Assignment3 : Named Entity Recognition

Arpan Mangal
M.Tech, CSA (14353)
Indian Institute of Science
arpanmangal@iisc.ac.in

Abstract

In this assignment, I try 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.
I need to build a sequence tagger that labels the given sentences in a tokenized test file. To accomplish this, I am using **MALLET**.

1 Pre-Processing

It does not require extensive pre-processing step. Pre-processing is done according to feature I want to extract, as and when on the fly.

2 Splitting

I have divided the randomly shuffled data into 70% Traning set, 10% Development set and 20% Test set.

3 Evaluation Metric

The goal is to predict correct label for each word. Since the distribution of labels i.e. 'D', 'T' and 'O', is highly skewed, the overall accuracy (correct predictions/total words) is not a good measure. In this case, F1-score will be a good metric to evaluate.

I have used 'Precision', 'Recall' and 'F1-score' as comparison metric but goal is to maximize F1-score.

4 Feature Selection

The main task of this assignment is to find features which can help in making good predictions. Based on the underlying task, I tried to use the following features.

1. **Orthographic features:** These are related to the orthography of the text like Uppercase, Lowercase, Title, Hypen, Alphanumeric, Bracks(whether contained in brackets), slash and e.t.c. Such features are very effective in boundary detection.The idea is that these feature might help in recognizing disease tokens, as quite often disease name occur in Uppercase, separated by hypens and all.

2. **Word Normalisation:** It attempts to reduce different form of words, such as nouns, adjective, verbs to their root form. Porter Stemmer has been used to reduce disease names to its root form.

3. **Part-of-Speech Tagging:** POS-tags are helpful in defining boundary of a phrase.

4. **Synonym:** I have used synonyms from WordNet as a feature. May be the synonym of the word is more common and identified easily.

5. **Glove Embedding:** It will give semantic context of the word. It turned out to very useful in the experiments.

6. **Category based on Mesh Number in MESH ontology:** I have performed unigram match of the word from Mesh Subject Heading present in the Mesh ontology. Heading which are mapped, their categories (extracted from Mesh Number) are used as a feature.

5 Training

Above mentioned features are added one by one and performance is observed on Validation Set. First of all, I added Orthographic features. I observed that the performance further degraded.

Features Used	Precision	Recall	F1-score
No Extra Feature	0.86	0.87	0.82
Orthographic(Or)	0.74	0.86	0.79
Word Normalization(WN)	0.86	0.87	0.83
Or + WN	0.74	0.86	0.79
POS tagging(POS)+ WN	0.87	0.88	0.85
Wordnet Synonyms (SYN) + POS + WN	0.87	0.88	0.86
Glove Embedding (GE) + SYN + POS + WN	0.89	0.9	0.89
MESH ontology + GE + SYN + POS + WN	0.9	0.91	0.9

Table 1: Performance Measure on Validation Set.

Without adding any feature, F1-score was 0.82, which degrades to 0.79 after adding this feature. Then I tried adding Word Normalization feature which improves F1-score from 0.82 to 0.83. To further confirm, I took combination of both Orthographic and Word Normalization feature and observed that performance degrades on validation set. It confirms that Orthographic features are not good in this application and therefore removed all together for further experiments.

After that I added POS tagging, Glove Embedding and Mesh Categories as features. Performance increases adding these features. Further to realize the importance of these feature, I took various combination of these features. It was observed that performance degrades in all cases. It justifies that all these features are helping in discriminating.

All the results are summarized in the following table.

6 Observations

I observed that adding Glove Embedding as a feature has a very impact on performance compared to other features. It can be seen from the above table. Glove Embedding improves the F1-score by 0.3. The reason for this is quite intutive. Since Glove Embedding captures the semantic meaning of the word, it should perform better, which is the case here.

From MESH ontology, I tried adding other features like MESH Subject Heading, MESH Number, and e.t.c. But none of them was useful, therefore not shown here.

7 Results

To test the performance on Test data, all the features are kept accept Orthographic feature. Results on Test set is as follows:

1. Precision: **0.90**
2. Recall : **0.91**
3. F1-score : **0.90**

8 Source Code

https://github.com/mangalarpan/Named-Entity-Recognition.git