NLP TERM PROJECT REPORT

NAMED ENTITY RECOGNITION IN BIOMEDICAL CORPUS

Team Darwinisms :

Anant Bhavsar (17MA20002)
Saurabh Singh (17MT30018)
Bharat Chaudhary (17MA20009)

PROBLEM STATEMENT

In recent years, deep contextual embeddings such as (BERT, FLAIR, ELMO) have proved to be competent enough to detect the exact spans of named entities (both in general English Domain and Biomedical Domain. In this task, we aim to address the gaps of detecting named entities using both context-independent word embeddings on the NCBI Disease, BC5CDR Disease and BC5CDR.

INTRODUCTION

Named Entity Recognition (NER) in textual documents is an essential phase for more complex downstream text mining analyses,being a difficult and challenging topic of interest among research community for a long time. In the domain of bio-medicine, entities can be chemicals, diseases, anatomies, pathways and genes/proteins, etc.which are named in bio-medical literature, which has been growing at an unprecedented speed (PubMed is a typical example). Resolving Biomedical NER successfully is prerequisite for extracting huge amount of biomedical knowledge deposited in the unstructured text-ual literature, transforming them into well-structured formats.Biomedical entities have their own diversities and characteristics of being named, causing the recognition of them in the literature more difficult

Named-entity recognition (NER) is a subtask of information extraction that seeks to locate and classify named entity mentions in unstructured text into predefined categories such as the person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc. In our task we have tried to identify diseases and chemicals in Biomedical Corpuses. We have used a Bidirectional LSTM CRF model for Sequence tagging, This model efficiently uses both past and future input features and also sentence level tag information. Using biomedical corpus BC5CDR and NCBI-disease for entity recognition. We have also used four pre trained word embeddings such as "*Glove, Pubmed-w2v, Pubmed-and-PMC-w2v and wikipedia-pubmed-and-PMC*" for addressing gaps in detecting named entities.

MOTIVATION

Recognition of biomedical named entities in the textual literature is a highly challenging research topic with great interest, playing as the prerequisite for extracting huge amount of high-valued biomedical knowledge deposited in unstructured text and transforming them into well-structured formats. Long Short-Term Memory (LSTM) networks have recently been employed in various biomedical named entity recognition (NER) models with great success.

MODEL ARCHITECTURE

Our Model is based on Bidirectional LSTM CRF. We have trained model on 5 cases, 4 using pre trained word embeddings and one case without any pre trained word embedding on three different corpuses.

Our model consists of three layers. The first layer is an embedding layer. The second layer is a word-based Bi-LSTM layer which combines a word lookup table into multiple word-based Bi-LSTM networks in order to obtain word embeddings. The third layer is a CRF layer which captures the relations among labels extracted from a CRF model.

The overall architecture of the Bi-LSTM-CRF method is shown in Fig. below, where the data flow is from bottom to top.

