Assignment -03 : Named Entity Recognition

1 Problem statement

To build an NER system for diseases and treatments. The input 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 respectively. Also, we need to perform ablation study by switching off subsets of features and see the degradation of performance. The final goal is to identify the most useful features (and the value of each feature) for this task.

2 Introduction and Methodology

Named-entity recognition (NER) is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations etc. Language models finds its application in various fundamental tasks such as assigning probability to a sentence and various other advanced tasks such as spelling correction, machine translation etc. NER finds a lot of appication in field of medicine and here we wish to identify disease and treatment from given set of documents. For this purpose, we have used conditional random field(CRF) which is a graph based discriminative method. Whereas a discrete classifier predicts a label for a single sample without considering "neighboring" samples, a CRF can take context into account; e.g., the linear chain CRF (which is popular in natural language processing) predicts sequences of labels for sequences of input samples.The linear chain CRF is the one in which prediction for the current token takes the dependency from the label for the previous token and a set of features for the current token.

For our purpose, we are using sk-learn CRF suite to implement linear chain CRF. And the algorithm used for parameter estimation is gradient descent using the L-BFGS method with elastic regularization.

3 DataSet

Training dataset consist of labeled sentences. 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. The data set given was divided into 80:20 train test split and the 10 fold cross validation was used to tune our hyper parameters on the training data.

4 Evaluation metric

Since, our given training datset is imbalanced, we use F1 score to evaluate the performance which is the harmonic mean of precision and recall. The F1 score is evaluated for each label individually and the average is taken to evaluate final F1 score.

5 Features used for CRF model

Identication of proper features and selecting the most important among them plays plays an important role in doing Named Entity Recognition using models like CRF . So , following are the features, we have experimented with in various subsets to evaluate their importance through the model performance.

Suffix:As we know, many diseases share the common suffix and thus, we pass last three characters and last two characters as two seperate suffix related features.

Prefix:Similarly, many diseases share the common prefix and thus, we pass first three characters as prefix related feature.

Istitle: This feature tells us whether the name starts with capital. It was observed that many diseases and treatment has their name started with capital.

Isalphanumeric: This feature is important because many diseases and treatments are alphanumberic.
(Isupper): This feature could be important sometimes because some of the diseases are cmpletely in upper case. For eg HIV.
POStag: Pos-tagger from nltk library has been used to assign tags to tokens in given dataset. It could play an important role in NER, and this can be the feature which differentiate between treatment and diseases.
Wordlength : The word length could be important to differentiate treatments and disease from others.
Wordstem: The stem of the word can provide useful information about the root word from which the word is derived. But this may even degrade the performance as the stem of disease and treatement maybe the same which reduces the differentiability.
No. of contexts in WordNet:WordNet is a lexical database for the English language.[1] It groups English words into sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members. So for every token, the no of synsets corresponding to the token are calculated which provides the number of contexts in which the corresponding token can be used. It has been observed that disease and treatment have very few contexts whereas others have many contexts. For eg. 'malaria' has only one context whereas 'run' has 13 contexts.
Wordwindow: For any given word for which entity has to be found, it feature can be sent as combination of above mentioned features alone or can be combined with the features of word occuring one position before and one position after to capture the context.
If wordwindow is 1 , it means that we are considering features of only the current word and if wordwindow is 3, it means we are combining the features of previous word and next word also.

6 Results and observations

Following are the classes of subset of features which we have taken initially and then incremented the subset by adding more feature one by one. Corresponding to each class, the F1-scores for each of the label seperately and average F1-score have been shown in table.
class 1:(Suffix, Prefix , Wordwindow) : Here, in this subset, we experimented with the mentioned features. We found out that taking only either of suffix or prefix reduces F-1 score by 10perc. and hence, we finally considered both. Also,whether we are considering features of only current word (wordwindow=1) or also that of previous and next word(wordwindow=3) does not make any difference to F1-score as we dont consider suffix and prefix as their features. Also, adding last two charactars as suffix in addition to last three charactars did not give any improvement.
Hence, final features considered after this class are : Suffix(last3), prefix and Wordwindow= 1 or 3. Following table summarizes labelwise F1 score and average F-1 score for this class:

Model Performance			
labels vs performance	Precision	Recall	F-1 score
Other(O)	.94	.97	.95
Disease(D)	.67	.61	.64
Treatment(T)	.70	.48	.57
Average			.718

