NAMED ENTITY RECOGNITION

Ritam Chattopadhyay, Kunal Sah

{ritamc,kunalsah}@iisc.ac.in

ABSTRACT

In this project we try to explore the domain of Named Entity Recognition(NER), in the lines of the idea proposed by Lample et al [1]. The paper gives the idea of BiLSTM-CRF based model for the NER task.

1. INTRODUCTION

Named entity Recognition is the process of identifying the tokens of sentences with a predefined set of tags. We use MIT Movie review dataset in our experiments, with 9775 sentence-tag pairs in test set and 2443 sentence-tag pairs in dev set. The dataset follows BIO tagging scheme.

2. TECHNICAL DETAILS

$\mathbf{X} = \{x_1, x_2, ..x_n\}$ is the sequence of input tokens
$\mathbf{y} = \{y_1, y_2, ...y_n\}$ is the sequence of tags.

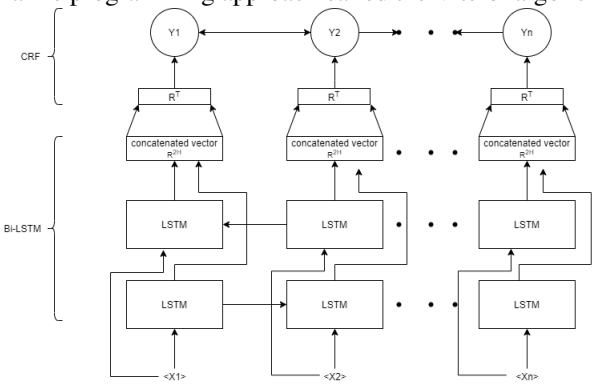
2.1. The Bi-LSTM

First component is the BI-LSTM, which takes in the sentence tokens and output vectors are passed through a linear layer to be projected in $\mathbf{R}^T$ space, T = Number of tags.

2.2. The CRF

The CRF or the Conditional Random Fields, has two matrices namely $\mathbf{A}$ and $\mathbf{P}$. The matrix $\mathbf{P}$ is obtained from the outputs of the Bi-LSTM and has dimension $n \times T$, n = Sequence length of the input. So, $\mathbf{P}_{ij}$ gives the score of the association of the j^{th} tag with the i^{th} token. The matrix $\mathbf{A}$ stores the bi-gram compatibility scores. So, $\mathbf{A}_{ij}$ = score of transition from i^{th} tag to the j^{th} tag. The score of a $(\mathbf{X}, \mathbf{y})$ pair is given by : $\mathbf{S}(\mathbf{X}, \mathbf{y}) = \sum_{i=1}^{n} \mathbf{P}_{\mathbf{x_i}, \mathbf{y_i}} + \sum_{i=0}^{n} \mathbf{A}_{\mathbf{y_i}, \mathbf{y_{i+1}}}$
Probability score for the tag sequence $\mathbf{y}$ for the inout token sequence $\mathbf{X}$, is given by $\mathbf{p}(\mathbf{y}|\mathbf{X}) = \frac{exp(\mathbf{S}(\mathbf{X}, \mathbf{y})}{\sum_{\tilde{y}} exp(S(X, \tilde{y}))}$.
During training phase, the log-likelihood of the observed pair $(\mathbf{X}, \mathbf{y})$ is maximized :
$\log(\mathbf{p}(\mathbf{y}|\mathbf{X}) = \mathbf{S}(\mathbf{X}, \mathbf{y}) - \log(\sum_{\tilde{y}} exp(S(X, \tilde{y})))$
During the decode phase, the selected output sequence is given by: $\mathbf{y} = argmax_{\tilde{y}} S(X, \tilde{y})$

Both these estimates can be efficiently computed by the dynamic programming approach called the Viterbi algorithm.



2.3. Parameters

The model has the following trainable parameters: the weight parameters of the Bi-LSTM, the bi-gram compatibility matrix $\mathbf{A}$ and the learnable word embeddings.

