Assignment 3 - building an NER system for diseases and treatments

Prakash Kumar Uttam
SR - 15247
EE, IISc
https://github.com/pkuttam/NER

1 Name-Entity-Recognition

Name entity recognition is a problem, where given text and corresponding features and labels for training, we will build a system which will identify the labels on new sentences. NER is challenging task because it is very much difficult to identify the labels for each word in never seen sentences.

2 dataset

We have provided dataset 'ner.text', where each word in each sentences has been labeled with **"O"** (ordinary), **"D"** (decease) and **"T"** (treatment) labels. These sentences are taken from medical contexts.

3 Building features

Different experiment has been carried out for building features for these words of sentences in datasets. The features are

1. word identity
2. word suffix
3. word prefix
4. word shape
5. word POS tag
6. word2vec and clustered to 3 clusters
7. end of sentence features
8. word category digit or alphabet

one of the feature vector is
{'+1:word.isupper()': False, 'BOS': True, 'word[-2:]': 'A', '+1:postag': 'JJ', 'postag': 'DT', 'bias': 1.0, '+1:word.lower()': 'new', 'word.istitle()': True, 'word.lower()': 'a', 'word2VecCluster': 2, 'word.isupper()': True, 'word.isdigit()': False, 'postag[:2]': 'DT', '+1:word.istitle()': False, 'word[-3:]': 'A', '+1:postag[:2]': 'JJ'}
The feature vector has assigned to all the word in the dataset.
Different experiment has been carried out with different combinations of the dataset.

3.1 Observations

- The Word2Vec features which are clustered into 3 cluster boost the precision and recall on the test dataset. Because, It is able to give context meaning to each of the word.
- pos tagging also contributes to the precision and recall because the decease and treatment are particularly nouns.

4 Results

The precision, Recall and f1-score on test data is given below

Table 1: precision, Recall and F1-score

	precision	recall	f1-score	support
O	0.95	0.98	0.96	11452
D	0.80	0.67	0.73	975
T	0.76	0.59	0.67	657
avg / total	0.93	0.93	0.93	13084

4.1 observation

- Precision recall for "D" and "T" are low because of it's support.i.e, only the 975 and 657 words tagged with "D" and "T" in the training dataset. However "O" is tagged 11452 so the precision and recall is higher.