CS-626 Assignment 2
Chunking

Shubham Nemani, 203050011,
Anish M M, 203050066
Pooja Verma, 203050072,

October 18, 2020

1 Maximum Entropy Markov Model

1.1 Results

1Iteration	1Log Likelihood	1Accuracy
1	-1.09861	0.132
2	-0.57883	0.851
3	-0.41365	0.928
4	-0.32340	0.939
5	-0.26891	0.943

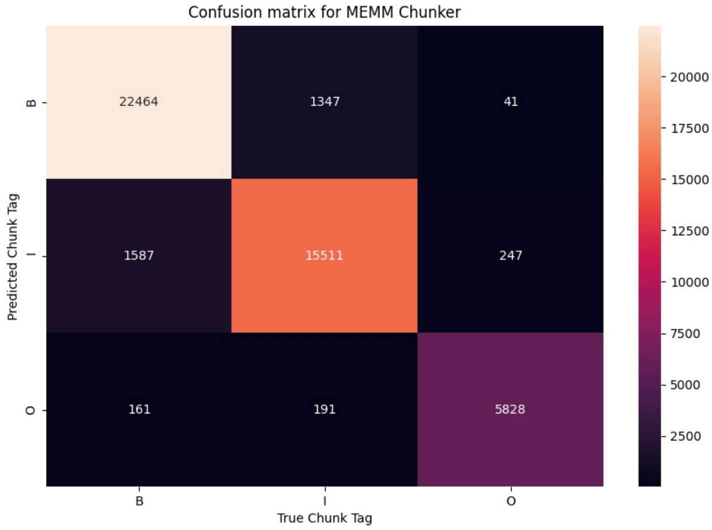
Table 1: Training values

1	1Precision	1Recall	1F1score
B	0.93	0.94	0.93
I	0.91	0.89	0.90
O	0.95	0.94	0.95
Overall	0.92	0.92	0.92

Table 2: Tag wise metrics

1.2 Error Analysis

- **Very poor performance with punctuation marks like " " " , ! ?**
  This is because these were found likely to be tagged as 'O' from the corpus, like – $P(O\ |?) = 1$, $P(O\ |!) = 1$, $P(O\ |;) = 0.99$, $P(O\ |\ ``) = 0.853$ but in some test sentences these were actually I but classifier tagged them O which results in next word to be wrongly tagged as B , since $P(B \mid O) = 0.858$



- **Ambiguity in comma , :**

  From the training corpus , comma is mostly tagged as O ($P(O \mid ,) = 0.946$). So , in test data it's almost every time classified as O , but some times true label is I .

Sentence -	Shopping	centers	in	Malaysia	,	Taiwan	,	Canada	and	Seattle
True	B	I	B	B	I	I	I	I	I	I
Predicted	B	I	B	B	O	B	O	B	O	B

  we can see since comma is predicted as O , thats why the word after it is tagged as B , also Malaysia , Taiwan , Canada and Seattle being NNP , have high probablity to be tagged as B.

sentence-	By	encouraging	massive	,	routine	,	voluntary	testing
true	B	B	I	I	I	I	I	I
predicted	B	B	I	O	B	O	B	I

- **In some cases may be true chunk tags are wrong like - in the sentence " in the Dirks"**

sentence	in	the	Dirks
True	B	O	O
Predicted	B	B	I

  since the is DT and Dirks is NNP , also starts with capital letter , so it is most likely to be chunked as NP – B I , since $P(O \mid NNP) = 0.0028$.

- **Ellipses in test sentences :** consider 2 types of usage for word 'that '

  1. **– that person**
     here that is refering to a particular person , so should be chunked as NP - B I
  2. **– that the person (ellipsis – the is dropped)**
     True- O B I
     Predicted- B (dropped) I
     here that is used as conjuction , so misclassfied.

- **Poor performance with interrogative sentences (starting with VBD):**

  sentences starting with 'Wh' words , are correctly classified as B , but sentences starting with 'Was' , 'Did' , 'Had' , are misclassified.

  True - O O O
  Predicted- B B B
  example

Did	Mr.	Loeb	
O	B	I	true
B	B	I	predicted

  this is because , occurences of 'Was' =2 , 'Did'=3 and 'Had'=3 , very less for a model to learn , on the other hand $P(B \mid was) = 0.996$, $P(B \mid did) = 0.927$ and $P(B \mid had) = 0.944$, so capital letter was ignored and labeled as B.

- It is observed that usage of chunk label O is not jusitfied because sometimes it is assigned to

  1. Conjuctions ,CC
  2. Punctuations
  3. Interrogative sentences (start word VBD)

2 Conditional Random Field

2.1 Results

1	1Precision	1Recall	1Fscore
B	0.96	0.96	0.96
I	0.94	0.94	0.94
O	0.95	0.96	0.96
Overall	0.95	0.95	0.95

