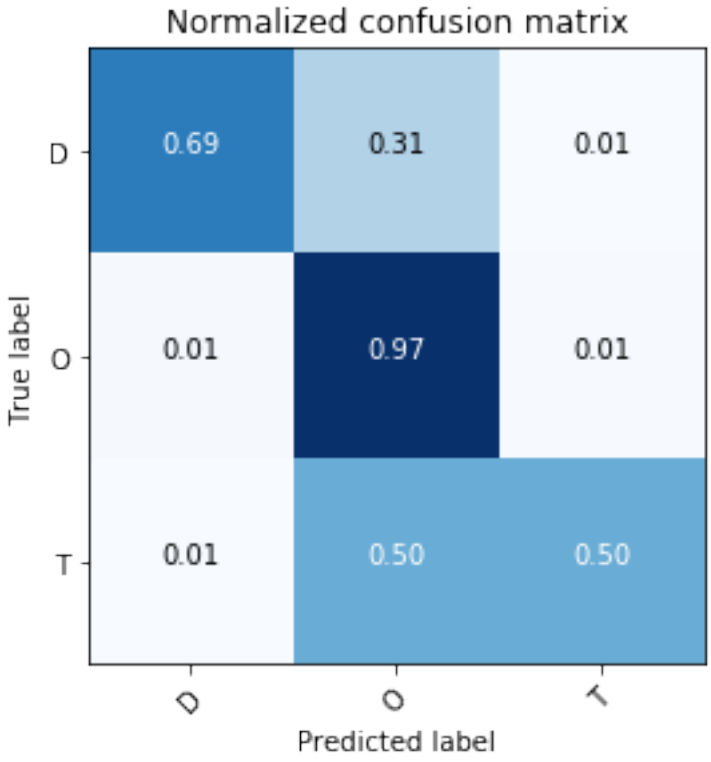
- Word shape: capital, has punctuation, number,etc

- wordnet : similarity hypernym,hyponym and synonyms of known diseases.
- stemming
- MESH : The MESH ontology contains information medical terms and their categories. There are multiple categories like drugs, disease,microbes, therapy,etc. Presence of the word in these categories is a feature.
- Higher window sizes

All these features gave less than 0.01 increase in the F1 score. The exact numbers and the results of other feature combinations are available in the file "features_vs_outfiles.txt" on github.

8 Analysis

The chosen features gave good improvement (from 0.85 to 0.92) in F1 score as expected. Some other features such as MESH, word shape failed to help.
Few different approaches were used to extract features from MESH (category wise BOW, substring match,etc) but they didnt work as per expectation. The reason is that many words belong to more than one category ("disease", "treatment", "anatomy", "drugs") so they were not helpful. A better way to use MESH would be to perform higher order ngram matches with the words in each category instead of just unigram matches.
Through the course of training, precision was traded off for recall in each of the categories. The features selected initially were quite brittle. The were almost acting as "rules". Hence the low recall. Adding GloVe vectors and the prefixes and suffixes significantly increased the F1 score. But overall, the model has higher precision than recall.
The confusion matrix for the final model :



9 Source code

The source code for the mentioned models can be found at https://github.com/gpsbhargav/Medical-NER

10 Software/Resources used

Mallet's CRF implementation was used. The feature engineering was done in python.

Features	**D P/R/F1**	**O P/R/F1**	**T P/R/F1**	**Avg F1**
Word	0.82/0.19/0.31	0.8/0.99/0.94	0.71/0.12/0.20	0.85
Word,POS	0.67/0.33/0.44/	0.90/0.98/0.94	0.61/0.19/0.29	0.86
word,pos,next word features	0.77/0.32/0.45	0.90/0.99/0.94	0.81/0.19/0.30	0.87
word,pos,next and previous word features,lemma	0.77/0.46/0.57	0.92/0.98/0.95	0.67/0.31/0.42	0.89
GloVe 50d	0.77/0.60/0.68	0.94/0.97/0.95	0.67/0.51/0.58	0.91
all above features + prefix and suffix	0.79/0.68/0.73	0.94/0.97/0.96	0.72/0.52/0.60	0.92

Table 1: Features and performances