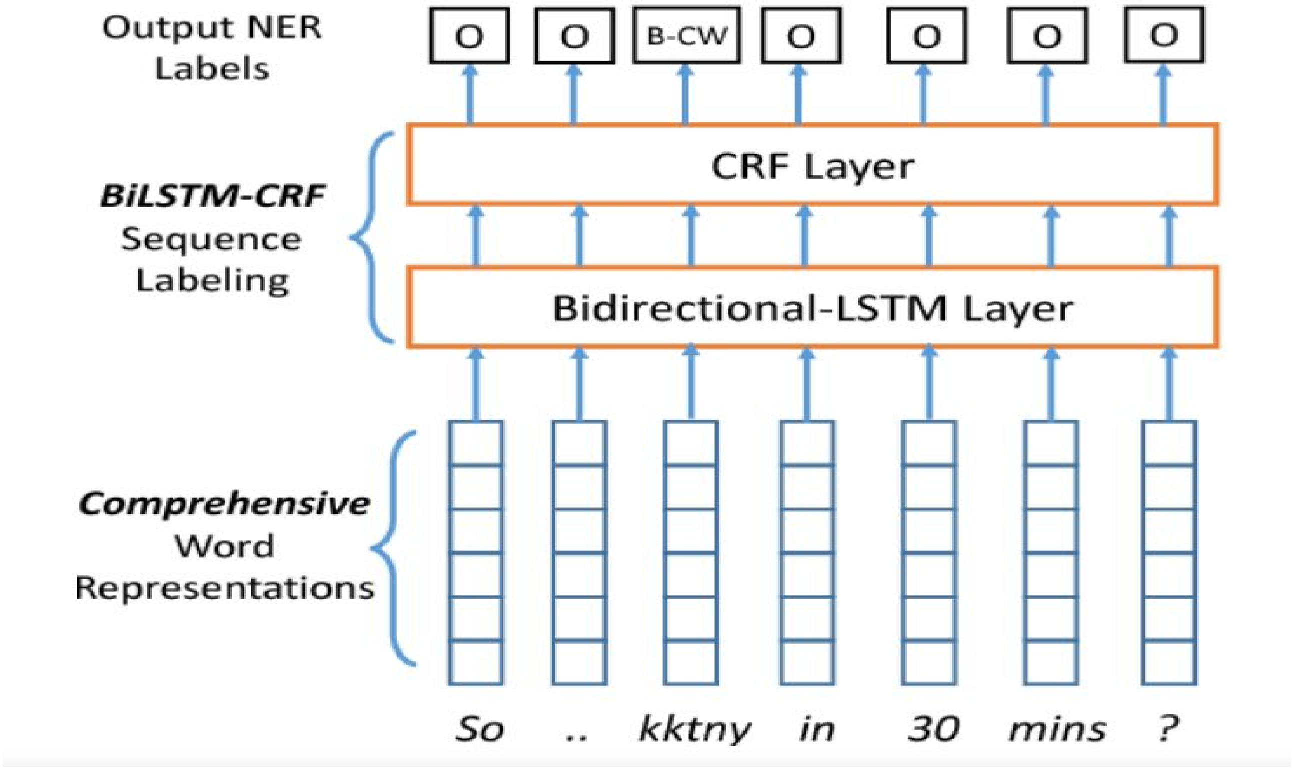


Figure: Bidirectional LSTM CRF Model

BiLSTM-CRF model can learn the information characteristics of a given dataset. Experiments on a publically available NCBI Disease Corpus as an evaluation standard dataset shows our approach achieves a 0.81 F1 measure, which outperforms a number of widely used baseline methods.

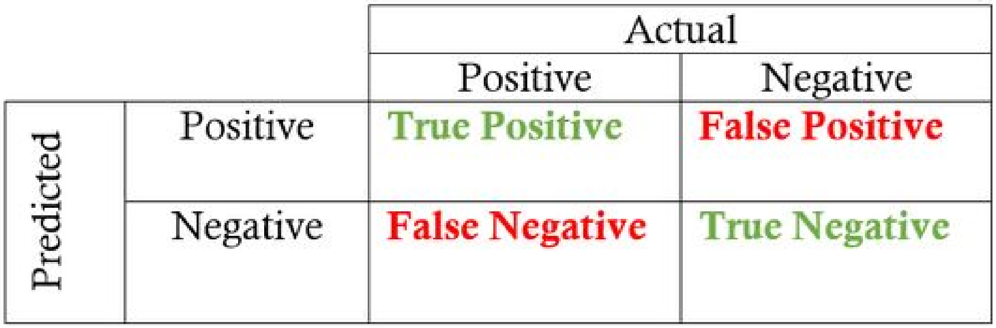
ERROR ANALYSIS

We have done error analysis by calculating Precision, Recall and F1 Score in both the cases exact and partial matches.

Partial match:- A detected Named Entity is counted as correct when any fragment composing the Named Entity is correctly detected.

Exact match:- A detected Named Entity is counted as correct only when all fragments composing the Named Entity are correctly detected.

Precision is a measure of exactness or quality, whereas recall is a measure of completeness or quantity. In simple terms, high precision means that an algorithm returned substantially more relevant results than irrelevant ones, while high recall means that an algorithm returned most of the relevant results.F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it’s better to look at both Precision and Recall.