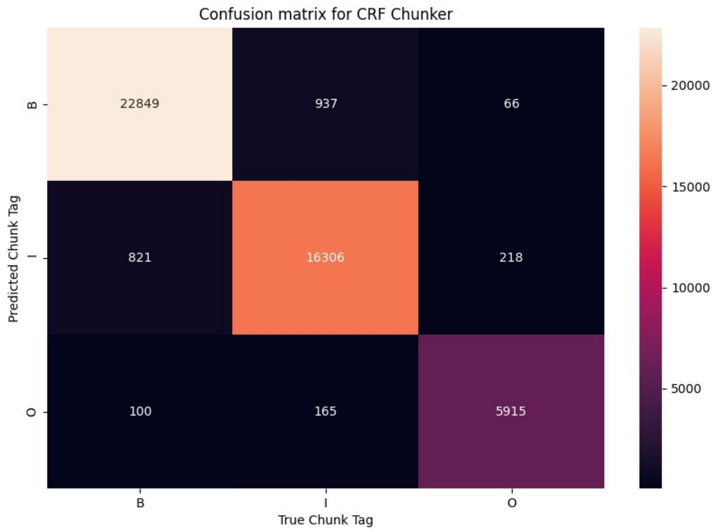
Table 3: Tag wise metrics

2.2 Error Analysis

- **Semantic analysis is required for some sentences like:**

the	longer	the	warranty	,	the	longer	the	cutomer	
O	O	B	I	O	O	O	B	I	true
B	I	B	I	O	B	I	B	I	predicted

  here , 'the longer' because of pos tags DT and VB , it is chunked as VP- DT VB Since , words like longer , shorter , smaller , larger are always labeled either as B or I in train data $P(O \mid these\_words) = 0$ so ,classifier is not able to label them O.



- **Very few occurences of some symbols like #**
  example –

betweeen	#	8	and	#	10	
B	I	I	I	I	I	true
B	B	I	O	B	I	predicted

  here , # being very less frequent , tagged as B inplace of I , also $P(B \mid \#)$ is twice $P(I \mid \#)$.

- **Miclassification of words like haven't , aren't.**

  Since these occurs in two parts , not as a single word :

	have	n't	are	n't	
example -	B	I	B	I	true
	B	O	B	O	predicted

  n't is miclassified as O inplace of I.

- **Some abnormal assignment of O tags in test sentences:**

  example –

won	a	$	100	million	Oregon	general	obligation	veterans	
B	O	O	O	O	O	O	O	B	true
B	B	I	I	I	I	I	I	I	predicted

  model is not able to learn these type of chunking

- Due to ambiguity in English for a word to be mostly tagged as noun and verb both, two chunk phrases NP VP are combined and predicted NP as whole.

  The classifier performance on punctuation tags is improved to some extent as comapred to MEMM, but the problem still exist.

  Like MEMM , CRF is also not able to learn usage of chunk label O.

Also the problem of comma being tagged as O inplace of I also exists in MEMM , because there are two interpretation of comma
example -

markets	,	investment	climate	and	management	practices .	
B	O	B	I	O	B	I	true
B	O	B	I	I	I	I	predicted

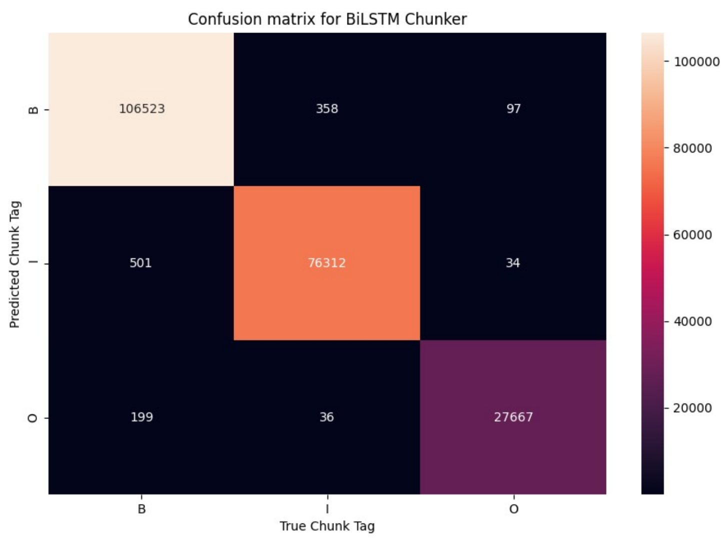
here , classifier chunked (‘investment climate and management prectices’) in one single chunk , inplace of combining it collectively with ‘markets’.

3 Bi-LSTM :

3.1 Results

1	1Precision	1Recall	1F1score
B	0.99	0.99	0.99
I	0.99	0.99	0.99
O	0.99	0.99	0.99
Overall	0.99	0.99	0.99

Table 4: Tag wise metrics



3.2 Error Analysis

- **Some sentences require semantic knowledge:**

The more ________, the more ________
True- O O O O
Predicted- B I ________, B I ________.

'the' being DT , 'the more' is chunked as PP – DT P also
$P(O \mid more) = 0.016$, very less as compared to that of B and I.

- **Like CRF , bilstm is also suffered from words like can't ,doesn't ,year's**

  ca n't , does n't
  True- O O O O
  Predicted- B I B I
  this is because $P(B \mid ca) = 1$, $P(B \mid does) = 0.952$

- **Wrong true lables in test sentences or abnormal test sentences**

  example - The Wall Street Journal
  True - O O O O
  Predicetd- B I I I

  It can be observed that classifier predictions are actually correct.

- **Bi-Lstm removed the problem of punctuations , O label with cunjuction .But O label with VBD problem still exists.**

  Example - What could imply
  True - B O B
  Predicted - B B I
  this is because , $P(B \mid could) = 0.979$, also 'imply' occur only 1 time in train and its pos-tag is VB and $P(I \mid VB) = 0.903$

  Its performance is very high as compared to MEMM and CRF.