3. RESULTS

Firstly we have implemented and experimented with the model proposed in the paper. The results can be summarized as follows:
Vocab size (Input tokens): $\mathbf{5969}$
Vocab size (Tag tokens) : $\mathbf{25}$
Weighted precision score: $\mathbf{0.872}$
Weighted recall score: $\mathbf{0.88}$
Weighted f1 score: $\mathbf{0.874}$
Micro F1 score : $\mathbf{0.879}$
Accuracy: $\mathbf{0.880}$
Classification Report:

tag	precision	recall	f1-score	support
1	0.91	0.96	0.93	14671
2	0.87	0.92	0.89	1107
3	0.87	0.81	0.84	812
4	0.86	0.82	0.84	802
5	0.82	0.70	0.76	670
6	0.84	0.84	0.84	692
7	0.87	0.93	0.90	565
8	0.68	0.63	0.66	500

tag	precision	recall	f1-score	support
9	0.62	0.38	0.47	415
10	0.62	0.56	0.59	461
11	0.91	0.92	0.92	490
12	0.89	0.69	0.78	424
13	0.81	0.77	0.79	435
14	0.81	0.81	0.81	379
15	0.88	0.74	0.80	435
16	0.89	0.80	0.84	220
17	0.86	0.65	0.74	219
18	0.80	0.41	0.55	99
19	0.52	0.32	0.39	85
20	0.65	0.27	0.39	62
21	0.58	0.40	0.47	53
22	0.21	0.07	0.11	56
23	0.10	0.03	0.04	40
24	0.83	0.67	0.74	30
25	0.00	0.00	0.00	8

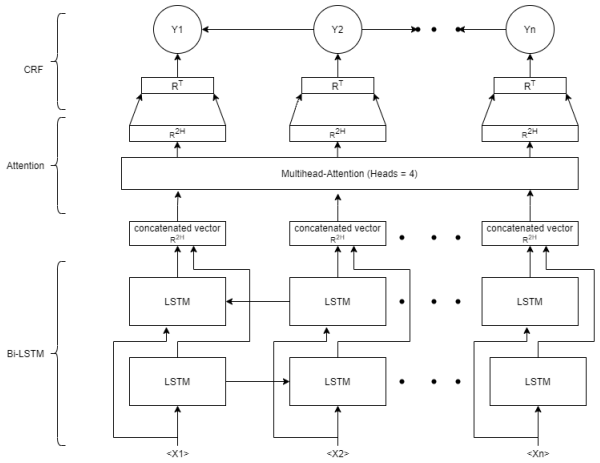


Fig: Bi-LSTM-4-Attetion-CRF Model

Among the models we proposed for the NER task, the Bi-LSTM-4-Attention-CRF (Heads = 4) gave the best Accuracy and F1 scores.

4. CONTRIBUTIONS

1. Implementation of the Bi-LSTM-CRF model proposed in the paper.
2. Implementation of a seq2seq encoder-decoder model with and without applying Bahdanau attention for the same task.
3. Applying Multi-head attention (with different values of heads) along with the Bi-LSTM-CRF network.
The approaches mentioned in points **2** and **3** are contributions from our end.
We provide our findings of the models proposed by us in a condensed tabular format:

Model	***F1 score***	***Accuracy***
Bi-LSTM-CRF Model (proposed in paper)	0.879	0.88
Bi-LSTM-1-Attention-CRF	**0.843**	**0.849**
Bi-LSTM-4-Attention-CRF	0.849	**0.849**
1-Attention-CRF	0.842	0.842
5-Attention-CRF	**0.843**	0.843
Seq2Seq Encoder Decoder Model	0.769	0.773
Encode-Decoder + Bahdanau Attention	0.786	0.786

5. RESOURCES

Dataset used : MIT Movie Review Dataset.
Libraries used : Python, Pytorch, SKLearn, Numpy, pytorch-crf.

6. REFERENCES

[1] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer, "Neural architectures for named entity recognition," *CoRR*, vol. abs/1603.01360, 2016.