		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

$$\textit{Precision} = \frac{\textit{True Positive}}{\textit{True Positive}+\textit{False Positive}}$$

$$\textit{Recall} = \frac{\textit{True Positive}}{\textit{True Positive}+\textit{False Negative}}$$

$$\textit{F1 Score} = 2 * \frac{\textit{Precision}*\textit{Recall}}{\textit{Precision}+\textit{Recall}}$$

Results and Discussion

Table 1: Partial Match

CORPUS	EMBEDDING	PRECISION	RECALL	F1 SCORE
NCBI-DISEASE-IOB	No Embedding	0.71	0.82	0.76
	Bioglove	0.71	0.74	0.72
	Pubmed-w2v	0.86	0.77	0.81
	Pubmed-and-PMC-w2v	0.81	0.78	0.80
	wikipedia-pubmed-and-PMC	0.57	0.56	0.57
BC5CDR-DISEASE-IOB	No Embedding	0.47	0.51	0.49
	Bioglove	0.79	0.74	0.76
	Pubmed-w2v	0.82	0.73	0.77
	Pubmed-and-PMC-w2v	0.83	0.71	0.77
	wikipedia-pubmed-and-PMC	0.62	0.49	0.55
BC5CDR-IOB	No Embedding	0.69	0.70	0.70
	Bioglove	0.82	0.71	0.76
	Pubmed-w2v	0.80	0.75	0.77
	Pubmed-and-PMC-w2v	0.87	0.65	0.75
	wikipedia-pubmed-and-PMC	0.73	0.78	0.76

Table 2: Exact Match

CORPUS	EMBEDDING	PRECISION	RECALL	F1 SCORE
NCBI-DISEASE-IOB	No Embedding	0.69	0.74	0.71
	Bioglove	0.68	0.65	0.67
	Pubmed-w2v	0.85	0.71	0.78
	Pubmed-and-PMC-w2v	0.8	0.73	0.77
	wikipedia-pubmed-and-PMC	0.49	0.40	0.44
BC5CDR-DISEASE-IOB	No Embedding	0.38	0.35	0.37
	Bioglove	0.78	0.68	0.73
	Pubmed-w2v	0.81	0.68	0.74
	Pubmed-and-PMC-w2v	0.82	0.65	0.73
	wikipedia-pubmed-and-PMC	0.60	0.33	0.42
BC5CDR-IOB	No Embedding	0.68	0.65	0.66
	Bioglove	0.81	0.67	0.73
	Pubmed-w2v	0.79	0.72	0.75
	Pubmed-and-PMC-w2v	0.87	0.61	0.72
	wikipedia-pubmed-and-PMC	0.72	0.74	0.73

MERIT AND DEMERIT ANALYSIS OF OUR MODEL

1) NCBI-DISEASE-IOB

In this Corpus we are getting the best performance by using *Pubmed-w2v* and *Pubmed-and-PMC-w2v* and the model is performing the worst in the case of wikipedia-pubmed-and-PMC.Model is giving best results on using 10-20 epochs, 100 units of Bi-LSTM and 50 units of Time distribution.*Pubmed-w2v* and *Pubmed-and-PMC-w2v* are trained exclusively trained on biomedical data so its performing better in capturing various relations between biomedical entities and would be able to provide embeddings for most of the words in our corpus thus better performance. Whereas *wikipedia-pubmed-and-PMC* is trained not specifically on biomedical data compared to the latter so it can be the case that it's missing embeddings for some of the domain specific terms and not able to detect some implicit relations between the entities, hence poor performance.

2) ***BC5CDR-DISEASE-IOB***

In this Corpus we are getting the best performance by using Pubmed-w2v and Pubmed-and-PMC-w2v and the model is performing the worst in the case of No pre-trained word embeddings. Model is giving best results on using 40-50 epochs, 100 units of Bi-LSTM and 50 units of Time distribution. Pubmed-w2v and Pubmed-and-PMC-w2v are trained exclusively trained on biomedical data so its performing better in capturing various relations between biomedical entities and would be able to provide embeddings for most of the words in our corpus thus better performance. Whereas in case of No pre-trained word embedding, embeddings are trained on our training set so it's a very small space to cover most of the words and capture essential relationships between various biomedical entities, hence poor performance.

3) BC5CDR-IOB(chemicals+disease)

In this Corpus we are getting satisfactory performance(f1 score > 0.7) in all of the embeddings. Model is giving best results on using 20-30 epochs, 100 units of Bi-LSTM and 50 units of Time distribution. In this corpus we are trying to detect both the diseases and the chemicals whereas in case of other two embeddings we are trying to identify diseases only.

Conclusion

Overall in our analysis we have found that best performance is observed by using *'Pubmed-w2v'* embedding in the case of all three of our corpuses. This maybe due to the reason that this embedding is trained specifically on the biomedical corpus and was able to capture all the specific relations that exist between different entities in this domain that other embeddings were finding difficult to capture.