Table 1: F-1 scores after considering best feature subset selection from class 1

class 2:(Suffix, Prefix , Wordwindow, Wordlength, Isupper): Here, we added to more features to observe any improvement. We observed that considering Wordwindow as 3 gave a marginal improvement of 1 perc than in case of Wordwindow as 1. Hence, it is always better to consider the features of previous and next word as well to train and predict. Also, both Wordlength and Isupper gave a small marginal improvement. Hence, we consider both and move forward to next class by adding more features. We could observe that still, our F1-scores is very low for treatments. Hence, final features considered after this class are : Suffix(last3), Prefix, Wordwindow=3, Wordlength and Isupper. Table 2 summarizes labelwise F1 score and average F-1 score for this class.
class 3:(Suffix, Prefix , Wordwindow, Wordlength, Isupper, Istitle, Isalphanumeric): Here, we added two more features to observe any improvement. We observed that considering only Istitle degrades the performance, where a Isalphanumeric leads to ver minor improvement.This might have happened because even words which starts after fullstop also have Istitle as true and this

Model Performance			
labels vs performance	Precision	Recall	F-1 score
Other(O)	.94	.96	.95
Disease(D)	.67	.63	.65
Treatment(T)	.69	.50	.58
Average			.728

Table 2: F-1 scores after considering best feature subset selection from class 2

feature rather reduces the discriminability among labels. Hence, we discard Istitle. Hence, final features considered after this class are : Suffix(last3), Prefix, Wordwindow=3, Wordlength, Isupper and Isalphanumeric. Table 3 summarizes labelwise F1 score and average F-1 score for this class.

Model Performance			
labels vs performance	Precision	Recall	F-1 score
Other(O)	.94	.97	.95
Disease(D)	.68	.62	.65
Treatment(T)	.70	.50	.58
Average			.728

Table 3: F-1 scores after considering best feature subset selection from class 3

class 4:(Suffix, Prefix , Wordwindow, Wordlength, Isupper, Isalphanumeric,POStag,No. of context in Wordnet): Here, we added two more features to observe any improvement.As POS tag is expected to improve differentiblity between treatment and disease and No. of context is expected to improve discriminablity of treatment and disease from other, we expect improvement in both the cases. Considering only POS tag as additional feature improves average F1-score to .743 and only No. of context improves it to .734. Combining both of them, we see an improvement of average F1-score to .751 which is a significant increase. Hence, final features considered after this class are : Suffix(last3), Prefix, Wordwindow=3, Wordlength, Isupper and Isalphanumeric, POStag and No. of context. Table 4 summarizes labelwise F1 score and average F-1 score for this class.

class 5:(Suffix, Prefix , Wordwindow, Wordlength, Isupper, Isalphanumeric,POStag,No. of context in Wordnet,Wordstem): Here, we added one more

Model Performance			
labels vs performance	Precision	Recall	F-1 score
Other(O)	.94	.97	.96
Disease(D)	.72	.67	.69
Treatment(T)	.72	.52	.60
Average			.751

Table 4: F-1 scores after considering best feature subset selection from class 4

features to observe any improvement. Stemmization of wors as an additional feature led to improvement in F1-score to .784. We observe, that there is a significant in recall of treatment which was not getting improved. This may be because many treatments may have common stems which improved the recall. Hence, final features considered after this class are : Suffix(last3), Prefix, Wordwindow=3, Wordlength, Isupper and Isalphanumeric, POStag,No. of context wordstem. Table 5 summarizes labelwise F1 score and average F-1 score for this class.

Model Performance			
labels vs performance	Precision	Recall	F-1 score
Other(O)	.95	.97	.96
Disease(D)	.76	.71	.74
Treatment(T)	.76	.59	.66
Average			.784

Table 5: F-1 scores after considering best feature subset selection from class 5

7 Conclusion

After all the observations, we conclude that among all the features we experimented on, Istitle was degrading the performance, Isalphanumeric had least impact on improvement whereas POStag and Wordstem improved F1-score significantly. Also, we conclude that it is always better to consider features of previous as well as next word along with current word to improve the model's performance. Best performace was F1 score of .784 on test data and final feature set considered by our model is (Suffix, Prefix , Wordwindow, Wordlength, Isupper, Isalphanumeric,POStag,No. of context in Wordnet,Wordstem).

Our accuracy was better in case of 'other' rather than 'treatment' or 'disease'. We had worst per-

formance in case of ‘treatment’ because it is hard to differentiate between treatment and others due to overlappings among them.

8 Github Link

Github Link : https://github.com/shubham-IISc/NERAssignment3