A03(NLU)-NER USING CRF

Chaikesh Chouragade
SR No. 14358
GITHUB LINK-https://github.com/chaikesh/NLU$_N$ER

ABSTRACT

In this assignment i have implemented NER(Named Entity Recognition) system for medical text dataset.NER is very important element in medical textmining and text analysis.Its applications include getting information on symptoms of newly evolving diseases, identifying the side-affects of the existing drugs and to get feedback on different kindsof treatment. This task is generally more challenging than common names recognition due to ambiguous naming conventions and long length strings of medical entities.

1 INTRODUCTION

Medical Entity Recognition is a crucial step towards efficient medical texts analysis.The task of a Medical Name Entity Recognizer is two fold. (i) Identification of entity in the sentences. (ii) Entity categorization.

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.There are several approaches to solve the NER problem such as CRF,Recurrent Neural Network.Here i have used the Conditional Random field (CRF) model which is a graph based discriminative model.The main task performed was to consider differennt subsets of features and choose the best model based on evaluation metric.

2 EVALUATION METRIC

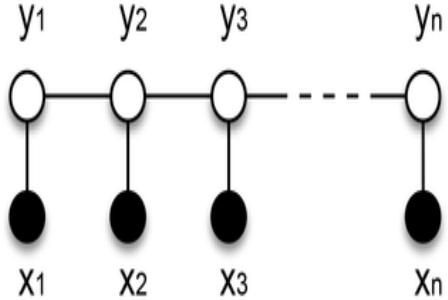
Since the data is unbalanced ,as most of the entities belong to Other(O) category,the models true performance cant be evaluated using only accuracy measure.Therefore different evaluation metrics are used .

F1-score which is the weighted harmonic mean of Precision and Recall is used.Precision is the percentage of the correct annotations and Recall is the percentage of the total NEs that are successfully annotated.

3 MODEL IMPLEMENTATION

A conditional random field is simply a conditional distribution with an associated graphical structure.It belongs to the sequence modeling family which takes context into account and predict sequence of labels for an input sequence.CRF finds its application in POS tagging, Named entity recognition,Shallow parsing,gene finding and an alternative to Hidden Markov Model(HMM) in some applications.
In CRF , corresponding to each token in input sequence we can give different kind of features as well as effectively incorporate both past and future contexts, therefore CRFs are considered to be the State of the art for structure predictions. There are different versions of CRFs which are used for sequence modelling,which are mentioned below...

1.the linear chain CRF in which the prediction for a token only takes the dependency from the label for the previous token in the sequence and the can include a rich set of features for the current word.



2.Dynamic conditional random fields are sequence models which allow multiple labels at each time step, rather than single labels as in linear-chain CRFs.

3.Relational Markov networks are a type of general CRF in which the graphical structure and parameter tying are determined by an SQL-like syntax.

4.Markov logic networks are a type of probabilistic logic in which there are parameters for each rst-order rule in a knowledge base.

For this task a Linear Chain CRF model is being used and the tool utilized to build such a model is sklearn-crfsuite. This is a python tool which helps to built a Linear Chain CRF. The algorithm used for parameter estimation of this model is LGBFS with elastic regularization.

4 FEATURES FOR CRF

Word2vecCluster:

Tokens are clustered (using k-means) into three category,for which we need word embedding. Word2vec was used for this purpose.in this feature we expect that the embeddings of Disease would be more similar to other disease rather than Treatment.

Capitalization and Digit information: For capital-small letters, it is a general notion that diseases start with capital letters or some pattern similar to that(eg: Hepatitis B, OEIS-complex) Various diseases come with name including digits such as Trisomy 13 so this feature is also selected..therefore both these features can be important.

Prefix and Suffix: The words prefix and suffix can provide significant information about the words label as various diseases share common suffix or prefix.The last and first three characters of the word were used for exracting this prefix and suffix information.

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.

Wordlength: word length is used with the intuition that symptoms span over more than one words and their average word length would help as their classifying feature. Drugs have the smallest word length.

Word Lemma: The lemma of a word can also provide significant information for capturing terms being derived from the same root. Such words usually represent the same thing.

Word context: The Current word and its contexts are very useful in recognition tasks.A window of 3 words is being used to capture the context effect for every word.So the feature for a particular word consists of the combination of features of its both side adjacent words and its own features.

POS-tag: Part of speech tags can complement the NER, and this can be the feature which differentiate between treatment and diseases. NLTK Pos-tagger is used for assinging tags to tokens in dataset.

5 MODELS WITH DIFFERENT SETS OF FEATURES

There are total of three models considered each with some subset of features. Although several more models were considered but only three are presented here.the main aim of performing this experiment was to decide which features are most important for the task of NER.

M1 Suffix, Prefix , Wordwindow,Wordlength, Capitalization and Digit information

M2 Suffix, Prefix , Wordwindow,Wordlength, Capitalization and Digit information,word lemma

M3 Suffix, Prefix , Wordwindow,Wordlength, Capitalization and Digit information,POS-tag,word2vecCluster

6 RESULTS

TAG (M1)	F1	PRECISION	RECALL
O	0.95	0.94	0.96
D	0.65	0.66	0.64
T	0.58	0.68	0.53

TAG (M2)	F1	PRECISION	RECALL
O	0.96	0.95	0.96
D	0.70	0.71	0.67
T	0.62	0.72	0.55

TAG (M3)	F1	PRECISION	RECALL
O	0.96	0.95	098
D	0.74	0.81	0.69
T	0.65	0.77	0.56

7 OBSERVATIONS

After analyzing the results of experiments performed following observation are important to note...

1.POS tag feature was significant in improving the performance.
2.Word2vecCluster was equally important featurein terms of performance.
3.word length and word lemma does not have any significant effect.
4. word context was found to give significant improvement.
5.Capitalization had degrading effect on performance,digit information had no significant improvements.

Average performance was observed for Disease(D) around 70-80 percent,which is intitutive as there is always overlap among Disease and Treatment tags. Best performance observed was for Other(O) 90-98 percent.Treatment(T) has worst performance,this may be due to the overlap among other and treatment